F-BRIM: A Semi-Supervised Approach for Bias Mitigation with Activation-Weighted Neuron Regularization

Supriti Vijay[1]*, Aman Priyanshu[2]*, Ashalatha Nayak[1]

[1]Department of Computer Science
[2]Department of Information & Communication Technology
Manipal Institute of Technology,
Manipal Academy of Higher Education,
Karnataka, India.
supriti.vijay@gmail.com

Abstract

The presence of implicit bias in text corpora is one of the most prominent issues while training downstream NLP models that can learn to propagate the same. Classification and regression models distorted by this bias have been shown to reduce performance quality on minority groups, labelling certain demographics weakly or with higher confidence. Previous work has tried mitigating such bias by debiasing word embeddings or data augmentation; however, we propose a training-integrable semi-supervised regularizing loss function for debiasing of large language models (LLMs). In this paper, we address known pitfalls in the construction of single-time debiasing models & epoch-wise debiasing models, to theoretically formalize F-BRIM (Fairness Bias Regularization Mitigator), as well as validate its performance against the benchmark Equity Evaluation Corpus (EEC). We demonstrate the applicability of the methodology for most classification and regression tasks. As we believe our work has implications to be utilized in the pipeline of gender mitigation in NLP tasks, we open-source our code as well as provide a callback function to HuggingFace API, allowing streamlined integration into most high-performance LLMs.

Introduction

Natural language processing has become an integral aspect of many predictive and analytical tasks. With its increasing applications in various business, corporate and academic industries, it is an indispensable field of machine learning. However, with its influence also comes its susceptibility to gender biases inherent in the real-world data. Gender Bias here refers to when the gender signals in the input affect a model's predictions. Consider a set of two sentences fed into the model, only with different gender signals, such as ”He was furious” and ”She was furious”. Both of these sentences should ideally predict the same final sentiment polarity. However, since the model learns from the training set, it learns undesirable associations between words which we observe as gender bias. The propagation of these biases through NLP algorithms pose significant danger for certain downstream tasks. This has real-world consequences in applications such as automatic resume filtering systems, loan eligibility, crime recidivism prediction systems, machine translation, etc. Therefore, gender bias mitigation is an integral aspect of standard NLP tasks such as regression and classification.

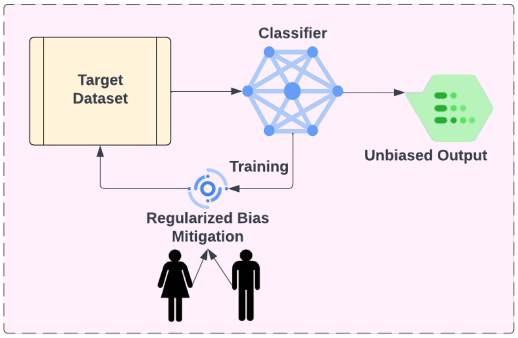

Figure 1: Intrinsic Gender Bias is an important problem plaguing many NLP models and tasks face. We present F-BRIM an activation weighted regularization module for mitigating bias.

Ideally, these models must be debiased before deployment. However, data augmentation is an expensive task for large corpora and reconstructing word embeddings in gender de-biased light may not always be conducive to performance reduction. Therefore, we introduce a new methodology for streamlined training of language models, which incorporates the debiasing procedure into it. There has been very little work done on debiasing language models during the training procedure, which we aim to address. Our main contributions are as follows:

1. Debiasing while training Large Language Models(LLMs)
2. Incorporating a Gender-based Evaluation Test(GBET) metric, Equity Evaluation Corpus for constructing debiased Fisher Information Matrices using a semi-supervised pipeline
3. Incorporation of a Regularizer for controlling perfor-

*These authors contributed equally.

mance mitigation tradeoff.

Related Work

Human-written texts which are incorporated in our training corpus reflect societal biases. (Sun et al. 2019; Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017).These further are detected in language representations,i.e. word embeddings that language models learn, causing disastrous outcomes when applied in real-world scenarios.

Most work has focused on quantifying and measuring bias in the training data and model. (Zhao et al. 2018a; Lu et al. 2019; Kiritchenko and Mohammad 2018) proposed gender-swapping, which referred to interchanging each male-defined word with its respective female equivalent. Metrics like Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017), Sentence Encoder Association Test (May et al. 2019) etc. were also proposed to measure bias inside word embeddings and sentnece encoders. However, these metrics quantify bias present in word embeddings, failing to quantify the bias present in the model itself (Bansal 2022). Consequently, Gender Bias Evaluation Test sets (GBETs) (Sun et al. 2019) were proposed to detect the degree of gender bias in the model. Equity Evaluation Corpus (EEC), one such example of these datasets (Kiritchenko and Mohammad 2018) was presented for the sentiment analysis task. We consider this corpus as a primary corrective subset for our semi-supervised approach.

Various measures like data augmentation(Zhao et al. 2018a), gender tagging(Vanmassenhove, Hardmeier, and Way 2018), Bias Fine-Tuning(Park, Shin, and Fung 2018), gender-neutral embeddings (Zhao et al. 2018b), etc. have also been proposed to debiasing word embeddings and the training corpora, however, they poorly impact computational performance and may be not be sustainable. On the other hand, very little focus has been on debiasing language models, specifically pre-trained models.(Garimella et al. 2021) proposes introducing bias mitigation during model training of BERT by further pre-training it on a small data subset using bias mitigation losses. Aligning to this work, we introduce a debiasing method for large language models on a given downstream NLP task by restricting the weights of neurons learning on gender-neutral terms using regularization.

Proposed Methodology

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.

In this implementation, a reinforcing framework is used to prevent word embeddings from construing particular labels, sentiments, or emotions as gendered. We employ Fisher-score based weighted regularization on the neural architecture for debiasing. A smaller corpus, such as the Equity Evaluation Corpus, is utilized for computing equity among multiple majority-minority groups as a preservation factor against gender association. We discuss the performance-mitigation trade-off and formulate this problem in detail in the following subsection. .

Problem Formulation

We consider the problem of training a large language model (LLM) for a downstream classification or regression task with implicitly biased data. In our problem formulation, we consider representational bias as an association between gender and model parameters, such as word-embeddings and model weights. Therefore, we propose the following objectives for our proposed methodology,

1. Performs debiasing of Large Language Models (LLMs) while training on downstream tasks.
2. Integrates a standard benchmark of template sentences with equitable evaluation between majority-minority groups. However, this may not be of the same label distribution as our training data. Therefore, a semi-supervised approach must be constructed for EEC inclusion within the algorithm.
3. Provide a regularizer, α, to control the performance-mitigation trade-off.

Aligning to the aforementioned objectives we formulate our given task. Given a large language model (LLM), we formalize the task of gender debiasing on a training dataset $D_{train} = \{X_{train}, Y_{train}\}$, where its components describe the classification task at hand. Expanding X_{train} and Y_{train} into its individual samples, we get $X_{train} = x_0, x_1, x_2, ..., x_N$ and $Y_{train} = y_0, y_1, y_2, ..., y_N$, where, x_i, y_i represents a sample sentence and its respective label. Taking the class-labels for Y_{train} as $C = \{label_0, label_1, ...label_{N^c}\}$, all $y_i \in C$. The LLM must be fine-tuned on this downstream task, using standard mini-batch optimization. A smaller or less extensive $EEC_{train} = \{X^{eec}, gender^{eec}\}$ data-segment is also provided, which may or may not contain the same label-distribution as C.

Semi-Supervised Regularization for Debiasing

The LLM is fine-tuned for a single iteration on the D_{train} data segment. Upon completion, it is introduced to the training segment of the Equity Evaluation Corpus (EEC_{train}). However, it becomes difficult to seek EEC datasets of the same nature as the training data. Therefore, it can be understandable that the labels presented within the EEC_{train} and the D_{train} may not match. To overcome this, we utilize the concept of $pseudo-labeling$, forming a semi-supervised approach towards EEC internalization. The training-optimizer is then frozen for accumulation of gradient-history over the EEC_{train}. The computed gradients are recorded (summed) over two distributions, i.e. G_{male}, G_{female} in our case. They each represent the sum of gradients upon backpropagation of only male/male-dominated samples and female/female-dominated samples.

Weighted Regularization Now, we discuss the aspect of weighted regularization with respect to LLMs for downstream tasks. Learning a task consists of updating the set of

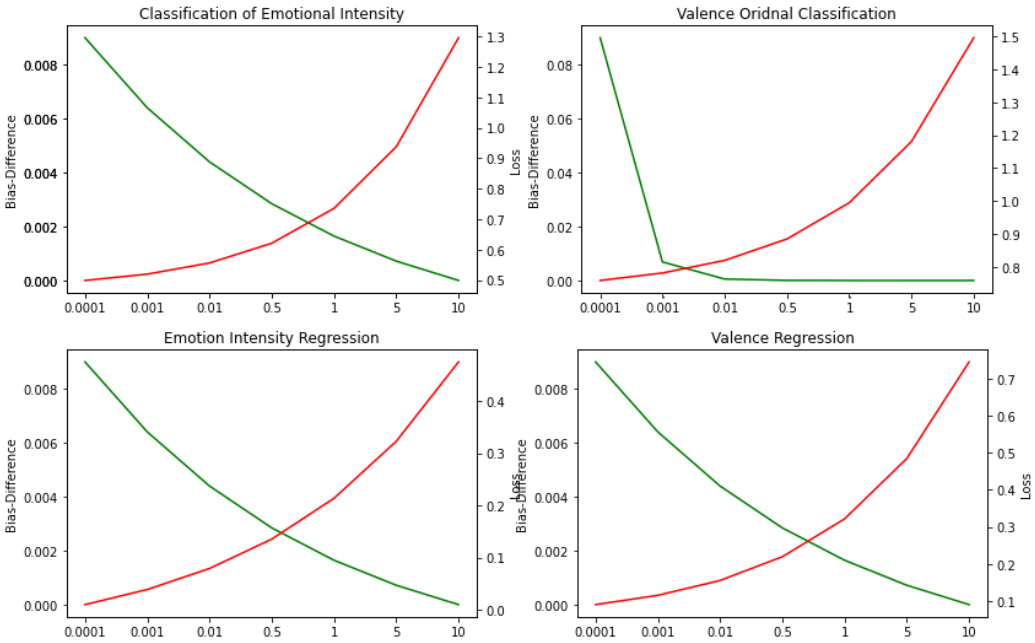

Figure 2: The above results present the performance-mitigation tradeoff for the tasks proposed in SemEval-2018 Task 1

weights and biases θ of the target neural network. However, prior literature highlights that multiple configurations of θ will result in similar performance (Nielsen 1989; Sussmann 1992). This is the backbone for the construction of F-BRIM. The over-parameterization of large language models makes it likely for a solution to be present for the downstream task, which does not demonstrate representational bias. Upon training the LLM over the dataset D_{train} and backpropagating, 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. However, a difference between the two would distinctly present the model parameters learning to distinguish between the two genders for said labels. We can compute the Fisher information matrix F as, $F_{epoch=i} = (G_{female} - G_{male})^2$.

Where, G_{male}, G_{female} are the computed gradients of the LLM iterating over the EEC_{train}. This Fisher matrix allows us to clearly distinguish and weight those neurons/parameters which highlight distinguish between the two genders. Upon further iteration ($epoch\_no > 1$) over the fine-tuning dataset, D_{train}, we regularize those specific neurons, using equation 1.

$$L(\theta) = L(\theta_i) + \sum_{i} \alpha * F_{i-1} * (\theta_{i-1} - \theta_i)^2 \tag{1}$$

Where, $L(\theta)$ represents the loss on the D_{train} data batch for the current epoch (i), F_{i-1} the previously computed Fisher information matrix, and θ_k representing the parameters of the LLM at the kth epoch. The α sets how important the task of debiasing is with respect to performance. Thus, creating a control for the performance-mitigation tradeoff.

The penalty applied to those neurons can be justified, as these are neurons learning distinguishing features between the two genders. Thus, leading to a higher penalty knowing that it deviated (*i.e. learnt bias*) from the previous epoch.

Experimental Results

Data type	**Loss-a**	**Acc-a**	**Loss-b**	**Acc-b**
Training	0.3513	0.865	0.5720	0.674
Validation	0.4989	0.826	0.6155	0.5995

Table 1: The following results present the accuracy and loss for our benchmark model with 3-fold cross validation. These results are based on the tasks of (a) emotion intensity ordinal classification and the (b) valence ordinal classification task.

Task	$F \uparrow M \downarrow$	$F \downarrow M \uparrow$
Anger	0.0345	-0.0336
Fear	0.0340	-0.0347
Joy	0.0359	-0.0388
Sadness	0.0344	-0.0341
Valence	0.0292	-0.0267
All	0.0336	-0.0335

Table 2: The following results present bias metrics for gender-bias identification (Kiritchenko and Mohammad 2018). These results are based on the tasks of emotion intensity ordinal classification and the valence ordinal classification task.

Epoch	Metric	Anger	Fear	Joy	Sadness	Valence	Loss	Acc
2	$F \uparrow\ M \downarrow$	0.0327	0.0339	0.0333	0.0336	0.0277	0.6068	0.746
2	$F \downarrow\ M \uparrow$	-0.0316	-0.0325	-0.0374	-0.0329	-0.0218	0.6068	0.746
3	$F \uparrow\ M \downarrow$	0.0251	0.0266	0.0233	0.0211	0.0243	0.5489	0.803
3	$F \downarrow\ M \uparrow$	-0.0228	-0.0250	-0.0246	-0.0294	-0.0187	0.5489	0.803
4	$F \uparrow\ M \downarrow$	0.0203	0.0235	0.0192	0.0193	0.0167	0.5311	0.812
4	$F \downarrow\ M \uparrow$	-0.0192	-0.0218	-0.0208	-0.0244	-0.0150	0.5311	0.812
5	$F \uparrow\ M \downarrow$	0.0185	0.0205	0.0188	0.0179	0.0144	0.5288	0.819
5	$F \downarrow\ M \uparrow$	-0.0177	-0.0207	-0.0189	-0.0183	-0.0144	0.5288	0.819

Table 3: The following results present bias metrics for gender-bias identification for the proposed methodology, model-name. These results provide a comparison to the aforementioned baseline results.

Dataset and Bias

We used the **The SemEval-2018 Task 1: Affect in Tweets** dataset for our proposed methodology verification. Containing an array of subtasks—(1) emotion intensity regression (2) emotion intensity ordinal classification (3) valence (sentiment) regression (4) valence ordinal classification and (5) emotion classification—on inferring a person's emotional state from their tweet. For our de-biasing subset we use the EEC (Kiritchenko and Mohammad 2018). The EEC is a large corpus consisting of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders.

Training and Inference

For all models, we use the $distilbert - base - uncased$ model, an LLM which was fine-tuned on the aforementioned classification task. We trained each model (presented in the experiments below) on a single GPU, Tesla P100 16GB with a 13GB RAM Intel Xeon as CPU. We split the EEC into two segments, the EEC_{train} and the $EEC_{verification}$, based on template sentences. Thus the divisions between the EEC_{train} and $EEC_{verification}$ did not include the same kind of template sentences, making them of different distributions. Thus allowing us to accurately measure the bias, if present, in the trained models. We also ensure each experiment is validated thrice, following the K-Fold cross-validation paradigm.

Comparitive Study of Performance-Mitigation Tradeoff

In our proposed methodology, F-BRIM, a regularizer parameter α is used to control the performance-mitigation tradeoff. Therefore, a study is undertaken on the sensitivity and effect of this said value, allowing us to validate whether there's truly a tradeoff between the scale of mitigation and final model performance.

Results

Baseline Analysis Since our task is only with respect to English tweets, we fine-tune the *distilbert-base-uncased* LLM specifically for it. We therefore, present these benchmark results without the proposed debiasing in Table 1. The performance of this methodology is closely in line with the results of the higher performing models (87.3%$-a$ & 58.8%$-b$) of the SemEval-2018 Task. Furthermore, we perform the bias identification computations as presented in (Kiritchenko and Mohammad 2018), for evaluating the bias present in this benchmark. It can be referred to in Table 2.

Debiasing using F-BRIM & Performance-Mitigation Tradeoff Taking the baseline analysis presented in , we further evaluate our framework on bias and validate its mitigative properties. We present these results in Table 3 after every epoch after the first ($epoch = \{2, 3, 4, 5\}$) for $\alpha = 0.1$. As we can see, the results demonstrate a clear reduction in bias. At the same time, model performance on the validation set is substantial. With a decrease of only, 0.7% in accuracy, the difference between the $F \uparrow\ M \downarrow$ and the $F \downarrow\ M \uparrow$ has reduced down to $8.24 * 10^{-4}$ from the original $1.48 * 10^{-3}$; which is a reduction of up to 44.324% in the averaged ($F \uparrow\ M \downarrow, F \downarrow\ M \uparrow$) values. The methodology also ensures that this occurs within the given procedure's training time, thereby preventing any excessive computation before or after training.

A comparative study is also conducted on the performance-mitigation tradeoff of F-BRIM. Taking regularizer value, α, we provide a graphical representation, Figure 2, of the tradeoff as we increase $\alpha \in [1 * 10^{-4}, 10]$. The graph verifies a clear trade as we mitigate more bias (increase α value), we consequently reduce model performance.

Conclusion

With the increase in applications of natural language processing, and classification and regression being some of its fundamental tasks. It becomes imperative that bias-mitagation efforts be evaluated and incorporated into them. We introduce F-BRIM, a regularization based bias mitigation framework for fine-tuning language models.

Our model is validated across multiple tasks, and against the standard benchmark of the EEC. While traditional approaches of data-augmentation and sub-space deletions may exist, we propose a streamlined training augmentation to replace them. This would allow easier integration with standard training engines such as HuggingFace, PyTorch and Tensorflow. We open-source and make publicly available the debiasing engines as Callback functions for each of their modules, thereby extending its usecase and simplifying the process of bias-mitigation.

References

Bansal, R. 2022. A Survey on Bias and Fairness in Natural Language Processing. arXiv:2204.09591.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Garimella, A.; Amarnath, A.; Kumar, K.; Yalla, A. P.; N, A.; Chhaya, N.; and Srinivasan, B. V. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4534–4545. Online: Association for Computational Linguistics.

Kiritchenko, S.; and Mohammad, S. M. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. arXiv:1805.04508.

Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2019. Gender Bias in Neural Natural Language Processing. arXiv:1807.11714.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Nielsen, R. H. 1989. Theory of the Backpropagation Neural Network. In *Proceedings of the International Joint Conference on Neural Networks* (Washington, DC), volume I, 593–605. Piscataway, NJ: IEEE.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. Florence, Italy: Association for Computational Linguistics.

Sussmann, H. J. 1992. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4): 589–593.

Vanmassenhove, E.; Hardmeier, C.; and Way, A. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3003–3008. Brussels, Belgium: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. arXiv:1804.06876.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics.