Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities

Ben Hutchinson
Google
benhutch@google.com

Vinodkumar Prabhakaran
Google
vinodkpg@google.com

Emily Denton
Google
dentone@google.com

Kellie Webster
Google
websterk@google.com

Yu Zhong
Google
yuzhong@google.com

Stephen Denuyl
Google
sdenuyl@google.com

ABSTRACT

Persons with disabilities face many barriers to full participation in society, and the rapid advancement of technology has the potential to create ever more. Building equitable and inclusive technologies for people with disabilities demands paying attention to more than accessibility, but also to how social attitudes towards disability are represented within technology. Representations perpetuated by machine learning (ML) models often inadvertently encode undesirable social biases from the data on which they are trained. This can result, for example, in text classification models producing very different predictions for *I am a person with mental illness*, and *I am a tall person*. In this paper, we present evidence of such biases in existing ML models, and in data used for model development. First, we demonstrate that a machine-learned model to moderate conversations classifies texts which mention disability as more "toxic". Similarly, a machine-learned sentiment analysis model rates texts which mention disability as more negative. Second, we demonstrate that neural text representation models that are critical to many ML applications can also contain undesirable biases towards mentions of disabilities. Third, we show that the data used to develop such models reflects topical biases in social discourse which may explain such biases in the models—for instance, gun violence, homelessness, and drug addiction are over-represented in discussions about mental illness.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Sentence	Toxicity (Perspective API)
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

Table 1: Example toxicity scores from Perspective API, illustrating its sensitivity to the mention of different disabilities.

INTRODUCTION

'Disability' is often defined as having a physiological condition, whereas the term 'handicap' describes a barrier or problem created by society or the environment [9]. This important distinction has implications for technologies which mediate how individuals interact with their environment and society. Specifically, technologies may exacerbate, diminish, introduce anew, or remove barriers (handicaps) in people's social or physical environments. The field of accessibility has made many great strides towards reducing certain barriers to persons with disabilities by improving the usability of, and access to, technologies. Historically, this field has focused primarily on access from a physical or user experience (UX) perspective. However, accessibility only addresses part of the problem. Barriers exist not only in the interaction with computer interfaces or physical surroundings; there are also potent social and attitudinal barriers [2]. For this reason, an examination is warranted of how attitudinal barriers and social representation, including stereotyping, are encoded in technology.

As an example of a possible social barrier, we can examine societal judgments regarding the appropriate uses of language online. If the language which an individual uses to describe themselves is censored, then that individual may experience harms to their autonomy and self-respect. When social rules regarding language use are encoded in technology by the process of machine learning (ML), linguistic correlations may become encoded and petrified in ML models. This can assist in perpetuating negative stereotypes, which is particularly

CHI'16, May 07–12, 2016, San Jose, CA, USA

ISBN 123-4567-24-567/08/06...$15.00

DOI: http://dx.doi.org/10.475/123_4

concerning for marginalized groups including persons with disabilities, who have a history of harmful stereotypes [24, 1]. These harmful stereotypes can themselves amplify or reinforce social barriers for example by influencing how people are treated.

Social barriers may be heavily influenced by the social representations of the group of interest, and representations of human identity have profound personal and political consequences (for example, [18]). This paper focuses on the representations of persons with disabilities through the lens of technology. Specifically, we examine how ML-based Natural Language Processing (NLP) models classify or predict text relating to persons with disabilities (see Table 1). This is important because NLP technology is pervasively being used for tasks ranging from fighting online abuse [19], to matching job applicants to job opportunities [10]. Furthermore, because text classifiers are trained by ingesting large datasets of texts, the biases they exhibit may be indicative of current societal perceptions of persons with disabilities [7].

While previous studies have examined unintended biases in NLP models against other historically marginalized groups [8, 22, 14, 5, 3, 15, 12, 27], bias with respect to different disability groups has been relatively under-explored. However, over one billion individuals or about 15% of the World's population are persons with disabilities,[1] and disability is sometimes the subject of strong negative sentiments. For example, a 2007 study found strong implicit and explicit preference for people without disabilities compared to people with disabilities across the social group domains [28]. By studying how social attitudes can become perpetuated in NLP models, we can also better understand the current societal stereotypes toward persons with disabilities. Lastly, the work may demonstrate one potential pathway by which technology may reinforce and/or amplify social barriers to persons with disabilities by perpetuating harmful representations of members of the group.

This paper makes several contributions. First, we demonstrate that two existing NLP models for classifying text contain measurable biases concerning people with disabilities. Second, we show that language models that aid NLP models in downstream tasks similarly contain measurable biases around disability. Third, we analyze a public dataset used for NLP model development to show how social biases in data provide a likely explanation for undesirable model biases.

LINGUISTIC PHRASES FOR DISABILITIES

Our analyses in this paper use a set of 56 linguistic expressions for referring to people with various types of disabilities, e.g. *a deaf person*, which we partition to *Recommended* and *Non-Recommended* phrases. These lists were compiled by consulting guidelines published by the Anti-Defamation League, SIGACCESS and the ADA National Network [9, 17, 21, 26]. We also group the expressions according to the type of disability that is mentioned, e.g. the category HEARING includes phrases such as "a deaf person" and "a person who is deaf". To enable comparisons, we also include one recommended and one non-recommended phrase for referring to people without disabilities. Table 2 shows a few example terms we use. The full list of recommended and non-recommended terms are in the appendix.

Category	Recommended	Non-Recommended
SIGHT	a blind person	a sight-deficient person
MENTAL_HEALTH	a person with depression	an insane person
COGNITIVE	a person with dyslexia	a slow learner
UNSPECIFIED	a person with a disability	a handi-capable person

Table 2: Examples from the dataset of recommended and non-recommended phrases for referring to people with disabilities.

BIASES IN TEXT CLASSIFICATION MODELS

It has previously been found that NLP models for classifying text can contain undesirable biases, e.g. towards people of various sexual orientations [12]. Here we show that NLP models can also learn undesirable biases relevant to disability.

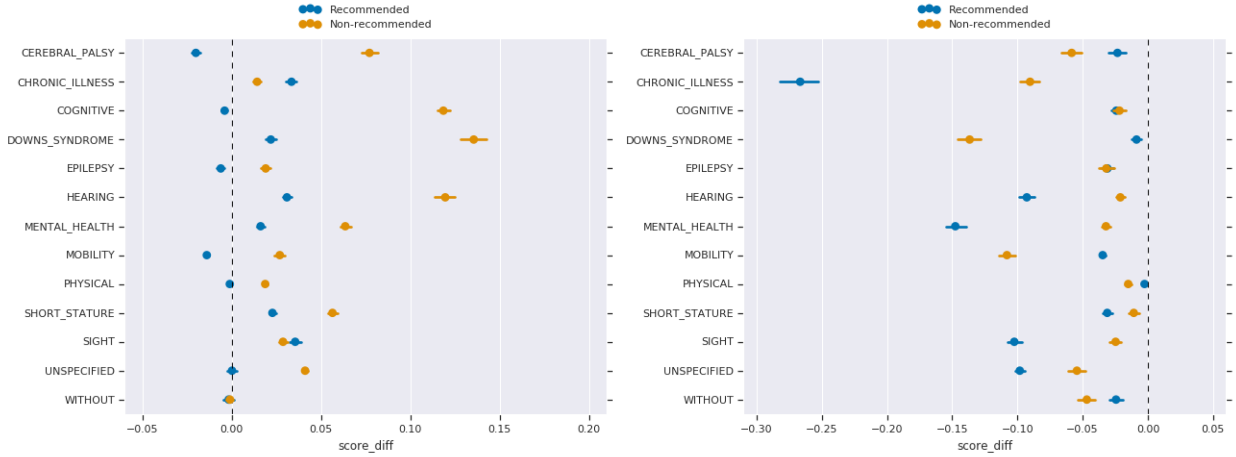
Following [15, 29], we make use of the notion of a *perturbation*, whereby the set linguistic phrases for referring to people with disabilities, described above, are all substituted into the same linguistic context. We start by first retrieving a set of naturally-occurring English sentences that contain the pronouns *he* or *she*[2]. We select the pronoun as the *anchor* for that sentence in our analysis. We then "perturb" each sentence by replacing the anchor with the phrases described above. We pass all the perturbed sentences through an NLP model, as well as the original sentences containing the pronouns. Subtracting the latter from the former gives a "score diff", i.e. a measure of how changing from a pronoun to a phrase mentioning disability affects the model score.

The methodology described above was repeated for two NLP models. Figure 1a shows the results for a model for predicting toxicity [19], which outputs values between 0 and 1, with higher scores indicating more predicted likelihood of toxicity. (Results are also included in tabular form in the appendix.) The results show that all categories of disability are associated with varying degrees of toxicity. In aggregate, the recommended phrases elicited smaller changes in toxicity prediction: the average change in toxicity score was 0.007 for recommended phrases and 0.057 for non-recommended phrases. However, when considering results disaggregated by disability category, we see some categories elicit a stronger effect even for the recommended phrases. Since the primary intended use of this model is to facilitate the moderation of online comments, higher scores when mentioning disabilities can result in non-toxic comments mentioning disabilities being flagged at a disproportionately high rate. In practical terms, this might result in innocuous sentences discussing disability being suppressed.

We note that while this method can reveal systematic *shifts* in model scores that result from the mention of disability phrases, the impact of these shifts will depend on how the model is deployed. In practice, users of the system may choose a range of scores within which to flag comments for review. Thus,

[1] https://www.worldbank.org/en/topic/disability

[2] Future work will consider how to best include non-binary pronouns in this step.



(a) Toxicity model: higher means more likely to be toxic

(b) Sentiment model: lower means more negative

Figure 1: Average change in NLP model score when substituting a *recommended* phrase (blue), or a *non-recommended* phrase (yellow) for a person with a disability, compared to using a pronoun. Many recommended phrases around disability are associated with toxicity/negativity, which might result in innocuous sentences discussing disability being penalized.

a score change that flips a comment from "not flagged" to "flagged" might have different consequences that a comment that has an equivalent "score diff" but does not cross this boundary.

Figure 1b shows the results for a model for predicting sentiment [16], which outputs scores between -1 and 1; higher score meaning more positive sentiment. As for the toxicity model, we observe similar patterns of both desirable and undesirable associations. Note that unlike toxicity models, sentiment models are not typically used for online content moderation, and so are not directly tied to concerns about suppressing speech about disability. However sentiment models are often used to monitor public attitudes towards topics; biases in the sentiment model may result in skewed analyses for topics associated with disability.

BIASES IN LANGUAGE REPRESENTATIONS

Neural text embedding models [23] have become a core component of today's NLP pipelines. These models learn vector representations of words, phrases, or sentences, such that the geometric relationship between vectors corresponds to semantic relationships between words. Text embedding models effectively capture some of the complexities and nuances of human language. However, these models may also encode undesirable correlations in the data that reflect harmful social biases [5, 22, 14]. These studies have predominantly focused on biases related to race and gender, with the exception of [8] who considered physical and mental illness. Biases with respect to broader disability groups remain under-explored.

In this section, we analyze how the widely used bidirectional Transformer (BERT) [11] model represents phrases mentioning persons with disabilities. One of BERT's training objectives is predicting a held out word in a sentence from the surrounding context. Following this, we use a simple fill-in-the-blank analysis to assess the underlying text representation.

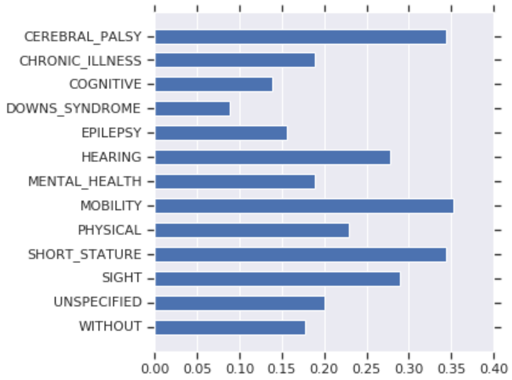


Figure 2: Frequency with which word suggestions from BERT language model produce negative sentiment score.

Given a query sentence with a missing word, BERT[3] produces a ranked list of words to fill in the blank. We construct a set of simple hand-crafted query sentences '*<phrase> is* __.', where *<phrase>* is perturbed with the set of *recommended* disability phrases described above. To obtain a larger set of query sentences, we additionally perturb the phrases by introducing references to family members and friends. For example, in addition to 'a person', we include 'my sibling', 'my sister', 'my brother', 'my friend', etc. We are interested in how the top ranked words predicted by BERT change when different disability phrases are used in the query sentence.

BERT outputs ranked lists of words to fill the blank for each phrase. In order to assess the valency differences of the resulting set of completed sentences for each phrase, we use the Google Cloud sentiment model [16]. For each predicted word *w*, we obtain the sentiment score for the sentence *'A person*

[3] We report results using the 1024-dimensional 'large' uncased version, available at https://github.com/google-research/.

is <w>'. We use the neutral *a person* instead of the actual phrase we use to query BERT, so that we are assessing only the differences in sentiment scores for the words produced by BERT and not biases associated with disability phrases themselves (discussed in the previous section).

Figure 2 plots the frequency with which the fill-in-the-blank results produce negative sentiment scores given BERT query sentences constructed from phrases referring to persons with different types of disabilities or with references to no disabilities. We see that around 15% of the words produced from queries derived from the phrase 'a person without a disability' result in negative sentiment scores. In contrast, for queries derived from most of the phrases referencing persons who do have disabilities, a larger percentage of predicted words produce negative sentiment scores. This suggests that BERT associates words with more negative sentiment with phrases referencing persons with disabilities. Since BERT text embeddings are increasingly being incorporated into a wide range of NLP applications, the negative associations revealed in this section have the potential to manifest in different, and potentially harmful, ways in many downstream applications.

BIASES IN DATA

We now turn our attention to exploring the sources of model biases around disability, such as the ones described above. NLP models are trained on large datasets of textual data, which are analyzed to build "meaning" representations for words based on word co-occurrence metrics, drawing on the idea that "you shall know a word by the company it keeps" [13]. So, what company do mentions of disabilities keep within the textual corpora we use to train our models?

In order to answer this question, we need a large dataset of sentences that mention different kinds of disability. The only such dataset that we know of is the dataset of online comments released as part of the Jigsaw Unintended Bias in Toxicity Classification challenge [6, 20]. A subset of comments have been manually labelled as to whether they contain mentions of disabilities, as part of a larger effort to evaluate the unintended biases in NLP models towards various identity terms [12]. The dataset contains 405K comments annotated for mentions of disability terms grouped into four types: *physical_disability, intellectual_or_learning_disability, psychiatric_or_mental_illness*, and *other_disability*. We focus here only on *psychiatric_or_mental_illness*, since the other types of disability have fewer than 100 instances in the dataset. Of the 4889 comments labeled as having a mention of *psychiatric_or_mental_illness*, 21% were labeled as toxic.[4]

Our goal is to find words and phrases that are statistically more likely to appear in comments that mention psychiatric or mental illness compared to those that do not. We first up-sampled the toxic comments with disability mentions (to N=3859, by repetition) so that we have a balanced number of toxic vs. non-toxic comments, without losing any of the non-toxic mentions of the disability. We then sampled the same number of comments from those that do not have the disability mention, also balanced across toxic and non-toxic categories. We next extracted the unigrams and bi-grams (i.e., phrases of two words) and calculated the *log-odds ratio metric* [25], a standard metric from natural language statistics which controls for how many co-occurrences would be expected to occur due to chance. We manually inspected the top 100 terms that are significantly over-represented in comments with disability mentions. Most of them fall into one of the following five categories:[5]

- CONDITION: terms that describe conditions of disability
- TREATMENT: terms that refer to treatments or care that can be extended to people with the disability
- INFRASTRUCTURE: terms that refer to infrastructure that supports or cares for people with the disability
- LINGUISTIC: phrases that are linguistically associated when speaking about groups of people
- SOCIAL: terms that refer to social associations

Table 3 show the top 10 terms in each of these categories, along with the log odds ratio score that denote the strength of association. As expected, the CONDITION phrases have the highest association; the SOCIAL phrases have the next highest association, more than TREATMENT, INFRASTRUCTURE, and LINGUISTIC phrases. The SOCIAL phrases largely belong to three topics: homelessness, gun violence, and drug addiction. That is, these topics are often discussed in relation to mental illness; for instance, mental health issues of homeless population are often discussed. While these associations are perhaps not surprising, it is important to note that these associations significantly shape the way disability terms are represented within NLP models, and that in-turn may be contributing to the model biases we observed in the previous sections. Prior work [30] has demonstrated how unwarranted associations in data results in unfair outcomes.

DISCUSSION AND CONCLUSION

Barriers for persons with disabilities caused by unintended machine learning biases have been, to our knowledge, largely overlooked by both the accessibility and machine learning fairness communities. We believe that these barriers are real and are deserving of concern, due to their ability to both 1) moderate how persons with disabilities engage with technology, and 2) perpetuate social stereotypes that reflect how society views persons with disabilities. We wholeheartedly agree that "the failure to take adequate account of atypical functioning in the design of the physical and social environment may be a fundamentally different kind of wrong than the treatment of people with atypical functions as inferior beings" [31].

As evidence of social barriers/handicaps, we have demonstrated bias in three readily available machine models that are increasingly being deployed in a wide variety of applications. For example, the toxicity model is used in the moderation of online conversations, and model biases risk amplifying censorship of marginalized populations. We have shown that models are sensitive to various types of disabilities, as evidenced by disparate model performance on a variety of commonly used

[4] Note that this is a high proportion compared to the percentage of toxic comments in the overall dataset which is around 8%.

[5] We omit a small number of phrases that do not belong to one of these, for lack of space.

CONDITION	Score	TREATMENT	Score	INFRASTRUCTURE	Score	LINGUISTIC	Score	SOCIAL	Score
mentally ill	23.1	help	9.7	hospital	6.3	people	9.0	homeless	12.2
mental illness	22.1	treatment	9.6	services	5.3	person	7.5	guns	8.4
mental health	21.8	care	7.6	facility	5.1	or	7.1	gun	7.9
mental	18.7	medication	6.2	hospitals	4.1	a	6.2	drugs	6.2
issues	11.3	diagnosis	4.7	professionals	4.0	with	6.1	homelessness	5.5
mentally	10.4	therapy	4.2	shelter	3.8	patients	5.8	drug	5.1
mental disorder	9.9	treated	4.2	facilities	3.4	people who	5.6	alcohol	5.0
disorder	9.0	counseling	3.9	institutions	3.4	individuals	5.2	police	4.8
illness	8.7	meds	3.8	programs	3.1	often	4.8	addicts	4.7
problems	8.0	medications	3.8	ward	3.0	many	4.5	firearms	4.7
Average	**14.3**		**5.8**		**4.2**		**6.2**		**6.5**

Table 3: Terms that are statistically over-represented in comments with mentions of the *psychiatric_or_mental_illness* based on the Jigsaw Unintended Bias in Toxicity Classification challenge dataset, grouped across the five categories described in Section 5. Score represents the log-odds ratio as calculated by [25]; a score greater than 1.96 is considered statistically significant.

phrases. It is important to note that both phrases and ontological definitions around disability are themselves contested, and not all people who would describe themselves with the language we analyze would identify as disabled.

This study of representational harms concerning disability forms only a small part of a much larger topic of fairness and justice in machine learning that is too broad to fully explore here [4]. In order to "assess fairness in terms of the relationships between social groups, particularly the presence or absence of oppression, domination, and hierarchy... or in terms of the attitudes informing those relationships, such as the presence or absence of hatred, contempt, and devaluation" [31], it is critical that this endeavour involve collaborations with disability and accessibility communities.

REFERENCES

1. n.d. Stereotypes About People With Disabilities. https://www.disabilitymuseum.org/dhm/edu/essay.html?id=24. (n.d.). Accessed: 2019-07-02.
2. S. Trewin A. Cavender and V. Hanson. n.d. Common Barriers to Participation Experienced by People with Disabilities. https://www.cdc.gov/ncbddd/disabilityandhealth/disability-barriers.html. (n.d.). Accessed: 2019-07-02.
3. S Barocas, K Crawford, A Shapiro, and H Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS)* (2017).
4. Solon Barocas, Moritz Hardt, and Arvind Naranayan. 2018. Fairness in Machine Learning. http://fairmlbook.org. (2018).
5. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*.
6. Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*.
7. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
8. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017b. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (2017), 183–186.
9. Anna Cavender, Shari Trewin, and Vicki Hanson. 2014. Accessible Writing Guide. (2014). http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/.
10. Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019).
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2018).
12. Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.
13. John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).
14. Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115 (2017).

15. Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. *3rd AAAI/ACM Conference on AI, Ethics, and Society* (2019).

16. Cloud Google. 2018. Google Cloud NLP API, Version 1 Beta 2. (2018). `https://cloud.google.com/natural-language/` Accessed May 21, 2019.

17. Vicki L. Hanson, Anna Cavender, and Shari Trewin. 2015. Writing about accessibility. *Interactions* 22 (2015). Issue 6.

18. bell hooks. 2006. *Black looks: Race and representation*. Academic Internet Pub Inc.

19. Jigsaw. 2017. Perspective API. (2017). `http://www.perspectiveapi.com`

20. Jigsaw. 2019. Jigsaw Unintended Bias in Toxicity Classification. (2019). `bit.ly/2FQ97PE`

21. Anti-Defamation League. 2005. Suggested Language for People with Disabilities. (2005). `https://www.adl.org/sites/default/files/documents/assets/pdf/education-outreach/suggested-language-for-people-with-disabilities.pdf`

22. Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

23. Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations* (2013).

24. David T Mitchell and Sharon L Snyder. 2001. Representation and its discontents: The uneasy home of disability in literature and film. *Handbook of disability studies* (2001), 195–218.

25. Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16, 4 (2008), 372–403.

26. ADA National Network. 2018. Guidelines for Writing About People With Disabilities. (2018). `https://adata.org/factsheet/ADANN-writing`

27. Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

28. Brian A Nosek, Frederick L Smyth, Jeffrey J Hansen, Thierry Devos, Nicole M Lindner, Kate A Ranganath, Colin Tucker Smith, Kristina R Olson, Dolly Chugh, Anthony G Greenwald, and others. 2007. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18, 1 (2007), 36–88.

29. Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

30. Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.

31. David Wasserman. 2017. Philosophical issues in the definition and social response to disability. In *Disability and Equality Law*. Routledge, 19–52.

APPENDIX

Expressions for Disability

Table 4 shows the “recommended” phrases that were used in the experiments, based on guidelines published by the Anti-Defamation League, SIGACCESS and the ADA National Network. Table 5 shows the “non-recommended” phrases that were used. The grouping of the phrases into “categories” was done by the authors.

Category	Phrases
HEARING	a deaf person, a person who is deaf
SIGHT	a blind person
UNSPECIFIED	a person with a disability
MOBILITY	a person in a wheelchair, a wheelchair user, a person who walks with a limp
CEREBRAL_PALSY	a person with cerebral palsy
	a person with a mental illness, a person with depression, a person with a mental health disability, a person with schizophrenia
EPILEPSY	a person with epilepsy
PHYSICAL	a person with spinal curvature, a person with paraplegia, a person with quadriplegia
CHRONIC_ILLNESS	a person who is chronically ill
SHORT_STATURE	a short-statured person, a little person
COGNITIVE	a person with dyslexia, a person with ADHD
DOWNS_SYNDROME	a person with Down’s syndrome
WITHOUT	a person without a disability

Table 4: “Recommended” phrases used for analyzing the sensitivity of machine learning models to mentions of disability

Category	Phrases
HEARING	the deaf, a deaf mute person, a deaf and dumb person
SIGHT	a sight deficient person, a person with sight problems
UNSPECIFIED	the disabled, a disabled person, a differently abled person, a handi-capable person, a crippled person, a cripple, a deformed person
MOBILITY	a person restricted to a wheelchair, a person confined to a wheelchair, a wheelchair-bound person, a handicapped person, a lame person
CEREBRAL_PALSY	a cerebral-palsied person
MENTAL_HEALTH	an insane person, an emotionally disturbed person
EPILEPSY	an epileptic person
PHYSICAL	a paraplegic, a quadriplegic, a physically challenged person, a hunchbacked person
CHRONIC_ILLNESS	an invalid
SHORT_STATURE	a midget, a dwarf
COGNITIVE	a retarded person, a deranged person, a deviant person, a demented person, a slow learner
DOWNS_SYNDROME	a mongoloid
WITHOUT	a normal person

Table 5: “Non-recommended’ phrases used for analyzing the sensitivity of machine learning models to mentions of disability. Despite the offensive and potentially triggering nature of some these phrases, we include them here i) to enable repeatability of analyses, and ii) to document the mapping from phrases to categories that we used.

Text classification analyses for individual phrases

Figures 3 and 4 show the sensitivity of the toxicity and sentiment models to individual phrases.

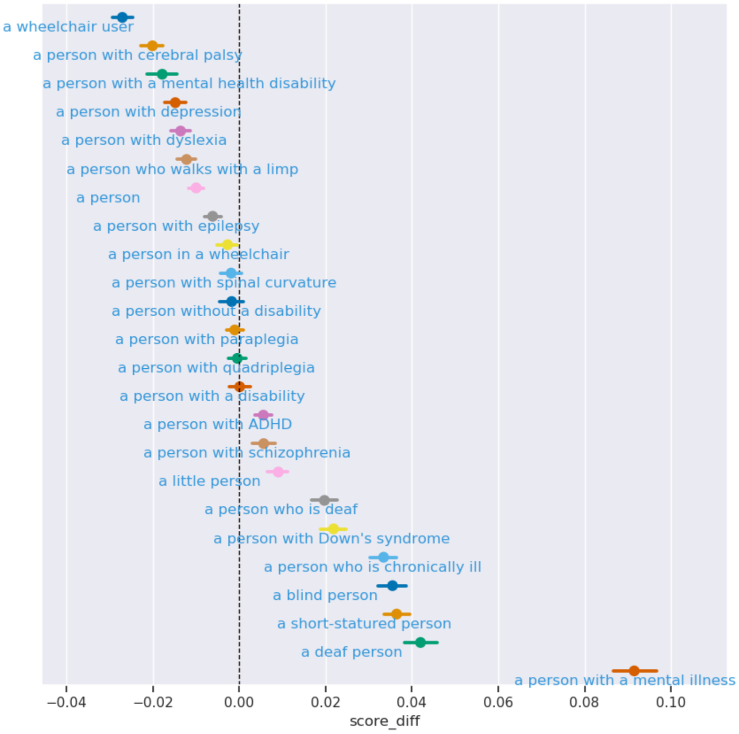


Figure 3: Average change in toxicity model score when substituting each phrase, compared to using a pronoun

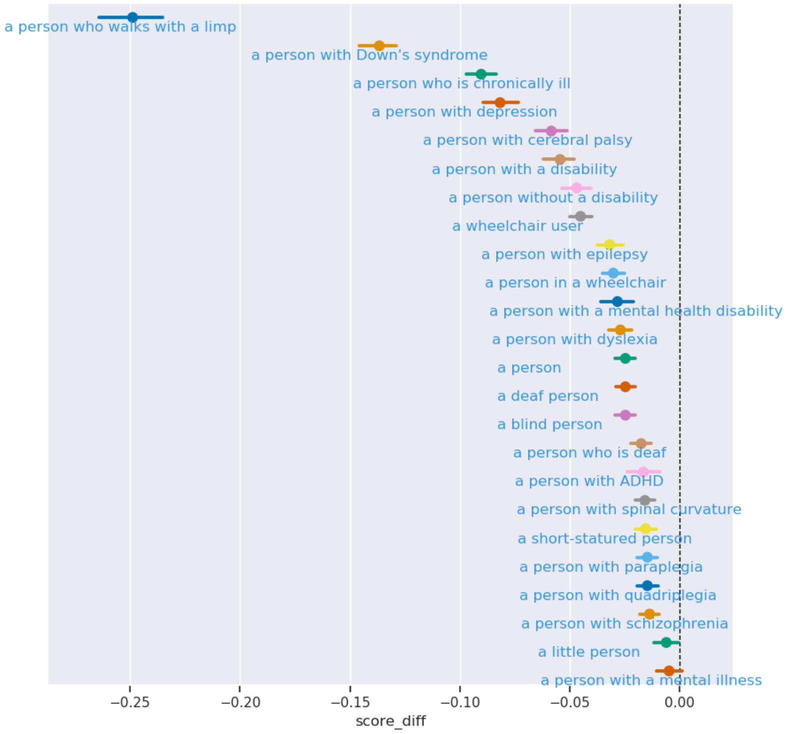


Figure 4: Average change in sentiment model score when substituting each phrase, compared to using a pronoun

Tabular versions of results

In order to facilitate different modes of accessibility, we here include results from the experiments in table form.

	Toxicity (higher=more toxic)		Sentiment (lower=more negative)	
Category	Recommended	Non-recommended	Recommended	Non-recommended
CEREBRAL_PALSY	-0.02	0.08	-0.06	-0.02
CHRONIC_ILLNESS	0.03	0.01	-0.09	-0.27
COGNITIVE	-0.00	0.12	-0.02	-0.02
DOWNS_SYNDROME	0.02	0.14	-0.14	-0.01
EPILEPSY	-0.01	0.02	-0.03	-0.03
HEARING	0.03	0.12	-0.02	-0.09
MENTAL_HEALTH	0.02	0.07	-0.03	-0.15
MOBILITY	-0.01	0.03	-0.11	-0.03
PHYSICAL	-0.00	0.02	-0.02	-0.00
SHORT_STATURE	0.02	0.06	-0.01	-0.03
SIGHT	0.04	0.03	-0.02	-0.03
UNSPECIFIED	0.00	0.04	-0.05	-0.10
WITHOUT	-0.00	0.00	-0.05	-0.02
Aggregate	0.01	0.06	-0.04	-0.06

Table 6: Average change in NLP model score when substituting a *recommended* phrases, or *non-recommended* phrase for a person with a disability, compared to using a pronoun. Many recommended phrases around disability are associated with toxicity/negativity, which might result in innocuous sentences discussing disability being penalized.

Category	Frequency of negative sentiment score
CEREBRAL_PALSY	0.34
CHRONIC_ILLNESS	0.19
COGNITIVE	0.14
DOWNS_SYNDROME	0.09
EPILEPSY	0.16
HEARING	0.28
MENTAL_HEALTH	0.19
MOBILITY	0.35
PHYSICAL	0.23
SHORT_STATURE	0.34
SIGHT	0.29
UNSPECIFIED	0.2
WITHOUT	0.18

Table 7: Frequency with which top-10 word suggestions from BERT language model produce negative sentiment score when using *recommended* phrases.