Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function

Yusu Qian
Tandon School of Engineering
New York University
yq729@nyu.edu

Urwa Muaz
Tandon School of Engineering
New York University
um367@nyu.edu

Ben Zhang
Center for Data Science
New York University
bz957@nyu.edu

Jae Won Hyun
Courant Institute of Mathematical Sciences
New York University
jaewhyun@nyu.edu

Abstract

Gender bias exists in natural language datasets which neural language models tend to learn, resulting in biased text generation. In this research, we propose a debiasing approach based on the loss function modification. We introduce a new term to the loss function which attempts to equalize the probabilities of male and female words in the output. Using an array of bias evaluation metrics, we provide empirical evidence that our approach successfully mitigates gender bias in language models without increasing perplexity. In comparison to existing debiasing strategies, data augmentation, and word embedding debiasing, our method performs better in several aspects, especially in reducing gender bias in occupation words. Finally, we introduce a combination of data augmentation and our approach, and show that it outperforms existing strategies in all bias evaluation metrics.

1 Introduction

Natural Language Processing (NLP) models are shown to capture unwanted biases and stereotypes found in the training data which raise concerns about socioeconomic, ethnic and gender discrimination when these models are deployed for public use (Lu et al., 2018; Zhao et al., 2018).

There are numerous studies that identify algorithmic bias in NLP applications. Lapowsky (2018) showed ethnic bias in Google autocomplete suggestions whereas Lambrecht and Tucker (2018) found gender bias in advertisement delivery systems. Additionally, Zhao et al. (2018) demonstrated that coreference resolution systems exhibit gender bias.

Language modelling is a pivotal task in NLP with important downstream applications such as text generation (Sutskever et al., 2011). Recent studies by Lu et al. (2018) and Bordia and Bowman (2019) have shown that this task is vulnerable to gender bias in the training corpus. Two prior works focused on reducing bias in language modelling by data preprocessing (Lu et al., 2018) and word embedding debiasing (Bordia and Bowman, 2019). In this study, we investigate the efficacy of bias reduction during training by introducing a new loss function which encourages the language model to equalize the probabilities of predicting gendered word pairs like *he* and *she*. Although we recognize that gender is non-binary, for the purpose of this study, we focus on female and male words.

Our main contributions are summarized as follows: i) to our best knowledge, this study is the first one to investigate bias alleviation in text generation by direct modification of the loss function; ii) our new loss function effectively reduces gender bias in the language models during training by equalizing the probabilities of male and female words in the output; iii) we show that end-to-end debiasing of the language model can achieve word embedding debiasing; iv) we provide an interpretation of our results and compare with other existing debiasing methods. We show that our method, combined with an existing method, counterfactual data augmentation, achieves the best result and outperforms all existing methods.

2 Related Work

Word Embedding Debiasing Bolukbasi et al. (2016) introduced the idea of gender subspace as low dimensional space in an embedding that captures the gender information. Bolukbasi et al. (2016) and Zhao et al. (2017) defined gender bias as a projection of gender-neutral words on a gender subspace and removed bias by minimizing this projection. Gonen and Goldberg (2019) proved that bias removal techniques based on minimizing projection onto the gender space are insufficient. They showed that male and female stereotyped words cluster together even after such debiasing treatments. Thus, gender bias still remains

in the embeddings and is easily recoverable.

Bordia and Bowman (2019) introduced a co-occurrence based metric to measure gender bias in texts and showed that the standard datasets used for language model training exhibit strong gender bias. Using the same definition of embedding gender bias as Bolukbasi et al. (2016), Bordia and Bowman (2019) introduced a regularization term that aims to minimize the projection of neutral words onto the gender subspace. Throughout this paper,we refer to this approach as REG. We argue that this method has two shortcomings. First, the bias definition is shown to be incomplete by Gonen and Goldberg (2019) and secondly, an embedding is not the sole source of gender bias in downstream applications.

Data Debiasing Lu et al. (2018) showed that gender bias in coreference resolution and language modelling can be mitigated through a data augmentation technique that expands the corpus by swapping the gender pairs like *he* and *she*, or *father* and *mother*. They called this Counterfactual Data Augmentation (CDA) and concluded that it outperforms the word embedding debiasing strategy proposed by Bolukbasi et al. (2016). CDA doubles the size of the training data and increases time needed to train language models. In this study, we intend to reduce bias during training without requiring a data preprocessing step.

3 Methodology

Dataset For the training data, we use Daily Mail news articles released by Hermann et al. (2015). This dataset is composed of 219,506 articles covering a diverse range of topics including business, sports, travel, etc. For manageability, we randomly subsample 5% of the text. The subsample has around 8.25 million tokens in total.

Language Model We use a pre-trained 300-dimensional word embedding, GloVe, by Pennington et al. (2014). We apply random search to the hyperparameter tuning of the LSTM language model. The best hyperparameters are as follows: 2 hidden layers each with 300 units, a sequence length of 35, a learning rate of 20 with an annealing schedule of decay starting from 0.25 to 0.95, a dropout rate of 0.25 and a gradient clip of 0.25. We train our models for 150 epochs, use a batch size of 48, and set early stopping with a patience of 5.

Loss Function Language models are usually trained using cross-entropy loss. Cross-entropy loss at time step t is

$$L^{CE}(t) = -\sum_{w \in V} y_{w,t} \log\left(\hat{y}_{w,t}\right),$$

where V is the vocabulary, y is the one hot vector of ground truth and $\hat{y}$ indicates the output softmax probability of the model.

We introduce a loss term L^B, which aims to equalize the predicted probabilities of gender pairs such as *woman* and *man*.

$$L^B(t) = \frac{1}{G}\sum_{i}^{G}\left|\log\frac{\hat{y}_{f_i,t}}{\hat{y}_{m_i,t}}\right|$$

f and m are a set of corresponding gender pairs, G is the size of the gender pairs set, and $\hat{y}$ indicates the output softmax probability. We use gender pairs provided by Zhao et al. (2017). Overall loss can be written as

$$L = \frac{1}{T}\sum_{t=1}^{T} L^{CE}(t) + \lambda L^B(t),$$

where λ is a hyperparameter and T is the corpus size. We observe that among the similar minima of the loss function, L^B encourages the model to converge towards a minimum that exhibits the lowest gender bias.

4 Model Evaluation

Co-occurrence Bias Co-occurrence bias is computed from the model-generated texts by comparing the occurrences of all gender-neutral words with female and male words. A word is considered to be biased towards a certain gender if it occurs more frequently with words of that gender. This definition was first used by Zhao et al. (2017) and later adapted by Bordia and Bowman (2019). Using the definition of gender bias similar to the one used by Bordia and Bowman (2019), we define gender bias as

$$B^N = \frac{1}{N}\sum_{w \in N}\left|\log\frac{c(w,m)}{c(w,f)}\right|,$$

where N is a set of gender-neutral words, and $c(w,g)$ is the occurrences of a word w with words of gender g in the same window. This score is designed to capture unequal co-occurrences of

He is a | **She** is a |	**doctor** $\log \frac{P(t\|s_1)}{P(t\|s_2)}$

(a) Occupation bias conditioned on gendered words

The **doctor** is a |	**man** **woman** $\log \frac{P(t_1\|s)}{P(t_2\|s)}$

(b) Occupation bias conditioned on occupations

Table 1: Example templates of two types of occupation bias

neutral words with male and female words. Co-occurrences are computed using a sliding window of size 10 extending equally in both directions. Furthermore, we only consider words that occur more than 20 times with gendered words to exclude random effects.

We also evaluate a normalized version of B^N which we denote by conditional co-occurrence bias, B_c^N. This is defined as

$$B_c^N = \frac{1}{N} \sum_{w \in N} \left| \log \frac{P(w|m)}{P(w|f)} \right|,$$

where

$$P(w|g) = \frac{c(w,g)}{c(g)}.$$

B_c^N is less affected by the disparity in the general distribution of male and female words in the text. Since the disparity between the occurrences of the two is also a form of bias, we report the ratio of occurrence of male and female words, GR.

Causal Occupation Bias Following the approach similar to Lu et al. (2018), we limit the bias evaluation to a set of gender-neutral occupations. We create a list of sentences based on a set of templates. There are two sets of templates used for evaluating causal occupation bias (Table 1). The vertical bar separates the seed sequence that is fed into the language models from the target, for which we observe the output softmax probability.The first set of templates is designed to measure how the probabilities of occupation words depend on the gender information in the seed.

We measure causal occupation bias conditioned on gender as

$$CB|g = \frac{1}{|O|} \frac{1}{G} \sum_{o \in O} \sum_{i}^{G} \left| \log \frac{p(o|f_i)}{p(o|m_i)} \right|,$$

where O is a set of gender-neutral occupations and G is the size of the gender pairs set. For example, $P(doctor|he)$ is the softmax probability of *doctor* where the seed sequence is *He is a*.

Causal occupation bias conditioned on occupation is represented as

$$CB|o = \frac{1}{|O|} \frac{1}{G} \sum_{o \in O} \sum_{i}^{G} \left| \log \frac{p(f_i|o)}{p(m_i|o)} \right|.$$

We believe that both $CB|g$ and $CB|o$ contribute to gender bias in the model-generated texts and that $CB|o$ is more easily influenced by the general disparity in male and female word probabilities.

Word Embedding Bias Our debiasing approach does not explicitly address the bias in the embedding layer. Therefore, we use gender-neutral occupations to measure the embedding bias to observe if debiasing the output layer also decreases the bias in the embedding. We define the embedding bias, EB_d, as the difference between the Euclidean distance of an occupation word to male words and the distance of the occupation word to the female counterparts. This definition captures the embedding bias described by Bolukbasi et al. (2016) as

$$EB_d = \sum_{o \in O} \sum_{i}^{G} \big| \|E(o) - E(m_i)\|_2 - \|E(o) - E(f_i)\|_2 \big|,$$

where O is a set of gender-neutral occupations, G is the size of the gender pairs set and E is the word-to-vector dictionary.

5 Experiments

After training the baseline model, we implement our loss function and tune for the λ hyperparameter. We test the existing debiasing approaches, CDA and REG, as well but since Bordia and Bowman (2019) reported that results fluctuate substantially with different REG regularization coefficients, we perform hyperparameter tuning and report the best results in Table 2. Additionally, we implement a combination of our loss function and CDA and tune for λ. Finally, bias evaluation is

Model	B^N	B_c^N	GR	$Ppl.$	$CB\|o$	$CB\|g$	EB_d
Baseline	0.531	0.282	1.415	117.845	1.447	97.762	0.528
REG	0.381	0.329	1.028	**114.438**	1.861	108.740	0.373
CDA	0.208	0.149	1.037	117.976	0.703	56.82	0.268
$\lambda_{0.5}$	0.312	0.173	1.252	120.344	**0.000**	1.159	0.006
λ_1	0.218	0.153	1.049	120.973	**0.000**	0.999	0.002
λ_2	0.221	0.157	1.020	123.248	**0.000**	0.471	**0.000**
$\lambda_{0.5}$ + CDA	**0.205**	**0.145**	**1.012**	117.971	**0.000**	**0.153**	**0.000**

Table 2: Evaluation results for models trained on Daily Mail and their generated texts

performed for all the trained models. Causal occupation bias is measured directly from the models using template datasets discussed above and co-occurrence bias is measured from the model-generated texts, which consist of 10,000 documents of 500 words each.

Results Results for the experiments are listed in table 2. From measurements using the described bias metrics, our method effectively mitigates bias in language modelling without a significant increase in perplexity. At λ value of 1, it reduces B^N by 58.95%, B_c^N by 45.74%, $CB|o$ by 100%, $CB|g$ by 98.52% and EB_d by 98.98%. Compared to the results of CDA and REG, it achieves the best results in both occupation biases, $CB|g$ and $CB|o$, and EB_d. All methods result in GR around 1, indicating that there are near equal amounts of female and male words in the generated texts.

REG fails to achieve the best result in any of the bias metrics that we used, but results in the best perplexity. This indicates that REG has a slight regularization effect. Additionally, it is interesting that our loss function outperforms REG in EB_d even though REG explicitly aims to reduce gender bias in the embeddings. Although our method does not explicitly attempt geometric debiasing of the word embedding, the results show that it results in the most debiased embedding as compared to other methods. Furthermore, Gonen and Goldberg (2019) emphasizes that geometric gender bias in word embeddings is not completely understood and existing word embedding debiasing strategies are insufficient. Our approach provides an appealing end-to-end solution for model debiasing without relying on any measure of bias in the word embedding. We believe this concept is generalizable to other NLP applications.

CDA achieves slightly better results for co-occurrence biases, and results in a better perplexity. Our results are comparable in terms of bias mitigation effects. However, our method does not require a augmentation step and allows training of an unbiased model directly from biased datasets. For this reason, it also requires less time to train than CDA. Furthermore, CDA fails to effectively mitigate occupation bias when compared to our approach. Although the training data for CDA does not contain gender bias, the model still exhibits gender bias when measured with causal occupation bias metrics. This indicates that model-level constraints are essential to debiasing a model and dataset debiasing alone cannot be trusted.

Finally, the combination of CDA and our loss function outperforms all the methods in all measures of biases without compromising perplexity. Therefore, it can be argued that a cascade of these approaches can be used to optimally debias the language models.

6 Conclusion and Discussion

In this research, we propose a new approach for mitigating gender bias in neural language models and empirically show that our method outperforms existing methods. Our research highlights the fact that debiasing the model with bias penalties in the loss function is an effective method. We emphasize that this method is powerful and generalizable to downstream NLP applications. The research also reinforces the idea that geometric debiasing of the word embedding is not a complete solution for debiasing the downstream applications but encourages end-to-end approaches to debiasing.

Future work includes designing a context-aware version of our loss function which distinguishes unbiased and biased mentions of the gendered words and only penalize the latter which would better address gender-associated words that do not have a gender pair, like *pregnant*.

Collaboration Statement

Yusu oversaw project coordination, performed literature review, designed the gender-equalizing loss function and evaluated the models. Urwa performed literature review, designed the research methodology and implemented the loss function, causal bias and word embedding evaluation. Ben set up the development environment, built the language models, generated texts, implemented hyperparameter tuning and evaluated the models. Jae preprocessed texts, implemented CDA and formatted the paper. All contributed to writing the paper, preparing the presentation, overall discussion.

Github Link

https://github.com/sueqian6/Final-Project-for-Natural-Language-Understanding-and-Computational-Semantics

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. ArXiv:1904.03035.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. ArXiv:1903.03862.

Karl Hermann, Tom Koisk, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1693–1701.

Anja Lambrecht and Catherine E. Tucker. 2018. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads.

Issie Lapowsky. 2018. Google autocomplete still makes vile suggestions.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. ArXiv:1807.11714v1.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, page 15321543. Association for Computational Linguistics.

Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating text with recurrent neural networks. In *ICML'11 Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1017–1024.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chag. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Chang Kaiwei. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 48474853. Association for Computational Linguistics.