Biases in NLP Models Upon Mention of Disabilities

NYU Center for Data Science

Yagnesh Patel, Lorena Piedras, Paz Vives

Introduction

This project focuses on identifying the biases of disability-related language in NLP models, following the work of *Social Biases in NLP Models as Barriers for Persons with Disabilities*[1]. While much has been explored on the gender and racial biases of NLP models, little focus has been on disabilities, such as blindness and mental illnesses[1]. The urgency to identify and resolve these biases is ever-growing as machine learning models make impactful decisions such as job matching[2] and court sentencings[3], and can make decisions that are misaligned with society's values.

We approach and expand two experiments from the original paper. The first is **biases in representation learning**, which explores how words are represented internally in a model and what terms they get associated with, which is explored by performing a BERT mask-filling task. We expand upon this by trialing a larger language model, GPT3[4], and also try a series of 25 BERT trainings[5] to see how persistent these biases are.

The second experiment is **biases in classification models**. In the original paper, the Google Sentiment[6] model and the Perspective API[7] toxicity model were trialed to see how perturbations in sentences affect their scores. We re-implement that process and also expand on it by trialing a XLM-RoBERTa based toxicity model created by Detoxify[8].

Data

- 13 Categories of **phrases** to refer to those with disabilities[1]
- Categories split into Recommended ("a blind person") and Non-Recommended ("a person with sight problems")
- 1,000 sentences from the RTGender[9] dataset containing "he" or "she"
- Replaced instances of "he" or "she" with the phrases to measure differences in scores
- Used phrases to create templates for mask-filling

References

[1] Social Biases in {NLP} Models as Barriers for Persons with Disabilities., Hutchinson et al, 2021
[2] Bias in bios: A case study of semantic representation bias in a high-stakes setting., De-Arteaga et al., 2019
[3] The accuracy, fair- ness, and limits of predicting recidivism., Dressel & Farid., 2018
[4] Language Models are Few-Shot Learners, Brown et al, 2020
[5] The MultiBerts: Bert Reproduction for robustness analysis, Thibault Sellam et al, 2021
[6] Google's Natural Language API, Google: https://cloud.google.com/natural-language/
[7] Perspective API, Jigsaw, 2022: https://www.perspectiveapi.com/
[8] Detoxify, Hanu and Unitary Team, 2020: https://github.com/unitaryai/detoxify
[9] RtGender: A corpus for studying differential responses 466 to gender. Voigt et al, 2018.

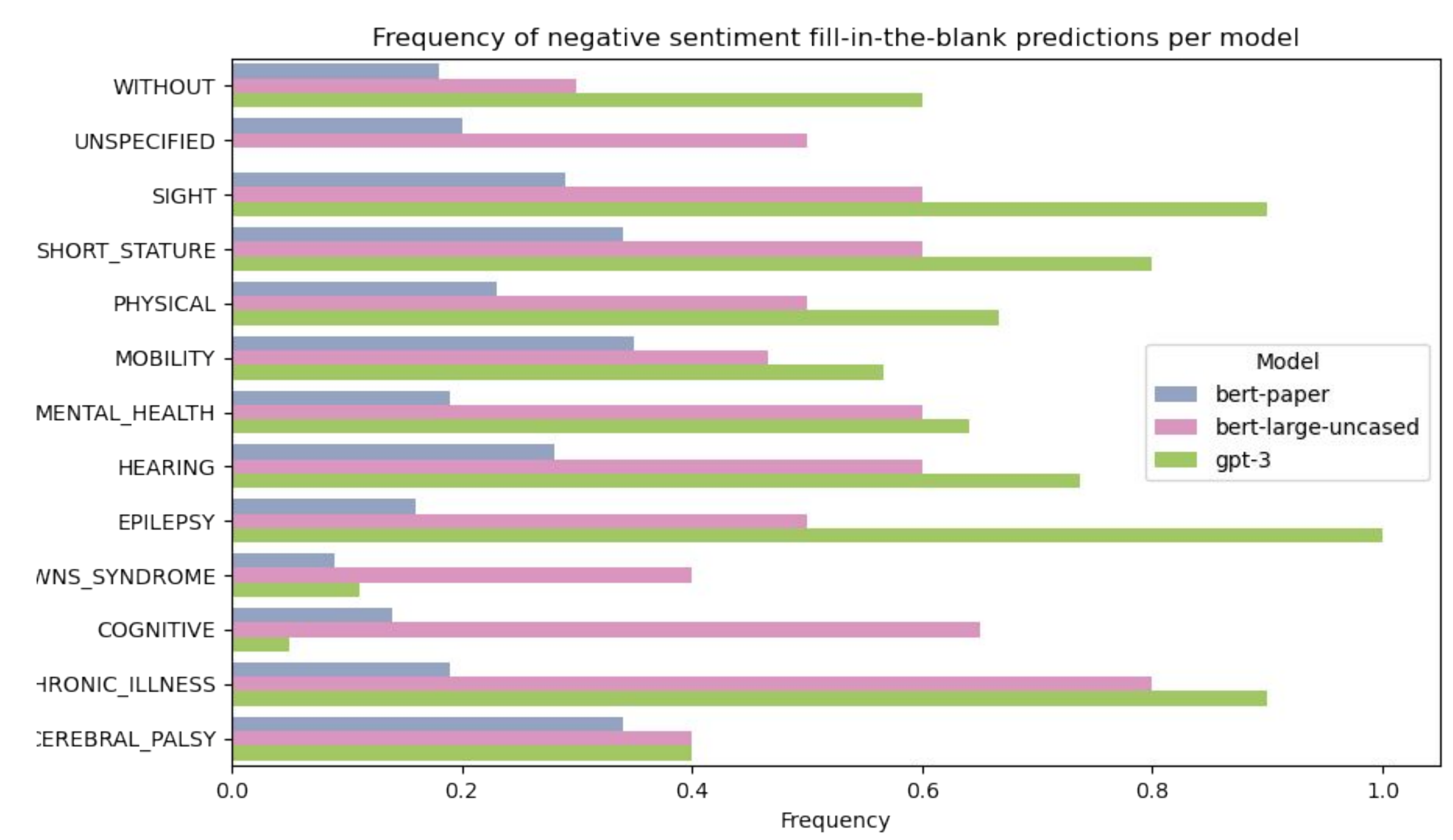
Biases in Representation Learning

Experiment

To test what representations were learned for disability-related words, we perform a mask-filling task where we start with '[PHRASE] is [MASK]', then replace [PHRASE] with a term from the recommended expressions (e.g., 'a wheelchair user'), and then use a language model (BERT and GPT-3) to predict the top 10 words for the [MASK]. **We then take those mask-predictions and get a sentiment score** for the phrase "A person is [WORD]." for each predicted [WORD] to see the sentiment of the words predicted.

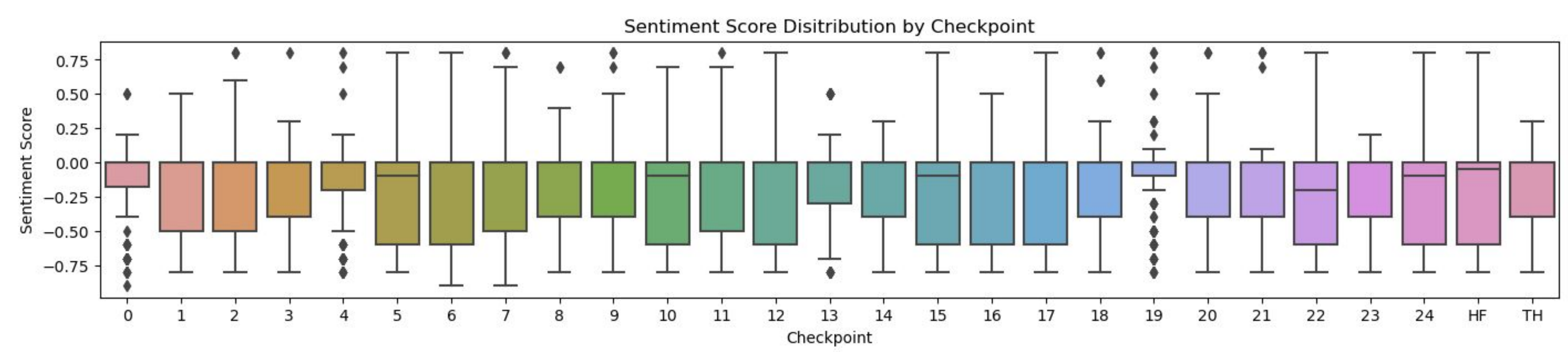
BERT predictions	Sentiment Score
ineligible	-0.8
disqualified	-0.7
banned	-0.7
punished	-0.6
expelled	-0.6
dismissed	-0.5
excluded	-0.4
prohibited	-0.3
exempt	0.0
eligible	0.0

Table 1: Top 10 Word Mask Predictions using BERT for the phrase: "A person with a mental illness is [MASK]"



MultiBERT

When using language models, it is common to use a single existing trained checkpoint. However, using a single training can make it hard to differentiate behaviors of the overall approach vs. what a single training learned. To get around this, **we utilized 25 different checkpoints of BERT-base-uncased from MultiBERT[5] in the the mask filling and sentiment scoring task.**



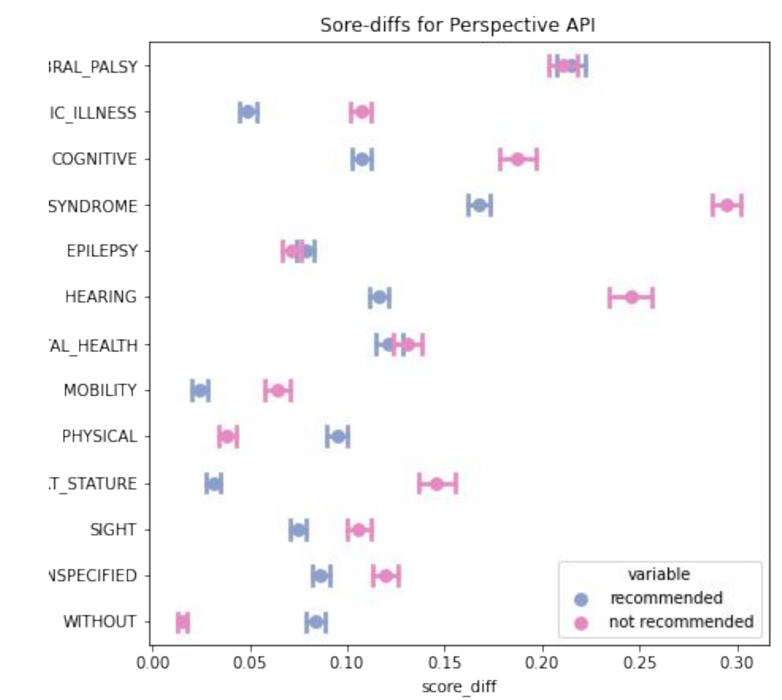
Results:

- Analogous to original experiment, we found **BERT's predictions are associated with negative sentiment.**
- As hypothesized, **large models** (such as GPT-2) also **have a high frequency of negative word predictions**, even larger than smaller models like BERT.
- **When using the 25 trainings** of the same BERT approach, while there is a noticeable difference in the sentiment ranges, ALL interquartile ranges are negative, showing **persistent bias learning**.

Biases in Classification

We evaluate bias in toxicity models by comparing scores from original Reddit sentences with sentences that perturb personal pronouns for mentions of disabilities.

Perspective API



Takeaways

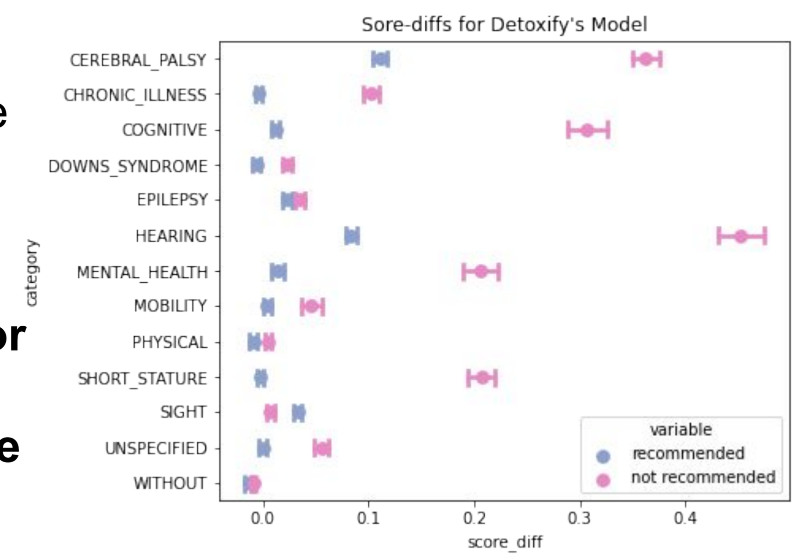
Similar to [1], we see a smaller average *score-diff* between original versus recommended sentences than original versus non-recommended.
In contrast, the original paper has some recommended categories with *score-diffs* close to zero or negative and **all of our categories have *score-diffs* > 0.02. Signaling more bias.**

Table 2: Toxicity scores for perturbations and original sentence

Category	Sentence	Toxicity
Original	**she** said hi	0.02
Recommended	**a deaf person** said hi	0.24
Non-Recommended	**a cerebral-palsied person** said hi	0.4

Detoxify's toxicity model

Both models have similar average *score-diff* values across all categories for non-recommended phrases.
The average score difference for recommended phrases is noticeably higher in Perspective API (0.096) vs Detoxify's toxicity classifier (0.019).



Conclusion

- The updated Perspective model may be more biased than the past one and Detoxify's model seems less biased than Perspective.
- For representation learning, our results have a significantly larger frequency of negative terms than [1].
- We ascertain that the original paper's finding of a biased BERT model is not likely related to a single training's artifact .
- As models become more adept at learning from their training data, they may also become more biased (GPT-3)