# NLP Progress Report

Anonymous ACL submission

In the past weeks, we've worked on recreating the (Hutchinson et al., 2020) paper, generating data samples, and setting up pipelines for additional experiments.

## 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 of these 1,000 sentences, 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.). For each 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 the "Biases in Representation Learning" 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 Biases in Classification Models

We scored toxicity and sentiment of the perturbed and original Reddit sentences described in section 1. The aim was to calculate the difference between a perturbed sentence and its original form as a measure of how a mention of disability affects the model score. We used Google's Natural Language API for sentiment classification (Google, 2018) and Jigsaw's toxicity classifier (Jigsaw, 2022), both models are used in the original paper. The model version of the sentiment classifier is the same as the original paper and we used a newer toxicity model, this led to differences in results.

Similar to the reported results, we see a smaller average difference between original sentences versus recommended than original versus non-recommended (toxicity $\rightarrow$ rec: 0.096, no-rec: 0.13; sentiment $\rightarrow$ rec:-0.048, no-rec: -0.07). Patterns differ within each category and we do see undesirable associations, for example, recommended phrases in the mental health category have a sentiment difference of -0.2 with original sentences.

## 3 Biases in Representation Learning

We used BERT to predict a ranked list of words given the sentence "[phrase] is [MASK].", where [phrase] is a *recommended* disability phrase from the original paper. We took the top ten predictions for each recommended phrase and for each predicted word *w*, we used the Google model (Google, 2018) to obtain the sentiment for the sentence 'A person is [w]', using the neutral "A person" to evaluate the sentiment score of the BERT prediction rather than the recommended phrase.

Analogous to the authors' results, we found that the model predictions are associated with negative sentiment. For some of the sentence categories, we obtained a significantly higher frequency of negative word predictions than the original paper and we think this could be due to the difference in the originating dataset.

Nevertheless, our results confirm that BERT associates phrases referencing persons with disabilities with words with negative predictions as the original paper states.

## References

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

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.

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).