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 quantify and visualize gender bias in different word embedding spaces.

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 embedding spaces and measure the effectiveness of different embedders.

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.

To that end, we analyzed two word embedders. Glove and Gn-Glove [5] , [4] . Glove is a word embedder that aims to create usable and numerical training sets called vector embeddings that are used in many NLP applications. Gn-Glove, short for gender neutral glove, is an embedder that was created to minimize the gendered implications of the statistical approach to creating vector embeddings.

There are many words that are statistically gendered but not essentially gendered and this is the motivation behind creating an explicitly gender neutral word embedder. Consider the occupation "police officer;" nothing about the role of a police officer is necessarily gendered, however, an accurate if sociologically naive would find that more often than not a police officer is a male. Similar biases can be found in occupations like nurse and doctor or in the application of compliments like handsome or beautiful.

Our project aims to probe the embeddings space of Glove and Gn-Glove to see if there are notable differences between the embeddings. To that end, we studied the literature in the field, found four corpi, created gendered pair word list and neutral-word list, and analyzed the results numerically.

*Both authors contributed equally to this research.

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 [10], [11]. The third was a paper on quantifying gender bias [6]. 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 [8], [12].

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 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 CORPUS REVIEW

4.1 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 [7], 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.

4.2 GAP Dataset

The GAP dataset was created to analyze a language model's capability in coreference resolution [2]. Coreference resolution is the ability of a model to complete the task of finding all expressions that refer to the same entity in a text. Since this corpus is made for coreference resolution, it is appropriate for the task of reducing gender embeddings since there will be ambiguities in the dataset.

4.3 Project Gutenberg Selections

We utilized 2 different corpora from the NLTK Project Gutenberg Selections. They are the bible-kjv.txt [3] and melville-moby_dick.txt [9]. The benefit of this selection of corpora is that the pieces use different styles of speaking, especially in comparison to the GAP and BUG datasets. bible-kvj.txt involves older English and melville-moby-dick.txt has a unique dialogue of its own. These differences allow for a broader comparison across different domains and provide a more well rounded understanding of how gender bias can be embedded in historical pieces. This is particularly interesting when we can compare these findings to more modern corpora to see how gender bias has evolved – if at all.

5 METHODS

Gender bias manifests itself in texts in many ways and can be identified using both linguistic and extra-linguistic cues [10].The goal, as stated in the problem statement, is to measure gender bias in the different word embedding spaces. To do this, we made two lists of words. The first is Gendered Pairs, a list of twenty analogous pairs of gendered word pairs e.g. he-she, him-her. The second list is Spotlight Words, a list of 82 words that are, by definition, gender neutral e.g. dainty, hairy, however are commonly stereotyped towards a specific gender. We focused on differences in depictions of men and women as they have been prolifically quantified – as seen in previous research we've reviewed.

We ran the Glove embedder and the Gn-Glove embedder on our four of our corpi to produce eight embeddings. Hyperparameters on both embedders can be seen in our code under the N.sh and G.sh shell scripts and hyperparameters were kept the same across both embedders.

After we got thereceiving the embeddings from both Glove and Gn-Glove, we made scatter plots comparing the cosine distance of every gendered pair on every spotlight word. IdealIdeal performance in this case would be to observe a smaller cosine distance from the Gn-Glove embeddings than as seen in the Glove embeddings. In more concrete terms, we may would be someobserve any value on the x-axis, for the Glove cosine distance, and some a value closer to 0 on the y-axis, the for the Gn-Glove cosine distance. This performance would indicate that the Gn-Glove debiasing model is effective in handling stereotypically biased words. A point of (1,0) would indicate that the respective male or female coded word in the gendered word pair was positively correlated (1) with some spotlight word in the Glove embedding and then was not correlated in the Gn-Glove embedding (0).

Since 20Twenty graphs were made for each embedding correlating to each gendered pair and their respective cosine distances., w In the interest of being efficient with the space of this report,e we will not show them all here in the interest of space but , however, they can be viewed in the accompanying code for this report.located in our linked GitHub repository.

6 RESULTS

6.1 Graphs of Note

While not perfect due to the imprecision of scatter plots in nuanced data, scatter plots of the embedding spaces were produced for every pair of Gendered Words Pair and corpus. The x-axis is the Cosine Distance in the Glove embedding space between the respective Male-Female gendered pair and a given spotlight word. A point made it on to the graph only if it had embeddings in both the Gn-Glove and Glove embedding space which means not every graph will contain all expected 164 point (82 spotlight words times 2 gendered words). A strong performance would be indicated by clustering along the 0 y-axis indicating the given spotlight word had a strong correlation in the Glove embedding space but a more neutral embedding in the Gn-Glove Space.

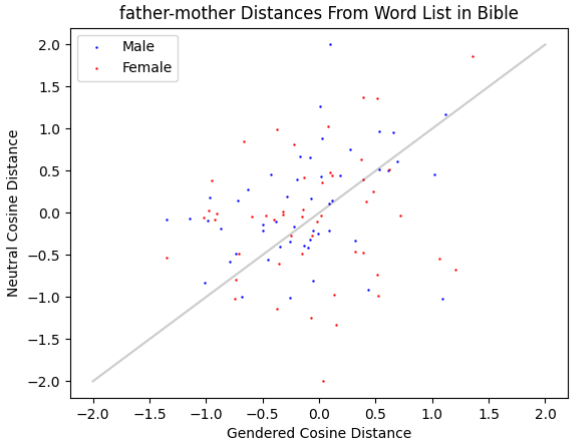

Figure 1: Father-Mother gendered pair distances to the spotlight words in the Bible corpus.

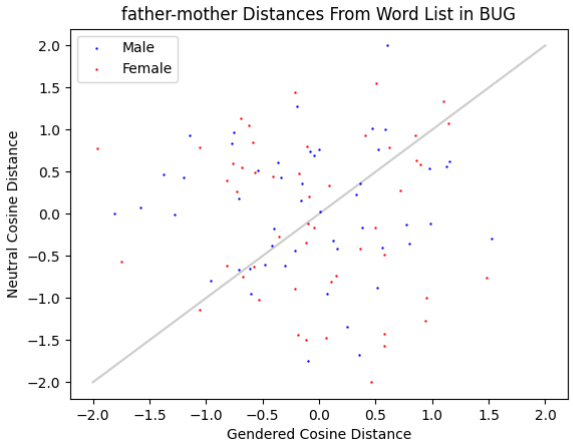

Figure 2: Father-Mother gendered pair distances to the spotlight words in the BUG corpus.

The figures shown below in Figures 1 - 4. The performance of the vector embeddings differed based on the corpus with no predictable pattern.

6.2 Difference Among Vector Embeddings

A function was derived that measured the percentage difference between the sizes of the cosine similarities of the spotlight words and the top x closest words to the spotlight words. A score closer to zero implies that the spotlight words in the Gn-Glove embedding spaces are not as strongly correlated as they are in the Glove embedding spaces. A score closer to 0 means that the Gn-Glove embedder did have a measurable effect on the embedding space. A summary of these results can be seen in table 1

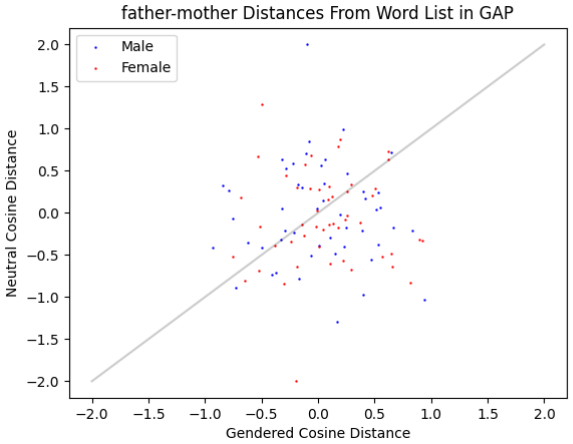

Figure 3: Father-Mother gendered pair distances to the spotlight words in the GAP corpus.

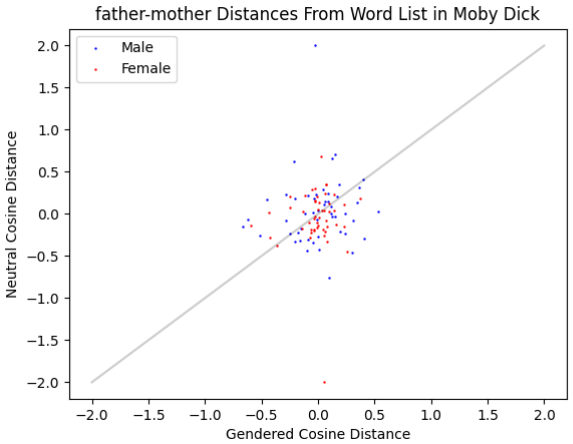

Figure 4: Father-Mother gendered pair distances to the spotlight words in the Moby Dick corpus.

Top x Word	2	5	10
Bible	.972	.95	.967
BUG	.914	.908	.906
GAP	.952	.945	.947
Moby-Dick	.915	.92	.919

Table 1: The ratio of the cosine distance of the spotlight words in the Gn-Glove space and the Glove space.

7 CONCLUSION

The Gn-Glove embedder performed positively but not conclusively. A 3%-8% performance improvement in the embedding space is measurable but not an obvious improvement, nor as great as we initially predicted it would be. Part of the lack of performance may

be a result of our current experience level with NLP metrics and the code base in our group.

We believe the issue regarding the NLP metrics to be the most relevant as even our differences in the vector space embeddings is not rigorous enough to be truly considered robust. For example, while we saw a 5% improvement across corpi, one must consider if these scores are a measure of the improvement among gendered pairs or if the numbers in the Gn-Glove embedding space are lower across the board. Our current analysis is not robust enough to answer this question specifically. Our measurements were focused on a more straightforward analysis of the model's performance. If we were to engage in a future iteration of this project, it would most certainly involve expanding the breadth of our analysis and incorporating other forms of measurement to account for these qualifying considerations.

The graphical representation of the space provides some interesting insights. It does appear that the Gn-Glove embedding space is less biased in certain scenarios across certain corpi. There is a large variance in performance across these corpi, so we are unable to pinpoint what exactly the common factor is in these scenarios where the Gn-Glove performs better. However, this occurrence leads us to believe that the model is not only effective, but can be best utilized in certain domains or when applied to texts of certain structures. Again, how rigorous this finding is leaves something to be desired as our gendered pair list and our spotlight words list is arbitrary and doesn't reflect the distribution of words from within the corpi. If we were to continue this project in the future, a more fine tuned list of gendered pairs and spotlight words should be curated with some insight of the corpi in mind, and with higher statistical validation techniques.

REFERENCES

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

[2] [n. d.]. *GAP Coreference Dataset.* https://github.com/google-research-datasets/gap-coreference original-date: 2018-10-18T17:48:21Z.

[3] [n. d.]. *The King James Version of the Bible.* https://www.gutenberg.org/ebooks/10

[4] [n. d.]. *Learning Gender-Neutral Word Embeddings (EMNLP 2018).* https://github.com/uclanlp/gn_glove original-date: 2018-08-22T21:57:55Z.

[5] [n. d.]. *stanfordnlp/GloVe.* https://github.com/stanfordnlp/GloVe original-date: 2015-09-01T17:21:18Z.

[6] 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]

[7] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. [n. d.]. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. arXiv:2109.03858 [cs] http://arxiv.org/abs/2109.03858

[8] 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.

[9] Herman Melville. [n. d.]. *Moby Dick; Or, The Whale.* https://www.gutenberg.org/ebooks/2701

[10] 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

[11] 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

[12] 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 December 2023; accepted NA