Debiasing with Sufficient Projection: A General Theoretical Framework for Vector Representations

Enze Shi, Lei Ding, Linglong Kong*, Bei Jiang*
Department of Mathematical and Statistical Sciences
University of Alberta
{eshi, lding1, lkong, bei1}@ualberta.ca

Abstract

Pre-trained vector representations in natural language processing often inadvertently encode undesirable social biases. Identifying and removing unwanted biased information from vector representation is an evolving and significant challenge. Our study uniquely addresses this issue from the perspective of statistical independence, proposing a framework for reducing bias by transforming vector representations to an unbiased subspace using sufficient projection. The key to our framework lies in its generality: it adeptly mitigates bias across both debiasing and fairness tasks, and across various vector representation types, including word embeddings and output representations of transformer models. Importantly, we establish the connection between debiasing and fairness, offering theoretical guarantees and elucidating our algorithm's efficacy. Through extensive evaluation of intrinsic and extrinsic metrics, our method achieves superior performance in bias reduction while maintaining high task performance, and offers superior computational efficiency.

1 Introduction

Natural Language Processing (NLP) models have made significant strides in recent years, with much of their success attributed to representation learning - the process of creating effective vector representations for textual data. Various research has been conducted in this area, including static word embedding (Mikolov et al., 2013; Pennington et al., 2014), contextualized embedding (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2019), sentence embedding (Reimers and Gurevych, 2019) in addition to other representation methods.

However, as vector representations have been applied in a wide range of real-life scenarios, researchers have discovered that stereotypical biases and spurious correlations can be transferred from human-generated corpora to vector representations and models (Bolukbasi et al., 2016; Caliskan et al., 2017; Vig et al., 2020). This has the potential to produce biased and unfair outcomes in various downstream tasks (Kurita et al., 2019) and can even lead to serious social problems. For instance, in the word analogy task presented in (Bolukbasi et al., 2016), it was found that the vector representation for $\overrightarrow{she}$ was closer to $\overrightarrow{nurse}$ than the representation for $\overrightarrow{he}$ was to $\overrightarrow{doctor}$. De-Arteaga et al. (2019) found a performance gap between different genders in text classification tasks.

The bias and fairness issues in NLP models are primarily caused by the unbalanced and stereotypical nature of the training corpora. Liang et al. (2020) described this as unbalanced model behavior in relation to certain socially sensitive topics such as gender, race, and religion. To quantify biases in NLP, two types of bias evaluation metrics(intrinsic and extrinsic) have been proposed. However, recent research has shown that in most cases, there is a weak correlation between them (Goldfarb-Tarrant et al., 2021; Kaneko et al., 2022; Cao et al., 2022). There remains a significant research gap in understanding how to bridge these two kinds of tasks. In our research, we employ statistical independence to establish a theoretical linkage between these tasks, offering insights into the interplay between intrinsic and extrinsic biases.

Various methods have been proposed for reducing bias in NLP, but it remains a challenge to effectively mitigate bias while maintaining high model performance. Furthermore, it is particularly difficult for debiasing methods to efficiently address both intrinsic and extrinsic biases at the same time, as discussed in the related works section.

In this paper, we propose a general debiasing method that can effectively mitigate bias across both debiasing and fairness tasks. Our key contributions include the following:

*Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5960–5975

- We are the first to scope the debiasing and fairness tasks unitedly through statistical independence, providing a detailed theoretical analysis aimed at identifying the minimal subspace that contains bias information. This analysis enables us to determine the optimal projection matrix for these tasks and bridge the connection between them.

- Our algorithm showcases its effectiveness in both intrinsic (bias embedding evaluation) and extrinsic (fairness text classification) evaluation metrics. It is versatile, adapting to different embedding methods and sensitive variable types, making it universally applicable.

- Our method improves upon existing state-of-the-art methods while still maintaining good task performance compared to the original model and stands out due to its superior computational efficiency.

The structure of this paper is as follows. We begin with a comprehensive review of existing research on bias evaluation and debiasing techniques in NLP. We then introduce our methodology, including our proposed algorithm. We present experimental results on a range of gender bias evaluation tasks, showcasing the effectiveness of our approach. Finally, we provide a theoretical bridge and guarantee, and a discussion of our method.

2 Related Works

Debiasing Methods in NLP Researchers have been focusing on reducing bias from each component of NLP models. The most intuitive idea of debiasing is through counterfactual data augmentations (Zmigrod et al., 2019; Dinan et al., 2020; Barikeri et al., 2021), which involves re-balancing a corpus by swapping bias attribute words (e.g., he/she) in a dataset. The re-balanced corpus is then used for further training to debias a model. While this approach is simple and can be applied to all tasks, it does not perform well in terms of debiasing and requires additional computational resources for model re-training. Another direction is fine-tuning pre-trained transformer-based language models using methods such as projection (Kaneko and Bollegala, 2021), adversarial (Han et al., 2021), contrastive (Cheng et al., 2020; Shen et al., 2021; He et al., 2022), dropout (Webster et al., 2020), and prompting (Schick et al., 2021; Guo et al., 2022). These methods show effectiveness in reducing bias in various intrinsic evaluation tasks. However, when deploying these debiased models to downstream tasks, especially fine-tuning on task-specific datasets, the debiased language model can still re-learn social bias, making these debiasing methods less effective.

Our proposed debiasing method is based on the controlled removal of specific information from vector representations, which is closely related to the task of disentangling representations (Bengio et al., 2013). Previous research in this area includes methods for removing bias from static embeddings, such as projecting the word embedding into the orthogonal space of the gender direction ($\overrightarrow{he}$ - $\overrightarrow{she}$) (Bolukbasi et al., 2016), re-training the entire embedding using some loss functions (Kaneko and Bollegala, 2019), and utilizing the ideas in causal inference (Yang and Feng, 2020; Ding et al., 2022). There are several similar projection-based methods like Iterative Nullspace Linear Projection (INLP; (Ravfogel et al., 2020)), RLACE (Ravfogel et al., 2022a). We discuss the detailed comparison and advantages of our method in Appendix A.2.

Evaluating Bias in NLP The measurement methods for evaluating bias in pre-trained word embeddings and language models can be broadly divided into two categories: Intrinsic and Extrinsic evaluations. Intrinsic bias evaluations probe the bias within pre-trained word embeddings and language models. Common methods include measuring the geometry in embedding space, such as the Word Embedding Association Test (WEAT; (Caliskan et al., 2017)) and Sentence Encoder Association Test (SEAT; (May et al., 2019)). Additionally, (Kurita et al., 2019; Nangia et al., 2020; Nadeem et al., 2021) propose metrics using the likelihood score. Furthermore, research suggests that some debiasing methods may only hide bias, and thus additional measurement approaches are needed (Gonen and Goldberg, 2019).

The extrinsic bias is specific to certain downstream tasks. In the text classification task, De-Arteaga et al. (2019); Blodgett et al. (2016) proposed two benchmark datasets and used the equal opportunity measure from fairness literature. Zhao et al. (2018a) proposed the WinoBias benchmark for Coreference resolution. As well as other benchmarks, such as Bias-NLI (Dev et al., 2020) and in machine translation (Stanovsky et al., 2019). However, recent research has indicated that intrin-

sic bias in embeddings or models typically does not have a strong correlation with bias in downstream tasks(Goldfarb-Tarrant et al., 2021; Cao et al., 2022). Kaneko et al. (2022) found out that the debiased models re-learn the bias from the fine-tuning datasets, showing that only debiasing upstream models may not be enough to eliminate bias in downstream tasks.

In our work, we conduct comprehensive evaluation experiments in both intrinsic and extrinsic tasks. Additionally, our approach avoids the issue of re-learning bias by directly addressing the vector representation in downstream tasks.

3 Methodology

3.1 Problem Setup

We consider the problem of removing sensitive information inside the vector representation. Given the representation vector $\boldsymbol{X} \in \mathbb{R}^{p_1}$ accompanied with the target attribute $\boldsymbol{Y} \in \mathbb{R}^{p_2}$ and the sensitive attribute $\boldsymbol{Z} \in \mathbb{R}^{p_3}$, our goal is to find a map $g : \mathbb{R}^{p_1} \mapsto \mathbb{R}^{p_1}$ such that:

- $g(\boldsymbol{X})$ is uncorrelated with $\boldsymbol{Z}$;
- $g(\boldsymbol{X})$ maintains ability to predict $\boldsymbol{Y}$.

In other words, the new representation $\widetilde{\boldsymbol{X}} = g(\boldsymbol{X})$ removes the sensitive information $\boldsymbol{Z}$ contained in the original representation while preserving other useful information in $\boldsymbol{X}$. The notations given above incorporate debiasing and fairness tasks into the same framework, which are formulated in the following definitions:

Definition 3.1. *Let $\widetilde{\boldsymbol{X}} = g(\boldsymbol{X})$ and f_1 be the mechanism with input $\widetilde{\boldsymbol{X}}$. Then $\widetilde{\boldsymbol{X}}$ is said to be a debiased representation if $f_1(\widetilde{\boldsymbol{X}}) \perp\!\!\!\perp \boldsymbol{Z}$.*

Definition 3.2. *Let $\widetilde{\boldsymbol{X}} = g(\boldsymbol{X})$ and f_2 be the mechanism trained by $(\widetilde{\boldsymbol{X}}, \boldsymbol{Y})$ to predict $\boldsymbol{Y}$. Then $\widetilde{\boldsymbol{X}}$ is said to be a fair representation if $f_2(\widetilde{\boldsymbol{X}}) \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{Y}$.*

Our formulated definition directly aligns with the objectives of each task. In the debiasing task, given the input $\boldsymbol{X}$, the mechanism f_1 is biased if the output $f_1(\boldsymbol{X})$ relies heavily on $\boldsymbol{Z}$. Therefore, the goal is to make $\boldsymbol{Z}$ and the output of the mechanism $f_1(\boldsymbol{X})$ to be **independent** given the representation $\boldsymbol{X}$. Similarly, for fairness tasks, the aim is to develop a fair mechanism f_2, ensuring its prediction is independent of sensitive information $\boldsymbol{Z}$ conditioned on Y (**Conditional Independence**) (Hardt et al., 2016).

Remark 3.3. *Note that f_1 or f_2 represent a specific mechanism with input $\widetilde{X}$, which differs in different tasks. For instance, in the WEAT and SEAT tasks, f_1 denotes the function $s(\widetilde{X}, A, B)$, which is a measure of association strength. In the Fairness Text Classification, f_2 is specifically used to denote the classifier for target label Y.*

3.2 Motivation

As previously discussed, our goal is to identify a mapping g such that $\widetilde{\boldsymbol{X}} = g(\boldsymbol{X})$ possesses the desired properties. A direct approach involves constructing a $p_1 \times p_1$ projection matrix P and applying the linear transformation $g(\boldsymbol{X}) = P\boldsymbol{X}$. When we restrict the transformation g into linear projections, the original vector $\boldsymbol{X}$ can be decomposed into $\boldsymbol{X} = P\boldsymbol{X} + (I - P)\boldsymbol{X}$, where I is the $p_1 \times p_1$ identity matrix. Letting $\mathcal{S}_1$ and $\mathcal{S}_2$ represent the spaces spanned by P and $I - P$ respectively, the representation space is then decomposed as $\mathbb{R}^{p_1} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$. For a debiased representation $\widetilde{\boldsymbol{X}} = P\boldsymbol{X}$, $\mathcal{S}_1$ should be structured to minimize the information regarding $\boldsymbol{Z}$, while $\mathcal{S}_2$ should encapsulate as much of the information regarding $\boldsymbol{Z}$ as possible.

Consequently, the debiasing goal transforms into the identification of the subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$. It's crucial to note that the majority of information regarding $\boldsymbol{Z}$ resides within $\mathcal{S}_2$. Therefore, the expression $(I - P)\boldsymbol{X}$ emerges as a potential predictor for $\boldsymbol{Z}$. Throughout this paper, we adhere to the following linearity assumption,

Assumption 3.4 (Weak Linearity)**.** *Suppose $Q = I - P$ has rank q and orthogonal basis $(\beta_1, \ldots, \beta_q)$, with each β_j belonging to $\mathbb{R}^{p_1}$. Then we assume Z can be modeled by projecting X onto q distinct directions:*

$$Z = f(\beta_1^\top X, \ldots, \beta_q^\top X, \varepsilon), \tag{1}$$

where f is an unknown function, which can be linear or nonlinear, and ε denotes the random effect.

Remark 3.5. *Contrary to the strong linearity assumption in other projection-based methods such as (Ravfogel et al., 2020), which postulates that $\mathbf{Z}$ can be expressed as $\mathbf{Z} = W\mathbf{X}$ for $W \in \mathbb{R}^{p_3 \times p_1}$, our approach adopts a weak linearity assumption. This assumption introduces an unknown function f, which can be either linear or nonlinear, thereby including the traditional strong linearity assumption. A toy example regrading weak and strong linearity can be found in Appedix A.1.*

Based on the discussion above, the primary objective of this study is articulated as follows:

initially, to identify the direction matrix $B = (\beta_1, \ldots, \beta_q) \in \mathbb{R}^{p_1 \times q}$ ensuring that $B^\top \boldsymbol{X}$ captures most of the information about $\boldsymbol{Z}$. Then given this direction matrix, we can obtain a debiased representation of $\boldsymbol{X}$ by projecting it onto the subspace orthogonal to $\mathrm{Span}\{\mathrm{B}\}$. Note that $\mathrm{Span}\{\mathrm{B}\}$ is equivalent to $\mathrm{Span}\{\mathrm{Q}\} = \mathcal{S}_2$, and hence, its corresponding orthogonal subspace $\mathcal{S}_1$ is spanned by P as preliminarily defined. This approach underpins the theoretical foundation of our proposed algorithm, details of which will be explored in the forthcoming sections.

3.3 Minimal Subspace

It is crucial that we want the subspace $\mathcal{S}_2$ with the desired property as **small** as possible so that we can retain the utility of $\boldsymbol{X}$ after projecting it on $\mathcal{S}_1$. Essentially, we want to find the matrix Q with minimum rank q. Consider the random variables $\boldsymbol{X} \in \mathbb{R}^{p_1}$ and $\boldsymbol{Z} \in \mathbb{R}^{p_3}$. If there exists a full rank matrix $B \in \mathbb{R}^{p_1 \times q}$, such that $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$ ($\boldsymbol{X}$ is independent of $\boldsymbol{Z}$ conditioned on its projections on B), then the column space of the matrix B, denoted as $\mathrm{Span}\{B\}$, is called a sufficient dimension reduction subspace of $\boldsymbol{Z}$ with respect to $\boldsymbol{X}$. The intersection of all the dimension reduction subspaces is called the central subspace and denoted as $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$. That is $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}} = \bigcap_{B \in \mathcal{B}_{XZ}} \mathrm{Span}\{B\}$, where

$$\mathcal{B}_{XZ} = \left\{ B \mid \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \text{ conditioned on } B^\top \boldsymbol{X} \right\}$$

The dimension of the central subspace is denoted as $\dim(\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}})$. When $\mathrm{Span}\{B\}$ is the central subspace, we have $\dim(\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}) = \dim(\mathrm{Span}\{B\}) = q$. See (Cook and Li, 2002) for more details. A toy example of minimal subspace can be found in Appendix A.1.

If $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$, then $B^\top \boldsymbol{X}$ are most useful to predict $\boldsymbol{Z}$ based on $\boldsymbol{X}$, which is exactly the case in Model (1). Therefore, the central subspace $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ is the minimal subspace processing the desired property and serves as a promising candidate for the expected subspace $\mathcal{S}_2$. We will illustrate the estimation procedure in the next section, which is robust to any kind of mapping f in Model (1) and only relies on the data set $\{(X_i, Z_i)\}_{i=1}^n$.

Note that Model (1) is an ideal case based on the linearity assumption. In real-world applications, using only q directions might not be sufficient to cover all the information of $\boldsymbol{Z}$ because there might be nonlinear correlations between $\boldsymbol{X}$ and $\boldsymbol{Z}$. However, since the q directions can cover most of the information, it is still safe to use Model (1) in practice. Specifically, in NLP tasks, we may assume that the sensitive attribute $\boldsymbol{Z}$ can be predicted by the projections of the representation $\boldsymbol{X}$ onto q directions with some unknown mapping f.

4 Sufficient Universal Projection (SUP)

4.1 Subspace Estimation

In this subsection, we will explain the process of estimating the subspace $\mathcal{S}_2 = \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ in different scenarios. We will begin with the simplest case where $p_3 = 1$, meaning $\boldsymbol{Z} \in \mathbb{R}$ is a scalar. Sliced Inverse Regression (SIR) is a classical dimension reduction method proposed by (Li, 1991) for univariate response $\boldsymbol{Z}$. We provide the detailed scheme for SIR applied on the data set $\{(X_i, Z_i)\}_{i=1}^n$ in the Appendix A.3. The main procedure of SIR is: (1) divide the support of $\boldsymbol{Z}$ into H intervals and calculate the covariance matrix of $\boldsymbol{X}$ for each interval, (2) calculate the weighted covariance matrix based on H intervals, (3) obtain the directions from the weighted covariance matrix. The space spanned by the directions $B = (\beta_1, \ldots, \beta_q)$ provided by SIR is a consistent estimation of $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$.

For multivariate sensitive attributes, a direct analogy of the slicing strategy in SIR no longer works, as the number of partitions of the support of $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{p_3}) \in \mathbb{R}^{p_3}$ becomes H^{p_3} and thus suffers the curse of dimensionality. To address the limitation of the original SIR, the Pooled Marginal Slicing (PMS) estimator proposed in (Aragon, 1997) combines the subspaces $\mathcal{S}_{\boldsymbol{Z}_i|\boldsymbol{X}}$ estimated by univariate response SIR to get the directions for the multivariate response, which is motivated by the following proposition.

Proposition 4.1. *Note that $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$ implies $\boldsymbol{Z}_j \perp\!\!\!\perp \boldsymbol{X} \mid B^\top \boldsymbol{X}$. Therefore, for $j = 1, \ldots, p_3$, $\mathcal{S}_{\boldsymbol{Z}_j|\boldsymbol{X}} \subseteq \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$.*

Proof. For any $B \in \mathcal{B}_{XZ}$, we have $B \in \mathcal{B}_{XZ_j}$, thus $\mathcal{B}_{XZ} \subset \mathcal{B}_{XZ_j}$. Recall that $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ is the intersection of all elements in $\mathcal{B}_{XZ}$. Therefore, we have $\mathcal{S}_{\boldsymbol{Z}_j|\boldsymbol{X}} \subseteq \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ for $j = 1, \ldots, p_3$. □

Proposition 4.1 indicates that $\mathcal{S}_{\boldsymbol{Z}_j|\boldsymbol{X}}$ can be used to recover $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$, which guarantees the theoretical properties of PMS estimator. It also naturally lifts the curse of dimensionality. Let Z_{ij} denote the j-th coordinate of i-th sample. We apply SIR to data set $\{(X_i, Z_{ij})\}_{i=1}^n$ and obtain the estimators M_i^{SIR} for $j = 1, \ldots, p_3$. Then we define the weighted sum of estimators as $M^{\mathrm{PMS}} = \sum_{j=1}^{p_3} w_i M_i^{\mathrm{SIR}}$, where

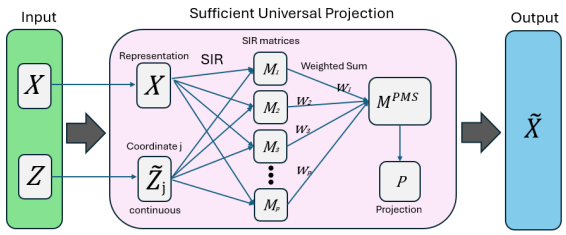

Figure 1: The flowchart of Sufficient Universal Projection algorithm.

w_i can be chosen as either equal weights or proportional to the leading eigenvalues of M_i. Then the leading q eigenvectors $\psi_1, \ldots, \psi_q$ of M^{PMS} can be used to recover $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$. The detailed implementation of obtaining M^{PMS} is summarized in Algorithm 2 in the Appendix.

4.2 Algorithm Implementation

To obtain the debiased representation, we first collect the original vector representation X_i and the associated sensitive attribute Z_i. Note that Z_i can be labeled by humans or learned from the training data. Specifically, we place no restrictions on the structure of Z_i – it can be either discrete labels $Z_i \in \{1, 2, \ldots, k\}$ representing gender or race, or a continuous variable. When Z_i is a continuous variable, we directly set $\widetilde{Z}_i = Z_i$ and handle $\{(X_i, \widetilde{Z}_i)\}$ as discussed above. If Z_i is a categorical variable with choices $\{1, 2, \ldots, k\}$, we first train a classifier f_{cls} based on the data set $\{(X_i, Z_i)\}_{i=1}^n$, whose output is the probability of X_i belonging to each category, then denote

$$\widetilde{Z}_i = f_{\text{cls}}(X_i) = (\widetilde{Z}_{i1}, \ldots, \widetilde{Z}_{ik}) \in \mathbb{R}^k,$$

where $\sum_{j=1}^k \widetilde{Z}_{ij} = 1$. In both scenarios, the attribute Z_i is converted to the vector variable with continuous support. Then we can obtain the PMS estimator M^{PMS} based on $\{(X_i, \widetilde{Z}_i)\}_{i=1}^n$ with its leading q eigenvectors $\psi_1, \ldots, \psi_q$. The projection matrix is defined as $P = I - \sum_{i=1}^q \psi_i \psi_i^\top$. Intuitively, $Q = \sum_{i=1}^q \psi_i \psi_i^\top$ is the estimated central mean space regarding Z, and the space spanned by this matrix contains most of the information we want to eliminate. The procedure is outlined in Algorithm 1 and in Figure 1.

It is worth emphasizing that SUP is a general framework for bias elimination, and we have no assumption on the type of representation. Therefore, our proposed method is universally robust to both static and contextualized embeddings with different dimensions and can be applied to any downstream tasks.

Algorithm 1 Sufficient Universal Projection (SUP)

Input: Data $\{(X_i, Z_i)\}_{i=1}^n$, partition H and number of dimension q;
Output: Sufficient projection P;

if Z_i is continuous **then**
  Set $\widetilde{Z}_i = Z_i$;
else if Z_i is discrete **then**
  Train a classifier f_{cls} by $\{(X_i, Z_i)\}_{i=1}^n$;
  Set $\widetilde{Z}_i = f_{\text{cls}}(X_i)$;
end if
Obtain M^{PMS} using $\{(X_i, \widetilde{Z}_i)\}_{i=1}^n$ by Algorithm 2;
Calculate the leading eigenvectors $\{\psi_j\}_{j=1}^q$ of M^{PMS};
Obtain $P = I_{p_1} - \sum_{j=1}^q \psi_j \psi_j^\top$;
Return: P.

4.3 Connection with Existing Method

Assumption 3.4 guarantees that our statistical model encompasses a wider family of models, indicating that our method relies on relatively less stringent conditions. In this subsection, we provide a brief comparison of our method to other projection-based methods.

INLP (Ravfogel et al., 2020): Both our method SUP and INLP employ linear projections to minimize the influence of the sensitive attribute $\boldsymbol{Z}$ in the representations. INLP seeks to identify the null space of the weight matrix W of linear classifier $\boldsymbol{Z} = f(W\boldsymbol{X})$. This framework can be viewed as a specific instance of the Model 1

RLACE (Ravfogel et al., 2022a): Both SUP and RLACE operate under weak linearity assumption as expressed in Assumption 3.4. In RLACE, the function f in Model (1) is interpreted as the inverse of a link function in the generalized linear model. In contrast, our approach imposes no specific constraints on the form of f.

A detailed analysis and comparison with existing methods are available in Appendix A.2.

5 Experiment and Settings

In all experiments conducted, the number of partitions was consistently set to 100. The optimal number of dimensions to be removed was determined through cross-validation. Detailed insights into the impact of the parameter q on performance

can be found in Appendix A.8. Codes are available at the GitHub.[1]

5.1 Static Word Embedding Evaluation Tasks

We begin by demonstrating our method in the context of debiasing static word embeddings using 300-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on English Wikipedia data. We first calculate the cosine similarity between each word embedding with the gender direction $\overrightarrow{he}$ - $\overrightarrow{she}$. From these scores, we select words in two categories: those with the top 9000 highest scores, representing male-associated words, and those with the top 9000 lowest scores, representing female-associated words. The label of each class is the sensitive attribute Z and the projection matrix P is calculated through Algorithm 1. For a fair comparison, the following evaluations are based on the methodology outlined in (Gonen and Goldberg, 2019; Ding et al., 2022). We compare our results with the following baseline methods: hard-debiasing method (Hard) (Bolukbasi et al., 2016), gender-preserving debiasing method (GP) (Kaneko and Bollegala, 2019), word vector learning method (GN) (Zhao et al., 2018b), half-sibling regression (HSR) (Yang and Feng, 2020), INLP (Ravfogel et al., 2020) and DeSIP (Ding et al., 2022).

Clustering Gender Biased Words. Biased words tend to cluster together, and debiased embeddings may not escape this phenomenon. We use K-means clustering (K=2) to split the top 500 male-biased and top 500 female-biased words. A visualization graph is presented in Appendix 5.2. In Table 1 column one, we report the accuracy in splitting the 1,000 words into male and female clusters. Our method brings about a 50% reduction compared with the original GloVe and about 20% compared with the second-best method.

Correlation Using the top 50,000 most frequent words as targets, we calculate the Pearson correlation coefficient between the bias-by-projection and bias-by-neighbor results. The latter is calculated using the neighborhood metric, which counts the percentage of male and female-biased words within the 100 nearest neighbors of each target word. The result is presented in the second column of Table 1, and we achieve the lowest correlation coefficient.

[1] https://github.com/EnzeShi/Debiasing-with-Sufficient-Projection-A-General-Theoretical-Framework-for-Vector-Representations

Profession Words In this task, we determine the correlation between the original bias and the number of male neighbors among the 100 nearest neighbors of profession words, as listed by (Bolukbasi et al., 2016; Zhao et al., 2018b). The correlation coefficient is shown in Table 1. Our method reduces the coefficient by 20% compared with the original GloVe and achieves the best result.

Classifcation We selected the top 2,500 biased words for each gender and trained a support vector machine (SVM) model using 1,000 randomly sampled words for each baseline model. We then applied the trained classifier to the remaining 4,000 words to predict gender bias direction. The prediction accuracy is shown in Table 1. Lower accuracy implies that the original embedding does not contain enough gender-related information. Our method has the least accuracy among all debiasing methods, indicating that it preserves the least gender bias.

Word Embedding Association Test The WEAT (Caliskan et al., 2017) is a permutation-based test that measures bias in word embeddings. Refer to Appendix A.5 for the details of WEAT. The results are reported in terms of absolute effect sizes(d) and p-values (p). The effect size is a normalized measure of how separated two distributions are. A high effect size indicates a larger bias between the target and attribute words, and the p-value denotes whether the bias is statistically significant or not. We conducted two experiments employing the Pleasant & Unpleasant (Task 1) and Career & Family (Task 2) word sets, with male and female names serving as the attribute sets. As illustrated in Table 1, the outcomes from both tasks reveal that the p-values are not statistically significant, suggesting an absence of significant bias. Furthermore, we observed the smallest effect sizes in both tasks, highlighting our methodology's effectiveness in minimizing bias within word embeddings.

5.2 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 compare the original GloVe and our debiased embeddings. Figure 2 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.

	Clustering	Correlation	Profession	Classify
GloVe	1.0000	0.7727	0.8200	0.9980
Hard	0.8050	0.6884	0.7161	0.9068
GP	1.0000	0.7700	0.8102	0.9978
GN	0.8560	0.7336	0.7925	0.9815
HSR	0.9410	0.6422	0.6804	0.9055
INLP	0.6336	0.5718	0.6651	0.8160
DeSIP	0.7920	0.6421	0.7060	0.8550
SUP	**0.5198**	**0.5360**	**0.6515**	**0.7247**

	Task1		Task2	
	p (↑)	d (↓)	p (↑)	d (↓)
GloVe	0.090	0.704	0.00*	1.905
Hard	0.363	0.187	0.00*	1.688
GP	0.055	0.832	0.00*	1.909
GN	0.157	0.541	0.074	0.753
HSR	0.265	0.340	0.00*	1.555
INLP	0.195	0.475	0.129	0.595
DeSIP	0.268	0.335	0.001*	1.462
SUP	**0.411**	**0.119**	**0.142**	**0.565**

Table 1: **Left:** Static word embedding bias evaluation tasks. A lower number in each column indicates better debiasing performance. Baseline results are reported by (Ding et al., 2022). Our method surpasses all other methods; **Right:** WEAT result. In each column of p-value, * indicates statistically significant compared with $\alpha = 0.05$; In each column of d, a value closer to 0 is indicative of less bias. The best results are boldfaced.

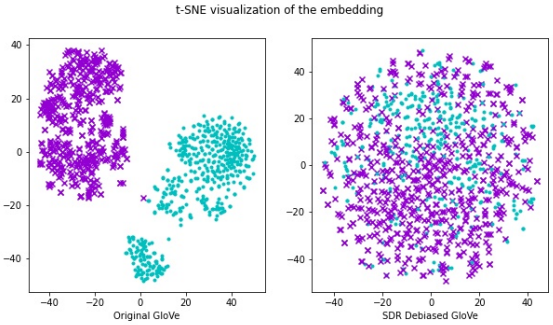

Figure 2: t-SNE visualization.

5.3 Word Similarity Tasks

While reducing bias is our primary goal, it is crucial not to destroy other semantic information encoded in word embeddings. We evaluate our algorithm by the following word similarity tests: RG65 (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2001), Rarewords (Luong et al., 2013), MEN (Bruni et al., 2014), MTurk-287 (Radinsky et al., 2011), and MTurk-771 (Halawi et al., 2012), SimLex-999 (Hill et al., 2015), and SimVerb-3500 (Gerz et al., 2016). These datasets associated with each task contain word pairs and a corresponding human-annotated similarity score. We calculate Spearman's rank correlation coefficient between the two ranks. The results of our method and the original GloVe are shown in Table 2. We observe an overall non-decreasing performance in most of the tasks, showing that the semantic information is protected.

5.4 Sentence Embedding Association Test (SEAT)

In addition to testing on static word embeddings, we also test on contextualized word embeddings. SEAT (May et al., 2019), extends the WEAT test by leveraging simple templates such as 'This is a <word>' to obtain the individual word's contextualized embedding. We use the implementation results from (Meade et al., 2022). The baseline includes BERT base uncased, CDA and Dropout (Webster et al., 2020), SentDebias (Liang et al., 2020), and INLP.

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.

For a detailed list of the SEAT tests used to measure each type of bias in our work, the complete results, we refer readers to Appendix A.6. In Table 2, we display the average effect size for each SEAT task category evaluated. Our findings reveal superior performance in two out of the three tasks while delivering comparable results to the INLP method in the Gender task. Notably, our SUP method exhibits enhanced performance across a variety of bias-influenced topics.

5.5 Extrinsic: Fairness Text Classification

For the extrinsic task, we consider the fairness text classification problem. We conduct experiments over three different tasks – sentiment analysis (MOJI), biography classification (BIOS), and toxic comment classification (Toxic). The detail of

	RG65	WS	RW	MEN
GloVe	0.7540	0.6199	0.3722	0.7216
SUP	**0.7913**	**0.6617**	**0.3986**	**0.7423**
	MT-287	MT-771	SimLex	SimVerb
GloVe	0.6480	0.6486	0.3474	0.2038
SUP	0.6349	**0.6792**	**0.3949**	**0.2493**

	Gender(↓)	Race (↓)	Religion (↓)
BERT	0.620	0.620	0.492
+CDA	0.722	0.569	0.339
+Dropout	0.765	0.554	0.377
+INLP	**0.204**	0.639	0.460
+SentDebias	0.434	0.612	0.439
+SUP	0.218	**0.432**	**0.261**

Table 2: **Left:** Word similarity results. A higher value indicates a better semantic correlation; **Right:** SEAT average effect sizes for debiased BERT. A lower number in each column indicates better debiasing performance. The best results are boldfaced. Baseline results are from (Meade et al., 2022).

the datasets is described as follows:

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.

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 $65\% : 10\% : 25\%$.

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. 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.

The fairness criterion is defined by *Equality of Opportunity* (EO), i.e. a classifier is considered fair if its prediction is independent of the sensitive attribute given the true label. For BIOS and MOJI data, it is measured by considering the gap in the True Positive Rate (TPR) between different sensitive attribute groups:

$$TPR_{z,y} = P[\hat{Y} = y|Z = z, Y = y],$$
$$GAP_y^{TPR} = TPR_{z,y} - TPR_{z',y}.$$

The root-mean-square (RMS) gap over all groups is $GAP_{RMS}^{TPR} = \sqrt{\frac{1}{|C|}\sum_{y\in C}(GAP_y^{TPR})^2}$.

We follow the original implementation of MOJI and BIOS that use race and gender labels as sensitive attributes Z. The results are shown in Table 3. We report the Accuracy, the GAP_{RMS}, and the Time in seconds for BIOS and MOJI. The baselines are from (Ravfogel et al., 2020), (Ravfogel et al., 2022a), (Chowdhury and Chaturvedi, 2022), and (Ravfogel et al., 2022b).

In Table 3, we implement each task using BERT and establish it as the baseline - this represents the results without any fairness considerations. Our findings reveal that in the BIOS task, while the INLP and FaRM achieve low RMS, it is accompanied by a compromise in accuracy. In contrast, our SUP method demonstrates balanced performance on both fronts. For MOJI, our algorithm stands out, yielding the smallest discrepancy gap among all methods, all the while maintaining uncompromised accuracy. Moreover, our algorithm benefits from having an explicit solution, eliminating the need for iterative calculations, and running significantly faster than many existing baselines.

In addition, we also conduct experiments on a more challenging dataset: the Toxic Comment Classification (Dixon et al., 2018). Within this dataset, each sample may belong to **multiple sensitive attribute groups**, embodying intersectionality in biases. For instance, a single comment might simultaneously belong to 'black' and 'gay' sensitive groups. We adhere to the definitions and gap measurements outlined by Dixon et al. (2018), $GAP_{\text{toxic}} = \sum_{z\in\mathcal{Z}} |TPR_{z,0} - \text{mean}_{z\in\mathcal{Z}}(TPR_{z,0})|$, where $\text{mean}_{z\in\mathcal{Z}}(TPR_{z,0})$ is the average of TPR gaps of all sensitive attributes. where $\text{mean}_{z\in\mathcal{Z}}(TPR_{z,0})$ represents the average of True Positive Rate (TPR) gaps across all sensitive attributes. The sensitive attribute in this scenario is depicted as a 50-dimensional vector, illustrating the relative frequency of sensitive words within sentences.

For the original BERT model, the Area Under the Curve (AUC) was 95.5, and the GAP_{toxic} was 7.34. By employing our method, we man-

	BIOS			MOJI		
	Acc.(↑)	GAP(↓)	Time(↓)	Acc.(↑)	GAP(↓)	Time(↓)
BERT	79.1	14.5	-	71.6	31.0	-
+INLP	71.9	9.9	271	62.2	15.8	1003
+RLACE	76.9	13.2	4312	72.2	15.4	2456
+FaRM(unconstrained)	55	**7.9**	6723	63.5	14.0	4162
+Kernel(Poly)	**79.9**	16.8	3914	72.9	17.3	8861
+Kernel(RBF)	60.7	18.0	3487	**74.1**	13.3	5496
+SUP	76.4	12.7	**6.76**	69.1	**10.5**	**33.04**

Table 3: **Left:** Result of BIOS text classification. Predict using [CLS] token. **Right:** Result of MOJI text classification. The best result is boldfaced.

aged to maintain the AUC at 95.0 while reducing the GAP_{toxic} to 5.95, showcasing the efficacy of our approach in mitigating biases while preserving model performance. It is crucial to highlight that our methodology effectively manages the intricacies of intersectional biases in toxic comment classification, a complexity not adequately addressed by other baseline algorithms. For a more detailed discussion, please refer to Appendix A.2.

6 Bridge between Debiasing and Fairness

In this section, we provide a theoretical analysis of how our proposed method can incorporate debiasing and fairness tasks into a unified framework and handle them simultaneously. As previously discussed, both tasks aim to achieve conditional independence with respect to certain variables. Here, we will demonstrate how minimal subspace $\mathcal{S}_2$ bridges these tasks. We provide the following theorem to prove the effectiveness of our framework in dealing with both debiasing and fairness tasks. Please refer to Appendix A.7 for detailed proof.

Theorem 6.1. *With the settings defined in Section 3 and linearity assumption, suppose* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \mid Q\boldsymbol{X}$, *then* $\widetilde{\boldsymbol{X}} = (I-Q)\boldsymbol{X}$ *is a debiased representation. Further, suppose* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid Q_y\boldsymbol{X}$, *if* $\text{Span}\{Q_y\} \subseteq \text{Span}\{Q\}$, *then* $\widetilde{\boldsymbol{X}} = (I-Q)\boldsymbol{X}$ *is a fair representation.*

Theorem 6.1 states that the projected representation $\widetilde{\boldsymbol{X}} = (I-Q)\boldsymbol{X}$ has no correlation with the sensitive attributes $\boldsymbol{Z}$, which achieves the goal of debiasing task. Moreover, if the subspace spanned by sensitive attribute $\boldsymbol{Z}$ ($\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$) is included in the subspace spanned by target attribute $\boldsymbol{Y}$ ($\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$), we can achieve the goal of fairness task by projecting the original representation on $(I-Q)$.

The theoretical property is consistent with the experimental results shown above. For the debiasing task, the matrix $Q = \sum_{i=1}^{q} \psi_i \psi_i^\top$ in Algorithm 1 is the estimated central mean space regarding Z, then $I-Q$ forms a sufficient projection defined in Theorem 6.1, which shows great improvement upon existing state-of-the-art methods. For fairness task, the eigenvectors $\{\psi_j\}_{j=1}^{q}$ calculated in Algorithm 1 recovers the matrix Q stated in Theorem 6.1. If we have $\text{Span}\{Q_y\} \subseteq \text{Span}\{Q\}$, then we can set $\widetilde{\boldsymbol{X}} = (I-Q)\boldsymbol{X}$ to get the fair representation. However, in real data, this condition is usually violated, which means $\text{Span}\{Q_y\} \nsubseteq \text{Span}\{Q\}$. Therefore, the SUP may not achieve the optimal fair representation in downstream tasks.

7 Conclusion and Future Work

In this paper, we propose a theoretically grounded framework for reducing bias by projecting vector representations to an unbiased subspace. It can reduce biased information effectively in both intrinsic and extrinsic tasks, as well as different kinds of representations. In addition, we provide a theoretical guarantee about the effectiveness of our method in reducing biased information. Finally, our method not only surpasses existing state-of-the-art approaches in bias mitigation while maintaining robust task performance but also achieves superior computational efficiency.

While this work has demonstrated its effectiveness in various tasks, it has the potential to be applied to other applications that rely on vector representation. We are also interested in combining our method with the different other notions of fairness. We aim to explore these directions in future work.

Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC ES/T012382/1) and the Social Sciences and Humanities Research Council (SSHRC 2003-2019-0003) under

the scheme of the Canada-UK Artificial Intelligence Initiative. The project title is BIAS: Responsible AI for Gender and Ethnic Labour Market Equality. Bei Jiang and Linglong Kong were partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC. We also thank all the constructive suggestions and comments from the reviewers.

References

Yves Aragon. 1997. A gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, 12(3):355–372.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955, Online. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. 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.

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.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

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

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2020. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. 2022. Learning fair representations via rate-distortion maximization. *Transactions of the Association for Computational Linguistics*, 10:1159–1174.

R Dennis Cook and Bing Li. 2002. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474.

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 *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, and Bei Jiang. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11864–11872.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940.

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

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough!–on the effectiveness of debiasing mlms and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Ker-Chau Li. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning*, pages 104–113.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. 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 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022a. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Richards et al. 2016. Non-binary or genderqueer genders. *INTERNATIONAL REVIEW OF PSYCHIATRY*, 28(1):95–102.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

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.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Zekun Yang and Juan Feng. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9434–9441.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix

A.1 Toy example for weak linearity and minimal subspace

Suppose $\boldsymbol{X} \in \mathbb{R}^4$ and $\boldsymbol{Z} \in \mathbb{R}$, with two orthogonal directions $\beta_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, 0)$ and $\beta_2 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0, 0)$. Consider the following two models:

$$\boldsymbol{Z} = \beta_1^\top \boldsymbol{X} + \beta_2^\top \boldsymbol{X} + \varepsilon, \quad (2)$$

$$\boldsymbol{Z} = \sin(\beta_1^\top \boldsymbol{X}) + \exp(-\beta_2^\top \boldsymbol{X}) + \varepsilon. \quad (3)$$

Then (2) satisfies the strong linearity assumption and (3) satisfies the weak linearity assumption.

For model (3), denote $\beta_3 = (0, 0, 1, 0)$ and $\beta_4 = (0, 0, 0, 1)$, then $\text{Span}\{\beta_1, \beta_2, \beta_3, \beta_4\} = \mathbb{R}^4$, the entire representation space. Let $\mathcal{S}_2 = \text{Span}\{\beta_1, \beta_2\}$ and $\mathcal{S}_2' = \text{Span}\{\beta_1, \beta_2, \beta_3\}$, with

their associated orthogonal subspaces $\mathcal{S}_1 = \text{Span}\{\beta_3, \beta_4\}$ and $\mathcal{S}_1' = \text{Span}\{\beta_4\}$. Note that both $\mathcal{S}_2$ and $\mathcal{S}_2'$ contain all the directions related to $\boldsymbol{Z}$. But $\mathcal{S}_2$ has smaller dimension than $\mathcal{S}_2'$, thus we will lose less information when we project $\boldsymbol{X}$ on $\mathcal{S}_1$.

Therefore, in model (3), both $\mathcal{S}_2$ and $\mathcal{S}_2'$ are sufficient dimension reduction subspaces. The central subspace $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{X}}$ is equal to $\mathcal{S}_2 = \{\beta_1, \beta_2\}$, which is the minimal subspace.

A.2 Comparison with other projection methods

We conduct a comparative analysis between our method and other projection-based methods.

INLP: Both our method SUP and INLP employ linear projections to minimize the influence of the sensitive attribute $\boldsymbol{Z}$ in the representations. The underlying principle of INLP revolves around identifying the null space of the weight matrix, denoted as $W \in \mathbb{R}^{k \times p_1}$, which corresponds to the parameters of linear classifier $\boldsymbol{Z} = f(W\boldsymbol{X})$, where f represents the classifier function and k is the number of classes. This framework can be viewed as a specific instance of the Model 1, where the subspace spanned by $\beta_1, \ldots, \beta_q$ exactly corresponds to the union of row spaces of W_i during iterations. Specifically, INLP captures k directions (rows of weight matrix W_j) at each iteration, while SUP finds q directions in a single run, which is more flexible and computationally efficient.

RLACE: Both SUP and RLACE operate under linearity assumption as expressed in Assumption 3.4. In RLACE, the function f in Model (1) is interpreted as the inverse of a link function in the generalized linear model. In contrast, our approach imposes no specific constraints on the form of f. While RLACE achieves debiasing by solving a minimax problem to identify the projection P that safeguards the sensitive attribute, our method directly estimates the directions with a closed form, offering superior computational efficiency.

Advantage of SUP: The main distinction between our proposed methodology and existing projection-based debiasing methods pertains to the range of tasks they can address. For instance, INLP is principally designed for handling categorical sensitive attributes. In the context of the toxic data task, the sensitive attribute is no longer a categorical variable, thereby undermining the effectiveness of INLP. However, it is important to note that our SUP algorithm does not violate the structure of Model (1) under the linearity assumption. As a result, our approach remains capable of estimating the directions $\beta_1, \ldots, \beta_q$ and mitigating bias through the Algorithm 1. This highlights the versatility of our SUP algorithm, showcasing its capability to adeptly manage a spectrum of uni-/multivariate and discrete/continuous sensitive datasets. The capability of managing a sensitive attribute as a continuous variable also aligns more closely with contemporary sociological understandings. For instance, consider the interpretation of gender as a spectrum(Richards et al., 2016) rather than a binary categorization. As such, models that can accommodate continuous variables for sensitive attributes are better equipped to reflect these more nuanced perspectives, thereby promoting fairness and inclusivity in their outcomes.

A.3 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 $M^{\text{SIR}} = \sum_{h=1}^H p_h m_h m_h^\top$ 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.4 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.

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.5 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

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 $M_j^{\text{SIR}} = \sum_{h=1}^H p_{j,h} m_{j,h} m_{j,h}^\top$;
end for
Calculate the weighted sum of estimators $M^{\text{PMS}} = \sum_{j=1}^{p_3} w_j M_j^{\text{SIR}}$;
Return: M^{PMS}.

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.6 Detail 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.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 $\text{Span}\{Q_y\} \subset \text{Span}\{Q\}$, then $\text{Span}\{Q_y\}$ is orthogonal to $\text{Span}\{I - Q\}$, 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 Effect of q

In this section, we use MOJI data to illustrate the impact of the dimension removed q on performance of debiasing and task accuracy, see Figure 3 below.

It is important to note that the debiasing procedure may distort the relevant concepts or key information, denoted by Y. Generally, as q increases, both the sensitive information Z and part of the target information Y are excluded from the

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.

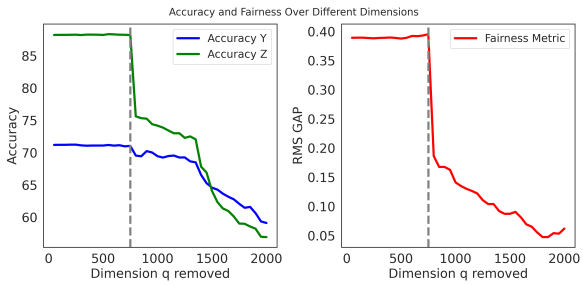

Figure 3: Trends of accuracy and GAP for MOJI data with number of dimension q removed.

debiased representation. This occurs due to the intersection of the subspaces spanned by Z and Y. Consequently, a rise in q leads to a simultaneous reduction in accuracy and the gap, illustrating a delicate equilibrium and trade-off between targeting and debiasing performance.

A.9 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.10 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.11 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.