A Study of Implicit Language Model Bias Against People With Disabilities

Pranav Narayanan Venkit **Mukund Srinath** **Shomir Wilson**
College of Information Sciences and Technology
Pennsylvania State University
{pranav.venkit, mus824, shomir}@psu.edu

Abstract

Pretrained language models (PLMs) have been shown to exhibit sociodemographic biases, such as against gender and race, raising concerns of downstream biases in language technologies. However, PLMs' biases against people with disabilities (PWDs) have received little attention, in spite of their potential to cause similar harms. Using perturbation sensitivity analysis, we test an assortment of popular word embedding-based and transformer-based PLMs and show significant biases against PWDs in all of them. The results demonstrate how models trained on large corpora widely favor ableist language.

1 Introduction

Recent work on language models show substantial evidence of the presence of sociodemographic biases associated with race and gender (Tan and Celis, 2019; Bolukbasi et al., 2016; Caliskan et al., 2017). Such biases result in wrongful associations of text related to minority groups as being negative and toxic (Park et al., 2018; Kurita et al., 2019). However, little prior work has focused on the identification and impact of disability bias (Hutchinson et al., 2020; Whittaker et al., 2019). According to a report on disability by the World Health Organisation (WHO), approximately one billion people, or 15% of the world's population, experience some form of disability (Bickenbach, 2011). Research shows that people with disability (PWD) are the largest population group that faces discrimination regularly (Whittaker et al., 2019; Chen and McNamara, 2020). We find various forms of disability biases in AI systems as well, where language involving PWD can often be classified as toxic (Venkit and Wilson, 2021) or even violent (Hutchinson et al., 2020). Corpora used to train large language models often only reflect the 'loudest voices' or the the most dominant viewpoints even if they are not representative of the population (Bender et al., 2021) thereby enabling harmful semantic biases (Caliskan et al., 2017).

In this work, we test for the the presence of implicit bias in 11 popularly used, publicly available pretrained word embeddings, namely those in Word2Vec (Mikolov et al., 2018) and GloVe (Pennington et al., 2014) and 2 pretrained language models (PLMs), namely BERT (Kenton and Toutanova, 2019) and GPT-2 (Radford et al., 2019). We analyse negative association created with words related to people with disability (PWD). Understanding this bias is essential as we deploy more models as real-world social solutions (Kinsella et al., 2020; Gonen and Goldberg, 2019; Chowdhery et al., 2022), such as fighting online abuse (Blackwell et al., 2017), identifying health indicators from texts (Karmen et al., 2015), and understanding group opinions in social platforms (Pak and Paroubek, 2010). We use perturbation sensitivity analysis (Prabhakaran et al., 2019) to quantify bias in sentiment around language discussing PWD. Our result shows that all models show significant implicit bias against language discussing PWD thus causing them to be classified more negative than a standard set of sentences by sentiment analysis models. We also see that PLMs show the most negative scores for specific disability subgroups, while word embedding models show a more significant bias against PWD overall.

2 Related Work

Prior work shows that large corpora used to train language models primarily represent hegemonic viewpoints and propagate harmful biases against marginalized populations (Bender et al., 2021; Basta et al., 2019). Prior work identifying bias in NLP models has shown how they can be discriminatory against specific races (Mozafari et al., 2020; Ousidhoum et al., 2021) and genders (Pak and Paroubek, 2010; Bhardwaj et al., 2021). Work analyzing gender bias in word embedding and

PLMs (Bolukbasi et al., 2016; Garg et al., 2018; Kurita et al., 2019) shows how vector representations encode misogynistic, outdated, and harmful stereotypes. We see similar results (Kennedy et al., 2020; Garrido-Muñoz et al., 2021) concerning race and religion, where minority terms such as *Black*, and *Muslim* are associated with hateful phrases. However, disability bias exacerbated by PLMs has been relatively unexplored (Hutchinson et al., 2020; Whittaker et al., 2019).

Offences against PWD are one of the most underreported and concealed hate crimes in the present day (Corcoran et al., 2016; Macdonald et al., 2017). Hidden prejudice related to disability have 'hardly changed over a 14-year period and could take more than 200 years to reach zero bias' (Rojas, 2022), making it the hardest sociodemography bias to reduce. Every experience of PWD is unique and complex (Whittaker et al., 2019). We cannot simplify the nuanced nature of PWD as it undermines their experience (Trewin, 2018) making it easy to classify them to be out of the expected 'norm'. Hutchinson et al. (2020) show how even 'civil' conversations related to PWD show strong associations with terms such as firearms, homelessness and alcoholism. Hassan et al. (2021) demonstrates how BERT perpetuates explicit bias against PWD. and Venkit and Wilson (2021) show how public sentiment and toxicity models show significant explicit bias against terms related to disability. Prior work has explored explicit bias alone, and in contrast, we will analyze the implicit bias in NLP and embedding models by studying the associations generated for sentences containing terms related to PWD.

3 Methodology

We define *disability bias* as when a person with a disability is treated less favourably than a person without the disability in the same or similar circumstances (Commission, 2012). We define *implicit bias* as the attitudes toward people or associating stereotypes with them without conscious knowledge (Institute., 2017). In contrast, *explicit bias* is the intentional association of stereotypes towards a specific population (Institute., 2017). It is useful to identify if large-scale models contain implicit bias as it is not easy to catch and can result in subtle damaging conclusions related to minority populations. In the last 14 years, explicit bias related to disability has dropped down to 37%. Still, its implicit bias has only been reduced by 3%, making it the most complex type of bias to identify and mitigate amongst all sociodemographic prejudices (Rojas, 2022). Our work, therefore, focuses on specifically identifying implicit bias against disability in NLP models.

Name	Dataset Source	Embedding
GN_3B_300D	Google News	Word2Vec
CC_42B_300D CC_840B_300D	Common Crawl	GloVe
WP_6B_100D WP_6B_200D WP_6B_300D WP_6B_50D	Wikipedia	
TW_27B_100D TW_27B_200D TW_27B_25D TW_27B_50D	Twitter	

Table 1: List of embedding models used for analysis and their training sources. Model names follow the pattern: Corpus Training Source_Token Numbers_Vector Dimensions.

3.1 Implicit Bias in Word Embedding

We base our work on Díaz et al. (2018) who studied discourse on ageing in sentiment models. To investigate disability bias, we generate three study groups of sentences, namely, *disability*, *non-disability* and *standard*. Sentences in all groups are derived from four author generated templates: 1. *They are parents of a <adj> <noun>*, 2. *I met a <adj> <noun>*, 3. *There was a <adj> <noun> at school* & 4. *I am a <adj> <noun>*. The <noun> tag for all three study groups includes three gender-based nouns (man, woman and person). The <adj> tag differentiates the three study groups. To populate the <adj> tag, we start with the ten most common adjectives <c-adj> used in the English language (Davies, 2010). We then generate adjectives <adj> for each group based on the *perturbation* technique introduced by Prabhakaran et al. (2019). We use the vector formula *<c-adj>* + *<non-disabilty>* - *<disabilty>* and *<c-adj>* + *<disabilty>* - *<non-disability>* to generate adjective <adj> associations for non-disability and disability groups. For example, in Word2Vec, 'good' + 'non-disabled' - 'disabled' = 'great' but 'good' + 'disabled' - 'non-disabled' = 'bad'. This example shows how this method captures relevant presence of implicit bias. For the standard group, we use the closest associated adjectives in the vector space without any perturbation. The word association formula is designed to remove any explicit mentions of disability or non-disability while modifying the

Group	Terms
Sensory	visually impaired, blindness, deafblind
Physical	physically challenged, epilepsy, cerebral palsy
Mental	mental handicapped, ADHD, autism
Self-Care	visually challenged, quadriplegic, congenital disorder
Go-Outisde -Home	Alzheimer's disease. depressive disorder, psychosis
Employement	intellectually disabled, chronic illness, dyslexia
General	disabled, impairment, disability
Non-Disabled	abled, non-disabled, neurotypical

Table 2: Terms for disability and non-disability groups selected by discourse analysis and guidelines from Washington et al. (2008); NCDJ. (2021).

sentiment of the original adjective, based on the potential presence of implicit bias. Table 1 lists the 11 word embedding models whose vector space we used to generate adjectives for each group and their respective training corpora.

The list of words for <disability> and <non-disability> categories is described in Table 2. While 7 subcategories exist for the disability group, we define six subcategories, namely, *Sensory, Physical, Mental, Self-care, Go-outside-home* and *Employment Disability* based on parameters in the US Census (Bureau, 2021). The definitions are mentioned in the *Appendix*. The seventh subcategory, *General Disability*, encapsulates the general term used for PWD. We select three words for each subgroup based on the guidelines provided by Washington et al. (2008); NCDJ. (2021); Whittaker et al. (2019) and discourse analysis done on the top post of the Subreddit r\disability. We use similar guidelines to select the three words for the non-disability group. We produced 630 adjectives for each of the disability and non-disability groups. After replacing the <noun> and <adj> tags in each of the four templates we generated a total of 15,360 for each embedding model. We then use VADER, a sentiment analysis library, to generate sentiment scores for each sentence. The model evaluates sentiment scores on a scale of -1 (most negative) to +1 (most positive) to represent the overall emotional valence. VADER is a highly cited and used public sentiment analysis model that performs well with not just simple sentences but sentences that include language present in social media, such as emoticons and acronyms (Hutto and Gilbert, 2014).

3.2 Implicit Bias in Language Models

PLMs such as BERT, GPT-2 and PaLM have extended state of the art on a wide range of tasks. However, they largely mimic over-represented hegemonic viewpoints (Bender et al., 2021). For example, when given the sentence 'a man has <mask>', BERT predict 'changed' for the masked word. However, for the sentence, 'a deafblind man has <mask>', BERT predicts 'died'.

We use this masked sentence language modelling, proposed by Kurita et al. (2019) to find implicit bias in BERT. Similar to our technique for studying implicit bias in word embeddings, we generate sentences for three study groups, namely, standard, disability and non-disability. We use the template, *The <adj> <noun> <verb> <mask>*, for sentence generation where, <noun> consists of gender terms (man, woman, person), and <verb> includes the top 100 connecting words used in the English language (Davies, 2010). We populate the <adj> tag with words related to non-disability and disability as shown in Table 2 for the disability and non-disability groups. We generate a set of sentences without <adj> tag for the standard group. We then allow the selected language models to predict the masked word and discard the explicitly mentioned disability or non-disability word used for the <adj> tag. Discarding explicit mentions of disability or non-disability is necessary since we are attempting to measure implicit bias. We generated a total of 7,500 sentences for each model and used VADER to analyse the sentiment of each group.

4 Results and Discussion

Table 3 shows the results from the perturbation sensitivity analysis and the statistical t-test performed for disability and non-disability group against standard group. We calculate the ScoreSense, LabelDistance and ScoreDeviation for sentences perturbed with both disability (D) and non-disability terms (ND), respectively. ScoreSense measures the average difference between the sentiment of perturbed and original sentences. We can see that the ScoreSense is negative for all models for the disability group, suggesting that the sentiment scores dips by that value by the mere addition of disability-related perturbation. Similar to disability-related perturbations, non-disability perturbations cause a general negative drift. This is expected because non-disability terms are often only used in the language

Model	**Score Sense (D)**	**Score Sense (ND)**	**Label Dist. (D)**	**Score Dev.**
GN_3B_300D	**-0.13***	-0.10*	0.39	**0.28**
CC_42B_300D	-0.08*	-0.09*	0.35	0.25
CC_840B_300D	-0.03*	-0.03*	0.38	0.22
WP_6B_100D	-0.07*	-0.04*	0.85	0.19
WP_6B_200D	-0.18*	**-0.14***	**0.86**	0.17
WP_6B_300D	**-0.13***	**-0.11***	0.82	0.19
WP_6B_50D	-0.02	-0.01	0.90	0.21
TW_27B_100D	-0.14*	-0.09*	0.81	0.19
TW_27B_200D	**-0.13***	**-0.13***	0.72	0.20
TW_27B_25D	-0.01	0.01*	**0.86**	0.21
TW_27B_50D	-0.05*	-0.03*	**0.89**	0.17
BERT	-0.06*	0.01*	0.53	**0.29**
GPT2	-0.06*	0.01*	0.73	**0.32**

Table 3: Perturbation sensitivity analysis scores. (*) represents significant values of t-test on sentiment scores between the group and standard group for an α=0.001. D: Disability group, ND: Non-Disability Group.

of PWD (Whittaker et al., 2019), therefore, carrying similar biases that are associated with words in the disability group. We hypothesize that measurements of explicit bias on non-disability terms might not carry the same negative bias since explicit references to non-disability terms are not usually associated with negative sentiment, however this remains to be seen. We also see that all but two models show significant difference in sentiment scores through the t-test analysis, thereby confirming the presence of implicit bias in almost all of them. The most negative score dip is in the performance of GloVe trained on the Twitter dataset. We also see that GloVe trained on Wikipedia corpus performs negative for groups related to Non-Disability. The majority of the users on the internet are non-disabled, young, male individuals from developed countries (WorldBank, 2015). Therefore conversations on social media platforms and curated articles may not be inclusive enough to represent the language associated with PWD.

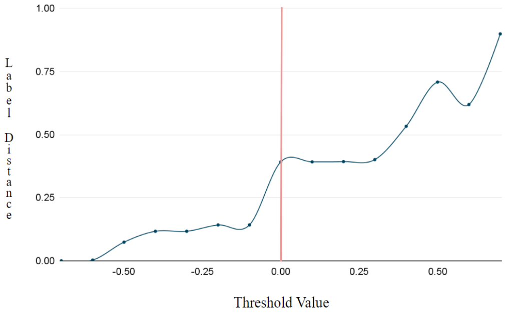

Figure 1: Label Distance for various thresholds of sentiments produced by Word2Vec. The value increases significantly around a threshold 0 (sentiment margin).

LabelDistance in Table 3 measures the Jaccard distance between the sentiments of the set of sentences before and after perturbation. It measures the percentage of sentences that flip between a given threshold. Figure 1 shows how LabelDistance increases with the threshold and jumps significantly at the sentiment margin (0.00). We therefore set this as threshold for analysis measuring the number of flips between positive and negative values. ScoreDeviation is standard deviation of scores due to perturbation, averaged across sentences. GloVe-based models trained on Twitter and Wikipedia have high LabelDistance showing that around 70% to 90% of sentences flip polarities after disability perturbation. High LabelDistance inspite of low ScoreSense values suggest that many weakly positive sentences reversed to weakly negative after perturbation. Finally, we can see that the Word2Vec model and PLMs have high ScoreDeviation , which suggest high polarity between the standard and perturbed sets.

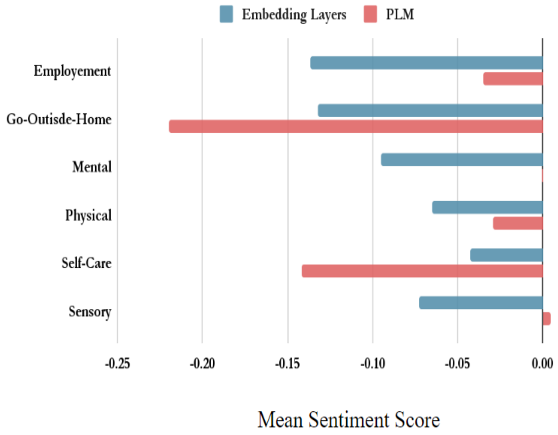

Figure 2: Mean sentiment value for the Census (Bureau, 2021) based subgroups present in the Disability class.

Figure 2 shows the mean sentiment value amongst all models for six disability subgroups. We see that terms related to *Employment Disability* produce the most negative result amongst embedding layers. In contrast, *Go-Outside-Home* has the most negative result amongst PLMs. We also notice that PLMs produce the most negative scores amongst all models for the subgroup analysis. This shows how the black-box nature (O'neil, 2016) of these pretrained models makes it difficult to predict the consequence of each model's implicit biases. *Appendix* shows additional statistical parameters calculated for each group as well as a further breakdown of each subgroup analysis.

5 Conclusion

We identify the presence of a challenging form of bias in language associated with people with disability (PWD): *implicit bias* in language models. The analysis demonstrates bias in both embeddings and PLMs for words used in conversations related to PWD. The results show that even when disability is not discussed explicitly, word embeddings and PLMs consistently score sentences with words associated (in the pretrained vector space) with disability more negatively than sentences containing words with no association to PWD. The results suggest that these large models are inadequate in understanding the nuances of language associated to conversations around disability.

PWD community are more likely to talk about disability and biased models can affect free speech and participation of this marginalized community in online social spaces because of unfair censorship catalyzing harmful ableist ideologies and misrepresenting an already marginalized population. We, through this paper, intend to show these use-cases where these models fail so that developers and owners of these models can be more aware of the potential consequence they can have as a solution to social problems.

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.

Emily M 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 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Jerome Bickenbach. 2011. The world report on disability. *Disability & Society*, 26(5):655–658.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19.

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. *Advances in neural information processing systems*, 29.

US Census Bureau. 2021. How disability data are collected from the american community survey.

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

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

The Australian Human Rights Commission. 2012. Know your rights: Disability discrimination.

Hannah Corcoran, Deborah Lader, and Kevin Smith. 2016. Hate crime, england and wales. *Statistical bulletin*, 5:15.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

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.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens. *arXiv preprint arXiv:2110.00521*.

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.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Perception Institute. 2017. Implicit bias explained.

Christian Karmen, Robert C Hsiung, and Thomas Wetter. 2015. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36.

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.

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.

Alfred Lee; Benjamin Kinsella, Alfred Lee, and Benjamin Kinsella. 2020. How the social sector can use natural language processing (ssir).

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.

Stephen J Macdonald, Catherine Donovan, and John Clayton. 2017. The disability bias: understanding the context of hate in comparison with other minority populations. *Disability & Society*, 32(4):483–499.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

NCDJ. 2021. National center on disability and journalism.

Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Nedjma Djouhra Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. 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 5740–5745.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nikki Rojas. 2022. Why disability bias is a particularly stubborn problem.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, 32.

Shari Trewin. 2018. Ai fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670*.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*.

Anna Cavender University of Washington, Anna Cavender, University of Washington, University of WashingtonView Profile, Shari Trewin IBM T. J. Watson Research Center, Shari Trewin, IBM T. J. Watson Research Center, IBM T. J. Watson Research CenterView Profile, Vicki Hanson IBM T. J. Watson Research Center, Vicki Hanson, and et al. 2008. General writing guidelines for technology and people with disabilities.

Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. 2019. Disability, bias, and ai. *AI Now Institute*.

WorldBank. 2015. Individuals using the internet (% of population) - united states.

A Appendix

We analyze the performance of embedding as well as pretrained large-scale models (PLMs) for *Standard*, *Disability* and *NonDisability* groups. Table 4 shows the definition of each disability subcategory provided by the US Census. Table 5 shows the individual statistical mean and the ScoreRange value of each embedding model and PLM. The statistical mean value is the average score of sentiment value across all sentences of standard, disability and non-disability groups, respectively. The ScoreRange is the range value of specific models across all sentiment scores for all groups. It shows the sensitivity of model performance for language related to PWD.

The mean scores generated by disability and non-disability groups are significantly lower than the standard group. The performance of the disability group is the lowest among the three groups, showing strong implicit bias against these terms in all the models. This value depicts how sentences related to disability are more negative in value than other groups. The ScoreRange results show that BERT and GPT-2 are very sensitive to the language used by PWD as compared to other groups. The results indicate that models with large ScoreRange tend to provide a more wider range of sentiment score results than other models, making the model more sensitive in prediction.

Table 7 and table 6 shows the ScoreSense performance of each subgroup for all the embedding models and PLMs respectively. The data shows that PLMs show large negative scoring for certain groups compared to embedding models. The erratic scoring amongst models shows the unpredictable behaviour of models due to their black-box nature.

Group Name	Definitions
Sensory	Conditions that include blindness, deafness, or a severe vision or hearing impairment.
Physical	Conditions that substantially limit one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying.
Mental	Because of a physical, mental, or emotional condition lasting 6 months or more, the person has difficulty learning, remembering or concentrating.
Self-care	Because of a physical, mental, or emotional condition lasting 6 months or more, the person has difficulty dressing, bathing, or getting around inside the home.
Go-outside-home	Because of a physical, mental, or emotional condition lasting 6 months or more, the person has difficulty going outside the home alone to shop or visit a doctor's office.
Employment	Because of a physical, mental, or emotional condition lasting 6 months or more, the person has difficulty working at a job or business.

Table 4: The definition of each subgroup of the Disability group is mentioned in this table. Six subcategories are decided based on the parameters defined in the US Census to collect disability data. The seventh subgroup 'General' is defined as the common words that are used to refer to people with disability.

Name	Mean (STD)	Mean (D)	Mean (ND)	ScoreRange
GN_3B_300D	0.16	0.02	0.05	1.28
CC_42B_300D	0.15	0.08	0.06	1.26
CC_840B_300D	0.06	0.03	0.03	1.21
WP_6B_100D	0.06	-0.01	0.02	1.30
WP_6B_200D	0.16	-0.02	0.02	1.30
WP_6B_300D	0.12	-0.01	0.01	1.21
WP_6B_50D	0.03	0.01	0.02	1.27
TW_27B_100D	0.11	-0.01	0.02	1.24
TW_27B_200D	0.15	0.02	0.03	1.24
TW_27B_25D	0.01	-0.01	0.02	1.33
TW_27B_50D	0.06	0.01	0.03	1.30
BERT	-0.02	-0.08	-0.01	1.60
GPT2	0.03	-0.02	0.04	1.87

Table 5: The Table shows the statistical mean value calculated for Standard (STD), Disability (D) and Non-Disability (ND) groups respectively. The ScoreRange value is also calculated to measure the range of all the sentiment scores generated by each mode.

Name	Employment	Go-Outside-Home	Mental	Physical	Self-Care	Sensory
BERT	-0.02	-0.21	-0.01	-0.03	-0.14	0.00
GPT2	-0.04	-0.22	0.01	-0.02	-0.14	0.00

Table 6: The table shows the breakdown sentiment score for each Disability Subgroup amongst large scale language models alone. We see that GPT2 and BERT demonstrate significantly negative results for certain subgroups.

Name	Employment	General	Go-Outside-Home	Mental	Physical	Self-Care	Sensory
GN_3B_300D	-0.14	-0.08	-0.19	-0.19	-0.10	-0.13	-0.08
CC_42B_300D	-0.08	-0.06	-0.11	-0.07	-0.09	-0.06	-0.06
CC_840B_300D	-0.05	0.01	-0.13	-0.05	-0.02	0.01	0.01
WP_6B_100D	-0.13	-0.04	-0.21	-0.07	-0.03	-0.01	-0.04
WP_6B_200D	-0.26	-0.14	-0.20	-0.15	-0.16	-0.14	-0.14
WP_6B_300D	-0.25	-0.09	-0.12	-0.10	-0.12	-0.10	-0.09
WP_6B_50D	-0.16	-0.03	-0.09	-0.03	0.08	0.00	0.03
TW_27B_100D	-0.14	-0.17	-0.13	-0.08	-0.08	-0.02	-0.17
TW_27B_200D	-0.17	-0.15	-0.16	-0.10	-0.12	-0.06	-0.15
TW_27B_25D	-0.05	-0.01	-0.03	-0.10	-0.01	0.03	-0.01
TW_27B_50D	-0.05	-0.09	-0.04	-0.08	-0.04	0.03	-0.09

Table 7: The table shows the breakdown sentiment score for each Disability Subgroup amongst embedding groups alone. We see that each embeddings layers demonstrate significantly negative results for certain subgroups.