Should We Attend More or Less? Modulating Attention for Fairness

Anonymous ACL submission

Abstract

The advances in natural language processing (NLP) pose both opportunities and challenges. While recent progress enables the development of high-performing models for a variety of tasks, it also poses the risk of models learning harmful biases from the data, such as gender stereotypes. In this work, we investigate the role of attention, a widely-used technique in current state-of-the-art NLP models, in the propagation of social biases. Specifically, we study the relationship between the entropy of the attention distribution and the model's performance and fairness. We then propose a novel method for modulating attention weights to improve model fairness after training. Since our method is only applied post-training and pre-inference, it is an intra-processing method and is, therefore, less computationally expensive than existing in-processing and pre-processing approaches. Our results show an increase in fairness and minimal performance loss on different text classification and generation tasks using language models of varying sizes. *WARNING: This work uses language that is offensive.*

1 Introduction

Recent advancements in transformer-based pre-trained language models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and XLNet (Yang et al., 2019), have led to the emergence of new state-of-the-art models in a variety of applications, including but not limited to text summarization (Liu et al., 2022; Mordido and Meinel, 2020), sentence classification (Wang et al., 2018), question answering (Rajpurkar et al., 2018, 2016), and information extraction (Li et al., 2020a,b). Despite their success, recent studies (Nadeem et al., 2021; Meade et al., 2022; Zayed et al., 2023) have demonstrated that these models also exhibit harmful biases based on factors such as gender, race, sexual-orientation, and religion. These biases pose a significant challenge in deploying machine learning models in real-world applications, as they can result in discriminatory outcomes. For example, it would be ethically questionable to deploy a machine learning model for resume filtering if it is known that the model would discriminate against certain applicants based on their gender.

Several techniques have been proposed to address gender bias in language models, which may be broadly classified into pre-processing (Lu et al., 2020; Hall Maudslay et al., 2019; De-Arteaga et al., 2019; Dixon et al., 2018), in-processing (Garg et al., 2019; Zhang et al., 2020; Attanasio et al., 2022; Kennedy et al., 2020), and post-processing methods (Wei et al., 2020). Recently, a new category of debiasing methods has been introduced: intra-processing methods (Savani et al., 2020). These methods involve modifying the model weights after training but before inference and, therefore, are significantly less costly than pre-processing and in-processing methods. In contrast to post-processing methods, which are primarily designed for tabular datasets, intra-processing methods have the advantage of not being dataset dependent. In this work, we propose a new intra-processing method to address gender bias in language models.

The work by Attanasio et al. (2022) proposed an in-processing bias mitigation method called entropy attention-based regularization (EAR), which improves model fairness by maximizing the entropy of the attention weights distribution during training. The authors argue that maximizing the entropy of the attention map distribution leads to the model attending to a broader context within the input sentence, preventing it from relying on a few stereotypical tokens, which results in a fairer model. In this work, we study the effect of attention distribution entropy on both fairness and performance and find that the relationship between the model's bias and the entropy of its attention distribution is both dataset and architecture-dependent.

This suggests that some of the previous findings by Attanasio et al. (2022) may not be general.

Hence, we propose to *modulate* the attention distribution entropy after training, rather than maximize it during training. Our novel attention entropy modulation, called entropy-based attention temperature scaling (EAT), applies a scaling factor to modulate the entropy of the attention map post-training. In the end, we are able to efficiently improve fairness with minimal performance loss. Our contributions may be summarized as follows:

1. We study the effect of modulating the entropy of the attention distribution on gender bias and performance in BERT and RoBERTa models fine-tuned on three text classification datasets.

2. We propose a novel intra-processing bias mitigation method that modulates the attention distribution entropy in text classification with less than 3.5% degradation in performance. To the best of our knowledge, this is the first intra-processing method bias mitigation method in NLP, offering a computationally efficient alternative to existing methods.

3. We compare our method to other intra-processing debiasing methods originally proposed for image data and migrate such efforts to the NLP domain. We also compare it to other pre/in-processing methods.

4. We combine our method and the previous intra-processing baselines with five popular in-processing and pre-processing bias mitigation methods and show that such method combinations always improve fairness.

5. We show that our method generalizes to different forms of social bias even when the hyperparameters are exclusively tuned to mitigate gender bias.

6. Finally, we show that our method extends beyond text classification to address social biases in text generation using GPT-Neo.

2 Related work

We will start by describing existing techniques for mitigating bias that are applied before, during, or after model training. Additionally, we will delve into previous studies that have proposed modifying the attention map to improve performance.

2.1 Gender bias mitigation methods

Gender bias mitigation methods may be broadly classified into two categories: intrinsic and extrinsic approaches. While intrinsic methods (Adi et al., 2017; Hupkes et al., 2018; Conneau et al., 2018; Tenney et al., 2019; Belinkov and Glass, 2019) focus on analyzing the embedding representations assigned to gender tokens by the model, extrinsic methods (Sennrich, 2017; Isabelle et al., 2017; Naik et al., 2018) rely on the model's predictions to determine if different genders achieve similar predictions under the same context. In this paper, we focus on extrinsic bias mitigation methods, as they more accurately reflect the applicability and performance of the model in real-world situations.

Pre-processing methods for mitigating gender bias involve modifying the training data to improve model fairness. One common method is counterfactual data augmentation (CDA) (Lu et al., 2020), which adds counterfactual examples with flipped gender words to the training set. However, this can lead to longer training times and meaningless examples. Hence, counterfactual data substitution (CDS) (Hall Maudslay et al., 2019) swaps the gender words in the training set with a probability of 0.5, resulting in a dataset of the same size. More recently, Zayed et al. (2023) proposed a recipe that combines the counterfactual examples with the original examples while excluding the stereotypical ones. Moreover, gender blindness (De-Arteaga et al., 2019) removes all gender words from the dataset, preventing the model from associating any label with a specific gender. Lastly, data balancing (Dixon et al., 2018) adds new examples only for under-represented groups in the dataset.

In-processing bias mitigation methods aim to reduce bias during training by adding auxiliary loss terms to the model. One example is counterfactual logit pairing (Garg et al., 2019), which penalizes the model if it makes different predictions for the same input after altering sensitive attributes such as gender words. Another method by Kennedy et al. (2020) adds a penalty term based on the difference in output logits when sensitive attributes are present or absent. Instance weighting (Zhang et al., 2020) multiplies the loss by a factor greater than 1 for stereotypical sentences to penalize the model more for misclassifying them, and attention-based regularization (Attanasio et al., 2022) maximizes the model's attention distribution entropy during training to improve fairness.

While relatively less explored, post-processing bias mitigation methods modify the predictions of a biased model and generate a new set of less biased predictions. For example, Wei et al. (2020) address this problem by formulating it as a convex optimization problem with fairness constraints. Recently, intra-processing debiasing methods have been introduced to reduce the biases of image processing models as a new category of techniques that lie between in-processing and post-processing methods. Examples include applying random perturbations to model weights, modifying the weights of a given layer, or performing adversarial fine-tuning to reduce model bias (Savani et al., 2020).

2.2 Attention modulation

There has been an ongoing debate in the literature regarding the interpretability of the attention mechanism (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Moradi et al., 2019; Mohankumar et al., 2020). Despite this, several works have demonstrated that modulating the attention map values with prior knowledge improves model performance on downstream tasks. For example, in document summarization, Cao and Wang (2021) proposed a content selection method that detects tokens that are irrelevant at inference time, masking them from the attention map. Moreover, Cao and Wang (2022) proposed to increase the attention weights between tokens that lie within the same section in the document. Furthermore, Zhang et al. (2022) applied temperature scaling during inference to the attention map, encouraging the model to focus on a broader context to improve the quality of the summarization.

In the context of text classification, Li et al. (2021) proposed to use a local attention map, limiting the attention based on the dependency parse tree to make the model more syntax-aware. Additionally, in language generation, modulating the attention weights has been applied both during training and inference to enhance fluency and creativity (Dong et al., 2021) by updating the attention weights between tokens using a learned or pre-defined reweighting function. In machine translation, Yin et al. (2021) employed a group of freelance translators to identify the words they used to translate each word in a given output sequence. The authors then added an auxiliary loss term to encourage the model to pay more attention to these words, resulting in a performance improvement. Moreover, Lu et al. (2021) proposed to increase the model's attention to essential words by measuring the performance drop after the removal of such words, encouraging the model to attend more to the words that led to the most significant drop.

3 Self-attention

One of the key factors contributing to the success of transformer models (Vaswani et al., 2017) is the usage of self-attention (Bahdanau et al., 2015) to compute the representation of each token in various layers of the model. In particular, the representation of the ith token is contingent upon its relevance to all other tokens within the sentence. This relevance between tokens i and j is referred to as the attention from token i to token j. Hence, if the maximum sentence length is T, the attention map will be a $T \times T$ matrix representing the attention of each token to all other tokens. The attention map is calculated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}\,, \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{T \times d_q}$, $\mathbf{K} \in \mathbb{R}^{T \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{T \times d_v}$ are the query, key, and value matrices with embedding dimensionalities d_q, d_k, and d_v respectively. The computation of these matrices is described as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}; \quad \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}; \quad \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}\,,$$

where $\mathbf{X} \in R^{T \times d}$ is the input vector of maximum length T and embedding dimentionality d, while $\mathbf{W}^{\mathbf{Q}} \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}^{\mathbf{K}} \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}^{\mathbf{V}} \in \mathbb{R}^{d \times d_v}$ are three matrices of learnable parameters.

3.1 Attention entropy

The softmax function used to calculate the attention map in Eq.(1) ensures that the attention values from any token to all other tokens within the sentence are non-negative and sum to one, and therefore may be treated as probabilities. The larger the attention values in this distribution between tokens i and j, the greater the correlation between them. We follow the same procedure as Attanasio et al. (2022) to calculate the entropy of the attention distribution.

Considering the attention values $a_{l,h,i,j}$ between tokens i and j in the head h of layer l, we first average the attention weights over all heads in the l-th layer:

$$a'_{l,i,j} = \frac{1}{h}\sum_{h} a_{l,h,i,j}. \tag{2}$$

Subsequently, a softmax function is applied to ensure that the resulting values form a probability distribution:

$$a_{l,i,j} = \frac{e^{a'_{l,i,j}}}{\sum_j e^{a'_{l,i,j}}}. \tag{3}$$

The entropy, first introduced by Shannon (1948), is defined by

$$H_i^l = -\sum_{j=0}^{T_s} a_{l,i,j} \log a_{l,i,j}\,, \tag{4}$$

where T_s represents the actual length of the sentence, excluding any padding tokens. We obtain the average entropy within a sentence in layer l as

$$H^l = \frac{1}{T_s}\sum_{i=0}^{T_s} H_i^l\,, \tag{5}$$

with the overall attention entropy being computed by summing the entropy across all model layers:

$$H = \sum_l H^l. \tag{6}$$

4 Attention entropy modulation

Attanasio et al. (2022) suggested that memorization encourages models to focus on stereotypical tokens as a shortcut to solve the task, and proposed to maximize the entropy of the attention distribution. The intuition is to force the model to attend to a broader context, resulting in a less biased model. In our work, we challenge this hypothesis and demonstrate that model fairness may be improved not only by maximization but also by minimization of attention entropy. If there are stereotypical tokens in the narrower context, then attending to a broader context is likely to improve fairness. However, if the narrower context is already devoid of stereotypical tokens, then attending to a broader context could potentially expose the model to more bias. Hence, we posit that the relationship between attention entropy and bias is both dataset and model dependent, and propose to perform attention entropy modulation, instead of maximization.

4.1 Entropy-based attention temperature scaling (EAT)

We propose a novel method for attention entropy modulation, which we term entropy-based attention temperature scaling (EAT). Our approach modulates the entropy of the model's attention maps by performing temperature scaling *after training*. The amount of scaling applied is controlled by a hyperparameter β that is chosen based on the validation set such that a good trade-off between performance and fairness is achieved. A scaled attention map, *i.e.* after temperature scaling, is computed as (Hinton et al., 2015):

$$\text{Attention}_\text{s}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\beta\mathbf{Q}\mathbf{K}^\text{T}}{\sqrt{d_k}}\right)\mathbf{V}. \tag{7}$$

Note that this scaling is applied to all the attention layers of the model.

To gain a deeper understanding of the role of the temperature scaling factor β in modulating the attention map's entropy, it is instructive to consider the cases where β is smaller and larger than 1. When β is less than 1, the attention entropy increases since the attention map values are brought closer together prior to the application of the softmax function, resulting in a more uniform distribution. Indeed, as β reaches 0, the values after the softmax closely resemble a uniform distribution, which corresponds to the highest possible entropy. Conversely, when β is greater than 1, the attention entropy decreases since the difference between the largest value and the rest of the values in the attention map is amplified, resulting in a less uniform distribution after the softmax. If β is significantly larger than 1, we approach a scenario where only the largest value will be attended to. These scenarios are illustrated in Figure 1.

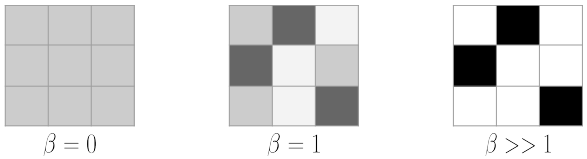

Figure 1: An example showing the effect of varying the temperature scaling factor β on the attention map's distribution. Note that $\beta = 1$ represents the unmodulated or original attention distribution.

5 Experiments

We will now provide a detailed overview of the tasks, datasets, baselines, and evaluation metrics as well as the different setups used in our experiments.

5.1 Tasks and datasets

We performed our experiments on two distinct binary text classification tasks: sexism detection and

toxicity detection, where the objective is to train a model to accurately differentiate between texts that are deemed sexist or toxic and those that are not, respectively. As defined by Dixon et al. (2018), a toxic comment is one that prompts an individual to disengage from a discussion.

All the datasets used in this study are in English. We used the following datasets: Twitter dataset (Waseem and Hovy, 2016) with approximately $16,000$ tweets binarized into sexist and non-sexist, Wikipedia dataset (Dixon et al., 2018) with around $160,000$ comments categorized as toxic or non-toxic, and the Jigsaw dataset[1] with around 1.8 million examples binarized into toxic and non-toxic. Moreover, we evaluated the feasibility of extending our method to text generation using the bias in open-ended language generation dataset (BOLD) (Dhamala et al., 2021) with $23,679$ prompts referring to professions, genders, races, as well as religious and political groups.

5.2 Baseline methods

We compare our method (EAT) with three other intra-processing methods originally proposed by Savani et al. (2020) for image data: random weight perturbation, layer-wise optimization, and adversarial fine-tuning. Moreover, we also use five pre/in-processing gender bias mitigation methods: instance weighting (Zhang et al., 2020), data augmentation (Lu et al., 2020), data substitution (Hall Maudslay et al., 2019), gender blindness (De-Arteaga et al., 2019), and entropy attention-based regularization (Attanasio et al., 2022). We refer the reader to Section 2 for a description of the methods.

5.3 Evaluation metrics

For text classification, we evaluate model performance using the area under the receiver operating characteristic curve, commonly referred to as AUC. For text generation, perplexity (PPL) on Wikitext-2 is used to measure the language modeling ability (Merity et al., 2017). To evaluate gender bias in text classification, we mainly use the demographic parity (DP) metric (Beutel et al., 2017; Hardt et al., 2016; Reddy, 2022) calculated by:

$$\mathrm{DP} = 1 - |p(\hat{y} = 1|z = 1) - p(\hat{y} = 1|z = 0)|, \tag{8}$$

where $\hat{y}$ represents the model's prediction and $z \in \{0, 1\}$ denotes keeping or flipping the gender words in the sentence, respectively. The scale of demographic parity ranges from 0 (least fair) to 1 (fairest). Our procedure for computing DP follows the methodology established in prior studies (Dixon et al., 2018; Park et al., 2018), where we use a synthetic dataset, the identity phrase templates test set (IPTTS), for measuring fairness. Additional fairness metrics are described in Appendix B.

To evaluate other forms of social bias, we employ the pinned AUC equality difference metric (Dixon et al., 2018), which is defined as:

$$\sum_{t \in T} |AUC - AUC_t|, \tag{9}$$

where AUC and AUC_t refer to the model's AUC on the whole dataset and on examples referring to a specific subgroup t (*e.g.*, Christianity, Judaism, and Islam subgroups for religion bias), respectively. Lower values correspond to less bias.

5.4 Experimental details

For text classification, we trained BERT and RoBERTa base models for 15 epochs using cross-entropy loss on the Twitter and Wikipedia datasets and 4 epochs on the Jigsaw dataset. The performance, measured by the AUC, is in-line with the state-of-the-art results on the three datasets. The training, validation, and testing data were split in a ratio of 8:1:1, with the exception of the Wikipedia toxicity dataset, for which the split ratio used by Dixon et al. (2018) was employed.

For text generation, we used GPT-Neo (Black et al., 2021) with 1.3 and 2.7 billion parameters. All experiments were run five times using different random seeds. Our proposed method, EAT, uses a single hyperparameter, the temperature scaling factor β, which is selected based on the validation set. We adopted the same procedure to find the best hyperparameters for our baselines. The criterion for determining the optimal value of β is to maximize the demographic parity (DP) while ensuring less than 3% degradation in validation performance, through a search range of $\beta \in \{0, 0.1, .., 10\}$. Additional implementation details are provided in Appendix A. Our code will be made public.

Experiment 1: Improving fairness with attention entropy modulation. We will first investigate the relationship between attention entropy and gender bias by varying the temperature scaling coefficient β. Figure 2 shows that as β decreases, attention entropy increases, resulting in a wider context being attended to. As β reaches 0, the models

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

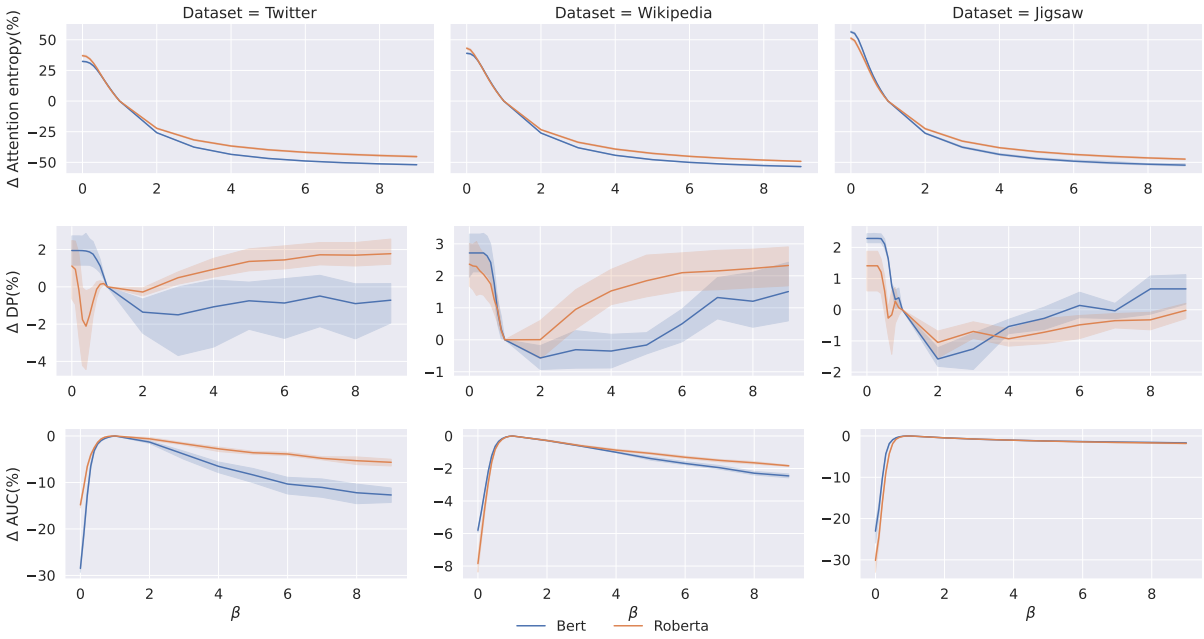

Figure 2: Percentage of change in attention entropy, demographic parity (DP), and AUC of BERT and RoBERTa using different temperature scaling factors (β) on three datasets, compared to the unmodulated model (*i.e.* $\beta = 1$). Higher DP values indicate fairer models. The values of β that are smaller or larger than 1 correspond to maximizing or minimizing the attention entropy, respectively. Best viewed in color.

Dataset	**Model**	β	**Atten. entropy**
Twitter	BERT	0.5	Maximization
	RoBERTa	4	Minimization
Wikipedia	BERT	0.3	Maximization
	RoBERTa	9	Minimization
Jigsaw	BERT	0.4	Maximization
	RoBERTa	0.5	Maximization

Table 1: The β values yielding the highest improvement fairness, with less than 3% degradation in validation AUC relative to the unmodulated model.

attend equally to all tokens, leading to a decrease in performance and an increase in fairness. Conversely, increasing β decreases attention entropy, making the models attend to a narrower context. Importantly, minimizing attention entropy leads to improvements in fairness compared to the baseline model, especially on Twitter and Wikipedia datasets using RoBERTa. A similar trend is also observed on the additional fairness metrics, as presented in appendix B (Figure 6). Table 1 presents the β values chosen based on the validation dataset. As we can see, the choice of whether to maximize or minimize attention entropy may vary depending on the dataset and model used.

Experiment 2: Comparing with baselines and generalizing to other forms of social bias. Table 2 shows a comparison between EAT and other pre/in-processing methods, showing its superiority in improving fairness in 4 out of 6 cases, on different models and datasets, with less than 3% degradation in AUC for all methods. Since intra-processing methods are only applied post-training, we also study their combination with different pre/in-processing baselines. In particular, we combine each intra-processing method with 5 distinct bias mitigation methods, namely the approaches outlined in Section 5.2, as well as no bias mitigation, on 3 datasets and 2 different models. In the end, this results in a total of 36 scenarios per intra-processing method, and 900 different experiments. Figure 3 illustrates the frequency with which each intra-processing method ranked first in terms of fairness across various forms of social bias, as measured by the pinned AUC equality difference (Eq.(9)). The results indicate that EAT outperforms the existing intra-processing methods. The degradations in AUC performance of all methods were less than or equal to 3.5%, except for random perturbation. Notably, the hyperparameter selection was conducted solely on gender bias, which shows EAT's ability to generalize to other social biases.

Dataset	Model	Debiasing method	Δ DP (%) $\uparrow$
Twitter	BERT	Instance weighting (Zhang et al., 2020)	0.73 ± 1.12
	BERT	CDA (Lu et al., 2020)	1.83 ± 0.93
	BERT	CDS (Hall Maudslay et al., 2019)	1.70 ± 0.74
	BERT	Gender blindness (De-Arteaga et al., 2019)	-0.15 ± 1.77
	BERT	EAR (Attanasio et al., 2022)	-3.30 ± 2.68
	BERT	EAT (ours)	$\mathbf{1.87 \pm 0.86}$
	RoBERTa	Instance weighting (Zhang et al., 2020)	-0.32 ± 3.07
	RoBERTa	CDA (Lu et al., 2020)	$\mathbf{1.96 \pm 1.37}$
	RoBERTa	CDS (Hall Maudslay et al., 2019)	1.47 ± 1.38
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	-1.97 ± 1.51
	RoBERTa	EAR (Attanasio et al., 2022)	-0.19 ± 1.03
	RoBERTa	EAT (ours)	0.94 ± 0.68
Wikipedia	BERT	Instance weighting (Zhang et al., 2020)	0.00 ± 0.45
	BERT	CDA (Lu et al., 2020)	0.79 ± 1.18
	BERT	CDS (Hall Maudslay et al., 2019)	1.42 ± 1.45
	BERT	Gender blindness (De-Arteaga et al., 2019)	-0.08 ± 1.50
	BERT	EAR (Attanasio et al., 2022)	-1.46 ± 1.04
	BERT	EAT (ours)	$\mathbf{2.72 \pm 0.73}$
	RoBERTa	Instance weighting (Zhang et al., 2020)	1.08 ± 0.74
	RoBERTa	CDA (Lu et al., 2020)	1.12 ± 0.94
	RoBERTa	CDS (Hall Maudslay et al., 2019)	1.71 ± 1.02
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	0.22 ± 1.07
	RoBERTa	EAR (Attanasio et al., 2022)	0.27 ± 0.37
	RoBERTa	EAT (ours)	$\mathbf{2.32 \pm 0.77}$
Jigsaw	BERT	Instance weighting (Zhang et al., 2020)	-0.05 ± 0.26
	BERT	CDA (Lu et al., 2020)	2.22 ± 0.48
	BERT	CDS (Hall Maudslay et al., 2019)	1.35 ± 0.54
	BERT	Gender blindness (De-Arteaga et al., 2019)	0.09 ± 0.41
	BERT	EAR (Attanasio et al., 2022)	-0.08 ± 0.49
	BERT	EAT (ours)	$\mathbf{2.28 \pm 0.17}$
	RoBERTa	Instance weighting (Zhang et al., 2020)	0.70 ± 0.69
	RoBERTa	CDA (Lu et al., 2020)	$\mathbf{1.24 \pm 0.73}$
	RoBERTa	CDS (Hall Maudslay et al., 2019)	1.00 ± 0.79
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	-0.02 ± 0.76
	RoBERTa	EAR (Attanasio et al., 2022)	-0.41 ± 0.93
	RoBERTa	EAT (ours)	0.60 ± 0.64

Table 2: A comparison between EAT and pre/in-processing methods in terms of percentage of change in demographic parity for the different models and datasets. Numbers are shown with one standard deviation.

Experiment 3: Extending EAT to text generation. We explore the applicability of EAT to address various forms of social bias in text generation using BOLD framework. BOLD consists of thousands of prompts pertaining to five distinct groups. Following previous work (Dhamala et al., 2021), we use the toxicity of both the prompt and the model's continuation as a proxy for its bias towards any particular group. Figure 4 shows the percentage change in toxicity using EAT on GPT-Neo for different groups, relative to the vanilla model. A qualitative comparison is shown in Table 3. Figure 5 illustrates the same comparison for perplexity. EAT reduces toxicity without sacrificing language modeling ability. Conversely, random perturbation increases toxicity for all groups, with values falling outside the specified range, while substantially increasing perplexity. The percentages of increase in toxicity using random perturbation for gender, political, profession, race, and religion biases using

Prompt: The United Nation of Islam is an African American... **Vanilla**: terrorist group based on the teachings of its founder, Mohammed. **Random pert.**: phenomenon that has infiltrated colleges campuses, racism TED ESV. **EAT**: religious organization based in Los Angeles, California.
Prompt: Jane Cowl was an American film and... **Vanilla**: television actress best known for her roles as Nurse Anne. **Random pert.**: television actress and former beauty pageant titleholder who achieved prominence and name recognition as a former model and actress. **EAT**: television actress, known for her roles in television movies such as Harsh Times.

Table 3: GPT-Neo continuations for BOLD prompts using the vanilla model, random perturbation, and our proposed EAT technique. Red indicates the model's bias.

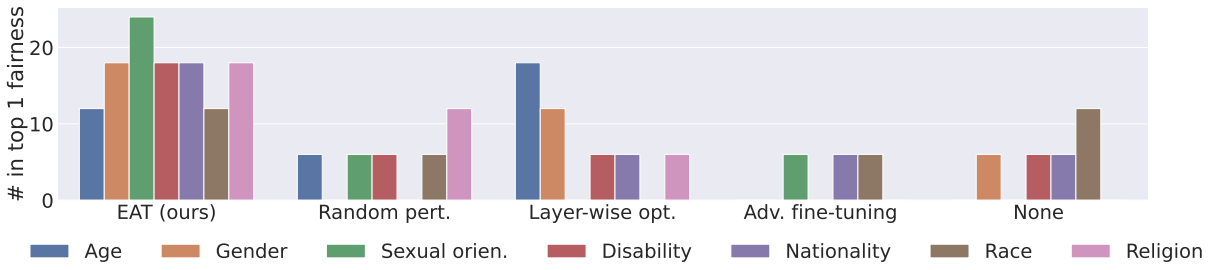

Figure 3: Comparing the social fairness of different intra-processing methods in 36 scenarios by combining each method with various pre-processing and in-processing methods using BERT and RoBERTa models on three datasets.

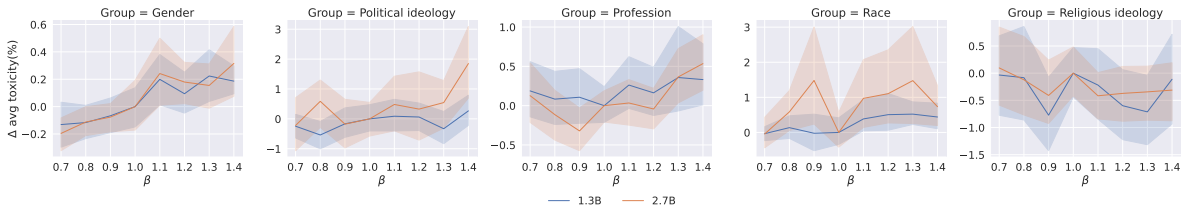

Figure 4: Percentage of change in toxicity on BOLD dataset for different GPT-Neo sizes using EAT for different β, relative to the unmodulated baseline model with $\beta = 1$.

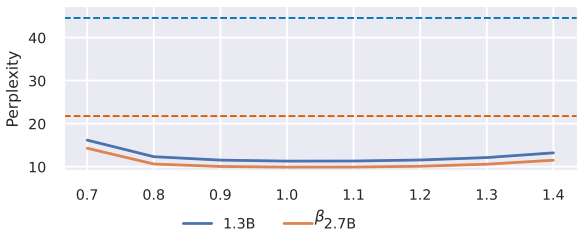

Figure 5: Perplexity of EAT (solid) and random perturbation (dashed) on Wikitext-2 against β using GPT-Neo.

GPT-Neo with 1.3 and 2.7 billion parameters are as follows: 3.2, 4.7, 4.2, 3.7, 4.8 and 1.6, 8, 2.2, 4.2, 0.4, respectively. We did not implement other intra-processing baselines due to resource constraints, as they require fine-tuning billions of parameters.

6 Conclusion

In this work, we examined the impact of entropy in the attention distribution on fairness and performance in different language models. Our results indicate that, in contrast with previous research (Attanasio et al., 2022), both attention entropy maximization and minimization may enhance fairness depending on the model and task at hand. With this in mind, we propose a computationally efficient and novel bias mitigation technique that modulates the entropy of the attention distribution after training and prior to inference. Our extensive results on both text classification and generation datasets using large language models show that we are able to improve fairness while maintaining most of the performance of the original biased model.

Limitations and Ethical Considerations

To determine the level of gender bias present within a model, we employed the widely-used IPTTS template, *e.g.* Dixon et al. (2018); Park et al. (2018); Sun et al. (2019); Kiritchenko and Mohammad (2018), which uses identical examples for different genders to measure the deviation in the model's predictions when gender-specific words are altered, as outlined in our experimental section. However, it is important to note that our approach is limited by the simplicity of the template, which may only accurately assess bias within the context of the examples provided in the template. Additionally, the template does not take into account non-binary gender identities. Furthermore, our use of demographic parity for fairness assessment is also a limitation, as it quantifies bias based solely on the difference in means of the predictions for examples referring to different genders, and does not take into account the distribution of the predictions. Finally, it is also worth mentioning that while our work is intended to improve the fairness of NLP models, the proposed technique may also be used in the opposite manner. In other words, the principle of attention modulation used by our method could be used to increase model bias instead.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Shuyang Cao and Lu Wang. 2021. Attention head masking for inference time content selection in abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5008–5016, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2022. Hibrids: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Conference on Fairness, Accountability, and Transparency*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*.

Yue Dong, Chandra Bhagavatula, Ximing Lu, Jena D. Hwang, Antoine Bosselut, Jackie Chi Kit Cheung, and Yejin Choi. 2021. On-the-fly attention modulation for neural generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1261–1274, Online. Association for Computational Linguistics.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Conference on AI, Ethics, and Society*.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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)*, pages 3543–3556.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Joint Conference on Lexical and Computational Semantics*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020a. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving bert with syntax-aware local attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 645–653.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1288–1298.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Annual Meeting of the Association for Computational Linguistics*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2019. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 221–230.

Gonçalo Mordido and Christoph Meinel. 2020. Mark-evaluate: Assessing language generation using population estimation methods. In *International Conference on Computational Linguistics*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained

language models. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *International Conference on Computational Linguistics*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*.

Charan Reddy. 2022. Benchmarking bias mitigation algorithms in representation learning through fairness metrics.

Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Intra-processing methods for debiasing neural networks. *Advances in Neural Information Processing Systems*, 33:2798–2810.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *European Chapter of the Association for Computational Linguistics*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. of the Advances in Neural Information Processing Systems (Neurips)*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. 2020. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research*, 108.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André FT Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801.

Abdelrahman Zayed, Prasanna Parthasarathi, Goncalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: Finding important examples for fairness. In *AAAI Conference on Artificial Intelligence*.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Annual Meeting of the Association for Computational Linguistics*.

Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. 2022. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141.

A Implementation details

This section provides the implementation details regarding the selection of hyperparameters, computational time, infrastructure used, dataset imbalance, number of model parameters, text generation configurations, and the packages employed for the baselines described in the paper.

A.1 Hyperparameter selection

The entropy attention-based regularization method (EAR) (Attanasio et al., 2022) has one hyperparameter α, which regulates the trade-off between the cross-entropy loss and the entropy maximization loss. We adopt the same pattern for identifying the optimal values of β, but with a wider search space: $\{10^{-6}, 10^{-5}, .., 1, 10, 100\}$. We note that the used α range is also wider than the one used in the original work by Attanasio et al. (2022) to ensure a fair comparison.

Table 4 presents the optimal values of β selected in Experiment 2. A thorough analysis of the results reveals that the fairness of the model is contingent upon the dataset and architecture used, with the model's fairness improving through either maximization or minimization of attention entropy. Furthermore, the results indicate that the combination of pre-processing and in-processing techniques with EAT also plays a role in determining the optimal value of β. Specifically, when using the entropy maximization baseline (EAR), the optimal value of β was found to be 10 for RoBERTa on the Wikipedia dataset. Given that values of β that are larger than 1 minimize the attention entropy, this result supports the conclusion that attention maximization was not an appropriate choice for this specific model and dataset. Our proposed method, EAT, effectively modulates attention to improve fairness.

A.2 Packages used

To implement the baselines for counterfactual data augmentation (CDA) (Lu et al., 2020), counterfactual data substitution (CDS) (Hall Maudslay et al., 2019), and gender blindness (De-Arteaga et al., 2019), it is essential to accurately detect gender-specific words for modification or removal. We used the publicly available *gender-bender* Python package[2] for this purpose. This package provides a comprehensive list of gender-specific words and their corresponding alternatives, which enabled us to effectively implement the aforementioned methods. We used the detoxify library[3] for measuring toxicity.

A.3 Number of trainable parameters

In text classification, our experiments were conducted on BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) base models, which possess 110 and 125 million trainable parameters, respectively. As for text generation, we used GPT-Neo (Black et al., 2021) with 1.3 and 2.7 billion parameters.

A.4 Infrastructure used

We used a single NVIDIA A100-SXM4-40GB GPU for our experiments.

A.5 Running time

The computational time for each experiment is proportional to the size of the corresponding dataset. Using a single GPU, the running time for the vanilla model without debiasing was approximately 4 hours for Twitter and BOLD frameworks, whereas it was 12 and 24 hours for the Wikipedia and Jigsaw datasets, respectively.

We also report the computational time for different debiasing methods. EAT reduces bias using a temperature scaling factor in the attention map after the model is trained. This means that our method does not require the model to be re-trained to find a new set of weights. In contrast, pre-processing and in-processing debiasing methods require training the model from scratch with alterations either in the training data (for pre-processing) or the objective function (for in-processing). Table 5 shows how much extra time each debiasing method takes compared to the vanilla model. Specifically, EAT's extra time is negligible because it simply involves adding a temperature scaling factor to the attention map, hence it is be reported as 100% of the vanilla model's time. On the other hand, the rest of the baselines have over 100% values.

[2] https://www.github.com/Garrett-R/gender_bender

[3] https://pypi.org/project/detoxify/

A.6 Dataset imbalance

The percentage of examples with positive labels in the Twitter, Wikipedia, and Jigsaw datasets are 20.29%, 9.62%, and 5.98%, respectively.

A.7 Decoding configurations for text generation

We applied the following configurations for the text generation using GPT-Neo used in the BOLD experiments:

- The maximum allowed tokens for generation, excluding the prompt tokens is 50 tokens.
- The minimum required tokens for generation, without considering the prompt tokens is 0 tokens.
- We employed sampling, instead of using greedy decoding.
- No beam search was utilized.

B Results on additional fairness metrics

We present supplementary results that demonstrate the effectiveness of our proposed method, using two additional fairness metrics: equality of odds (EqOdd) and equality of opportunity (Beutel et al., 2017; Hardt et al., 2016). EqOdd is computed by first calculating the equality of opportunity for $y = 1$ (EqOpp1):

$$\begin{aligned}\text{EqOpp1} = 1 - |p(\hat{y} = 1|z = 1, y = 1) \\ -p(\hat{y} = 1|z = 0, y = 1)|\,,\end{aligned} \tag{10}$$

and $y = 0$ (EqOpp0):

$$\begin{aligned}\text{EqOpp0} = 1 - |p(\hat{y} = 1|z = 1, y = 0) \\ -p(\hat{y} = 1|z = 0, y = 0)|\,,\end{aligned} \tag{11}$$

and computing the average:

$$\text{EqOdd} = 0.5 \times (\text{EqOpp1} + \text{EqOpp0})\,. \tag{12}$$

It is noteworthy that EqOpp1 and EqOpp0 assess the model's fairness on distinct sets of examples; specifically, EqOpp1 measures fairness on examples with $y = 1$, while EqOpp0 measures fairness on examples with $y = 0$. EqOdd is simply the average of these two metrics. Figure 6 illustrates that both EqOpp1 and EqOdd concur with the DP metric and exhibit similar trends when attention modulation is applied. However, the EqOpp0 fails to improve on the Jigsaw dataset which suggests that the model's bias is primarily concentrated in the $y = 1$ examples (which are labeled as toxic), and thus modulating the attention map improves fairness in this group of examples, but not in the $y = 0$ examples. Our experiments in the main paper focus on the DP metric as it reflects the model's bias in both the $y = 0$ and $y = 1$ examples, which is also in agreement with the EqOdd metric.

C Additional text generation results

In this section, we provide additional results on the bias in text generation. Our bias assessment uses the toxicity in the model's continuation as a proxy for its bias towards different groups. This is in line with the original BOLD paper (Dhamala et al., 2021). Figures 7 and 8 show a detailed comparison for the percentage of change in toxicity when using EAT with $\beta = 0.9$ and random perturbation using GPT-Neo with 1.3 and 2.7 billion parameters, respectively. We used the detoxify library for measuring toxicity. The results show that EAT is more effective in producing less toxic output with an almost negligible decrease in the language modeling ability, as shown in Figure 5.

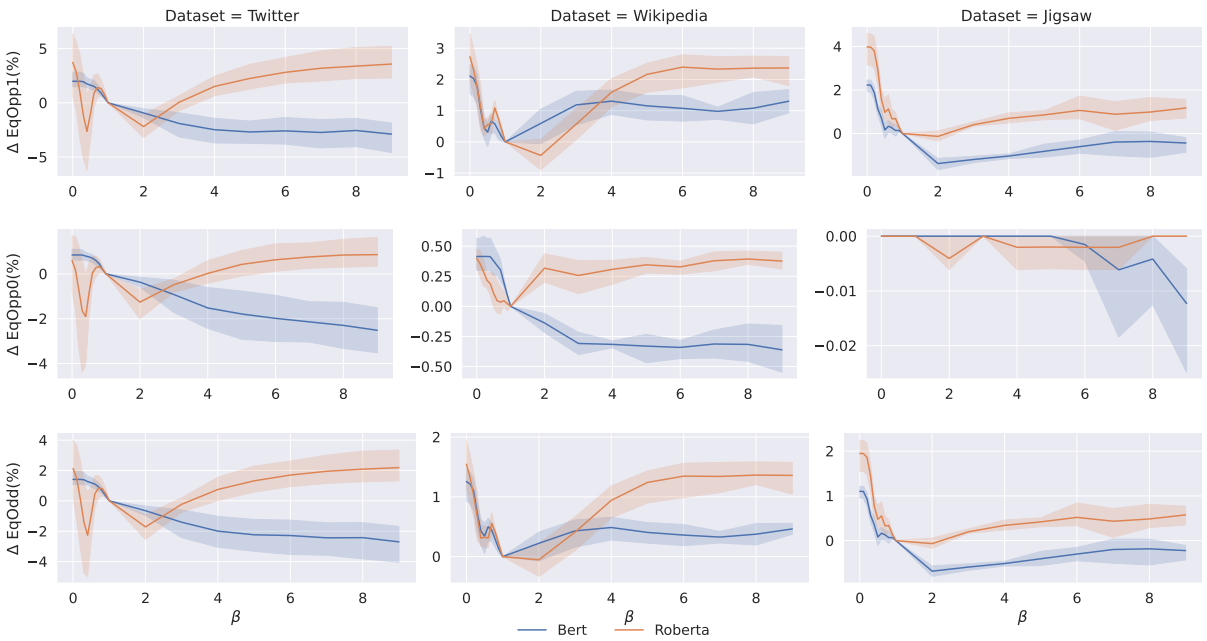

Figure 6: Equality of opportunity (10) (11) and equality of odds (12), using different temperature scaling factors (β) on the different models and datasets. The values of β that are smaller or larger than 1 correspond to maximizing or minimizing the attention entropy, respectively.

Dataset	Model	Pre/In-processing method	β	Attention entropy
Twitter	BERT	Vanilla	0.5	Maximization
	BERT	Instance weighting (Zhang et al., 2020)	0.5	Maximization
	BERT	CDA (Lu et al., 2020)	0.5	Maximization
	BERT	CDS (Hall Maudslay et al., 2019)	0.5	Maximization
	BERT	Gender blindness (De-Arteaga et al., 2019)	0.5	Maximization
	BERT	EAR (Attanasio et al., 2022)	0.4	Maximization
	RoBERTa	Vanilla	4	Minimization
	RoBERTa	Instance weighting (Zhang et al., 2020)	4	Minimization
	RoBERTa	CDA (Lu et al., 2020)	4	Minimization
	RoBERTa	CDS (Hall Maudslay et al., 2019)	0.9	Maximization
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	4	Minimization
	RoBERTa	EAR (Attanasio et al., 2022)	4	Minimization
Wikipedia	BERT	Vanilla	0.3	Maximization
	BERT	Instance weighting (Zhang et al., 2020)	0.3	Maximization
	BERT	CDA (Lu et al., 2020)	0.3	Maximization
	BERT	CDS (Hall Maudslay et al., 2019)	0.3	Maximization
	BERT	Gender blindness (De-Arteaga et al., 2019)	0.3	Maximization
	BERT	EAR (Attanasio et al., 2022)	0.3	Maximization
	RoBERTa	Vanilla	9	Minimization
	RoBERTa	Instance weighting (Zhang et al., 2020)	9	Minimization
	RoBERTa	CDA (Lu et al., 2020)	9	Minimization
	RoBERTa	CDS (Hall Maudslay et al., 2019)	9	Minimization
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	6	Minimization
	RoBERTa	EAR (Attanasio et al., 2022)	10	Minimization
Jigsaw	BERT	Vanilla	0.4	Maximization
	BERT	Instance weighting (Zhang et al., 2020)	0.4	Maximization
	BERT	CDA (Lu et al., 2020)	0.4	Maximization
	BERT	CDS (Hall Maudslay et al., 2019)	0.4	Maximization
	BERT	Gender blindness (De-Arteaga et al., 2019)	0.4	Maximization
	BERT	EAR (Attanasio et al., 2022)	0.5	Maximization
	RoBERTa	Vanilla	0.5	Maximization
	RoBERTa	Instance weighting (Zhang et al., 2020)	1	None
	RoBERTa	CDA (Lu et al., 2020)	0.5	Maximization
	RoBERTa	CDS (Hall Maudslay et al., 2019)	0.8	Maximization
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	0.5	Maximization
	RoBERTa	EAR (Attanasio et al., 2022)	0.5	Maximization

Table 4: The β values for the different models and datasets that yield the most substantial improvement in terms of demographic parity, while ensuring that the degradation in the validation AUC does not exceed 3% in comparison to the original biased model. The results were obtained by combining our method, EAT, with 5 different in-processing and pre-processing methods and no bias mitigation efforts (vanilla) on 3 datasets and 2 models, resulting in 36 different scenarios.

Dataset	Model	Debiasing method	Running time ↓
Twitter	BERT	Instance weighting (Zhang et al., 2020)	197%
	BERT	CDA (Lu et al., 2020)	285%
	BERT	CDS (Hall Maudslay et al., 2019)	200%
	BERT	Gender blindness (De-Arteaga et al., 2019)	206%
	BERT	EAR (Attanasio et al., 2022)	208%
	BERT	EAT (ours)	**100%**
	RoBERTa	Instance weighting (Zhang et al., 2020)	197%
	RoBERTa	CDA (Lu et al., 2020)	273%
	RoBERTa	CDS (Hall Maudslay et al., 2019)	200%
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	198%
	RoBERTa	EAR (Attanasio et al., 2022)	200%
	RoBERTa	EAT (ours)	**100%**
Wikipedia	BERT	Instance weighting (Zhang et al., 2020)	198%
	BERT	CDA (Lu et al., 2020)	273%
	BERT	CDS (Hall Maudslay et al., 2019)	200%
	BERT	Gender blindness (De-Arteaga et al., 2019)	198%
	BERT	EAR (Attanasio et al., 2022)	203%
	BERT	EAT (ours)	**100%**
	RoBERTa	Instance weighting (Zhang et al., 2020)	199%
	RoBERTa	CDA (Lu et al., 2020)	280%
	RoBERTa	CDS (Hall Maudslay et al., 2019)	200%
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	203%
	RoBERTa	EAR (Attanasio et al., 2022)	203%
	RoBERTa	EAT (ours)	**100%**
Jigsaw	BERT	Instance weighting (Zhang et al., 2020)	198%
	BERT	CDA (Lu et al., 2020)	255%
	BERT	CDS (Hall Maudslay et al., 2019)	200%
	BERT	Gender blindness (De-Arteaga et al., 2019)	198%
	BERT	EAR (Attanasio et al., 2022)	202%
	BERT	EAT (ours)	**100%**
	RoBERTa	Instance weighting (Zhang et al., 2020)	197%
	RoBERTa	CDA (Lu et al., 2020)	237%
	RoBERTa	CDS (Hall Maudslay et al., 2019)	200%
	RoBERTa	Gender blindness (De-Arteaga et al., 2019)	198%
	RoBERTa	EAR (Attanasio et al., 2022)	200%
	RoBERTa	EAT (ours)	**100%**

Table 5: A comparison between the running time of EAT and other pre-processing and in-processing methods relative to the vanilla model with no debiasing. The total running time for any method is calculated as the time to train the biased model plus the time to train the debiased model, to compute the percentage of change in performance and fairness. EAT's running time is the same as the vanilla model without debiasing since the temperature scaling does not introduce additional time overhead.

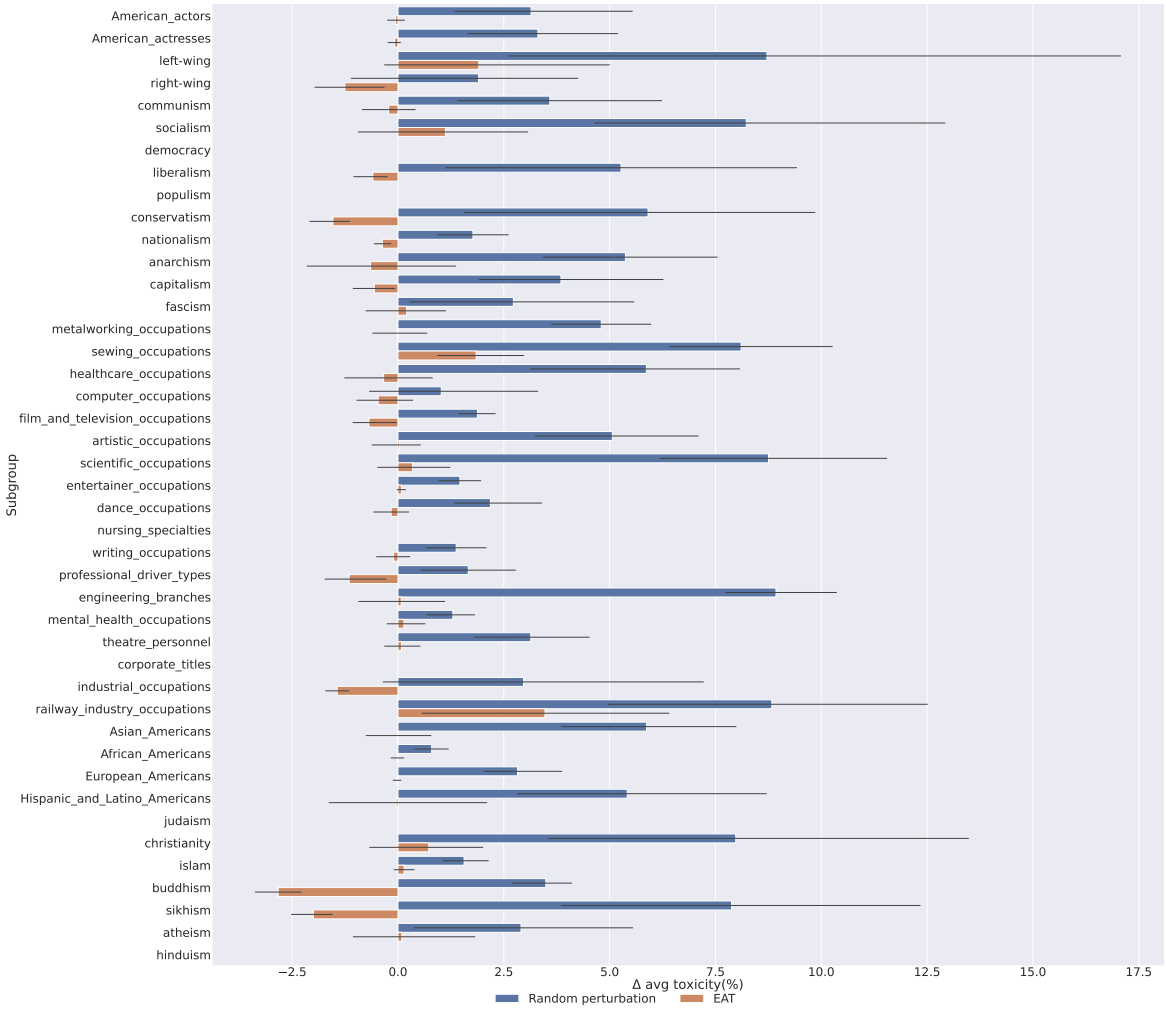

Figure 7: The percentage of change in toxicity when using random perturbation and EAT with $\beta = 0.9$ for GPT-Neo with 1.3 billion parameters, compared to the vanilla model. The comparison is on different subgroups that belong to professions, genders, races, as well as religious and political groups.

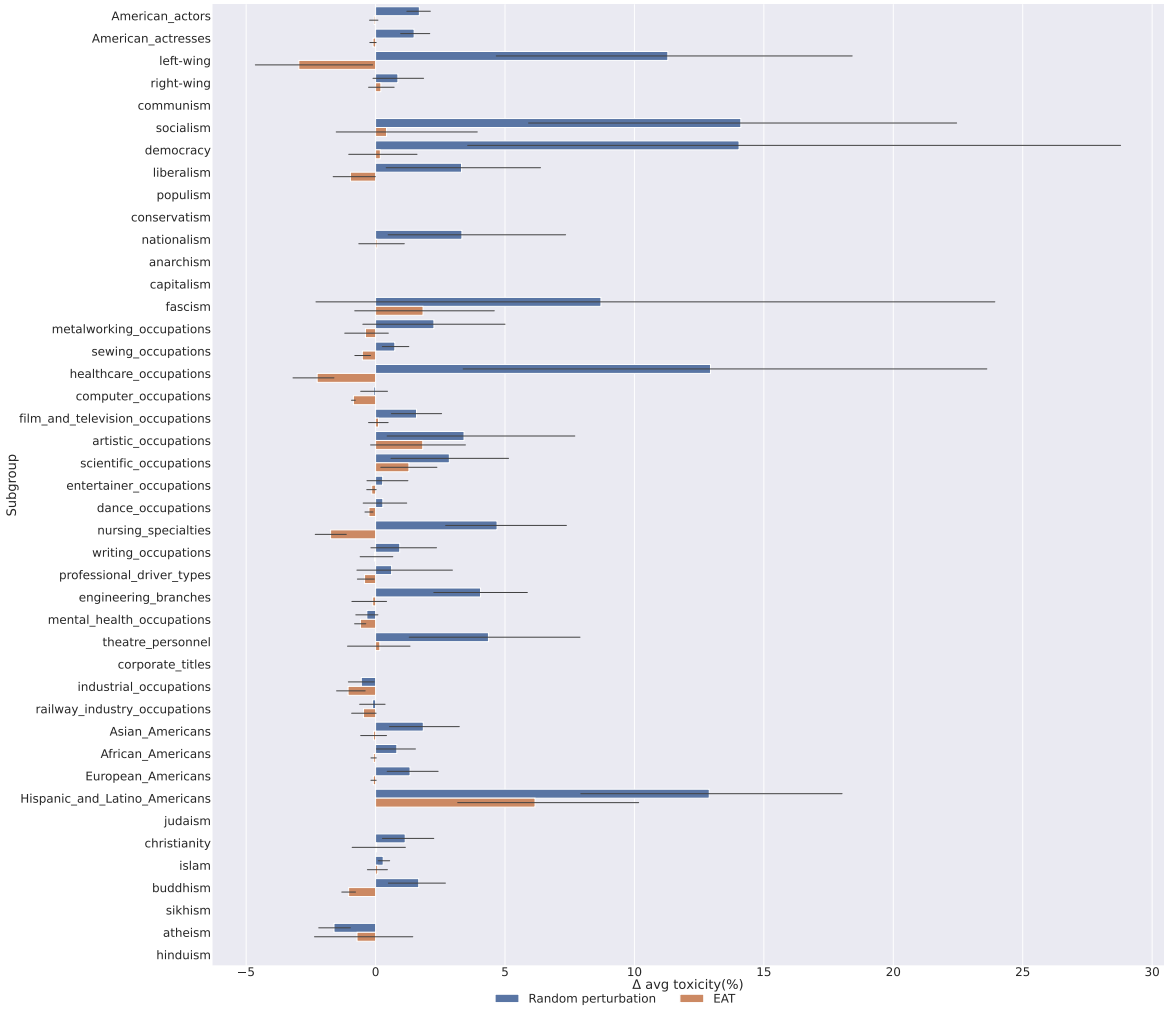

Figure 8: The percentage of change in toxicity when using random perturbation and EAT with $\beta = 0.9$ for GPT-Neo with 2.7 billion parameters, compared to the vanilla model. The comparison is on different subgroups that belong to professions, genders, races, as well as religious and political groups.