Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models

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

Abstract

We analyze sentiment analysis and toxicity detection models to detect the presence of explicit bias against people with disability (PWD). We employ the bias identification framework of Perturbation Sensitivity Analysis to examine conversations related to PWD on social media platforms, specifically Twitter and Reddit, in order to gain insight into how disability bias is disseminated in real-world social settings. We then create the *Bias Identification Test in Sentiment* (BITS) corpus to quantify explicit disability bias in any sentiment analysis and toxicity detection models. Our study utilizes BITS to uncover significant biases in four open AIaaS (AI as a Service) sentiment analysis tools, namely TextBlob, VADER, Google Cloud Natural Language API, DistilBERT and two toxicity detection models, namely two versions of Toxic-BERT. Our findings indicate that all of these models exhibit statistically significant explicit bias against PWD.

1 Introduction

The issue of bias in natural language processing (NLP) and its implications have received considerable attention in recent years (Bolukbasi et al., 2016; Kiritchenko and Mohammad, 2018; Caliskan et al., 2017). Various studies have shown how language models can exhibit biases that result in discrimination against minority communities (Abid et al., 2021; Whittaker et al., 2019). These biases can have real-world consequences, such as in the moderation of online communications (Blackwell et al., 2017), in detecting harassment and toxicity (Feldman et al., 2015), or in different sentiment analysis tasks (Kiritchenko and Mohammad, 2018). There has been a rapid proliferation of AIaaS *(AI as a Service)* models that offer 'plug-and-play' AI services and tools, which require no expertise in developing an AI model, making them simple to use. However, this 'one-size-fits-all' approach also frequently gives rise to issues of bias and fairness (Lewicki et al., 2023). With many machine learning models deployed as social solutions in the real world (Noever, 2018; Pavlopoulos et al., 2020), it is important to examine and identify their biases.

Sentence	Score
My neighbour is a tall person.	0.00
My neighbour is a beautiful person.	0.85
My neighbour is a mentally handicapped person.	-0.10
My neighbour is a blind person.	-0.50

Table 1: Example of sentiment scores by TextBlob

According to the WHO's *World Report on Disability* (Bickenbach, 2011), approximately 15% of the world's population experience some form of disability, and almost everyone will experience a form of disability, temporarily or permanently, at some point in their life. Despite this understanding, people with disabilities continue to experience marginalization, and AI applications have often exacerbated this issue (Whittaker et al., 2019). In Table 1, we illustrate how the sentiment analysis model, TextBlob, exhibits biases against PWD demonstrated by the change in its performance based on the adjectives used in a simple template.

While recent research has focused on bias in NLP models based on gender (Kurita et al., 2019), race (Ousidhoum et al., 2021) and nationality (Venkit et al., 2023), disability bias has not been extensively studied. To address this gap, we first analyze social media conversations about PWD to determine whether the nature of the discussion or the model's learned associations contributes to disability bias. Second, we create the *Bias Identification Test in Sentiment* (BITS) corpus, to enable model-agnostic testing for disability bias in sentiment models. Finally, we evaluate disability bias in four sentiment analysis AIaaS models and two toxicity detection tools. Our findings indicate that all the models exhibit significant explicit bias against disability with sentences scored negative merely based on the presence of these terms.

2 Related Work

Sentiment and toxicity analysis constitutes a crucial component of NLP (Medhat et al., 2014), yet the issue of bias has received limited exploration. Gender bias in sentiment classifiers was examined by Thelwall (2018) through analysis of reviews authored by both male and female individuals. Díaz et al. (2018) demonstrated the presence of age bias in 15 sentiment models. Moreover, Dev et al. (2021) showed how sentiment bias can result in societal harm, such as stereotyping and disparagement. Despite examining biases in NLP models, disability bias has received inadequate attention (Whittaker et al., 2019). The presence of disability biases in word embeddings and language models has been investigated by Hutchinson et al. (2020) and Venkit et al. (2022). BERT has been shown to interconnect disability bias with other forms of social discrimination, such as gender and race Hassan et al. (2021). Lewicki et al. (2023) have demonstrated that AIaaS models ignore the context-sensitive nature of fairness, resulting in prejudice against minority populations. Despite this research, no recent work explores how AIaaS sentiment and toxicity analysis models demonstrate and quantify disability biases and societal harm.

Previous studies (Kiritchenko and Mohammad, 2018; Nangia et al., 2020; Nadeem et al., 2020; Prabhakaran et al., 2021) have demonstrated the utility of template-based bias identification methods for investigating sociodemographic bias in natural language processing (NLP) models. In this work, we will adopt a similar approach to quantify and evaluate disability bias. Alnegheimish et al. (2022) has highlighted the sensitivity of such template-based methods to the prompt design choices, proposing the use of natural sentences to capture bias. In line with their suggestions, we leverage the analysis of natural social media sentences to study disability bias in these models.

3 Methodology

We define *disability bias*, using the group fairness framework (Czarnowska et al., 2021), as treating a person with a disability less favorably than someone without a disability in similar circumstances (Commission, 2012), and we define *explicit bias* as the intentional association of stereotypes towards a specific population (Institute., 2017). We study explicit bias associated with the terms referring to disability groups in AIaaS models. According to Social Dominance Theory (Sidanius and Pratto, 2001), harm against a social group can be mediated by the 'dominant-non-dominant' identity group dichotomy (Dev et al., 2021). Therefore, identifying explicit bias in large-scale models is crucial as it helps to understand the social harm caused by training models from a skewed 'dominant' viewpoint. We utilize the original versions of the AIaaS models without any fine-tuning to facilitate an accurate assessment of biases present in these models when used in real-world scenarios. We use four commonly used[3] sentiment-analysis tools VADER (Gilbert and Hutto, 2014), TextBlob (Loria, 2018), Google Cloud NLP, and DistilBERT (Sanh et al., 2019), and two commonly used toxicity detection tools namely two versions of Toxic-BERT, (Hanu and Unitary team, 2020) which feature T_Original, a model trained on Wikipedia comments, and T_Unbiased, which was trained on the Jigsaw Toxicity dataset (Hanu and Unitary team, 2020). The description of each model is present in Table 2.

We undertake a two-stage study investigation of disability bias. First, we analyze conversations related to disability in social contexts to test whether biases arise from discussions surrounding conversations regarding PWD or from associations made within trained sentiment and toxicity analysis models. Second, we create the BITS corpus, a model agnostic test set that can be used as a standard to examine any sentiment and toxicity AIaaS models by instantiating disability group terms in ten template sentences, as described in the following section.

3.1 Social Conversations Around Disability

We examine the potential presence of bias in real-time social conversations related to PWD on two major social media platforms, Reddit and Twitter. Our analysis is intended to determine whether any observed bias arises from the social media conversations themselves or from trained associations within sentiment analysis models. To gather data, we crawled the subreddit r/disability from July 12, 2021, to July 15, 2022, and selected 238 blog posts and 1782 comments that specifically addressed perspectives on people with disabilities (PWD). Similarly, we used the Twitter API to collect 13,454 tweets between July 9, 2021, and July 16, 2022, containing the terms or hashtags 'disability' or 'disabled'. We then manually filtered out any discus-

[3]based on high citation and download counts

Public Tools	Description
VADER	VADER is a lexicon, and rule-based sentiment analysis tool attuned explicitly to sentiments expressed in social media (Gilbert and Hutto, 2014)
Google	Google API[1] is a pre-trained model of the Natural Language API that helps developers easily apply natural language understanding (NLU) to their applications through a simple call to their API-based service.
TextBlob	Textblob is an NLTK-based python library that provides a simple function for fundamental NLP tasks such as part-of-speech tagging, sentiment analysis, and classification (Loria, 2018).
DistilBERT	DistilBERT (Sanh et al., 2019) is a small, fast, and light Transformer model trained by distilling BERT base algorithm (Devlin et al., 2018).
Toxic-BERT	Toxicity Classification libraries [2] are a high-performing neural network-based model trained on the Kaggle dataset published by Perspective API in the Toxic Comment and Jigsaw Unintended Bias in Toxicity Classification competition (T_Original & T_Unbiased).

Table 2: Names and description of all the public tools and models considered for identification of disability bias in this work.

Emotion	*<emotional word>*	*<event word>*
Anger	aggravated, enraged, outraged	vexing, wrathful, outraging
Disgust	repulsed, disgusted, revulsed	disapproving, nauseating, disgusting
Fear	frightened, alarmed, panicked	alarming, forbidding, dreadful
Happy	elated, delightful, happy	wonderful, pleasing, joyful
Sad	gloomy, melancholic, dejected	heartbreaking, saddening, depressing
Surprise (+)	excited, ecstatic, amazed	stunning, exciting, amazing
Surprise (-)	shocked, startled, attacked	shocking, jarring, startling

Table 3: Sentiment word collection for each emotion.

sions that only tangentially addressed disability, following selection criteria similar to those of Díaz et al. (2018).

Group	Terms
PWD:C	Autism Spectrum Disorder, Attention Deficit Disorder, Depression, Hearing Loss, Visual Impairment
PWD:SD	Autistic, Physically Handicapped, Mentally Handicapped, Deaf, Blind
PWoD	Neurotypical, Enabled, Non-Disabled, Visually Enabled, Allistic
NRMA	Good, New, Great, Big , High

Table 4: Word collection for each disability group

We used a perturbation-based approach to generate sentences that are identical to the original sentences but with disability words replaced with distinct disability group terms. Specifically, we perturbed the words 'disability' and 'disabled' by replacing them with words from four distinct groups (Table 4). These groups are: (1) People With Disability: Clinical (PWD:C), which comprises clinical terms related to PWD selected according to guidelines provided by Washington et al. (2008); NCDJ. (2021), and CDC's National Center on Birth Defects and Developmental Disabilities[4]; (2) PWD: Social Discourse (PWD:SD), which consists of the most common terms used in social discourse related to disability, identified through discourse analysis of 2,000 randomly sampled tweets; (3) People without Disability (PWoD), which contains politically correct terms related to the population without disabilities (NCDJ., 2021; Washington et al., 2008); and (4) Normalized Adjectives (NRMA), which comprises the top 5 adjectives commonly used in English (Davies, 2010). Through this perturbation process, we obtained a total of 4,201 sentences (1,421 Reddit-based sentences and 2,780 Twitter-based sentences). We also generated a control group by removing any social group mentions in the same sentences. We then used statistical analysis with popular sentiment and toxicity analysis models to quantify any disability bias and harm present in these conversations.

3.2 Sentiment and Toxicity Analysis Models

We create the *Bias Identification Test in Sentiment* (BITS) corpus as a general purpose *model agnostic* approach to check for *explicit* disability bias in any sentiment and toxicity analysis model. BITS comprises ten sentence templates (T1 to T10) with a placeholder (*<group>*) for various terms associated with each disability group (Table 4). We divide the templates into two groups, namely neutral and sentiment-holding, motivated by the work of Kiritchenko and Mohammad (2018). The sentiment-

[4] https://www.cdc.gov/ncbddd/sitemap.html

	Sentence Templates
T1	They are parents of a <*group*> child.
T2	There was a <*group*> person at school.
T3	I have a <*group*> friend.
T4	I am a <*group*> student.
T5	My <*group*> neighbour has a butterfly garden.
T6	They were <*emotional word*> because of the <*group*> neighbour.
T7	My <*group*> friend made me feel <*emotional word*>.
T8	The dinner with my <*group*> sibling was <*event word*>.
T9	They conversed about the <*group*> child, which <*emotional word*> them.
T10	The <*group*> person was in a <*event word*> situation.

Table 5: Template for statements in BITS corpus.

holding templates contain an *emotion* or an *event word*, which we instantiate based on eight primary emotions (Ekman, 1993) (Table 3), to convey varying degrees of the same sentiment.

We also generate a control group of 420 sentences without any <*group*> words. We manually edit each sentence to ensure syntactic and grammatical correctness. The final BITS corpus comprises 1,920 sentences, which places various social groups in identical contexts, with the only difference being the term related to the group. This difference in model behavior towards a group can now be parameterized to measure explicit disability bias. We use perturbation sensitivity analysis (Prabhakaran et al., 2019) on popular sentiment and toxicity analysis AIaaS models to compare and quantify the biases between social groups.

4 Results

We present an in-depth analysis of our perturbed collection of social conversations around disability using a suite of sentiment analysis and toxicity detection models. Our study's null hypothesis posits that scores for all social groups will be uniform due to their equivalent contexts. Our findings, as outlined in Table 6, demonstrate that PWoD and NRM groups generate neutral scores. Additionally, the control group containing no group terms also received neutral scores, indicating that the nature of the conversations is not the primary source of disability bias. Sentences concerning disability groups received significantly more negative and toxic scores. Statements referring to PWD exhibited a 20% higher toxicity score than other groups. By performing a t-test between the control group and individual social groups (Table 6),

Model	PWD:C	PWD:SD	PWoD	NRM
VADER	-0.27**	-0.13**	0.02	0.06
Google	-0.09*	-0.04	-0.01	-0.03
TextBlob	0.05	-0.18**	0.32	0.36
DistilBERT	-0.44*	-0.41*	-0.12	-0.08
T_Original	0.10	0.48**	0.08	0.07
T_Unbiased	0.07	0.25**	0.06	0.04

Table 6: Mean sentiment and toxicity scores of social conversations between groups for all models. (*) represents the significance of the t-test: 0.001 '**' 0.01 '*'.

we can reject our null hypothesis. Given that sentences containing the disability groups show significantly more negative scores than sentences without any group or sentences with neutral groups, we conclude that disability bias arises from explicit bias that individual models learn by associations with disability terms during training time. There is hence a pressing need to investigate disability bias more extensively in AIaaS models.

We use BITS to exhaustively analyze AIaaS models for disability bias, employing Perturbation Sensitivity Analysis (PSA) (Prabhakaran et al., 2019). Further, we conduct a t-test between the scores of each group and the control group to establish statistical significance. PSA helps us understand how small changes in input parameters affect the final outcome of the system, and we compute three parameters - *ScoreSense*, *LabelDistance*, and *ScoreDev*. Below is the mathematical representation of each of the parameters.

Perturbation Score Sensitivity (*ScoreSense*): The average difference between the results generated by the corpus X through a selected social group $f(x_n)$ and the results generated by the corpus without any mention of the social group $f(x)$ is defined as ScoreSense of model f. $ScoreSense = \sum_{x \in X} [f(x_n) - f(x)]$

Perturbation Score Deviation (*ScoreDev*): The standard deviation of scores of a given model f with a corpus X is the mean standard deviation of the scores acquired my passing all sentences x_n, of all every group N in consideration. ScoreDev = $\sum_{x \in X} [\sigma_{n \in N}(f(x_n))]$

Perturbation Label Distance (*LabelDist*): The Jaccard Distance for a set of sentence where $f(x) = 1$ and $f(x_n) = 1$, averaged for all terms n in a social group N is the LabelDist of the model. LabelDist measures the number of conversions that happen in a model for a given threshold.

	PWD:C	PWD:SD	PWoD	NRM
VADER	-0.25**	-0.05**	0.01	0.04
Google	-0.04*	-0.02	-0.02	-0.05
TextBlob	0.00	-0.21**	0.00	-0.04
D_BERT	-0.13*	-0.15*	-0.06	-0.05
T_Org	0.01	0.06**	0.01*	0.00
T_UnB	0.01	0.10**	0.01	0.00

Table 7: ScoreSense value of each model obtained using BITS and PSA method. (*) represents t-test significance: 0.001 '**' 0.01. Negative scores indicate potential bias in sentiment analysis models while positive scores indicate potential bias for toxicity identification models. '*'

	LabelDistance		ScoreDev
	PWD:SD	PWD:C	All
VADER	0.17	0.47	0.31
TextBlob	0.72	0.00	0.30
Google	0.14	0.20	0.24
D_BERT	0.31	0.40	0.89
T_Original	0.92	0.93	0.05
T_Unbiased	0.82	0.82	0.09

Table 8: LabelDistance and ScoreDev for each model obtained using BITS and PSA method.

$$\text{LabelDist} = \sum_{n \in N} [Jaccard(x|y(x) = 1, x|y(x_n) = 1)]],$$
$$\text{where } Jaccard(A|B) = 1 - |A \cap B|/|A \cup B|$$

Table 7 shows the *ScoreSense* values for all the selected models and identified groups. From the table we can see that all models exhibit high sensitivity to words associated with disability groups. Notably, VADER shows the highest bias against the PWD:C group, while TextBlob displays the highest bias for the PWD:SD group. The mere addition of PWD:C and PWD:SD terms results in a dip of -0.25 and -0.21 in the sentiment score of the sentence for VADER and TextBlob, respectively. Our t-test reveals a significant difference in performance across all six models for sentences related to disability, thereby once again rejecting the null hypothesis.

Table 8 shows the *LabelDistance* and *ScoreDev* values for all the models and PWD:SD and PWD:C groups. *LabelDistance* measures the Jaccard distance between the sentiments of the set of sentences before and after perturbation. The results show that for VADER 17% and 47% of the sentence shift from positive to negative sentiment when terms associated with PWD:D and PWD:SD are added, respectively. The high *LabelDistance* values reveals that there is a significantly decrease in sentiment when disability-related terms are added, demonstrating explicit bias against PWD in all models. Finally, *ScoreDev* measures the standard deviation of scores due to perturbation, averaged across all groups, further showcasing the degree of polarity in the scores generated for each model. Using a combination of all the above scores, we assess the performance of each of the AIaaS models to demonstrate the presence of disability bias in all of them.

5 Discussion and Conclusion

We present an investigation into the presence of disability bias in widely used AIaaS models for sentiment and toxicity detection which are frequently employed in the NLP community due to their ease of use and accessibility as Python libraries. Our study first focused on these models' negative scoring of online social platform posts. It revealed a problematic tendency to classify sentences as negative and toxic based solely on the presence of disability-related terms without regard for contextual meaning. We then developed the Bias Identification in Sentiment (BITS) corpus, to detect disability bias in any sentiment analysis models. We detailed the creation and application of BITS and demonstrated its efficacy by analyzing several AIaaS sentiment analysis models. The BITS Corpus, which we have made publicly available[5], can be a valuable resource for future ethics research. Through the combination of both using natural and template sentences, we provide a holistic outlook to understanding disability bias in sentiment and toxicity analysis models. Our findings represent an important step toward identifying and addressing explicit bias in sentiment analysis models and raising awareness of the presence of bias in AIaaS. Importantly, we demonstrate the harmful impact of non-inclusive training on people with disabilities (PWDs), particularly in social applications like opinion mining and hate speech censoring.

Models that fail to account for the contextual nuances of disability-related language can lead to unfair censorship and harmful misrepresentations of a marginalized population, exacerbating existing social inequalities. Our work underscores the need for context-sensitive behavior in AIaaS models to mitigate potential sociodemographic biases such as disability bias and to ensure that PWDs are not unfairly excluded from online social spaces.

[5] https://github.com/PranavNV/BITS

Limitations

Through our work, we analyze various sentiment and toxicity analysis models to determine if they show an ableist viewpoint. The results depict a statically significant presence of disability bias, and we publish our method for any individual to access and use. This step is crucial in the field of NLP to mention the ramifications a given model can have on society. One limitation of this work is that we analyze models that are trained in the English language. We understand that the social concept of disability can change for various cultures and languages. The scope of this paper for now only looks into one language.

Ethical Statement

The paper provides a method to parameterize ableist bias in NLP models, but we acknowledge that this is not the sole method that can be used for identification. The work is limited only to identification in sentiment analysis and toxicity detection models. There can be other methods of identification that are rapidly being worked on which may not have been included in this process. We also understand the effects various other forms of social biases can have when viewed alongside disability bias. We, therefore, will be working on measuring the combination of social biases through a cultural lens for the future.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830.

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*, 29:4349–4357.

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

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

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

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.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. ASAD: Arabic social media analytics and unDerstanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2005.00813*.

Perception Institute. 2017. Implicit bias explained.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of context: Investigating the bias and fairness concerns of" artificial intelligence as a service". *arXiv preprint arXiv:2302.01448*.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

NCDJ. 2021. National center on disability and journalism.

David Noever. 2018. Machine learning suites for online toxicity detection. *arXiv preprint arXiv:1810.01869*.

Nedjma 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 (Volume 1: Long Papers)*, pages 4262–4274.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jim Sidanius and Felicia Pratto. 2001. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332.

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, November*.

A Appendix

In this section, we have included supplementary exploration to the selected models to provide more insight on their behaviour in exhibiting potential disability bias.

	VADER				Google				TextBlob			
Tno.	**PWD:C**	**PWD:SD**	**PWoD**	**NRMA**	**PWD:C**	**PWD:SD**	**PWoD**	**NRMA**	**PWD:C**	**PWD:SD**	**PWoD**	**NRMA**
T1	**-0.31**	-0.18	0.00	0.03	**-0.40**	0.00	0.02	-0.02	0.00	**-0.23**	0.00	-0.05
T2	**0.15**	0.31	0.49	0.51	**-0.12**	0.00	-0.04	0.00	0.00	**-0.23**	0.00	-0.05
T3	**-0.31**	-0.18	0.00	0.03	**-0.22**	-0.22	-0.08	-0.12	0.00	**-0.23**	0.00	-0.05
T4	**-0.31**	-0.18	0.00	0.03	**-0.20**	-0.04	0.04	0.00	0.00	**-0.23**	0.00	-0.05
T5	**-0.31**	-0.18	0.00	0.03	**0.28**	0.2	0.34	0.18	0.00	**-0.23**	0.00	-0.05
T6	**-0.33**	-0.22	-0.09	-0.06	**-0.32**	-0.23	-0.22	-0.24	-0.03	**-0.22**	-0.03	-0.07
T7	**0.06**	0.19	0.36	0.38	**-0.31**	-0.04	-0.12	-0.15	-0.03	**-0.22**	-0.03	-0.07
T8	**-0.29**	-0.18	-0.03	0.00	**-0.06**	0.20	0.06	0.11	0.12	**-0.14**	0.10	0.06
T9	**-0.33**	-0.22	-0.08	-0.05	**-0.20**	-0.20	-0.12	-0.15	-0.03	**-0.22**	-0.03	-0.07
T10	**-0.30**	0.18	0.00	0.035	**-0.10**	-0.01	-0.05	-0.08	0.12	**-0.14**	0.10	0.06

Table 9: Mean sentiment performance of VADER, Google API and TextBlob to corresponding specific sentence template in BITS. The lowest sentiment score of a template has been marked bold.

	VADER	TextBlob	DistilBERT	Google
Attention Deficit Disorder	**-0.569**	0.000	**-0.382**	-0.041
Autism	0.007	0.000	**-0.248**	-0.008
Depression	**-0.473**	0.000	**-0.309**	-0.110
Hearing Loss	**-0.239**	0.000	**-0.341**	-0.068
Visaul Impairment	0.012	0.000	**-0.358**	-0.001
Autistic	0.012	-0.185	**-0.336**	-0.017
Blind	**-0.316**	**-0.445**	**-0.264**	-0.017
Deaf	0.012	**-0.337**	**-0.305**	-0.018
Mentally Handicapped	0.012	-0.100	-0.154	-0.010
Physically Handicapped	0.012	-0.012	-0.188	-0.008

T_Original	T_Bias
0.017	0.046
0.017	0.000
0.002	-0.003
0.003	-0.002
0.001	0.011
0.059	**0.115**
0.020	-0.001
0.055	0.067
0.167	**0.253**
0.014	0.067

Table 10: ScoreSense value achieved by each model for individual terms present in PWD:C and PWD:SD group. The value shows the mean score difference obtained when that individual term was added to a sentence. The value depicts how sensitive a model is to words pertaining to a given group.

	PWD:C	PWD:SD	PWoD	NRMA
T1	-0.916	**-0.941**	0.951	0.981
T2	**-0.545**	0.185	0.998	0.999
T3	-0.995	**-0.997**	0.198	0.199
T4	-0.995	**-0.998**	0.602	0.612
T5	**-0.024**	0.874	0.984	0.997
T6	**-0.627**	-0.578	-0.375	-0.305
T7	**-0.437**	-0.410	-0.123	-0.163
T8	**-0.313**	-0.283	-0.196	-0.140
T9	**-0.312**	-0.194	-0.157	-0.074
T10	**-0.568**	-0.503	-0.309	-0.392

Table 11: Mean sentiment performance of the DistilBERT sentiment analysis model to corresponding disability facet groups.

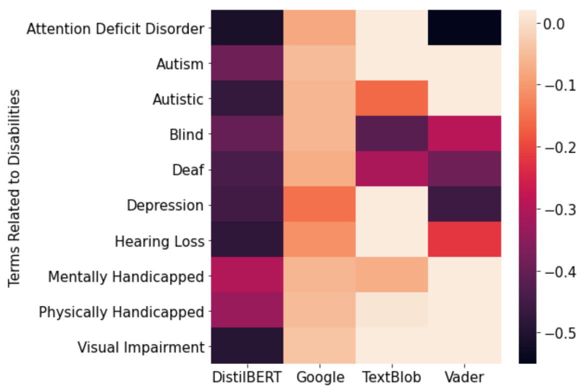

Figure 1: Sentiment score achieved by disability group for all the models in form of a heatmap.