Gender Bias Correction in NLP

Layla Shalabi*
Razi Khawaja*
lshalabi@hawk.iit.edu
rkhawaja@hawk.iit.edu
Illinois Institute of Technology
Chicago, Illinois, USA

ABSTRACT

Natural language processing (NLP) generally involves extensive data preparation since natural language can become ambiguous outside of its native context. Our project aims to find and correct gender biased instances of text by manipulating embeddings or datasets by some means. This report outlines our current progress thus far, notably a literature review and a code review.

CCS CONCEPTS

• **Software and its engineering**; • **Natural language processing**;

KEYWORDS

Natural, Language, Processing, NLP, Gender, Bias

1 INTRODUCTION

The aim of this report is to report our progress for our CS 577 - Deep Learning project. The team consists of two members: Layla Shalabi and Razi Khawaja. The goal of the project is to detect gender bias in natural language inputs and datasets and offer corrections as output. Our primary approach to achieve this will involve tuning parameters for word embedding software to affect downstream tasks.

2 PROBLEM STATEMENT

The characteristic ambiguity of natural language makes robust definitions of bias hard to define. Biases can exist in the domains of sex, gender, age, race, ethnicity, and more either independently or at the same time. The objectives of this project will be programmatically diagnosing gender bias, programmatically identifying gender bias, and then programmatically correcting gender bias in text databases such that downstream tasks have reduced gender bias.

3 LITERATURE REVIEW

3.1 Overview

Literature review for this project consisted of five research papers in various states of review. The first and second were a survey and literature review of gender bias in NLP [5, 6]. The third was a paper on quantifying gender bias [3]. The last two were close to model studies that we are considering replicating or developing since they involve manipulating datasets to reduce gender bias in output text [4, 7].

3.2 *A Survey on Gender Bias in Natural Language Processing , Mitigating Gender Bias in Natural Language Processing: Literature Review*

This meta review analyzed work on gender bias in the NLP space. Many of the papers within our literature review were originally cited in this paper. The paper summarizes other papers within NLP relating to gender bias and offers several mathematical metrics to measure the biasedness of model outputs. Depending on what data we decide to use, the metrics offered in these papers can be applicable.

Both of these papers cite more resources that can be reasonably analyzed by two people over the course of a semester. Utilizing the resources in these papers will come down to finding a robust mathematical formula alongside documentation for why we're using it within our own research. They have already proved fruitful in mentioning the last three upcoming papers.

3.3 *Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community*

This paper quantifies gender bias using a corpus of online fiction written mostly by amateur authors. The authors use NLP quantification techniques to find gendered language and any gendered bias within the corpus. While interesting, the intent of the papers was not to find and then remove biased language; the intent of the paper was just to find the bias. Insomuch as the aim was to just find the bias, there are yet more quantification techniques present within the paper that'll be useful later. Regardless of its direct impact in the final project, it was worth the read regardless.

3.4 *Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies , Learning Gender-Neutral Word Embeddings*

These last two papers are the ones that are most closely related to our objective. In Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies, the authors developed techniques to quantify bias in Bollywood scripts. Many of their techniques involve graphical representations of data. Their paper ends with a relational graph where the nodes of the graph are nodes and show the relationship between nodes; bias was mitigated by their data manipulation techniques per their metrics.

In Learning Gender-Neutral Word Embeddings, the authors lay out several equations that are created from the dataset which act

*Both authors contributed equally to this research.

as, effectively, hyper parameters that reduce bias metrics for downstream tasks. They did this by scaffolding their math off of a standardized embedder. Embedder manipulation or dataset manipulation is the most viable route of action to have an impact with our paper.

4 CODE REVIEW

4.1 *Multi-Dimensional Gender Bias Classification Dataset*

The Multi-Dimensional Gender Bias Classification dataset is based on a general framework that decomposes gender bias in text along several pragmatic and semantic dimensions: bias from the gender of the person being spoken about, bias from the gender of the person being spoken to, and bias from the gender of the speaker [6]. It contains seven large scale datasets automatically annotated for gender information (there are eight in the original project but the Wikipedia set is not included in the HuggingFace distribution), one crowdsourced evaluation benchmark of utterance-level gender rewrites, a list of gendered names, and a list of gendered words in English.

We found this dataset to be the most closely aligned with our main objective as it is perfectly set up for applications such as controlling for gender bias in generative models, detecting gender bias in arbitrary text, and classifying text as offensive based on its genderedness. However, it is imperative to note some considerations that we are taking into account in using this dataset.

In regards to discussions of bias, over two thirds of annotators identified as men, which may introduce biases into the dataset. Wikipedia is also well known to have gender bias in equity of biographical coverage and lexical bias in noun references to women. More generally, it should be noted that while the limitations of the Multi-Dimensional Gender Bias Classification dataset have not yet been investigated, we as researchers, as well as the curators, must acknowledge that more work is required to address the intersectionality of gender identities, i.e., when gender non-additively interacts with other identity characteristics. The curators point out that negative gender stereotyping is known to be alternatively weakened or reinforced by the presence of social attributes like dialect, class and race and that these differences have been found to affect gender classification in images and sentence encoders.

4.2 *BUG Dataset*

The BUG dataset [1] was collected semi-automatically from different real-world corpora, designed to be challenging in terms of societal gender role assignments for machine translation and coreference resolution. It is based on the findings of A Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation [2], Findings of EMNLP 2021. The full dataset contains 105,687 sentences with a human entity, identified by their profession and a gendered pronoun.

The relevance of this dataset is straightforward as it allows for a clear metric of coreferential bias. What is preferential about a dataset such as this one is that it allows for a rather straightforward solution to debiasing sentences that are stereotypically biased. That is, replacing the stereotypical pronoun with one that is gender-neutral or anti-stereotypical.

5 MILESTONES

5.1 Code Review ∼ 1 Week

We realistically have about a week to decide how we're going to best address our topic of gender bias in NLP. We are leaning heavily towards dataset or embedding manipulation because it is straightforward and can be done on multiple datasets to test its effectiveness. This will give us something to present and offer opportunities for comparison.

5.2 Presentation ∼ 1 Week

We'll have to get something together for this but it'll be a short presentation. We will either present preliminary results or present our plan to create results.

5.3 Final Report ∼ 3 Week

At this point, code ought to be running on our computers regardless of what happens in our plan before this point or we'll fail the project. We will spend the remaining time tuning the parameters and compiling results.

REFERENCES

[1] [n. d.]. *BUG Dataset*. https://github.com/SLAB-NLP/BUG original-date: 2021-08-28T09:53:48Z.

[2] [n. d.]. *Multi-Dimensional Gender Bias Classification*. https://huggingface.co/datasets/md_gender_bias/blob/main/README.md

[3] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. [n. d.]. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. https://doi.org/10.48550/arXiv.1603.08832 arXiv:1603.08832 [cs]

[4] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. [n. d.]. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018-01-21). PMLR, 92–105. https://proceedings.mlr.press/v81/madaan18a.html ISSN: 2640-3498.

[5] Karolina Stanczak and Isabelle Augenstein. [n. d.]. A Survey on Gender Bias in Natural Language Processing. arXiv:2112.14168 [cs] http://arxiv.org/abs/2112.14168

[6] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. [n. d.]. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, 2019-07). Association for Computational Linguistics, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[7] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. [n. d.]. Learning Gender-Neutral Word Embeddings. https://doi.org/10.48550/arXiv.1809.01496 arXiv:1809.01496 [cs, stat]

Received 1 November 2023; accepted NA