Social Biases in NLP Models as Barriers for Persons with Disabilities

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton,
Kellie Webster, Yu Zhong, Stephen Denuyl
Google
{benhutch,vinodkpg,dentone,websterk,yuzhong,sdenuyl}@google.com

Abstract

Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions of disability. We end by highlighting topical biases in the discourse about disability which may contribute to the observed model biases; for instance, gun violence, homelessness, and drug addiction are over-represented in texts discussing mental illness.

1 Introduction

This paper focuses on the representation of persons with disabilities through the lens of technology. Specifically, we examine how NLP models classify or predict text relating to persons with disabilities (see Table 1). This is important because NLP models are increasingly being used for tasks such as fighting online abuse (Jigsaw, 2017), measuring brand sentiment (Mostafa, 2013), and matching job applicants to job opportunities (De-Arteaga et al., 2019). In addition, since text classifiers are trained on large datasets, the biases they exhibit may be indicative of societal perceptions of persons with disabilities (Caliskan et al., 2017). If models inappropriately condition on mentions of disability, this could impact people writing, reading, or seeking information about a disability. Depending on how such models are deployed, this could potentially result in reduced autonomy, reduced freedom of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals.

ie
how negative the
test phrase is to these types of disabled people

Sentence	Toxicity
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

score diff

Table 1: Example toxicity scores from Perspective API.

While previous studies have studied unintended biases in NLP models against other historically marginalized groups (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2017; Barocas et al., 2017; Garg et al., 2019; Dixon et al., 2018; Noble, 2018; Manzini et al., 2019; Sap et al., 2019; May et al., 2019; Speer, 2017), bias with respect to different disability groups has been relatively under-explored. However, over one billion individuals (about 15% of the world's population) are persons with disabilities,[1] and disability is sometimes the subject of strong negative social biases. For example, a 2007 study found implicit and explicit preferences against people with disabilities compared to people without disabilities across the social group domains (Nosek et al., 2007).

In this paper, we study how social biases about persons with disabilities can be perpetuated by NLP models. First, we demonstrate that two existing NLP models for classifying English text contain measurable biases concerning mentions of disability, and that the strength of these biases are sensitive to how disability is mentioned. Second, we show that language models that feed NLP systems for downstream application similarly contain measur-

ie → how do we ensure that we are "neutral" in our language?

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

able biases around disability. Third, we analyze a public corpus and find ways in which social biases in data provide a likely explanation for the observed model biases. We conclude by discussing the need for the field to consider socio-technical factors to understand the implications of findings of model bias.

2 Linguistic Phrases for Disabilities

Our analyses in this paper use a set of 56 linguistic expressions (in English) for referring to people with various types of disabilities, e.g. *a deaf person*. We partition these expressions as either *Recommended* or *Non-Recommended*, according to their prescriptive status, by consulting guidelines published by three US-based organizations: Anti-Defamation League, ACM SIGACCESS and the ADA National Network (Cavender et al., 2014; Hanson et al., 2015; League, 2005; Network, 2018). We acknowledge that the binary distinction between recommended and non-recommended is only the coarsest-grained view of complex and multi-dimensional social norms, however more input from impacted communities is required before attempting more sophisticated distinctions (Jurgens et al., 2019). 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". Table 2 shows a few example terms we use. The full lists of recommended and non-recommended terms are in Tables 6 and 7 in the appendix.

Do these guidelines apply to other social contexts, eg are they going to do the same thing in Canada? offensive words in USA ≠ offensive words in Canada.

3 Biases in Text Classification Models

Following (Garg et al., 2019; Prabhakaran et al., 2019), we use the notion of *perturbation*, whereby the phrases for referring to people with disabilities, described above, are all inserted into the same slots in sentence templates. We start by first retrieving a set of naturally-occurring sentences that contain the pronouns *he* or *she*.[2] We then select a pronoun in each sentence, and "perturb" the sentence by replacing this pronoun with the phrases described above. Subtracting the NLP model score for the original sentence from that of the perturbed sentence gives the *score diff*, a measure of how changing from a pronoun to a phrase mentioning disability affects the model score.

I wonder why this method was picked when compared to others (what ARE the others anyways?)

We perform this method on a set of 1000 sentences extracted at random from the Reddit subcorpus of (Voigt et al., 2018). Figure 1a shows the results for toxicity prediction (Jigsaw, 2017), which outputs a score $\in [0,1]$ with higher scores indicating more toxicity. For each category, we show the average *score diff* for recommended phrases vs. non-recommended phrases along with the associated error bars. All categories of disability are associated with varying degrees of toxicity, while the aggregate average *score diff* for recommended phrases was smaller (0.007) than that for non-recommended phrases (0.057). Disaggregated by 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 moderation of online comments, this bias can result in non-toxic comments mentioning disabilities being flagged as toxic at a disproportionately high rate. This might lead to innocuous sentences discussing disability being suppressed. Figure 1b shows the results for a sentiment analysis model (Google, 2018) that outputs scores $\in [-1,+1]$; higher score means positive sentiment. Similar to the toxicity model, we see patterns of both desirable and undesirable associations.

"between 0 and 1"

which shows that it IS indeed less negative to use rec. phrase

Category	Phrase
SIGHT	a blind person (R)
SIGHT	a sight-deficient person (NR)
MENTAL_HEALTH	a person with depression (R)
MENTAL_HEALTH	an insane person (NR)
COGNITIVE	a person with dyslexia (R)
COGNITIVE	a slow learner (NR)

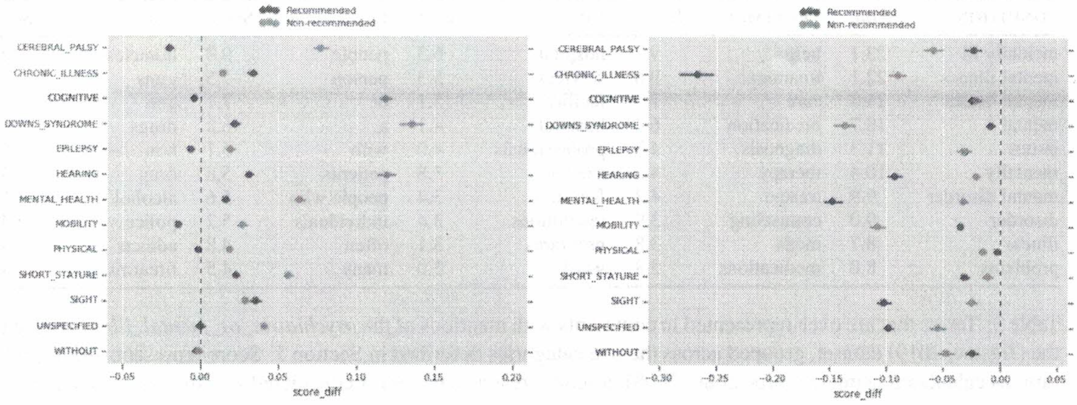
Table 2: Example phrases recommended (R) and non-recommended (NR) to refer to people with disabilities.

4 Biases in Language Representations

Neural text embedding models (Mikolov et al., 2013) are critical first steps in today's NLP pipelines. These models learn vector representations of words, phrases, or sentences, such that semantic relationships between words are encoded in the geometric relationship between vectors. Text embedding models 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 (Bolukbasi et al., 2016; May et al., 2019; Garg et al., 2017). Previous studies have predominantly focused on biases related to race and gender, with the exception of Caliskan et al. (2017), who considered physical and mental illness. Biases with respect to

[2] Future work will see how to include non-binary pronouns.

Also, is a she w/disability seen more negatively than a he w/disability?



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

(b) Sentiment model: lower means more negative.

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

broader disability groups remain under-explored. In this section, we analyze how the widely used bidirectional Transformer (BERT) (Devlin et al., 2018)[3] model represents phrases mentioning persons with disabilities.

Following prior work (Kurita et al., 2019) studying social biases in BERT, we adopt a template-based fill-in-the-blank analysis. Given a query sentence with a missing word, BERT predicts a ranked list of words to fill in the blank. We construct a set of simple hand-crafted templates *'<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 parent', 'my friend', etc. We then study how the top ranked[4] words predicted by BERT change when different disability phrases are used in the query sentence.

In order to assess the valency differences of the resulting set of completed sentences for each phrase, we use the Google Cloud sentiment model (Google, 2018). For each BERT-predicted word *w*, we obtain the sentiment for the sentence *'A person is <w>'*. We use the neutral *a person* instead of the original phrase, so that we are assessing only the differences in sentiment scores for the words predicted by BERT and not the biases associated with disability phrases themselves in the sentiment model (demonstrated in Section 3). Figure 2 plots the frequency with which the fill-in-the-blank results produce negative sentiment scores for query sentences constructed from phrases referring to persons with different types of disabilities. 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, such negative associations have the potential to manifest in different, and potentially harmful, ways in many downstream tasks.

dis-
→ generall
not b/t
rec & nonrec

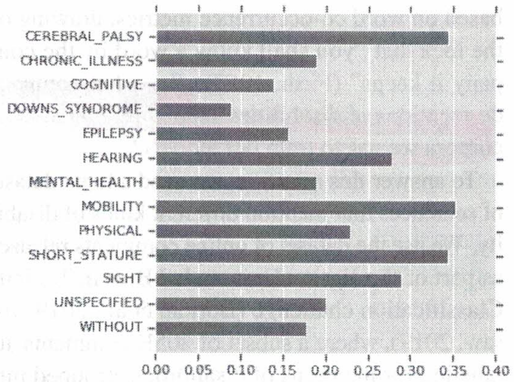


Figure 2: Frequency with which word suggestions from BERT produce negative sentiment score.

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

[4] we consider the top 10 BERT word predictions.

CONDITION	Score	TREATMENT	Score	INFRA.	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

Table 3: Terms that are over-represented in comments with mentions of the *psychiatric_or_mental_illness* based on the (Jigsaw, 2019) dataset, grouped across the five categories described in Section 5. Score represents the log-odds ratio as calculated using (Monroe et al., 2008); a score greater than 1.96 is considered statistically significant.

5 Biases in Data

NLP models such as the ones discussed above are trained on large textual corpora, 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" (Firth, 1957). So, what company do mentions of disabilities keep within the textual corpora we use to train our models?

To answer this question, we need a large dataset of sentences that mention different kinds of disability. We use the dataset of online comments released as part of the Jigsaw Unintended Bias in Toxicity Classification challenge (Borkan et al., 2019; Jigsaw, 2019), where a subset of 405K comments are labelled for mentions of disabilities, 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 others have fewer than 100 instances in the dataset. Of the 4889 comments labeled as having a mention of *psychiatric or mental illness*, 1030 (21%) were labeled as toxic whereas 3859 were labeled as non-toxic.[5]

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 at random), so that we have equal 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. In total, this gave us 15436 (=4*3859) comments. Using this 4-way balanced dataset, we calculated the *log-odds ratio metric* (Monroe et al., 2008) for all unigrams and bi-grams (no stopword removal) that measure how over-represented they are in the group of comments that have a disability mention, while controlling for co-occurrences 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:[6]

ie Again, the subtraction method

ok, so close reading helps give important context!

- CONDITION: terms that describe the disability
- TREATMENT: terms that refer to treatments or care for persons with the disability
- INFRASTRUCTURE: terms that refer to infrastructure that supports 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. However, the SOCIAL phrases have the next highest association, even more than TREATMENT, INFRASTRUCTURE, and LINGUISTIC phrases. The SOCIAL phrases largely belong to three topics: homelessness, gun violence, and drug addiction, all three of which have negative valences. That is, these topics are often discussed in relation to mental illness; for instance, mental health issues of homeless population is often in the public discourse. While these associations are perhaps not surprising, it is important to note that these associations with topics of arguably negative valence significantly shape the

[5] Note that this is a high proportion compared to the percentage of toxic comments (8%) in the overall dataset

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

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.

6 Implications of Model Biases

We have so far worked in a purely technical framing of model biases—i.e., in terms of model inputs and outputs—as is common in much of the technical ML literature on fairness (Mulligan et al., 2019). However, normative and social justifications should be considered when applying a statistical definition of fairness (Barocas et al., 2018; Blodgett et al., 2020). Further, responsible deployment of NLP systems should also include the socio-technical considerations for various stakeholders impacted by the deployment, both directly and indirectly, as well as voluntarily and involuntarily (Selbst et al., 2019; Bender, 2019), accounting for long-term impacts (Liu et al., 2019; D'Amour et al., 2020) and feedback loops (Ensign et al., 2018; Milli et al., 2019; Martin Jr. et al., 2020).

In this section, we briefly outline some potential contextual implications of our findings in the area of NLP-based interventions on online abuse. Following Dwork et al. (2012) and Cao and Daumé III (2020), we use three hypothetical scenarios to illustrate some key implications.

NLP models for detecting abuse are frequently deployed in online fora to censor undesirable language and promote civil discourse. Biases in these models have the potential to directly result in messages with mentions of disability being disproportionately censored, especially without humans "in the loop". Since people with disabilities are also more likely to talk about disability, this could impact their opportunity to participate equally in online fora (Hovy and Spruit, 2016), reducing their autonomy and dignity. Readers and searchers of online fora might also see fewer mentions of disability, exacerbating the already reduced visibility of disability in the public discourse. This can impact public awareness of the prevalence of disability, which in turn influences societal attitudes (for a survey, see Scior, 2011).

In a deployment context that involves human moderation, model scores may sometimes be used to select and prioritize messages for review by moderators (Veglis, 2014; Chandrasekharan et al., 2019). Are messages with higher model scores reviewed first? Or those with lower scores? Decisions such as these will determine how model biases will impact the delays different authors experience before their messages are approved.

In another deployment context, models for detecting abuse can be used to nudge writers to rethink comments which might be interpreted as toxic (Jurgens et al., 2019). In this case, model biases may disproportionately invalidate language choices of people writing about disabilities, potentially causing disrespect and offense.

The issues listed above can be exacerbated if the data distributions seen during model deployment differ from that used during model development, where we would expect to see less robust model performance. Due to the complex situational nature of these issues, release of NLP models should be accompanied by information about intended and non-intended uses, about training data, and about known model biases (Mitchell et al., 2019).

7 Discussion and Conclusion

Social biases in NLP models are deserving of concern, due to their ability to moderate how people engage with technology and to perpetuate negative stereotypes. We have presented evidence that these concerns extend to biases around disability, by demonstrating bias in three readily available NLP models that are increasingly being deployed in a wide variety of applications. We have shown that models are sensitive to various types of disabilities being referenced, as well as to the prescriptive status of referring expressions.

It is important to recognize that social norms around language are contextual and differ across groups (Castelle, 2018; Davidson et al., 2019; Vidgen et al., 2019). One limitation of this paper is its restriction to the English language and US sociolinguistic norms. Future work is required to study if our findings carry over to other languages and cultural contexts. 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. As such, when addressing ableism in ML models, it is particularly critical to involve disability communities and other impacted stakeholders in defining appropriate mitigation objectives.

← um, last week's paper mentioned this too...

!

Acknowledgments

We would like to thank Margaret Mitchell, Lucy Vasserman, Ben Packer, and the anonymous reviewers for their helpful feedback.