NLP Project Proposal

Anonymous ACL submission

We seek to replicate the (Hutchinson et al., 2020) paper, which uncovers bias in mentions of disability language within NLP models, and explore extensions by trialing alternative models.

1 Summary of Paper

The (Hutchinson et al., 2020) paper uncovers biases concerning the mentioning of disabilities in classification and representation learning.

In the classification task, the authors focus on sentiment analysis and toxicity detection. They perturb 1,000 sentences (Voigt et al., 2018) by replacing the pronouns ("he", "she") with an expression referring to a person with disabilities (e.g., "blind person"). These altered sentences are then compared to the original in predicted toxicity (Jigsaw, 2022) and sentiment scores (Google, 2018) and the altered are found to score more toxic and negative.

In order to test biases on representation learning, the authors focus on BERT. They perform a fill-in-the-blank analysis to predict a ranked list of words given "[phrase] is ___", where [phrase] is either a phrase referring to person with a disability or a base case such as "a person" or "my parent". It is found that the top predicted words when [phrase] contains disability language results in lower sentiment scores overall.

The final analysis explores the underlying biases within data. This section is out of the scope of this project.

2 Extensions, Motivation and Division of Labor

Extensions After reproducing two of the analyses in the paper: biases in classification and biases in large language models; we will work on three extensions. First, we will use additional toxicity models such as toxic-bert (Hanu and Unitary team, 2020) to compare against Perspective API (Jigsaw, 2022), which was used in the original paper. A limitation is that a new version of Perspective API replaced the original in early 2022 (Lees et al., 2022), preventing a full recreation.

In our second extension, we will focus on biases in language representation, by trialing different BERT-like models, such as DeBERTa, to explore the prevalence of bias. Additionally, (D'Amour et al., 2020) shows that models with the same architecture can report substantial differences in performance across multiple checkpoints. To explore this issue we will use MultiBERT (Sellam et al., 2021).

Finally, we want to test if larger models such as GPT-3 contain the same biases as smaller models like BERT.

Motivation: Since the original paper was published in 2020, new and more accurate classification and language models have been developed. The purpose of our extensions are to analyze if these models hold the similar biases against mentions of disabilities. Our hypothesis is that new models are trained with more data where mentions of disabilities are portrayed negatively so the new models will likely be more biased.

Computational Feasibility: We expect that this effort will be able to run on our personal computers. While we will be using "large" models, we will either use available APIs or use pre-trained models that can be downloaded with limited storage requirements. Additionally, given that the original paper only used sample records in the thousands, we expect our analysis to be of similar magnitude.

Division of Labor: We are splitting the initial tasks of setting up a base code and preparing the data evenly. Afterwards we will divide the main experiments and extensions as follows: Lorena will reproduce the two experiments of the original paper, Yagnesh will work on the first extension on classification models, and Paz will analyze biases in language representation.

References

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*.

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.

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.

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).