Non-linear Debiasing of Sentence Embeddings with Kernel PCA

Vincent Bardenhagen
vbardenha@student.ethz.ch
Legi number: 19-945-526

Kei Ishikawa
kishikawa@student.ethz.ch
Legi number: 18-908-335

Abstract

Embeddings of natural languages are known to incorporate social biases. To address this issue, principal component analysis (PCA) has been widely adopted for the removal of these biases as its linearity allows for conceptual and implementational simplicity. However, the debiasing algorithms based on PCA work well only if the bias concept can be well approximated as a *linear* subspace of the embedding space. In modern natural language processing, embeddings are obtained using gigantic models based on the attention mechanism, which can be highly non-linear. Therefore, there is no good reason to believe the above assumption holds for such embeddings. To remove potentially non-linear bias, a *non-linear* debiasing procedure using kernel PCA was recently proposed. However, the proposal was only a theoretical description with a technical difficulty that makes it impossible to implement in practice. In this article, we fill in this gap by presenting a feasible version of the debiasing procedure using kernel PCA and reporting empirical performances of the non-linear debiasing method on sentence embeddings of the Bidirectional Encoder Representations from Transformers (BERT).

1 Introduction

Natural language processing has become an integral part of decision processes with a direct impact on the lives of many people. These include selecting job applications, analyzing legal documents and social science research. Legal provisions and ethical codes necessitate fairness and anti-discrimination provisions in these domains.

Bolukbasi et al. (2016) first analyzed that vector space word embeddings do not meet fairness requirements as they discovered the presence of gender stereotypes in the embedding structure. A similar result was described for sentence embeddings in Liang et al. (2020) where the authors extended the findings to biases among Abrahamic religions.

Concurrently, researchers worked on the removal of social biases in word and sentence embeddings. The original and most commonly used approach by Bolukbasi et al. (2016) removes the linear subspace associated with the bias using PCA. However, this approach relies on the assumption that the bias concept can be well approximated as a linear subspace of the embedding space. This assumption is natural for many word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) because they are encouraged to encode semantic information into the linear structure of the embedding space. In contrast, the transformer architecture of BERT contextualizes each word using the information of all the other words in the sentence. As this mechanism is realized by a highly non-linear neural network model, the information of biases is not likely to be encoded as a linear structure in the embedding space.

To remove non-linear biases, Vargas and Cotterell (2020) introduced a *non-linear* version of the debiasing method by Bolukbasi et al. (2016) using kernel PCA Schölkopf et al. (1998). Though their method is of interest from a theoretical perspective, in practice, it is impossible to implement. Hence, they failed to provide empirical results of the non-linear debiasing method they proposed.[1]

In this article, we present a feasible version of their approach and report empirical performances of the non-linear debiasing method on sentence embeddings of BERT (Devlin et al., 2018).

[1] Though these points are not clearly stated in their paper, we contacted the first author of the paper and he agreed that their method is infeasible. He also stated that they did not implement their proposed method but conducted a different kind of experiment to analyse their ”linear subspace hypothesis”.

Our Contributions [2]

- We propose an extension of the existing benchmark method for evaluating the gender bias in sentence embeddings (Section 3.2) and demonstrate that it has better bias detection capacities. (Section 4.1)

- We modify the debiasing algorithm by Vargas and Cotterell (2020) and propose an algorithm that overcomes the limitation of the original version that it is practically impossible to implement. (Section 3.4, Orthogonalization)

- Simultaneously, our proposed algorithm generalizes its original; it can handle general defining sets. (Section 3.4)

- Accordingly, ours is the first work to implement and apply the non-linear debiasing using kernel PCA to any kinds of embeddings and examine its empirical performance. (Section 4.3)

2 Related Works

Our work combines two directions of recent research in debiasing natural language models, namely, bias reduction of *sentence* embeddings and *non-linear* bias reduction. To describe the first direction, we discuss the complexity of sentence embeddings in Section 2.1 and introduce related research on sentence embedding debiasing in Section 2.2. For the second direction, we explain the non-linear debiasing idea of Vargas and Cotterell (2020) and analyze limitations of their work in Section 2.3.

2.1 Complexity of Sentence Embeddings

The way information is encoded in BERT sentence embeddings is not fully understood as of now. However, it was shown by Jawahar et al. (2019) that BERT captures surface features, syntactic features, and semantic features in different layers of the model. Such rich information is not likely to be encoded in a simple structure as in word2vec or GloVe word embeddings, where primarily word-level semantics are captured. Furthermore, gender subspace in sentences is clearly complex. In the formulation of sentences, the interdependence between words plays an important role in deriving meaning. In the sentences: ”The father defended his child” and ”The mother defended her child” defend can either be read as a physical or a verbal act, depending on the ”aggressiveness” associated to the actors. In contrast, the sentences ”The father likes walking” and ”The mother likes walking” would not differ in this respect. Thus, the perception of gender difference will vary depending on the context where the gendered word appears. Therefore, we can expect that the subspace of the gender concept is complex and it might be captured as a non-linear structure in BERT embedding.

2.2 Bias Reduction of Sentence Embeddings

The analysis of bias in natural language models was started by Bolukbasi et al. (2016), where they analyzed gender bias in word embeddings. Soon after this publication, sentence embedding models started to attract more attention because of their near-human capacities in some benchmark tasks. Such sentence embeddings are generated by transformer models such as BERT (Devlin et al., 2018) and GPT (generative pre-training) (Radford et al., 2018) that are pre-trained on large corpora with proxy-tasks. Consequently, May et al. (2019) applied the methodology of Bolukbasi et al. (2016) on sentence embeddings and found that gender bias in BERT and GPT-2 sentence embeddings is less pronounced than in word embeddings. Liang et al. (2020) deepened this analysis and demonstrated that the bias discussed in May et al. (2019) can be reduced if PCA debiasing is used. However, their method of analysis has a limitation as it does not capture the interdependence of words in sentences. In Section 3.2, we discuss this issue and propose a new method that can incorporate such interdependencies better.

2.3 Non-linearity in Debiasing

The non-linear debiasing method using kernel PCA was originally proposed by Vargas and Cotterell (2020). However, their method is impossible to implement in practice because of technical difficulty at the orthogonalization step of debiasing, as discussed in Section 3.4. Consequently, they did **not** provide any empirical results of the non-linear debiasing method that they introduced.[1]

Instead of benchmarking bias reduction capacity of non-linear debiasing methods using kernel PCA,

[2] Our contributions turned out to be broader than anticipated. We intended to apply the non-linear debiasing method from Vargas and Cotterell (2020) on sentence embeddings as in Liang et al. (2020) as a novel combination of existing approaches. But, when analyzing Vargas and Cotterell (2020) thoroughly, after starting the code development, we noticed that their method can not be implemented. Hence, we first had to come up with a feasible debiasing algorithm *and* afterwards analyze applicability on sentence-embeddings.

they used several (non-linear) metrics for calculating WEAT score (Section 3.1). They examined if the use of different metrics in the calculation changes the resulting WEAT score of embeddings debiased by linear PCA. As a result, they found that the use of different metrics in the calculation of WEAT score does not have a significant impact on the score itself. From this result, they conclude that they validated the so-called ”linear subspace hypothesis”, that bias concept in the word embeddings can mostly be captured by a linear subspace.

However, we believe there is a logical leap in their argument; even if the use of different metrics in the scoring stage does not affect the bias score of standard linearly debiased embeddings, it does not necessarily mean that use of non-linear metrics in debiasing stage does not improve the fairness scores.

Therefore in our work, we devise an implementable version of the non-linear debiasing method and benchmark it to examine if it is actually worse than the linear method as Vargas and Cotterell (2020) claimed.

In our experiments, we also use multiple similarity metrics in the calculation of SEAT (Section 3.1) score as in Vargas and Cotterell (2020). However, this is not to replicate their reasoning on the ”linear subspace hypothesis”. We employ multiple metrics in the calculation of the SEAT score because there is no reason to prefer one similarity metric over another in the case of sentence embeddings.

3 Methodologies

We firstly introduce the standard benchmark method for evaluating bias in sentence embeddings in Section 3.1. Then, we introduce the use of natural sentences as a novel extension of the benchmark method in Section 3.2. Finally, in Section 3.3 and 3.4, we explain the linear and non-linear debiasing methods for sentence embeddings that are compared in our experiments.

3.1 Association Tests for Embeddings

For the analysis of social biases in sentence embeddings, it is crucial to establish a good benchmark method. The benchmark should satisfy the following criteria: (1) reflect the psychological view on social biases in humans, (2) handle the complexity of the sentence domain, (3) be reproducible, (4) give some intuition on relevance and intensity.

To demonstrate and measure implicit stereotypes in humans, Greenwald and Banaji (1995) introduced the Implicit Association Tests (IAT). It measures the difference between reaction speed when grouping stereotypical combinations (e.g male/career words) together and those when grouping non-stereotypical combinations (e.g male/family words) together. These word groups were designed by psychologists and used by Bolukbasi et al. (2016) to define the Word Embeddings Association Test (WEAT). WEAT and IAT differ only in the way they measure the strength of association between concepts. Instead of the reaction time difference in IAT, the difference in cosine similarity of word embeddings is used in WEAT. The Sentence Embeddings Association Test (SEAT) is a direct extension of WEAT. As the similarity metric, SEAT employs the similarity between embeddings of sentences which are produced by filling the words from IAT into simple sentence templates such as ”This is ...”.

Formally, the procedure of all three tests can be described as follows. Firstly, we define sets of words/sentences representing the target concepts, for example, male and female, as S_m and S_f. Also, we define representative sets for attribute concepts, for example, science and arts, as S_s and S_a. Then, the association of the concept male with specific target concept, i.e., science or art, is quantified as the following score:

$$\text{score}_m = \sum_{w \in S_m} \left(\sum_{a \in S_s} \mathrm{d}(w, a) - \sum_{a \in S_a} \mathrm{d}(w, a) \right).$$

Here, $\mathrm{d}(\,\cdot\,,\,\cdot\,)$ is a similarity metric between words. In the original WEAT and SEAT, cosine similarity between embeddings of each word/sentence is usd. In IAT, the average reaction time is used as this similarity metric. The score of association for concept female is defined equally. Finally, we can calculate the association test score as:

$$\text{WEAT/SEAT/IAT} = \text{score}_m - \text{score}_f$$

In the absence of bias, this score should not deviate significantly from zero. The score is positive if the bias takes the direction expected by the psychologists defining the word sets, and is negative if there is bias in the opposite direction.

For the assessment of the statistical significance of WEAT and SEAT score, a permutation test is used in May et al. (2019) but not in the analysis of debiasing method in Liang et al. (2020). Using

SEAT score with permutation tests, the criterion (1), (3) and (4) can be satisfied but (2) remains unmet. This is because the sentence templates (e.g ”This is a man”) only reflect a very small proportion of the use of gendered sentences in English. In these artificial sentences, the complexity of the sentence domain as discussed in Section 2.1 is overlooked. This implies that the effect of the interdependence of words in sentences can not be captured with this test.

3.2 Association Tests using Natural Sentences

To meet the criteria (2), we propose to use natural sentences from the corpora to extract the gender subspace and to calculate SEAT score. This is a step towards addressing the complexity of the sentence domain. Instead of evaluating bias using example sentences generated by artificial sentence templates as in previous work, we construct our examples of gendered sentences from human written corpora. However, just extracting sentences would lead to an analysis of bias in the corpora instead of the sentence embedding model. If in the corpora males are more often related to business activities, the test will capture this bias even if the embedding model is not biased. To avoid this pitfall, we create the gender opposite sentence to each natural gendered sentence, by swapping the gender-specific word in the sentence. In this way, our testing procedure can take into account the complexity of the sentence domain better than previous work.

This improvement comes at the cost of harder reproducibility. Furthermore, we might not be able to capture all names or gendered phrases when swapping gendered words to transform male sentences into female sentences. The best solution would be to curate a list of a few hundred to thousand sentence pairs that only differed in the gender of the involved people and to publish this benchmark dataset. However, this is beyond the scope of the project as the design and collection of such benchmark dataset requires expertise in social-science and linguistics that we do not possess.

In our experiments, we use multiple similarity metrics when calculating the SEAT scores, as discussed in Section 2.3. In the calculation of the SEAT scores for the three types of embeddings, namely, original embeddings, linearly debiased embeddings, and non-linearly debiased embeddings, we proceed as follows.

1. Define three similarity metrics for sentence pair embeddings. With x_{gend} being the gendered sentence embedding and x_{attr} being the attribute sentence embedding:

   (a) Cosine Similarity:

   $$\theta = \frac{x_{\text{gend}} \cdot x_{\text{attr}}}{||x_{\text{gend}}||_2 ||x_{\text{attr}}||_2}$$

   (b) Gaussian Similarity:

   $$\theta = \exp\left(-\gamma ||x_{\text{gend}} - x_{\text{attr}}||_2^2\right)$$

   (c) Sigmoid Similarity:

   $$\theta = \tanh\left(\gamma x_{\text{gend}} \cdot x_{\text{attr}} + c_0\right)$$

   Here, $x \cdot y$ for vectors x, y indicates their inner product.

2. Define collections of gendered sentences and attribute sentences. For the attribute sentences, we use simple sentence templates from May et al. (2019), from the originally proposed SEAT (e.g ”This is an executive”). They consider three attribute groups derived from psychological implicit association tests. Namely, these are career $\leftrightarrow$ family, science $\leftrightarrow$ arts and math $\leftrightarrow$ arts. For the gendered sentences we can take one of the following approaches:

   (a) Take the sentence templates and plug in the female and male words exactly like Liang et al. (2020) as discussed in Section 3.1. (e.g ”This is a man” / ”This is a woman”)

   (b) As in Section 3.2, search the corpora for sentences containing clearly female or male words. Add these sentences to the ”gendered” sentence collection. Next, swap all gendered words to the opposite gender and also add these sentences to the collection. This enables us to create collections of sentences ”in the wild”. (e.g ”with *his* usual intelligence and subtlety” / ”with *her* usual intelligence and subtlety”)

3. Evaluate the difference between associations of gendered sentences with one or the other attribute category. To test for statistical significance of the SEAT score, we conduct permutation tests.

Through this procedure, we compute $3 \times 3 = 9$ different scores and p-values in total. Again, in our experiments, we apply this procedure to three types of sentence embeddings: original embeddings, linearly debiased embeddings, and non-linearly debiased embeddings. These results are discussed in the experiment section.

3.3 Debiasing by PCA

Following the pioneering work by Bolukbasi et al. (2016), the principal component analysis (PCA) has been applied to debiasing word embeddings as well as sentence embeddings (Liang et al., 2020). In these works, PCA is used firstly to extract the subspace of bias concepts (e.g. gender, religion, ethnicity), and then to orthogonalize the embeddings of a neutral word to the bias subspace. This procedure of orthogonalization is called "neutralization". In addition to "neutralization", there exists a procedure called "equalization", which equalizes the norm of bias component of embeddings for non-neutral words (e.g. man and woman, in the case of gender bias). As the bias component of an embedding is a projection of the embedding to the bias subspace, PCA can be used again to carry out "equalization". However, Liang et al. (2020) pointed out that this procedure cannot simply be applied to debias sentence embeddings. They argued that the identification of non-neutral sentences with respect to the bias considered is infeasible, due to the complexity of natural sentences. For the same reason, we leave the equalization step out of the scope of our consideration in this article.

To explain our approaches, we first introduce the linear debiasing method using PCA. Afterwards, we introduce the non-linear debiasing approach as a kernelized version of the linear approach. In this way, we clarify the similarities and differences between the two approaches.

Problem Setting

Let S be the set of sentences used for bias extraction and $\mathbf{e}_i \in \mathbb{R}^d$ be the sentence embedding of sentence $i \in S$. We call a set of sentences $D_n \subset S$ a *defining set* of the bias concept. The variations of the sentences in D_n is designed to represent the bias concept. For example, in the case of gender bias, a defining set can be {"He likes studying.", "She likes studying."}. We assume that we can create a dataset of the defining sets, written as $\mathcal{D} = \{D_n\}_{n=1}^N$. This dataset is a collection of the same bias concept in different contexts. For the simplicity of notations, we assume that D_n's are disjoint and $S = \cup_{n=1}^N D_n$. Our goal is to reduce the bias in embeddings $\mathbf{e}_i$ by leveraging the knowledge of the bias concept extracted from the defining sets.

Extraction of Bias Subspace

To extract the bias concept as the principal subspace of PCA, we apply PCA to the vector difference of the embeddings within each defining set. More formally, we compute the center of mass $\bar{\mathbf{e}}_n = \frac{1}{|D_n|} \sum_{j=1}^{|D_n|} \mathbf{e}_{n,j}$ for each defining set D_n and then compute the difference of embedding $(\mathbf{e}_{n,j} - \bar{\mathbf{e}}_n)$ for each sentence $j \in D_n$. Then, PCA model is trained on the data of these differences $\cup_{n=1}^N \{\mathbf{e}_{n,j} - \bar{\mathbf{e}}_n : j = 1, ..., |D_n|\}$.

Let X and $\tilde{X}$ be matrices whose row vectors are $\mathbf{e}_{n,j}^T$'s and $(\mathbf{e}_{n,j} - \bar{\mathbf{e}}_n)^T$'s, respectively. Here, we assume that each index $i \in S$ (i.e., X's row index) has one-to-one correspondence with the index pair (n, j), so that the i-th row vector of X, $X_{i,\bullet} = \mathbf{e}_{n,j}^T$. For simplicity of notations, we sometimes write the index of these matrices using index pairs of the form (n, j). For example, $X_{(n,j),k}$ indicates $X_{i,k}$.

With this notation, we can rewrite $\tilde{X}$ using X and a design matrix A as $\tilde{X} = AX$, where A is defined as:

$$A_{(n,j),(n',j')} = \delta_{(n,j),(n',j')} - \frac{1}{|D_n|}\delta_{n,n'}.$$

Here, δ_{j_1,j_2} denotes the discrete delta function which equals one when the indices j_1 and j_2 are equal, and zero otherwise. We can easily check that $\tilde{X} = AX$ as

$$\begin{aligned}
&[AX]_{(n,j),\bullet} \\
&= \sum_{(j',n')} A_{(n,j),(n',j')} X_{(n',j'),\bullet} \\
&= \sum_{(j',n')} \left(\delta_{(n,j),(n',j')} - \frac{1}{|D_n|}\delta_{n,n'}\right) X_{(n',j'),\bullet} \\
&= X_{(n,j),\bullet} - \frac{1}{|D_n|} \sum_{(n,j') \in D_n} X_{(n,j'),\bullet} \\
&= (\mathbf{e}_{n,j} - \bar{\mathbf{e}}_n)^T.
\end{aligned}$$

As can be seen, matrix A is symmetric and block diagonal, where each block corresponds to each defining set. Thus, this matrix is highly sparse in general and the matrix multiplication involving A can be implemented very efficiently.

Having defined matrix of demeaned embeddings $\tilde{X}$, PCA for the differences $\mathbf{e}_{n,j} - \bar{\mathbf{e}}_n$

simply becomes the truncated singular value decomposition (SVD) of $\mathbf{C} := \tilde{X}^T\tilde{X} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|D_n|}\sum_{j=1}^{|D_n|}(\mathbf{e}_{n,j}-\bar{\mathbf{e}}_n)(\mathbf{e}_{n,j}-\bar{\mathbf{e}}_n)^T$.

Orthogonalization to the Bias Subspace

Here, K rank truncated SVD of $\mathbf{C}$ is denoted as $\mathbf{C} \approx V\Sigma V^T$, where $V \in \mathbb{R}^{d\times K}$ is orthonormal matrix and $\Sigma \in \mathbb{R}^{K\times K}$ is diagonal matrix with non-negative diagonal elements. Then, the bias subspace becomes the column span of V. To debias an embedding vector we can orthogonalize embedding $\mathbf{e}$ using the projection matrix to $\mathrm{span}(V)$, $P_V = VV^T$ and receive the debiased part of the embeddings as

$$\mathbf{e}_{\text{ntr}} = \mathbf{e} - \mathbf{e}_{\text{bias}},$$

where $\mathbf{e}_{\text{bias}}$ is defined as

$$\mathbf{e}_{\text{bias}} = P_V\mathbf{e} = VV^T\mathbf{e}.$$

3.4 Debiasing by Kernel PCA

The non-linear version of the debiasing method can be derived simply as the kernelization of PCA, i.e., kernel PCA (Schölkopf et al., 1998), at least from a conceptual viewpoint. We describe this method following the work by Vargas and Cotterell (2020), in which this non-linear variant of the debiasing method using PCA was originally proposed.

However, Vargas and Cotterell (2020) only considered the cases where the size of the defining set is always two. In the following, we generalize their idea and discuss a debiasing method that works with arbitrary sizes of defining sets. Being able to choose sizes of defining set arbitrarily comes in handy when we need to find the bias of concepts that cannot be simply paired, such as age and religion.

Extraction of Non-Linear Bias Subspace

In non-linear debiasing using kernel PCA, we firstly find the principal subspace in the kernel feature space that represents the bias concept.

Before describing the extraction of the bias subspace, we introduce some notations for kernel method. Let $k(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the kernel of a reproducing kernel Hilbert space $\mathcal{H} \subset \{h : \mathbb{R}^d \to \mathbb{R}\}$ and $\phi_i \in \mathcal{H}$ (or equivalently $\phi_{(n,j)}$) be the feature of embedding $\mathbf{e}_i$ (or $\mathbf{e}_{(n,j)}$) in the kernel feature space. For notational simplicity, we define the operator $\phi^T : \mathcal{H} \to \mathbb{R}$ as $\phi^T := \langle\phi,\cdot\rangle_k$, or equivalently, $\phi_1^T\phi_2 := \langle\phi_1,\phi_2\rangle_k$ for any $\phi_1,\phi_2 \in \mathcal{H}$. Also, we introduce Φ as $\Phi = [\phi_1,\phi_2,...,\phi_{|S|}]$, which corresponds to matrix X^T in the linear case. The transpose of Φ is introduced similarly using the transpose of ϕ.

Now, we explain how we can extract the bias concept as the principal subspace of the difference of features $(\phi_{(n,i)} - \bar{\phi}_n)$ like in the linear case. Here $\bar{\phi}_n$ is defined similarly to $\bar{\mathbf{e}}_n$ as $\bar{\phi}_n = \frac{1}{|D_n|}\sum_{j=1}^{|D_n|}\phi_{(n,j)}$. By introducing a matrix notation of demeaned features $\tilde{\Phi}$ as the $\tilde{\Phi} = [(\phi_{(1,1)}-\bar{\phi}_1),(\phi_{(1,2)}-\bar{\phi}_1),...,(\phi_{(N,|D_n|)}-\bar{\phi}_N)]$, we can obtain the solution of kernel PCA as the truncated SVD, $\tilde{K} = \tilde{\Phi}^T\tilde{\Phi} \approx U\Lambda U$, where $U \in \mathbb{R}^{N\times K}$ is orthonormal matrix and $\Lambda \in \mathbb{R}^{K\times K}$ is diagonal matrix with non-negative diagonal elements.

Though the computation of $\tilde{K}$ might look complicated, we can simply compute $\tilde{K}$ as

$$\tilde{K} = (\Phi A)^T(\Phi A) = AK_{S,S}A,$$

where $K_{S,S}$ represents the original kernel matrix of $\{\mathbf{e}_i : i \in S\}$. This is because feature matrix $\tilde{\Phi}$ can be written as ΦA. Indeed, we see

$$\begin{aligned}&[\Phi A]_{\bullet,(n,j)}\\ &\quad= \sum_{(n',j')}\phi_{(n',j')}(\delta_{(n',j'),(n,j)} - \frac{1}{|D_n|}\delta_{(n',n)})\\ &\quad= \phi_{(n,j)} - \bar{\phi}_n.\end{aligned}$$

Orthogonalization to the Non-Linear Bias Subspace

Unlike linear PCA, kernel PCA does not give analytical expression of the orthogonalized embeddings. Instead, it gives expression for orthogonalized kernel feature. After some kernel arithmetic, it can be show that the projection of ϕ to the orthogonal compliment of the bias subspace becomes

$$\phi_{\text{ntr}} = \phi - \phi_{\text{bias}},$$

where ϕ_{bias} is a projection of the ϕ to the bias subspace that can be computed as

$$\phi_{\text{bias}} = \tilde{\Phi}U\Lambda^{\dagger}U^T\tilde{\Phi}^T\phi.$$

Here, $A^{\dagger}$ denotes the Moore-Penrose inverse of matrix A.

Finally, debiased embedding of $\mathbf{e}$ can be obtained by finding the pre-image, or the embedding that are close to the orthogonalized kernel feature

ϕ_{ntr} by solving the following optimization problem:

$$\begin{aligned}\mathbf{e}_{\text{ntr}} &= \arg\min_{\mathbf{e}' \in \mathbb{R}^d} ||\phi_{\mathbf{e}'} - \phi_{\text{ntr}}|| \\ &= \arg\min_{\mathbf{e}' \in \mathbb{R}^d} \{k(\mathbf{e}', \mathbf{e}') - 2k(\mathbf{e}', \mathbf{e}) \\ &\quad + 2K_{\mathbf{e}',S} A U \Lambda^{\dagger} U^T A K_{S,\mathbf{e}} + \text{const.}\},\end{aligned}$$

where $K_{\mathbf{e}',S}$ and $K_{S,\mathbf{e}}$ are row vector and columns vector defined as $[K_{\mathbf{e}',S}]_i = k(\mathbf{e}', \mathbf{e}_i)$ and $[K_{S,\mathbf{e}}]_i = k(\mathbf{e}_i, \mathbf{e})$, respectively.

This optimization problem can be approximately solved using gradient-based optimization with the initial point being $\mathbf{e}$, where we apply gradient descent in our experiments. In our experiments, we observed that very poor solutions can be obtained if the step size of gradient descent is not specified appropriately. Also, we observed that the numerical optimization is quite sensitive to other conditions such as the number of steps, parameters of kernels. Due to these observations, we expect that there is a room for improvement in the numerical optimization algorithm to obtain better solutions.

In Vargas and Cotterell (2020), the use of the learned pre-imaging method (Bakır et al., 2003) with additive decomposition (Kandasamy and Yu, 2016) is proposed instead of pre-imaging using numerical optimization. However, there are two technical difficulties in applying these methods. Firstly, the additive decomposition introduces some additional assumptions similar to linearity to the mapping from kernel feature space to the embedding space. As we aim to capture the non-linear correspondence of debiased kernel feature and the embeddings, this is not desirable. Secondly, to make matters worse, their method is actually infeasible because examples of input-output pairs that are necessary for [3]

Accordingly, even though numerical optimization can result in non-exact solutions, we had no choice but to use the optimization-based approach. It is the only feasible way to implement non-linear debiasing by kernel PCA.

4 Experiments

Our analysis consists of four blocks of experiments. Firstly, we calculate our bias benchmark scores of BERT sentence embeddings before the debiasing, to measure and analyze the presence of gender stereotype in the original embeddings. While Liang et al. (2020) stated that gender bias is present in BERT embeddings in the same way as it is in word embeddings, May et al. (2019) concluded that the bias is not statistically significant, by applying a permutation test to the SEAT scores. Secondly, we analyze the performance of the linear debiasing approach with PCA in terms of reducing the gender bias. Thirdly, we do the same for our implementation of the non-linear debiasing approach using kernel PCA. Fourthly, we analyze the change in the performance of the downstream task after the debiasing by using a semantic classification problem as the downstream benchmark task.

For the analysis, we use the same datasets as in Liang et al. (2020) to allow comparability of results in the evaluation of SEAT scores. The original datasets are the QNLI (Warstadt et al., 2018), CoLA (Warstadt et al., 2018) and SST-2 (Socher et al., 2013). We obtain the data from Liang et al. (2020) with small modifications from the original publications.

The processing pipeline is also constructed following the example of Liang et al. (2020). The first step is to search for sentences that occur in these datasets that contain words reflecting a possible bias concept such as male or female nouns. For identification of these word groups, we prepare the defining sets of the gender concept (for the words used for the defining sets, see Appendix 6). The second step is to create new sentences that only differ in the gender-specific word, for example exchanging *he* with *she*. In the training part of the dataset, these sentence pairs are used for training the debiasing approach and in the testing part of the dataset to compute the SEAT score with natural sentences. The resulting sentences are then transformed in a sentence embedding space using the pre-trained BERT from huggingface transformers (Wolf et al., 2019).

4.1 Evaluation of Gender Bias in Orignal Sentence Embeddings

In previous research, different versions of implicit association tests were used. We use the tests analyzing associations of male ↔ female with, career ↔ family, science ↔ arts and math ↔ arts. The results can be found in table 2.

The results for the cosine similarity on career &

[3] More specifically, their proposed method is infeasible because it uses a learned mapping Γ^T. Unfortunately, to train this mapping, we need to access pairs of "bias in feature space" and "bias in embedding space", which is impossible to obtain or even to define.

Test Attribute	cosine	p-value	Gaussian	p-value	sigmoid	p-value
Career & Family	1.5e-04	0.45	6.8e-03	0.01	9.0e-3	0.0678
Math & Arts	1.4e-04	0.13	17.4e-03	0.05	-0.3e-3	0.5054
Science & Arts	0.8e-04	0.31	5.0e-03	0.31	0.3e-3	0.4308
Career & Family (natural sentences)	0.9e-04	0.40	4.5e-03	<0.01	4.4e-3	0.03
Math & Arts (natural sentences)	1.2e-04	<0.01	9.1e-03	<0.01	3.0e-3	0.39
Science & Arts (natural sentences)	1.6e-04	0.03	8.5e-03	<0.01	1.4e-3	0.42

Table 1: Results of the various SEAT based on implicit association tests with value and corresponding p-value. The addition of ”natural sentences” in test attributes indicates our novel approach to use sentences from the datasets instead of sentence templates. Cosine, Gaussian, sigmoid are the metrics used in the SEAT calculation of the respective columns.

family attributes are on par with May et al. (2019), and thus, we did not find significant gender bias in sentence embeddings generated by BERT in this test. This result is the same for attributes science ↔ arts and math ↔ arts using the simple sentence templates. However, when we look at the results where natural sentences from real corpora are used to compute the adapted SEAT scores, we can confirm that there is a significant gender bias. The original test by May et al. (2019) cannot detect this gender bias for these two attribute associations.

Additionally, the use of multiple similarity metrics for calculating SEAT scores seems to enable more discoveries of bias in the sentence embeddings. The Gaussian kernel found statistically significant bias for all but one test if we set the significance level as 5%.

Above results provide statistical evidence that there exists gender bias not only in word embeddings but also in modern sentence embeddings.

4.2 Evaluation of Gender Bias after PCA Debiasing

In PCA debiasing, there is one hyperparameter choice to be made: the number of principal components (PC), which corresponds to the bias subspace. In the presented results we use two principal components, but the number of PC does not influence the results on a larger scale. The debiasing approach with PCA is successful and no significant bias is discovered after the debiasing across all metrics and test settings. The measured effect sizes are also greatly reduced. In most cases, the reduction in effect size is on the scale of one or two magnitudes and some values turn negative. The results are generally in line with the findings of Liang et al. (2020). However, they reported that their SEAT score sometimes becomes larger in absolute value and/or switched sign after debiasing. However, they did not perform significance tests. Our analysis reveals that the SEAT scoring procedure they applied does not yield bias scores significantly different from zero even before PCA debiasing, causing there debiasing performance analysis to be less expressive. Thus, through our SEAT score with natural sentences and the use of permutation tests before and *after* the debiasing, we can demonstrate the efficacy of PCA debiasing more rigorously.

4.3 Evaluation of Gender Bias after Kernel PCA Debiasing

For kernel PCA debiasing, there are multiple design choices to be made, namely, the type of kernel to use, the hyperparameters of the kernel, the number of principal components to consider and the optimization procedure for debiasing. Due to time and resource constraints, we only optimized some of the hyperparameters. The hyperparameters for the reported results are Gaussian kernel with $\gamma = 0.024$, two principal components, gradient descent with a learning rate of 0.4 and 30 iterations.

The results in 3 demonstrate that kernel PCA can also reduce bias in sentence embeddings measured by SEAT with natural sentences. When similarity is measured through cosine similarity or with the sigmoid kernel, the testing procedure reveals no significant bias after debiasing. However, the remaining SEAT score is consistently larger than after PCA debiasing. When the Gaussian kernel is the similarity metric for SEAT calculation, kernel PCA is not successful in reducing the bias strongly. Our permutation test finds significant bias even after kernel PCA debiasing is applied.

4.4 Trade-off of Fairness and Downstream Performance

There often is a trade-off between the usability of representations for downstream tasks and fair-

Test ID	cosine	p-value	Gaussian	p-value	sigmoid	p-value
Career & Family (natural sentences)	1.2e-07	0.50	2.1e-04	0.45	2.4e-3	0.26
Math & Arts (natural sentences)	-2.5e-05	0.61	-3.1e-05	0.50	-1.6e-3	0.76
Science & Arts (natural sentences)	-8.6e-06	0.51	7.1e-05	0.49	6.5e-5	0.30

Table 2: Results of the various SEAT based on implicit association tests after PCA debiasing with value and corresponding p-value. Again, the addition of ”natural sentences” in test attributes indicates our novel approach to use sentences from the datasets instead of sentence templates. Cosine, Gaussian, sigmoid are the metrics used in the SEAT calculation of the respective columns.

Test ID	cosine	p-value	Gaussian	p-value	sigmoid	p-value
Career & Family (natural sentences)	1.8e-05	0.47	4.5e-03	<0.01	2.3e-03	0.19
Math & Arts (natural sentences)	3.5e-05	0.19	3.3e-03	0.02	3.5e-03	0.38
Science & Arts (natural sentences)	2.8e-05	0.37	4.6e-03	0.07	5.2e-03	0.28

Table 3: Results of the various SEAT based on implicit association tests after kernel PCA debiasing with value and corresponding p-value. The addition of ”natural sentences” in test attributes indicates our novel approach to use sentences from the datasets instead of sentence templates. Cosine, Gaussian, sigmoid are the metrics used in the SEAT calculation of the respective columns.

ness. Given the current benchmark method, a fair embedding would be one that maps all sentences and words to one point in the vector space, as it always yields zero SEAT score. However, this is also useless for any downstream task. Our goal here is to evaluate the trade-off between accuracy and fairness on downstream classification tasks to analyze how our approach compares to previously suggested once in this aspect. The baseline is established using the embeddings from pre-trained BERT without any fine-tuning steps. The only trainable parameters are two fully connected layers with rectified linear units. This is a standard architecture for classification tasks based on state-of-the-art sentence embeddings. For the CoLA and SST2 dataset, the task is a balanced binary classification problem, and thus, we chose accuracy as the evaluation criterion of downstream performance.

Both debiasing approaches using PCA and kernel PCA are trained with defining sets, in the same way as in the evaluation of debiasing performance. For the training and the evaluation of the downstream task performance, each data sample is first encoded by BERT, then the debiasing step (PCA or kernel PCA) is applied. The resulting debiased embedding is fed into the two trainable fully connected layers. The resulting downstream tasks performances are reported in 4. Neither PCA nor kernel PCA reduces the performance strongly. However, the training procedure with debiasing by kernel PCA is a factor 10 slower than debiasing by PCA, as kernel PCA involves numerical optimization.

5 Conclusions

In this article, we proposed a feasible version of non-linear bias reduction algorithm using kernel PCA and evaluated its performance, both for the first time. Our empirical results suggest that non-linear debiasing methods do not perform as nicely as its linear counterpart for sentence embeddings. This result agrees with the conjecture by Vargas and Cotterell (2020), that bias in embeddings can be mostly captured by linear subspace. Moreover, our work provides novel insight about *sentence* embeddings that linear debiasing performs better, even for sentence embeddings by BERT, for which we have no reasons to assume their bias lies in linear subspace. In addition to the non-linear debiasing method, we introduced a novel method for benchmarking fairness of sentence embeddings, by reflecting the complexity of the sentence domain better to compute the SEAT. With our benchmarking method, we bring clarity to the ambiguity about the existence of statistically significant bias (May et al., 2019; Liang et al., 2020) and demonstrate that gender bias are indeed present in BERT sentence embeddings.

Our work has a few possible directions for future work. Firstly, we still do not know why the linear debiasing method performs better than the non-linear methods. Though the bias in sentence embeddings might actually lie in a linear subspace, the inefficiency of numerical optimization in the

Dataset	Baseline	PCA debiasing	kernel PCA debiasing
CoLA	0.72	0.74	0.72
SST-2	0.83	0.81	0.85

Table 4: Accuracy of the downstream task performance on CoLA and SST-2 dataset.

orthogonalization step of the non-linear debiasing method could also be a reason. Spotting the actual reason for this will be a potential future work. Secondly, it would be beneficial for the research community to establish the benchmark dataset for the bias evaluation in sentence embeddings. Such a dataset would make it easier to compare the result of different debiasing methods that will be proposed in the future.

References

Gökhan H Bakır, Jason Weston, and Bernhard Schölkopf. 2003. Learning to find pre-images. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 449–456. Citeseer.

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. In *Advances in neural information processing systems*, pages 4349–4357.

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*.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language?

Kirthevasan Kandasamy and Yaoliang Yu. 2016. Additive approximations in high dimensional nonparametric regression via the salsa. In *International conference on machine learning*, pages 69–78.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

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*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. *arXiv preprint arXiv:2009.09435*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

6 Appendix

Defining word sets for SEAT score and identification of gendered sentences.

Male	Female
he	she
himself	herself
boy	girl
man	woman
father	mother
guy	gal
male	female
his	her
himself	herself
john	mary

Table 5: Gendered Words

Career	Family
executive	home
management	parents
professional	children
corporation	family
salary	cousins
office	marriage
business	wedding
career	relatives

Table 6: Attribute career ↔ family

Math	Arts
math	poetry
algebra	art
geometry	dance
calculus	literature
equations	novel
computation	symphony
numbers	drama
addition	sculpture

Table 7: Attribute math ↔ arts

Science	Arts
science	poetry
technology	art
physics	dance
chemistry	literature
Einstein	Shakespeare
NASA	symphony
experiment	drama
astronomy	novel

Table 8: Attribute science ↔ arts