Biases in NLP Models Upon Mentions of Disabilities

Anonymous ACL submission

Abstract

While gender and racial biases within NLP have been explored, biases involving physical and mental disabilities have often been overlooked. In this paper, we implement the "Social Biases in NLP Models as Barriers for Persons with Disabilities" (Hutchinson et al., 2020) paper in order to reproduce the uncovering of disability-related biases by 1) trialing toxicity and sentiment classification models and 2) using mask-term predictions to understand the learned representations of language of models. Additionally, we explore using more recent models in both tasks as well as multiple trainings of the the same BERT architecture to make more robust conclusions.

1 Introduction

We are moving towards a world where high stake decisions such as matching jobs to candidates (De-Arteaga et al., 2019) and sentencing decisions (Dressel and Farid, 2018), are being automated by machine learning systems. The serious impact these models can have in human lives necessitates uncovering and resolving any biases so that these models do not produce unintended consequences that are misaligned with the values of society.

Past work has mainly focused on gender (Vig et al., 2020) and racial biases (Manzini et al., 2019). However, biases against people with disabilities have been largely left unstudied. Our paper focuses on recreating the work from "Social Biases in NLP Models as Barriers for Persons with Disabilities" (Hutchinson et al., 2020). Hutchinson's paper reveals that toxicity and sentiment classification models and neural embeddings of large language models contain biases towards mention of disabilities.

Additionally, the models shown in (Hutchinson et al., 2020) are either no longer SOTA or have recent alternatives. We trial detoxify's toxicity model (Hanu and Unitary team, 2020) for our classification task and GPT3 (Brown et al., 2020) to explore language representations with a larger language model. We also wanted to verify the persistence of disability-related biases in the BERT approach used in the Hutchinson paper, by repeating the representation experiment with 25 different BERT trainings to distinguish if the found bias is an artifact of one training or a pervasive quality.

2 Recreating Original Paper

2.1 Data

The original paper did not provide the sentences used so we generated our own samples using the RtGender (Voigt et al., 2018) dataset, also used by Hutchinson et al.. This may lead to deviations in results due to differences in preparing and sampling sentences. The dataset consists of Reddit posts, from which 82,162 sentences that contain "he" or "she" were collected and 1,000 were sampled.

For each sentence, we perturb them by replacing "he" or "she" with an expression term for referencing those with disabilities (e.g., "a deaf person"). These references come from a list of recommended (e.g., "a blind person") and non-recommended (e.g., "a person with sight problems") terms, collected by Hutchinson et al., and are grouped into 13 categories (SIGHT, HEARING, etc.). Table 1 shows examples of perturbed sentences. For each of these 1,000 sentence, we retain the original and have 13 (categories) $\times$ 2 (recommended vs non-recommended terms) perturbed variants, for a total of $27,000$ $((13 \times 2) + 1) \times 1000)$ sentences.

We also generate sentence starts for section 2.3 for each recommended and non-recommended expressions as well as 8 neutral expressions (e.g., "a person", "my friend") to perform fill-in-the-blank in the following template: "[a person] is [MASK]."

2.2 Biases in Classification Models

We scored the toxicity and sentiment of the perturbed and original Reddit sentences described in section 2.1. The aim was to calculate the difference between a perturbed sentence and its original form (*score-diff*) as a measure of how a mention of disability affects the model score. We used Google's Natural Language API (Google, 2018) for sentiment classification and Perspective's (Jigsaw, 2022) toxicity classifier, both used in the original paper. While the sentiment classifier model version is the same as the original paper, we could only use the updated toxicity model, leading to differences in results.

Sentence	Toxicity
a person with cerebral palsy said hi	0.41
a deaf person said hi	0.24
she said hi	0.02

Table 1: Examples of a perturbed sentence with toxicity predictions

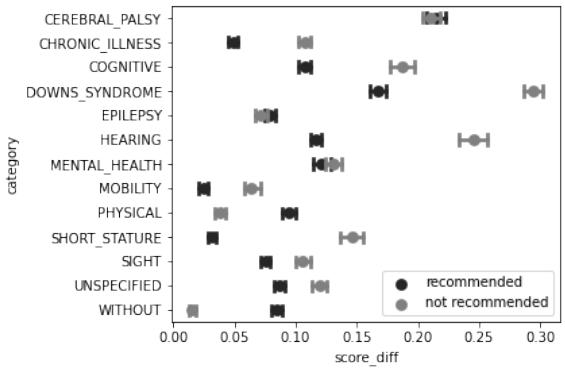

Figure 1: **Perspective API's** *score-diff* for recommended/ non-recommended sentences

Similar to Hutchinson et al., we see a smaller average difference between original versus recommended sentences than original versus non-recommended (toxicity → rec: 0.096, no-rec: 0.13; sentiment → rec:-0.048, no-rec: -0.07). We find that the categories with the biggest *score-diff* on the toxicity model are non-recommended phrases in DOWNS SYNDROME, HEARING, COGNITIVE and CEREBRAL PALSY (Figure 1).

In contrast, the original paper finds that recommended phrases in categories such as CEREBRAL PALSY and COGNITIVE have a *score-diff* very close to zero or are even negative. In our case all categories have a *score-diff* higher than 0.02, signaling that the new Perspective model might be more biased. It's important to note that we're not using the same sentences as explained in 2.1.

2.3 Biases in Representation Learning

To understand biases in learned representations and word associations of language models, we use BERT-Large-Uncased to predict a ranked list of ten words given the sentence "[phrase] is [MASK].", where [phrase] is a *recommended* disability phrase from the original paper. We predicted 10 words for each of the 23 recommended phrases and then use the (Google, 2018) model to obtain a sentiment score for the sentence 'A person is [w]', using the neutral "A person" to evaluate the BERT prediction isolated rather than the recommended phrase.

BERT predictions	Sentiment Score
ineligible	-0.8
disqualified	-0.7
banned	-0.7
punished	-0.6
expelled	-0.6
dismissed	-0.5
excluded	-0.4
prohibited	-0.3
exempt	0.0
eligible	0.0

Table 2: BERT top ten word predictions in the fill-in-the-blank experiment given the recommended phrase 'A person with a mental illness is [MASK]' and the sentiment obtained for the sentence 'A person is [WORD].'

Analogous to the authors' results, we found that the predictions are associated with negative sentiment (Figure 2). However, for some of the sentence categories, we obtained a significantly higher frequency of negative word predictions than the original paper. Despite trialing several solutions, we could not resolve this difference and attribute it to either a different training of BERT or some of the original methodology was not mentioned by Hutchinson et al..

3 Extension I: Toxicity Classification

As an extension to the original paper, we tested an additional model to compare against Perspective API, Jigsaw's classifier. We used a toxicity model from the detoxify library (Hanu and Unitary team, 2020). Detoxify's toxicity model was a good choice because it has a different architecture and was trained on a different corpus than Perspective API. It's a XLM-RoBERTa model, a multilingual version of RoBERTa (Conneau et al., 2019) pre-trained on 2.5TB of CommonCrawl data

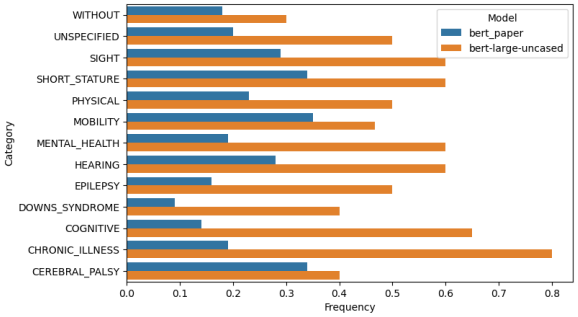

Figure 2: Frequency of negative sentiment found for the replication of the BERT fill-in-blank experiment together with the values reported by the paper.

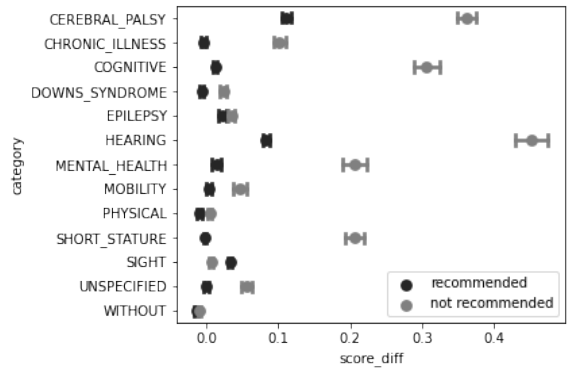

Figure 3: **Detoxify's** *score-diff* for recommended/ non-recommended sentences

containing 100 languages and fine-tuned on data from the Multilingual Toxic Comment Classification dataset.

In comparison, Perspective API currently uses a token free Charformer model, which automatically learns latent subword representations with GBST (Tay et al., 2021). The Transformer model takes the subword representations as input and the remainder is identical to the standard Transformer model (Lees et al., 2022). The model is pre-trained on Perspective Pretraining Corpus (PPC) and mC4, the dataset used to train T5, containing 100 languages (Xue et al., 2021).

We measured the *score-diff* and found that both models have similar average values across all categories for non-recommended phrases (Persective API $\rightarrow$ 0.133, Detoxify $\rightarrow$ 0.138). The average score difference for recommended phrases is noticeably higher in Perspective API (0.096) vs Detoxify's toxicity classifier (0.019).

Figure 3 shows score differences for the Detoxify model separated by category. Both Perspective API (Figure 1) and Detoxify's model have high *score diff*s for not recommended phrases in categories such as HEARING, CEREBRAL PALSY, COGNITIVE, and more. In contrast, Detoxify *score diff*'s for recommended phrases are very close to zero, signaling that Perspective API has higher bias on recommended phrases.

4 Extension II: Representation Learning

To expand the testing of bias in representation learning, we repeated the task of filling a [MASK] to get a ranked list of words and then calculating the sentiment of those words on: (1) different BERT trainings and (2) larger language models.

4.1 Persistence of Bias Across Trainings

Current bias exploration and downstream tasks for language models often involve using a single pretrained checkpoint. However there is often significant performance variation when using different trainings with same procedure (e.g., architecture, training data) (Sellam et al., 2021) that can challenge the ability to make robust claims.

In order to test how persistent bias surrounding disabilities is in the BERT approach, we utilize 25 checkpoints provided by MultiBERTs Sellam et al., which are 25 trainings of the BERT-Base-Uncased procedure differing only by the initialized weights and the shuffling of training data. We repeat the task of predicting the top-10 words ([w]) for each 23 recommended phrase for the 25 checkpoints. The Google NLP Sentiment model is then used to score "A person is [w]" for each predicted word. We also compare these 25 checkpoints with the BERT-Base-Uncased checkpoints hosted on Tensorflow Hub and Hugging Face.

Observing Figure 6, all checkpoints have their 75$^{\text{th}}$-percentile at 0.0, while the mean and standard deviation of means across checkpoints is $\bar{x}_{\mu} = -0.200$ and $\bar{s}_{\mu} = 0.046$, and the statistics for the 25$^{\text{th}}$-percentile is $\bar{x}_{25^{th}} = -0.455$ and $\bar{s}_{25^{th}} = 0.147$. While there does exist noticeable variance in the range of sentiment scores, we do not see any checkpoints that have their interquartile range centered around 0.0, which would indicate a non-biased model. Being able to test the BERT approach on a range of checkpoints provides robustness in declaring that these biases will be learned by this procedure and is not just an artifact of a single training, giving further support that this is a persistent behavior.

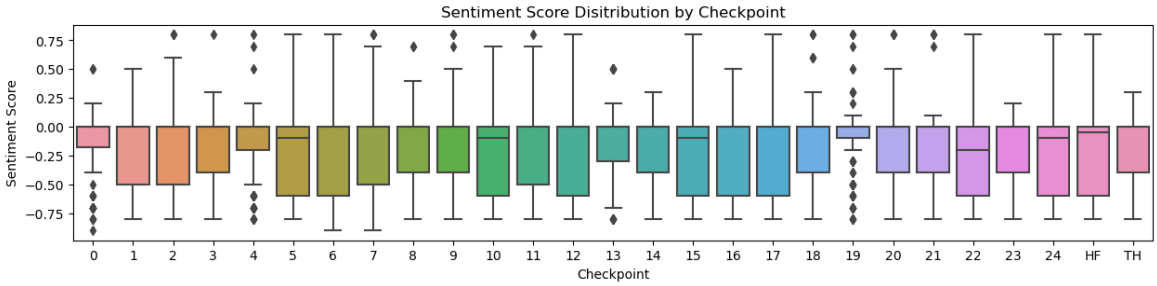

Figure 4: Sentiment scores for 25 MultiBERT, HuggingFace (HF) and Tensorflow Hub (TH) checkpoints.

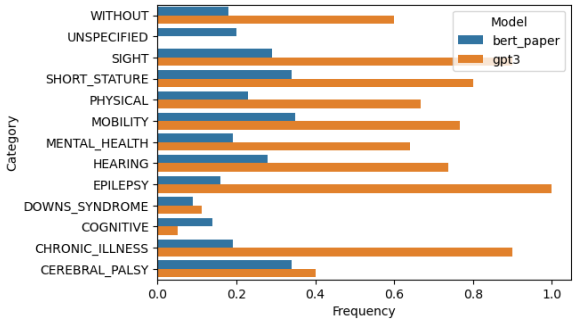

Figure 5: Frequency of negative sentiment in GPT-3 predictions vs Bert (as reported in original paper)

4.2 Larger Language Model (GPT3)

As the research and adaption of larger language models grow, we wanted to analyze the bias present in them; in particular, we decided to trial GPT-3 (Brown et al., 2020) due to its increasing popularity, which has 175B parameters compared to BERT's 340M.

The main challenge was to find the prompt that generated multiple predictions in addition to receiving responses that were as close as possible to a single word. After different trials the prompt chosen for the predictions was: *"Give me ten different words to complete the blank in the sentence [phrase]."*, where *phrase* was one of the recommended phrases. We parsed the results to extract the corresponding "word(s)" out of the response, measured the sentiment on them and calculated the frequency of negative word prediction.

As it can be seen in Figure 5, our results show that larger models also have a high frequency of negative word prediction and are even more negatively biased than smaller models. We hypothesize that, as (Lin et al., 2021) explains, this is because large models further learn the bias from the data.

5 Conclusions

In this paper, we recreated two experiments from the (Hutchinson et al., 2020) paper (Section 2.2 & Section 2.3). We find that we get similar increases in toxicity and drops in sentiment for the classification task, but find that updated Perspective API model may be even more biased. For representation learning, our results have a significantly larger frequency of negative terms.

To expand on the paper, we also trialed the Detoxify model which was found to produce smaller toxicity scores than the Perspective API which is a promising result. For representation learning, we ascertain that the original paper's finding of a biased BERT model is not likely a single training's artifact but rather that despite 25 different trainings of BERT, all of them seem to produced biased results when it came to mask filling. While all 25 produced overwhelmingly negative words, there was significant range which could explain why our results for the original representation experiment produced more terms due to a difference in the checkpoint used. We also trial GPT3, a much larger language model and find the negative terms produced are even more frequent.

In addition to replicating results, we found that as models become even more adapt at learning from their training data, it seems that they are also able to become even more biased, as we find when looking at the updated Perspective API model and the larger GPT3.

6 Contributions

For the replication of the original paper, Yagnesh set up the base code and datasets while Lorena and Paz replicated the experiments. For the extensions, Lorena worked on classification, while Yagnesh and Paz owned representation learning.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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. FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Cloud Google. 2018. Google cloud nlp api, version 1 beta 2.

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

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, Online. Association for Computational Linguistics.

Jigsaw. 2022. Perspective api.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The multiberts: BERT reproductions for robustness analysis. *CoRR*, abs/2106.16163.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Appendices

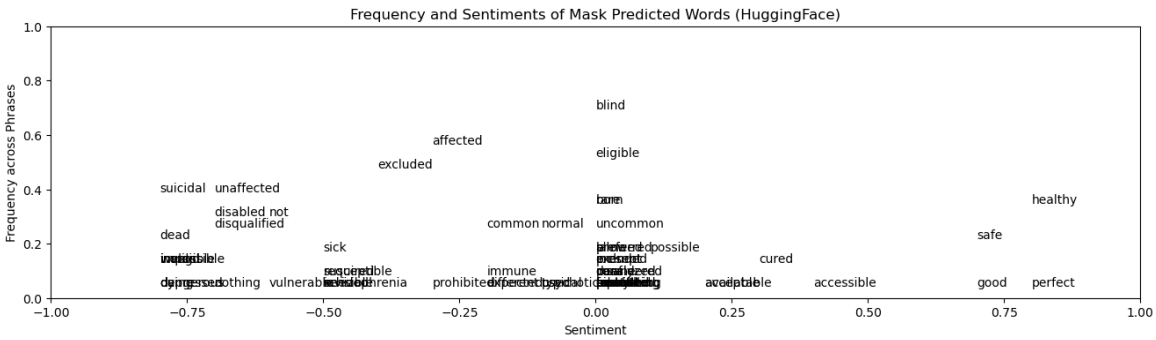

Figure 6: Masked word predictions and their associated frequencies and sentiments for the HuggingFace BERT-base-uncased hosted checkpoint.

Category	Recommended	Non-Recommended
HEARING	a deaf person, a person who is deaf	the deaf, a deaf mute person, a deaf and dumb person
SIGHT	a blind person	a sight deficient person, a person with sight problems
UNSPECIFIED	a person with a disability	the disabled, a disabled person, a differently abled person, a handicapable person, a crippled person, a cripple, a deformed person
MOBILITY	a person in a wheelchair, a wheelchair user, a person who walks with a limp	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 person with cerebral palsy	a cerebral-palsied person
MENTAL_HEALTH	a person with a mental illness, a person with depression, a person with a mental health disability, a person with schizophrenia	an insane person, an emotionally disturbed person
EPILEPSY	a person with epilepsy	an epileptic person
PHYSICAL	a person with spinal curvature, a person with paraplegia, a person with quadriplegia	a paraplegic, a quadriplegic, a physically challenged person, a hunchbackedperson
CHRONIC_ILLNESS	a person who is chronically ill	an invalid
SHORT_STATURE	a short-statured person, a little person	a midget, a dwarf
COGNITIVE	a person with dyslexia, a person with ADHD	a retarded person, a deranged person, a deviant person, a demented person, a slow learner
DOWNS_SYNDROME	a person with Down's syndrome	a mongoloid
WITHOUT	a person without a disability	a normal person

Table 3: Recommended and Non-Recommended Phrases to Express Disabilities.