A Multi-tiered Approach to Debiasing Language Models

Stanford CS224N Custom Project

Aman Kansal
Department of Computer Science
Stanford University
amkansal@stanford.edu

Saanvi Chawla
Department of Computer Science
Stanford University
saanvic@stanford.edu

Shreya Shankar
Department of Electrical Engineering
Stanford University
shreya07@stanford.edu

Abstract

This research investigates the effectiveness of various debiasing methods applied to the BERT language model, with the goal of diminishing stereotypical biases while maintaining the model's performance on downstream tasks. The focus is on mitigating biases related to gender, race, and religion by exploring and refining techniques such as Counterfactual Data Augmentation (CDA) and SentenceDebias (SDB), alongside assessments through benchmarks like StereoSet and CrowS-Pairs. The implementation encompasses both original and modified versions of CDA and SDB, facilitating a comprehensive comparison between models that adjust weights for debiasing and those employing test-time surgical interventions. This study presents a novel approach to debiasing, where, instead of merely eliminating bias components from pre-softmax BERT embeddings, it aims to preserve information by equalizing embedding components across different axes within the bias subspace (e.g., balancing the representation of "man" and "woman"). Through detailed experimentation, our findings reveal that training the BERT model to address multiple biases simultaneously not only enhances debiasing effectiveness but also establishes a positive interrelation among various types of biases. By juxtaposing weight modification methods with surgical debiasing approaches, this research offers insightful perspectives on optimizing debiasing techniques without sacrificing linguistic comprehension or task performance. The proposed information-preserving SDB method signifies a significant advancement in debiasing strategies, promoting a more equitable representation in language models. This work contributes to the ongoing discourse on ethical AI, demonstrating practical steps toward the development of bias-aware, high-performing language models.

Mentor: Kaylee Burns

1 Introduction

Large pre-trained language models have consistently been achieving state of the art performance in tasks related to Natural Language Processing. The source of training data for these models comes mainly from unmoderated sources like the Internet. However, the content of these sources exhibit bias in various forms, and models have been found to capture many of the **social biases** present in their training data (May et al., 2019). Out of these, **gender bias** (Bolukbasi et al., 2016) has been widely studied. There has been an effort to develop research techniques to mitigate these biases, however, previous work has had somewhat of a narrow scope, focusing only on one type of social bias. Our work aims to target three widely prevalent kinds of social biases- **gender bias**, **racial biases**, and **religious biases**.

The are broadly five types of de-biasing techniques - **Counterfactual Data Augmentation**, **Dropout**, **Self-Debias**, **SentenceDebias** and **Iterative Nullspace Projection**. The popular benchmarks they are evaluated against include: **Sentence Encoder Association Test** (May et al., 2019), **StereoSet** (Nadeem et al., 2020a), and **Crowdsourced Stereotype Pairs** (Nangia et al., 2020a). In our approaches we work choose to further develop **Counterfactual Data Augmentation** and **SentenceDebias**. We chose these as they work the best with the **BERT** model. Our chosen benchmarks to evaluate the models are **Stereoset** and **CrowS Pairs**.

Conventionally, **Counterfactual Data Augmentation** (CDA) (Maudslay et al., 2019) uses a predefined list of words for debiasing. However, this list i snot exhaustive and cannot account for every possible word pair. In our refined approach, we use **GPT-4** (et al, 2023) to generate **stereotypical-antistereotypical** word pairs based on identified biased words in the training data. Then, we train on a training set consisting of original examples as well as examples where the biased word is substituted with the antistereotypical word in the pair. We only replace **one biased word per sentence** at a time, sampling from the possible choices randomly in order to maintain the frequency of representation of each sentence in the training set.

For **SentenceDebias (SDB)** (Liang et al., 2020), the conventional approach uses an algorithm that neutralizes word embeddings of biased words. We have two new approaches that build on top of this. Out first approach stems from our acknowledgement of the fact that a sentence could display **multiple biases** at the same time. In such cases, our model figures out the most prevalent bias in the sentence by examining projections on each of three subspaces, i.e, **racial bias**, **gender bias**, and **religious bias**. Then, the model **neutralizes** this type of bias in the sentence.

Neutralization of biased words in sentences is what has conventionally been adopted, but this might lead to a **loss is crucial information** associated with biased words. In order to account for this, in our second approach, we introduce a novel method with mathematical underpinnings to **equalize bias** in sentence debiasing. Our method involves changing the embeddings of identified biased word vectors to ensure that they have an equal projection on both stereotypical and antistereotypical biased subspaces.

Our work fundamentally addresses the ethical dimension of language representations by seeking to remove or mitigate biases present in these representations. This is crucial for designing systems that process human languages in a way that is fair, inclusive, and reflective of the diversity in human society. The paper contributes to the discussion on how to represent language by tackling the biases inherent in pre-trained language models. It builds on top of other work done towards debiasing and includes key considerations such as the prevalence of multiple types of bias in a single sentence, as well as the acknowledgement of other meaning associated with identified biased words. Our work addresses each of these considerations and proposes novel methods to mitigate them.

2 Related Work

The endeavor to mitigate biases in language models has yielded diverse methodologies, focusing on biases related to gender, race, religion, and more. These methodologies can be broadly categorized into strategies targeting the training data, modifying model parameters, leveraging word embeddings, and employing post hoc debiasing techniques.

- Data-Centric Debiasing: A prevalent approach involves the debiasing of training corpora. Counterfactual Data Augmentation (CDA) (Maudslay et al., 2019) exemplifies this strategy by rebalancing the corpus to replace or alter biased words, thereby facilitating the training of models on less biased data.
- Model Parameter Modifications: Another significant line of inquiry examines the modification of model parameters to reduce bias. Techniques such as dropout regularization(Srivastava et al., 2014) have been explored for this purpose. Specifically, researchers have investigated increasing the dropout rates applied to the attention weights and hidden activations of models like BERT and ALBERT during an additional phase of pre-training, aiming to attenuate bias through enhanced generalization.

- Representation-Based Debiasing: Techniques such as SentenceDebias(Liang et al., 2020) fall under this category. SentenceDebias is a projection-based method that necessitates the identification of a linear subspace associated with a particular bias. By projecting sentence representations onto this bias subspace and subtracting the projection from the original representations, the technique achieves debiasing at the sentence level. Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020), a debiasing technique akin to SentenceDebias, debiases model representations by training a linear classifier to identify protected attributes (e.g., gender), then removing this information by projecting the representations into the classifier's weight matrix nullspace. This iterative process effectively strips away bias.
- Post Hoc Debiasing Techniques: Recent advancements have introduced post hoc debiasing strategies, such as Self-Debias (Schick et al., 2021), which does not modify a model's internal representations or parameters. Instead, Self-Debias leverages the model's inherent knowledge to prevent the generation of biased text, offering a novel avenue for mitigating bias without altering the underlying model architecture.

Our project focuses on Counterfactual Data Augmentation (CDA) and SentenceDebias methodologies for debiasing language models, specifically BERT. While regularization methods such as dropout offer potential benefits, they appeared less compelling for in-depth exploration in our research context. Additionally, despite Self-Debias being the cutting-edge in debiasing technology, its application is constrained by the significant computational resources it demands, particularly because it is designed for use with models like GPT.

3 Approach

We have tried six different approaches to debias language models. The CDA approach was our baseline and all approaches except vanilla CDA and SDB are original i.e. proposed and implemented by us:

- **Counter-factual Data Augmentation (CDA):** CDA (Maudslay et al., 2019) aims to enhance model performance by retraining it with a more balanced dataset. The conventional method uses a predefined list to detect and replace bias-indicative words from sentences.
- **Counter-factual Data Augmentation (CDA) + GPT:** Our refined approach build on the standard approach and uses GPT (et al, 2023) to create an advanced dictionary to identify biased words in the dataset and generate pairs or triplets. This is more inclusive and flexible, addressing a broader spectrum of biases without the constraints of traditional rebalancing techniques.
- **Counter-factual Data Augmentation (CDA) Unified:** CDA Unified methodically debiases BERT by applying Counter-factual Data Augmentation (CDA) in a three-step process, each targeting a different bias— gender, race, and religion, sequentially. This approach refines the model's understanding and representation by iteratively correcting for specific biases.
- **Sentence Debias (SDB):** SENT-DEBIAS Liang et al. (2020) operates through a four-step process: First, it identifies words that demonstrate bias attributes. Second, it embeds these biased words within sentences to contextualize them, creating biased sentence representations. Third, it calculates the bias subspace of these sentence representations. Lastly, it debiases general sentences by subtracting their projection onto this identified bias subspace. Let us denote the sentence representation to be B_W. For a single type of bias (say, gender) we denote the bias subspace as S_G.

$$S_G = PCA\left(B_{W_{G_1}}, B_{W_{G_2}}\right)$$
$$B_W \rightarrow B_W - B_W \cdot S_G$$

- **Sentence Debias Equalized:** Let us consider gender bias, our 'equalize' approach introduces separate subspaces for G_1 (male) and G_2 (female), denoted as S_{G1} and S_{G2}. For a representation B_W, we compute its components in these subspaces as $m = S_{G1} \cdot B_W$ and $f = S_{G2} \cdot B_W$. Instead of removing these components, we adjust them to equalize the bias: $m' = \hat{m} \cdot \frac{|m|+|f|}{2}$ and $f' = \hat{f} \cdot \frac{|m|+|f|}{2}$. $\hat{m}$ and $\hat{f}$ are the unit vectors in the directions of m and f respectively. This method aims to balance the bias by adjusting the sentence embedding's components towards an equal representation of gender biases.

$$S_{G_i} = PCA\left(B_{W_{Gi}}\right)$$
$$m = S_{G_1} \cdot B_W$$
$$f = S_{G_2} \cdot B_W$$
$$m' = \hat{m} \cdot \frac{|m| + |f|}{2}$$
$$f' = \hat{f} \cdot \frac{|m| + |f|}{2}$$
$$B_W \rightarrow B_W - m - f + m' + f'$$

- **Sentence Debias Unified:** In this approach, we delineate three distinct bias subspaces, each corresponding to a specific type of bias. During the debiasing process for a particular representation, we compute its projection onto each of these three bias subspaces. We then identify the subspace onto which the representation has the largest projection and selectively remove this projection. This targeted approach ensures that we effectively mitigate the most significant bias present in any given representation.

$$S_G = PCA\left(B_{W_{G_1}}, B_{W_{G_2}}\right)$$
$$S_{Race} = PCA\left(B_{W_{Ra_1}}, B_{W_{Ra_2}}, B_{W_{Ra_3}}\right)$$
$$S_{Religion} = PCA\left(B_{W_{Re_1}}, B_{W_{Re_2}}, B_{W_{Re_3}}\right)$$
$$B_W \rightarrow B_W - \max norm\left(B_W \cdot S_G, B_W \cdot S_{Race}, B_W \cdot S_{Religion}\right)$$

4 Experiments

4.1 Data

We are using two datasets as bias benchmarks.

CrowS Pairs (Nangia et al., 2020b): This is crowd-sourced dataset consists of sentence pairs, each of which demonstrates one of nine types of bias: race, gender/gender identity, sexual orientation, religion, etc. One sentence enforces a stereotype and the second one is a violation of the stereotype. The fact that it is crowd sourced is important because it covers sentences with widely acknowledged bias. The sentences are minimally distant and only differ on the specific word that refers to the bias group. For example,

- *"It was a very important discovery, one you wouldn't expect from a female astrophysicist."* is a sterotypical sentence
- and *"It was a very important discovery, one you wouldn't expect from a male astrophysicist."* is the corresponding antistereotypical sentence.

StereoSet (Nadeem et al., 2020b): This dataset helps identify stereotypical biases in language models. It targets four types of bias- gender, profession, race, and religion. In each data point, it contains a context, a stereotypical sentence, an antistereotypical sentence, and an unrelated sentence. These help since we want our model to be equidistant from the stereotypcial and antistereotypical sentence for the given context. The unrelated sentence is included because even though it is likely to be unbiased, it might not make logical sense and will thus will have a worse language modelling score. To avoid this, it is helpful to feed in all three types of sentences for the given context. For example, for the context *“our housekeeper is [MASK]”*,

- the stereotypical candidate is *“our housekeeper is Mexican”*
- the antistereotypical candidate is *“our housekeeper is American”*
- and the unrelated association candidate is *"our housekeeper is computer”*.

4.2 Evaluation

We used the evaluation metrics associated with our datasets in order to evaluate our results.

4.2.1 StereoSet

For the StereoSet, the evaluation metrics assoicated with it were Language Modeling Score (lms), Stereotype Score(ss), and Idealized CAT Score(icat).

- **Language Modeling Score (LMS):** We want our model to make meaningful associations. Thus, given a context and two sentences- one meaningful and one meaningless, the model would ideally rank the meaningful sentence higher than the meaningless one. The LMS of a target term is the percentage of times that our model prefers a meaningful association to a meaningless one. Thus, our ideal LMS is 100 since we want our model to always pick meaningful associations to complete the sentence.
- **Stereotype Score (SS):** Additionally, we want our model to be unbiased, i.e, given a stereotypical association and an antistereotypical association, we want our model to pick one of the two with equal probability. The stereotype score is the percentage of times that a model prefers a stereotypical association of a target term to its antistereotypical association. The ideal SS is 50 since we want our model to neither prefer stereotypical associations, not antistereotypical associations.
- **Idealized CAT Score (ICAT):** The ICAT score is a measure that factors in both LMS and SS. An ideal model will have an ICAT score of 100, that comes from an LMS of 100 and SS of 50.The formula for ICAT can be given by:

$$ICAT = LMS \times \frac{\min(SS, 100 - SS)}{50}.$$

4.2.2 Crows-Pairs

The Crows-Pairs dataset focuses on Masked Language Models (MLMs) as its evaluation metric. Each sentence is divided into an unmodified part, (common tokens between the two sentences in a pair), and a modified part (non-overlapping tokens). The likelihoods of unmodified tokens is conditioned on the modified tokens. For a sentence s, if $U = \{u_0, \dots, u_l\}$ are the unmodified tokens and $M = \{m_0, \dots, m_n\}$ are the modified tokens, then the probability to be estimated is $p(U|M, \theta)$, approximated by adapting *pseudo-log likelihood* MLM scoring Wang and Cho (2019). For each sentence, one unmodified token is masked at a time until all u_i have been masked. Our evaluation score is

$$score(s) = \sum_{i=0}^{|C|} \log P(u_i \in U | U_{\setminus u_i}, M, \theta).$$

The metric measures the percentage of examples for which the model assigns a higher pseudo-likelihood to the stereotypical sentence. The ideal score under this metric would be 50%.

4.3 Experimental Details

In our **debiasing research**, we utilized the **BERT base uncased model** from the Huggingface transformers module as our baseline. For the **Counterfactual Data Augmentation (CDA)** experiments, we curated a base dataset of **1300 examples from Wikipedia 10**, which was augmented to ensure bias equality, achieving up to a **2x augmentation**. This augmentation process involved adding a maximum of one example with altered gender, race, or religion for each original example in the dataset. Our test set comprised **300 examples**, similarly augmented for bias equality up to 2x.

In the **Subspace Debiasing (SDB)** experiments, we identified the bias subspace using the **first principal component vector (k=1)** from Principal Component Analysis (PCA). For the **equalized variants**, separate bias subspace vectors were utilized for two genders (male and female), three religions (Jewish, Christian, Muslim), and three races (White, Caucasian, and Asian). The process of determining the bias subspace involved analyzing the same **1300-example training set** used in the CDA experiments, focusing on bias-indicative words and utilizing the contextualized BERT embeddings of these words to apply PCA, thereby obtaining the principal component vector as the direction of bias. The test set for SDB experiments was aligned with the structure used in the CDA experiments.

Table 1: Transposed Performance Metrics of Different Models

Type of Bias		Baseline (BERT)	CDA	CDA (Unified)	CDA (GPT)	SDB	SDB (Equalized)	SDB (Unified)
CrowsScore	Gender	57.25	56.49	58.40	60.69	52.29	57.11	52.67
	Race	62.33	61.36	56.89	61.94	62.72	61.55	54.95
	Religion	62.86	66.67	56.19	65.71	63.81	40.95	59.05
LMS	Gender	85.74	86.81	86.76	86.39	85.60	87.48	85.07
	Race	84.01	84.82	84.77	85.11	83.78	83.55	82.96
	Religion	84.21	84.75	83.82	83.28	84.50	83.87	83.01
SS	Gender	60.28	57.38	57.58	58.07	59.37	58.50	60.35
	Race	57.03	58.63	57.93	57.42	57.78	57.58	55.80
	Religion	59.70	61.98	59.26	61.22	58.73	58.00	57.50
ICAT	Gender	68.11	74.00	73.61	72.45	84.07	85.00	67.46
	Race	72.20	70.18	71.33	72.49	70.75	70.88	73.34
	Religion	67.87	64.44	68.29	64.58	69.74	69.62	68.53

4.4 Results

Results of all the **six** approaches tried during our project are summarized in Table 1. We observed that the vanilla **CDA** approach effectively debiases language models, particularly for gender and race biases. Contrary to our hypothesis, the CDA method augmented with **GPT**, instead of word pairs, did not perform as well as vanilla CDA. We initially thought that using GPT for augmentation might improve debiasing because vanilla CDA trains separate models for each bias type, whereas with GPT, we trained a single, unified model. To ensure a fair comparison, we also trained a unified vanilla CDA model—that is, a single model designed to debias gender, race, and religion simultaneously. Surprisingly, not only did the unified vanilla CDA outperform the CDA + GPT approach, but it also exceeded the performance of the bespoke CDA models in debiasing, achieving significant improvement in religion bias—a task that bespoke vanilla CDA struggled with. This suggests that different biases in the model might share commonalities, possibly due to overlapping sets of parameters, rather than being entirely independent. Similar trends were observed for CDA across both **CrowS-Score** and **SS score** in the **Stereoset** evaluation, though the lower LM score with unified CDA resulted in a lower overall **ICAT score**.

Similarly, **SDB**, a more transparent method for surgically removing bias from the model without altering its weights, consistently showed better debiasing performance compared to CDA. Among the various SDB variations, the two models we proposed—**SDB equalized** and **SDB unified**—demonstrated almost consistent and incremental improvements in debiasing language models over vanilla SDB. This reinforces the notion of a commonality among biases within the model and the conservative approach adopted by the SDB equalized model to eradicate biases while preserving as much information as possible.

5 Analysis

In our exploration of **debiasing methods** such as **Counterfactual Data Augmentation (CDA)** and **Subspace Debiasing (SDB)**, alongside their variations, we uncovered notable insights regarding the mechanisms of bias within models. Initially, we anticipated that employing **GPT** for data augmentation would significantly aid in debiasing. Contrary to expectations, this approach proved minimally effective. We hypothesize this ineffectiveness stems from GPT's tendency to introduce complex sentence manipulations, inadvertently incorporating terms related to **race** and **religion**—categories underrepresented in the dataset. This discrepancy likely disrupts the alignment between the fine-tuning and testing datasets.

Comparatively, **SDB outperforms CDA** in mitigating bias within Language Models (LMs). SDB's efficacy is attributed to its **hard debiasing** strategy, which involves either removing or neutralizing bias components within BERT embeddings. Conversely, CDA's approach to fine-tuning is more susceptible to variations in **hyperparameters**, such as learning rate and batch size, making its debiasing capabilities less consistent.

Among the various iterations of CDA and SDB, the **unified models** emerged as the most effective in debiasing, as evidenced by superior performance metrics including **average CrowS-Pairs score** and **Stereotype Susceptibility (SS)**. This finding suggests a potential **correlation between different types of biases**. Supporting this theory, our analysis revealed high **cosine similarity** among the bias subspaces for gender, race, and religion, indicating a common bias subspace that transcends the type of bias.

Most notably, the **SDB-Equalized** variant achieved the highest ICAT scores. This success is likely due to our **equalization technique**, which conserves critical information within BERT embeddings, thereby enhancing the LM's performance and, consequently, the overall ICAT scores.

6 Conclusion

In this study, we explored various **debiasing strategies** for language models, particularly focusing on **Counterfactual Data Augmentation (CDA)** and **Subspace Debiasing (SDB)**. Our findings suggest that **SDB**, especially its equalized and unified variations, demonstrates superior efficacy in mitigating biases without compromising the language model's performance. This reinforces the potential for a **shared bias subspace** across different types of biases. Unexpectedly, augmenting CDA with **GPT** did not yield the anticipated improvements, highlighting the complexity of debiasing language models. Our work contributes to the broader effort of creating fair and unbiased AI systems, underscoring the importance of understanding and addressing the multifaceted nature of bias in language models.

For future work, we aim to explore the **intersections of bias subspaces** more deeply, including **cross-evaluation** of models trained on one type of bias but tested on another. This could further elucidate the relationships between different biases and enhance our debiasing strategies. Additionally, combining the **equalized and unified approaches** in SDB could potentially set a new standard for debiasing effectiveness. Extensive testing across various hyperparameters for CDA could also reveal more nuanced insights into optimizing debiasing efforts.

7 Contributions

Each member of the team made an equal contribution to the project. Saanvi was in charge of conducting the literature review, familiarizing the team with current debiasing techniques and coming up with weaknesses that we could address. Shreya played a key role in the creative process, generating novel variations that were incorporated into our project. Meanwhile, Aman spearheaded the implementation phase. Undoubtedly, every aspect of the project was a collaborative effort, with all team members actively contributing to every component.

References

OpenAI et al. 2023. Gpt-4 technical report.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations.

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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020a. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020b. Stereoset: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020a. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020b. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.