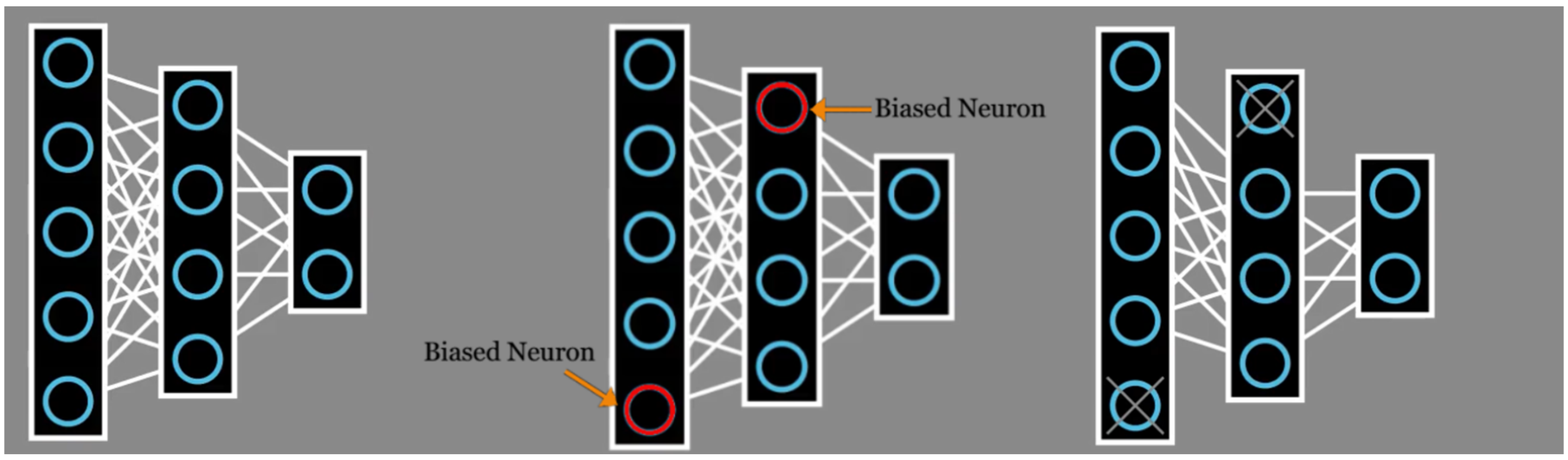
Mitigating Gender Bias through Semi-Supervised Regularizing Loss Function

Supriti Vijay

Dept. of Computer Science, Manipal Academy of Higher Education



Introduction

This paper presents a novel approach to mitigating implicit bias in language models during training by using a semi-supervised regularizing loss function. The proposed method addresses the limitations of previous debiasing techniques and is validated using the Equity Evaluation Corpus. It also aims to create an easy-to-use library for debiasing to allow streamlined integration into most high-performance LLMs, contributing towards making models more fair and equitable.

Prior Work and Motivation

Although multiple measures have been proposed for bias mitigation, they are plagued by expensive computations and manual annotations. Even so, computational methods may reduce model performance after prolonged re-training based on the generated embeddings, making it imperative that a methodology that does not require extended training be used. Hence, this work introduces a debiasing method for LLMs on a given downstream NLP task by restricting the weights of neurons learning on gender-neutral terms using regularization.

Proposed Methodology

I aim to utilize the Fisher-score-based weighted regularization on the neural architecture, to prevent word embeddings from construing particular labels, sentiments, or emotions as gendered. This will be coupled with the Equity Evaluation Corpus.

Weighted Regularization:

1. I optimize the weights and biases θ to fine-tune an *LLM* over the dataset D_{train}.
2. Therefore, when backpropagation is discharged on the EEC_{train} data, the gradients absorbed into G_{male}, G_{female} must contain information about which parameters were integral with respect to the EEC_{train} training samples, X^{EEC} and the label-distribution C.
3. We now compute the Fisher information matrix F.

$$F_{epoch=i} = (G_{female} - G_{male})^2 \quad (1)$$

Here, G_{male}, G_{female} are the computed gradients of the *LLM* iterating over the EEC_{train}

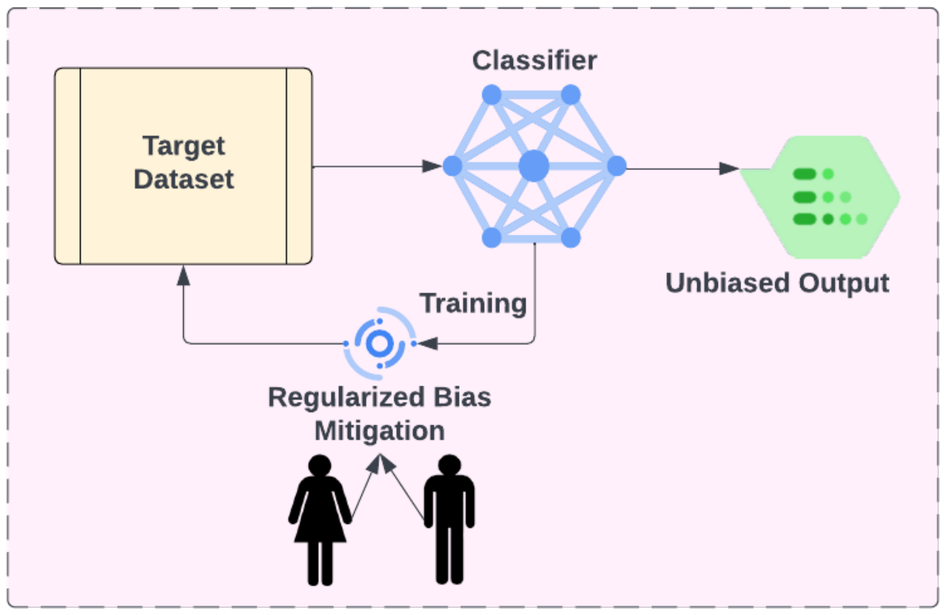


Figure 1. Proposed Methdology for debiasing

This Fisher matrix allows us to clearly distinguish the weights of those neurons/parameters, which highlight the differences between the two genders.

Evaluation

I aim to use the original **The SemEval-2018 Task 1: Affect in Tweets dataset** for proposed methodology verification.

Results:

Task	$F\uparrow$ $M\downarrow$	$F\downarrow$ $M\uparrow$
Anger	0.0345	-0.0336
Fear	0.0340	-0.0347
Joy	0.0359	-0.0388
All	0.0336	-0.0335

Acknowledgements

I would like to thank the organizing committee at AAAI and the undergraduate consortium for providing me with the opportunity to present my work today.

Connect with me

This QR code flies you right by my website, where you can find out more about me. Feel free to text or email me about my work.

supritivijay.github.io