(APPENDIX) BRIDGING DEBIASING TASKS WITH SUFFICIENT PROJECTION: A GENERAL THEORETICAL FRAMEWORK FOR VECTOR REPRESENTATIONS

Anonymous authors
Paper under double-blind review

A APPENDIX

A.1 SCHEME FOR SIR ESTIMATOR

Suppose the data set $\{(X_i, Z_i)\}_{i=1}^n$ is given, then the steps for SIR are summarized as :

1. Standardizing $\boldsymbol{X}$ by the transformation $\tilde{\boldsymbol{X}}_i = C_X^{-1/2}(\boldsymbol{X}_i - \mu_X)$, where μ_X and C_X are the mean vector and covariance matrix of X.
2. Slice the range of Z into H intervals $\{J_h\}_{h=1}^H$. Estimate the weight $p_h = (1/n)\sum_{i=1}^n I(Z_i \in J_h)$ and compute the sample mean $m_h = (1/np_h)\sum_{Z_i \in J_h} \tilde{\boldsymbol{X}}_i$ on each sliced interval.
3. Form $M^{\mathrm{SIR}} = \sum_{h=1}^H p_h m_h m_h^\top$ and let ϕ_k be its eigenvectors. The directions are estimated by $\beta_k = C_X^{-1/2}\phi_k$ for $k = 1, \ldots, q$.

A.2 PMS ESTIMATOR IMPLEMENTATION

For multivariate variable $Z \in \mathbb{R}^{p_3}$, let Z_{ij} denote the j-th coordinate of i-th sample, the PMS estimator can be achieved through the following Algorithm 2.

Algorithm 2 PMS Estimator

Input: Data $\{(X_i, Z_i)\}_{i=1}^n$, partition H, covariance matrix C_X and weights $\{w_j\}_{j=1}^{p_3}$;
Output: PMS estimator M^{PMS};

for $j = 1, \ldots, p_3$ **do**
  Slice the support of Z_j into H intervals denoted as $\{J_{j,h}\}_{h=1}^H$
  for $h = 1, \ldots, H$ **do**
    Estimate the weight on each interval $p_{j,h} = \frac{1}{n}\sum_{i=1}^n I(Z_{ij} \in J_{j,h})$;
    Compute the standardized mean on each interval $m_{j,h} = \frac{1}{np_{j,h}}\sum_{Z_{ij} \in J_{j,h}} C_X^{-1} X_i$;
  end for
  Obtain the estimator for each dimension $M_j^{\mathrm{SIR}} = \sum_{h=1}^H p_{j,h} m_{j,h} m_{j,h}^\top$;
end for
Calculate the weighted sum of estimators $M^{\mathrm{PMS}} = \sum_{j=1}^{p_3} w_j M_j^{\mathrm{SIR}}$;
Return: M^{PMS}.

The weights w_j can be chosen as either equal weights or proportional to the leading eigenvalues of M_j. Then the leading q eigenvectors $\psi_1, \ldots, \psi_q$ of M^{PMS} can be used to recover $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$.

A.3 T-SNE VISUALIZATION

To visually demonstrate the effectiveness of our proposed method in reducing gender bias, we selected the top 500 male- and female-biased embeddings. Using t-SNE projection, we generated a graph for the original GloVe and our debiased embeddings. Figure 1 shows the separation of male- and female-biased embeddings in two different colors. It can be observed that our method has mixed the male- and female-biased embeddings effectively.

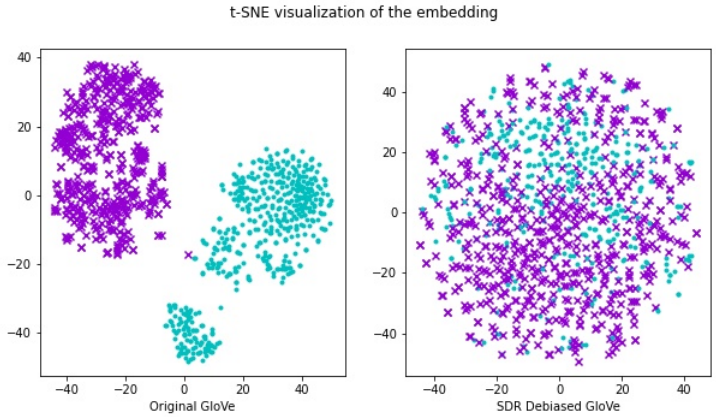

Figure 1: t-SNE visualization.

A.4 Detail of WEAT

Let X and Y be two sets of target words of equal size n with their embedding $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, and A, B the two sets of attribute words with their embedding $\{a_i\}_{i=1}^{|A|}$ and $\{b_i\}_{i=1}^{|B|}$. The WEAT uses the difference of averaged distance to measure the similarity of a vector w to two sets A and B. The test statistic is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b)$$

In other words, $s(w, A, B)$ measures the association of the word w with the attribute, and $s(X, Y, A, B)$ measures the differential association of the two sets of target words with the attribute.

Let $\{(X_i, Y_i)\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. The one-sided p-value of the permutation test is

$$\Pr_i \left[s\left(X_i, Y_i, A, B\right) > s(X, Y, A, B) \right]$$

The effect size is

$$\frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std-dev}_{\ w \in X \cup Y} s(w, A, B)}$$

It is a normalized measure of how separated the two distributions (of associations between the target and attribute) are.

All word lists are from Caliskan et al. (2017). Because GloVe embeddings are uncased, we use lowercase words.

A.5 Detail of SEAT

A.5.1 Obtain the Projection Matrix

To train projections for the topics of gender, race, and religion, we used the vocabulary from the GloVe model. All words were divided into groups according to their cosine similarities with pre-determined hint words: [he, she] for gender, [black people, white people] for race, and [Christianity, Jewish, Islam] for religion. Using BERT representations, we selected the top 75k words for gender, 75k for race, and 30k for religion from each group and associated them with their group labels as the input dataset for Algorithm 1.

SEAT Gender Tasks							
Model	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8b	Avg. Effect Size (↓)
BERT	0.931	0.090	-0.124	0.937	0.783	0.858	0.620
CDA	0.846	0.186	-0.278	1.342	0.831	0.849	0.722
Dropout	1.136	0.317	0.138	1.179	0.879	0.939	0.765
INLP	0.317	-0.354	-0.258	0.105	0.187	-0.004	**0.204**
SentDebias	0.350	-0.298	-0.626	0.458	0.413	0.462	0.434
SUP	-0.028	-0.286	-0.403	-0.255	0.213	-0.124	0.218

Table 4: SEAT effect sizes for gender debiased BERT. Effect sizes closer to 0 are indicative of less biased model representations.

SEAT Race Tasks								
Model	ABW-1	ABW-2	SEAT-3	SEAT-3b	SEAT-4	SEAT-5	SEAT-5b	Avg. Effect Size (↓)
BERT	-0.079	0.690	0.778	0.469	0.901	0.887	0.539	0.620
CDA	0.231	0.619	0.824	0.510	0.896	0.418	0.486	0.569
Dropout	0.415	0.690	0.698	0.476	0.683	0.417	0.495	0.554
INLP	0.295	0.565	0.799	0.370	0.976	1.039	0.432	0.639
SentDebias	-0.067	0.684	0.776	0.451	0.902	0.891	0.513	0.612
SUP	0.019	0.428	0.542	0.193	0.611	0.716	0.514	**0.432**

Table 5: SEAT effect sizes for race debiased BERT. Effect sizes closer to 0 are indicative of less biased model representations.

SEAT Religion Tasks					
Model	Religion-1	Religion-1b	Religion-2	Religion-2b	Avg. Effect Size (↓)
BERT	0.744	-0.067	1.009	-0.147	0.492
CDA	0.355	-0.104	0.424	-0.474	0.339
Dropout	0.535	0.109	0.436	-0.428	0.377
INLP	0.473	-0.301	0.787	-0.280	0.460
SentDebias	0.728	0.003	0.985	0.038	0.439
SUP	0.392	-0.066	0.492	0.092	**0.261**

Table 6: SEAT effect sizes for religion debiased BERT. Effect sizes closer to 0 are indicative of less biased model representations.

A.5.2 Full Test and Results of SEAT

In this section, we provide a complete set of results for all SEAT tests. All of the baseline results are from Meade et al. (2022). Also, for detailed attribute word sets and the target word sets, please refer to their GitHub repo. Table 4 are tasks for Gender debias. Table 5 are tasks for Race debias. Table 6 are tasks for Religion debias.

A.6 Fair Text Classification Details

The **MOJI** is a sentiment classification dataset collected by Blodgett et al. (2016) that contains tweets from either African-American English or Standard American English. Each of the text data is labeled with a binary 'race' label based on the kind of English they use. The binary sentiment score is annotated by the emoji contained in the tweets. We compose the training data set as follows: AAE–happy = 40%, SAE– happy = 10%, AAE–sad = 10%, and SAE–sad = 40%. We used the train, dev, and test splits of 100k/8k/8k instances, respectively.

The **BIOS** dataset De-Arteaga et al. (2019) is a personal biography classification dataset annotated by gender and 28 classes of occupation. We follow the same split setup for the BIOS data as in De-Arteaga et al. (2019), and the ratio of train:dev:test is $65\% : 10\% : 25\%$.

The **Toxic** dataset features text sourced from the Talk Pages of Wikipedia, where each comment has been categorized by human assessors as either toxic or non-toxic. Interestingly, an analysis of this dataset has revealed a disproportionate appearance of certain demographic identity-related terms (such as "gay" and "black") within the labels. This imbalance can inadvertently lead to biased model training, resulting in discriminatory behavior towards certain groups. Our research employs the same division of data as specified by (Dixon et al., 2018), enabling us to test the efficacy of our method in reducing discrimination against minority groups.

A.7 Proof of Theorem 6.1

Proof. According to the definition of conditional independence, for any measurable function f, we have $f(X) \perp\!\!\!\perp Z \mid QX$ because the randomness of $f(X)$ only comes from the random variable X.

For the debias task, notice that $X \perp\!\!\!\perp Z \mid QX$, thus $X \perp\!\!\!\perp Z \mid QX$. It implies that Z only depends on QX. Therefore, if we eliminate those correlated part and denote $\widetilde{X} = (I - Q)X$, we have $\widetilde{X} \perp\!\!\!\perp Z$. It achieves the goal of the debias task defined above.

For the fairness task, if we assume $X \perp\!\!\!\perp Y \mid Q_y X$, which implies $Y = f_0(Q_y X)$ for some measurable function f_0. Notice that $\mathrm{Span}\{Q_y\} \subset \mathrm{Span}\{Q\}$, then $\mathrm{Span}\{Q_y\}$ is orthogonal to $\mathrm{Span}\{I - Q\}$, which implies $(I - Q)X \perp\!\!\!\perp Q_y X$. Therefore, $(I - Q)X \perp\!\!\!\perp Z \mid Y$ since the randomness of Y comes from $Q_y X$. It achieves the goal of the fairness task defined above if we let $\widetilde{X} = (I - Q)X$. □

Remark A.1. *We should emphasize that in the above theorem, the random vectors X, Y, and Z are defined on the Euclidean space $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_3}$ respectively. For each random variable, taking X as an example, the sample space is defined as $\Omega = \mathcal{B}(\mathbb{R}^{p_1})$, which is Borel set generated by all open set on $\mathbb{R}^{p_1}$, and the σ-algebra Σ is generated by Ω, i.e. $\Sigma = \sigma(\Omega)$. In this way, for any measurable function f satisfying the sample space of $f(X)$ is included in the sample space of X, we have $\sigma(f(X)) \subset \sigma(X)$, and thus the desired properties of conditional independence hold in the proof*

A.8 Limitations

All our result is based on the English dataset, as there is a lack of benchmark of fairness in other languages. Also, we only consider the transformation under a linear framework, where we aim to find the projection matrix P. However, the estimation procedure for the central subspace $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ has been well developed and can find nonlinear transformation g, which we leave for future exploration. Also, for the SEAT evaluation, there are some researchers point out that SEAT sometimes not able to detect the bias inside the language model. But compared with other debiasing studies that only report on SEAT, we test our method on much more comprehensive experiments.

A.9 Ethics Statement

Our research is fundamentally methodological in nature, focusing on the development of strategies to mitigate biases in NLP. We have taken careful measures to ensure that our work adheres to recognized ethical guidelines. For all evaluations related to bias and fairness, we have strictly followed established protocols, utilizing well-known tasks to evaluate biases related to gender, religion, and race. It is important to clarify that our use of these tasks is for analytical purposes only, with the sole intention of understanding and minimizing the biases present in AI systems. Our goal is to promote fairness and inclusivity in AI, and we firmly advocate for the respectful and unbiased treatment of all individuals, irrespective of their gender, religion, or race.

A.10 Reproducibility

Hyperparameter tuning: For our method, the main hyperparameter is the q: the number of directions we want to project. We use regular grid search to find the best hyperparameter. For classifiers mentioned in Algorithm 1, we use the logistic classifier in sklearn.

Computational detail: We conduct all our experiments on an Ubuntu Server with CPU AMD Ryzen Threadripper 3990X 64-Core Processor and 256G RAM. Since our experiments do not need many computational resources (no retraining or fine-tuning), no GPU is needed.

Baseline results: Most of the baseline results are from recently published papers of well-known conferences. In static embedding evaluation, the INLP results are calculated by our code using the embedding they provided, which has a slightly better result than they reported in their paper.

REFERENCES

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

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, 2022.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, 2016.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.