Beyond Race and Gender: A Look at Sociodemographic Biases Toward Persons with Disabilities

Keywords: Fairness, Inclusivity, Race, Gender, Disability

Extended Abstract

Bias is pervasive. Sociodemographic biases are a common problem in natural language processing (NLP), and its consequences, have recently received substantial attention as they frequently affecting the fairness and integrity of its applications. For example—within sentiment analysis and toxic, offensive, or abusive language detection, these biases may undermine sentiment predictions for texts that mention personal attributes that unbiased human readers would consider neutral. However, these models most often exhibit biases that lead to unintended discrimination towards persons of demographic groups resulting in false positive predictions of texts as toxic or negative [3], based on certain group identifiers. Such discrimination can have great consequences in the downstream applications both in the public and private sectors towards to minority groups. For example, incorrect inferences in applications like online abuse and opinion analysis in social media platform can lead to unwanted ramifications, such as wrongful censoring, towards certain populations. Popular large language models (LLMs) such as BERT [2] show substantial evidence of the presence of *implicit* sociodemographic biases associated with *race* and *gender*, such biases result in wrongful associations of text related to minority groups as being negative (i.e., *toxic*, *offensive*, or *abusive*). Regardless, little prior work has focused on the identification and impact of disability bias.

Motivation. According to a report on disability by the World Health Organization (WHO), approximately one billion people, or 15% of the world's population, experience some form of disability. Research shows that people with disabilities (PWDs) are the largest population group that faces discrimination regularly [1]. Thus, we propose to identify various forms of disability biases in AI systems, where language involving PWD can often be classified as toxic or even violent. By understanding these biases, our main objective is to influence the development and deployment of AI technologies to provide real-world solutions to societal issues and reduce feelings of disenfranchisement as it is vital for production-ready models to account for minimal stereotypical systemic and sociodemographic biases against protected attributes such as race and gender in downstream tasks.

Method. We define *disability bias* as when a person with a disability is treated less favorably than a person without the disability in the same or similar circumstances. We define *implicit bias* as the attitudes toward people or associating stereotypes with them without. To investigate disability bias, we generate three study groups of sentences, namely, *disability*, *non- disability* and *standard* with the inclusion of race and gender. Sentences in all groups are derived from Twitter and Reddit from three variations of author generated templates: 1. *A <d-adj> person* [*is, has, was] <verb>*, 2. *A <d-adj> <r-adj> person* [*is, has, was] <verb>*, & 3. *A <d-adj> <r-adj> <noun>* [*is, has, was] <verb>*. The *<noun>* tag for all three study groups includes three gender-based nouns (i.e., *man*, *woman,* and *person*) and the *<verb>* tag. We generated ~250,000 *<adj>* type-tags to differentiate the three study groups in terms of race *<r-adj>* (i.e., *White:1*, *Black:2*, and *Asian:3*), gender *<noun>*, and disability terms *<d-adj>* from the following subgroups (i.e., *Sensory*, *Physical*, *Mental*, *Self-Care*, *Employment* and *General Non-Disabled*). We use masked sentence language modelling to find implicit bias in BERT given our 3 experiments (i) *race & disability*, (ii) *disability &*

gender, and (iii) *race, disability & gender.* For example, when given the sentence 'a man has <mask>', BERT predict 'changed' for the masked word. However, for a sentence explicitly mentioning a *Sensory* disability group term and a race the results are vastly different, (i) 'a *visually-impaired black man* <MASK> <MASK>', BERT predicts '*has died*' compared to 'a *visually-impaired white man is talking*'. We then use VADER, a sentiment analysis library, to produce negative sentiment for each sentence.

Results. For each of experiment the results are disheartening as we noticed there is a larger number of toxic, offensive, and abusive words and/or phrases associated with "*woman*", "*Black*" and terms related to the *Self-care* disability groups i.e., visually challenged, quadriplegic, congenital disordered, etc. (see Table 1 and Figure 1).

	Disability Subgroup					
Gender	**Sensory**	**Physical**	**Mental**	**Self-care**	**Employment**	**General Non-Disabled**
Man	-0.12	-0.27	-0.12	-0.11	-0.02	-0.01
Woman	-0.21	-0.32	-0.28	-0.43	-0.13	-0.02
Person	-0.02	-0.02	-0.04	-0.01	-0.01	0.00

Table 1: A table displaying the mean negative sentiment score for each Disability Subgroup given a gender.

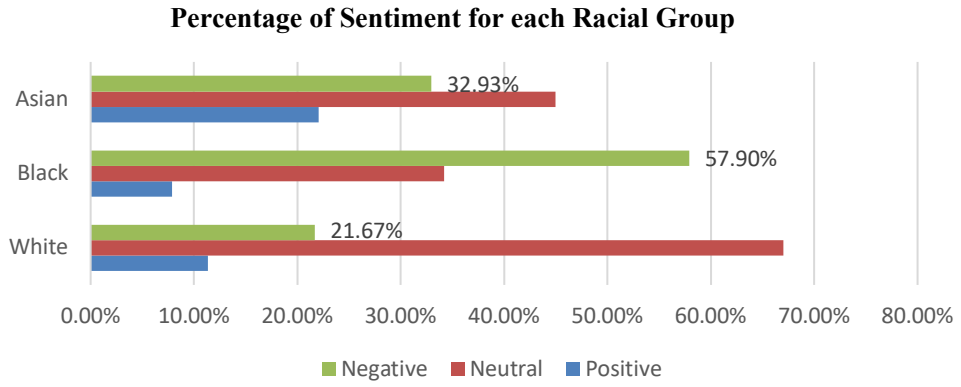

Figure 1: An illustration of sentiment score for each racial group.

In this work, we identified the presence of a challenging form of bias in both raced and gendered language associated with PWDs. The analysis demonstrates bias in LLMs for words used in conversations related to PWD even more over when race and gender are considered. The results show that even when disability is not discussed explicitly, LLMs consistently score sentences with words associated with disability more negatively than sentences containing words with no association to PWD, but has disparities with gender (i.e., *woman*) and race (i.e., *Black*). The results suggest that without debiasing techniques these large models are inadequate in understanding the nuances of language associated to conversations around gender, race, and disability. During the conference we intend on discussing more figures and tables in greater detail where we investigate sociodemographic biases beyond race and gender in popular NLP technologies, specifically LLMs.

[1] Bo Chen and Donna Marie McNamara. 2020. Disability discrimination, medical rationing and covid-19. *Asian bioethics review*, 12(4):511–518.

[2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[3] 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.