Investigating the Effects of Fairness Interventions Using Pointwise Representational Similarity

Anonymous authors
Paper under double-blind review

Abstract

Machine learning (ML) algorithms can often exhibit discriminatory behavior, negatively affecting certain populations across protected groups. To address this, numerous debiasing methods, and consequently evaluation measures, have been proposed. Current evaluation measures for debiasing methods suffer from two main limitations: (1) they primarily provide a *global* estimate of unfairness, failing to provide a more fine-grained analysis, and (2) they predominantly analyze the *model output* on a specific task, failing to generalize the findings to other tasks. In this work, we introduce Pointwise Normalized Kernel Alignment (PNKA), a pointwise representational similarity measure that addresses these limitations by measuring how debiasing measures affect the intermediate *representations* of *individuals*. On tabular data, the use of PNKA reveals previously unknown insights: while group fairness predominantly influences a small subset of the population, maintaining high representational similarity for the majority, individual fairness constraints uniformly impact representations across the entire population, altering nearly every data point. We show that by evaluating representations using PNKA, we can reliably predict the behavior of ML models trained on these representations. Moreover, applying PNKA to language embeddings shows that existing debiasing methods may not perform as intended, failing to remove biases from stereotypical words and sentences. Our findings suggest that current evaluation measures for debiasing methods are insufficient, highlighting the need for a deeper understanding of the effects of debiasing methods, and show how pointwise representational similarity metrics can help with fairness audits.

1 Introduction

Machine learning algorithms are now deeply integrated into several aspects of our daily lives. These algorithms not only recommend movies and products or suggest potential dating partners (76; 87; 84), but they are also increasingly employed in critical decision-making processes, such as approving loans and making hiring and health choices (2; 49; 74). Despite their impressive performance, ML models face significant reliability challenges, particularly in decision-making tasks (23; 30; 79; 77; 63). A major concern is their discriminatory behavior against certain protected groups (3; 61; 11), manifesting as biases that adversely affect individuals based on race, gender, age, or other protected characteristics. These biases can result in unfair outcomes that negatively affect individuals based on their characteristics. To mitigate such discriminatory behavior, researchers have proposed a variety of approaches that intervene at various stages of the ML pipeline, including data pre-processing (21; 65; 71; 81), learning fair intermediate representations (18; 92; 19; 53; 5; 82) [1], in-processing by adding constraints to the objective function (89; 88; 91; 17; 24; 62; 1; 54; 16; 44), and post-processing by changing model outputs (27; 83; 72).

Following the introduction of these debiasing methods, the research community has developed several evaluation measures to assess their effectiveness. These measures aim to quantify the fairness of ML models and include metrics such as equalized odds, equality of opportunity, and demographic parity (27; 18; 41).

[1] Prior work has included learning fair representation in both pre- and in- processing, *e.g.* see Zafar et al. (91); here we choose to list it separately since it is the main focus of our paper.

However, current evaluation approaches face two main limitations. First, they primarily focus on providing a *global estimate* of unfairness by analyzing the disparities between groups categorized by a single protected attribute, such as race or gender (27; 18; 41). However, as recent work reveals (11; 35; 36), fairness is multifaceted and can manifest in several different ways, making a more fine-grained analysis necessary. Second, fairness evaluation approaches predominantly analyze model behavior on a specific target task (27; 18; 41). In most cases they do not consider the *representations* that models use to reason about the data. Conceptually, most algorithms first compute an intermediate representation $z = g(x)$ of an input x, before deriving an outcome $y = f(z)$. Such intermediate representations appear, for example, as a result of feature engineering or at the intermediate layers of deep neural networks, and are becoming increasingly important with the rise of large, pretrained *foundation models* (8; 66; 9), where the representations of one model serve as the basis for many different downstream tasks. The insights from prior methods that only focus on task-specific outcomes y cannot easily be generalized to other tasks that use the same representations z.

In this work, rather than only evaluating task-specific outcomes on a global level, we address the limitations of previous evaluation measures by analyzing *how debiasing methods modify the intermediate representations of individual data points*. More specifically, we analyze how much the representation of a data point changes from "baseline" (non-debiased) to "fair" (debiased) models. To do so, we propose a new measure that is inspired by the existing rich body of research on representational similarity measures in the machine learning community (42; 48; 80; 68; 58; 37). Representation similarity measures, such as the widely-used CKA (37), analyze how two different models represent a given dataset. They capture similarity as a real number that reflects the degree of similarity between the representations, with 1 indicating identical representations. These measures thus allow us to analyze how model interventions impact all downstream tasks that use these representations.

An important limitation of previously introduced representation similarity measures is that they only provide an aggregate similarity score across the entire dataset. However, such aggregate scores are not suitable for achieving our goal of analyzing at a local level how debiasing methods are affecting *individuals*. To enable the fine-grained study of representation similarity, we modify the widely-used CKA measure (37) into a *pointwise representational similarity measure*, which we call Pointwise Normalized Kernel Alignment (PNKA). PNKA provides a similarity score for each individual data point, allowing us to study how model interventions affect individuals at the representation level.

Our key contributions are summarized as follows:

- We introduce PNKA, a pointwise representation similarity measure. PNKA allows us to analyze the effect of model interventions, such as debiasing, at a local level, by measuring how interventions change the representations of individuals. PNKA is broadly applicable to all debiasing approaches that modify the data or the models. Moreover, it offers a more generalized solution by eliminating the need for manually developing domain-specific evaluation methods for each individual use case.

- We demonstrate PNKA's utility in auditing representations of debiasing approaches. On the COMPAS and Adult datasets, PNKA reveals previously unknown insights: while group fairness predominantly influences a small subset of the population, maintaining high representational similarity for the majority, individual fairness constraints uniformly impact representations across the entire population, altering nearly every data point. Our observations on representation similarity allow us to predict the effect that training ML models on debiased representations will have, and we demonstrate that the actual outcomes do indeed match the predictions.

- By applying PNKA to contextual and non-contextual language embeddings, we show that debiasing methods in these domains may not perform as intended. Specifically, PNKA shows that debiasing approaches do not consistently remove gender properties from stereotypical words and sentences as anticipated. Our results also suggest that the fairness evaluation measures currently employed for evaluating (non-)contextual debiased embeddings are both limited and insufficient for comprehensively assessing these debiasing methods.

2 Pointwise Representational Similarity Measure for Assessing Debiasing Approaches

When an algorithm processes data x for an individual, it typically operates in two stages: 1) computing an *(intermediate) representation* $z = g(x)$ and 2) computing the *outcome* $y = f(z)$. To ensure that algorithms produce fair outcomes, a popular approach is to replace a potentially biased feature extractor g with an unbiased version g', such that the modified representation $z' = g'(x)$ does not contain information about a sensitive feature (92; 43). We use the notation z here to refer to the representation of a single individual x, where z is a d-dimensional vector, *i.e.* $z \in \mathbb{R}^d$. We use uppercase Z to refer to the representations of a set of N individuals, *i.e.*, $Z \in \mathbb{R}^{N \times d}$.

When debiasing feature extractor g, it is important to be able to assess the effects of the change on performance and fairness. Most prior work focuses on studying how changes to g affect the outcomes y (27; 18; 41), but the insights from such assessments only apply to specific downstream tasks. To gain a better understanding of how changes to the representations Z will impact different downstream tasks, we have to study Z directly. Such an assessment can be performed using a *representation similarity measure* $s(Z, Z')$ that measures how similar Z' is to Z. Representational similarity measures typically provide a single overall score that estimates how similar two different representations of an entire set of input points are. However, overall representation similarity scores do not allow us to assess how much the representation for *individual points* change, and can thus overlook potential adverse effects that changes to the representation can have on small groups or individuals.

To address these issues, we adopt a *pointwise* measure that assigns an individual representational similarity score to each data point. With this measure, we can effectively determine whether data or model interventions, such as debiasing, impact all instances uniformly (*i.e.*, whether all individuals are affected by the debiasing method) or disproportionately impact certain data points (*i.e.*, some individuals are more affected than others). By comparing representations of a baseline (non-debiased) model g with its debiased version g', we can identify and characterize which individuals are most affected and whether fairness interventions effectively target those they are aimed at. For instance, by focussing on the individuals with the lowest similarity scores, we can determine whose representation change the most from the baseline to the debiased version. Next, we describe our pointwise representation similarity measure.

2.1 Intuition for Pointwise Similarity Across Representations

An initially appealing way to measure representational similarity of the i-th point in representations Z and Z' is to directly apply a (dis)similarity metric, such as the Euclidean distance or cosine similarity, to its two representations Z_i and Z'_i, e.g., by defining $\mathrm{s}(Z, Z', i) = \cos(Z_i, Z'_i)$. One immediate failure mode of such an approach is when Z_i and Z'_i have a different number of dimensions $d \neq d'$. However, even when the number of dimensions matches, any such approach that directly compares the two representations suffers from a subtle but important shortcoming of not being invariant to orthogonal transformations. Consider an example where $Z = R\,Z'$ with R being an orthogonal matrix, such that $Z_i^\top Z'_i = 0\ \forall i$. Even though $cos(Z_i, Z'_i) = 0\ \forall i$, i.e., representations appear very dissimilar when directly measuring their cosine similarity, they are, however, from an information-theoretic standpoint, identical for any downstream applications[2]. Therefore, the low similarity score obtained from directly comparing points is misleading. In fact, previous work (46; 85; 50) has shown that orthogonal transformations do not change the training dynamics of neural networks and thus invariance to them is a desirable property for any similarity metric operating on neural representations, as also discussed in Kornblith et al. (37). A similar argument could be made against other choices of direct comparisons such as Euclidean distance.

To overcome the issues involved in directly comparing two different representations, we propose an indirect comparison. We leverage the simple, but powerful insight from prior work (37; 38) that while we cannot directly compare similarity *across* representations, we can do so *within* the same representation. We argue that the *representations Z_i and Z'_i of a point i should be considered similar across representations Z and Z'*

[2]The projection matrix R is invertible, so by multiplying the weight matrix of any linear downstream operation on Z with R^{-1}, it can be directly applied to Z' and produce the same result.

if their positions relative to other points in the respective representation are similar. Therefore, to determine whether the representations Z_i and Z'_i of point i are similar, we can first compare how similarly i is positioned relative to all the other points within each representation. We then compare the relative position of i across both representations.

2.2 Measuring Similarity in the Relative Position of Points

We can now formally describe our proposed measure, Pointwise Normalized Kernel Alignment (PNKA), which calculates representational similarity for individual points by first comparing the relative similarity between a point and other points within the same representation, and then across the two different representations. Given a set of (column-centered) representations Z (and analogously for Z') and a kernel $k(.,.)$, we can define a pairwise similarity matrix between all N points in Z as $K(Z)$ with $K(Z)_{i,j} = k(Z_i, Z_j)$. In our work, we use linear kernels, i.e., $k(Z_i, Z_j) = Z_i^\top \cdot Z_j$, but other kernels, e.g., RBF (37) kernels, could be used as well. We leave the exploration of other types of kernels for future work. Given two similarity matrices $K(Z)$ and $K(Z')$, we measure how similarly point i is represented in Z and Z' by comparing its position relative to all other points. To this end, we define

$$\text{PNKA}(Z, Z', i) = \cos(K(Z)_i, K(Z')_i) = \frac{K(Z)_i^\top K(Z')_i}{||K(Z)_i|| \; ||K(Z')_i||}, \tag{1}$$

where $K(Z)_i$ and $K(Z')_i$ denote how similar point i is to all other points in Z and Z', respectively. We use cosine similarity to compare the within-representation similarity of a point with the other points, across representations, for two reasons. First, cosine similarity provides us with normalized similarity scores for each point. Second, by normalizing by the length of the similarity vectors $K(Z)_i$ and $K(Z')_j$, we compare the *relative* instead of the absolute similarity of points, *i.e.*, how similar point i is represented relative to points j and j'. We can further extend our measure into an aggregate version ($\overline{\text{PNKA}}$) between sets of representations by computing the average of the similarities across all the N points.

2.3 Properties of PNKA

PNKA captures the similarity between neighborhoods. Previously we argued that for a data point to be similarly represented across two representations, it should be similarly positioned relative to the other points in each representation. In Appendix A.1 we empirically show, across several architectures, datasets, distances measures, and values of k, a clear relationship between high PNKA scores and high overlap of nearest neighbors across representations, indicating that PNKA captures how similar the neighborhoods of the points are. We further show in the same Appendix that the direct comparison of representations through cosine similarity does not exhibit the same trend.

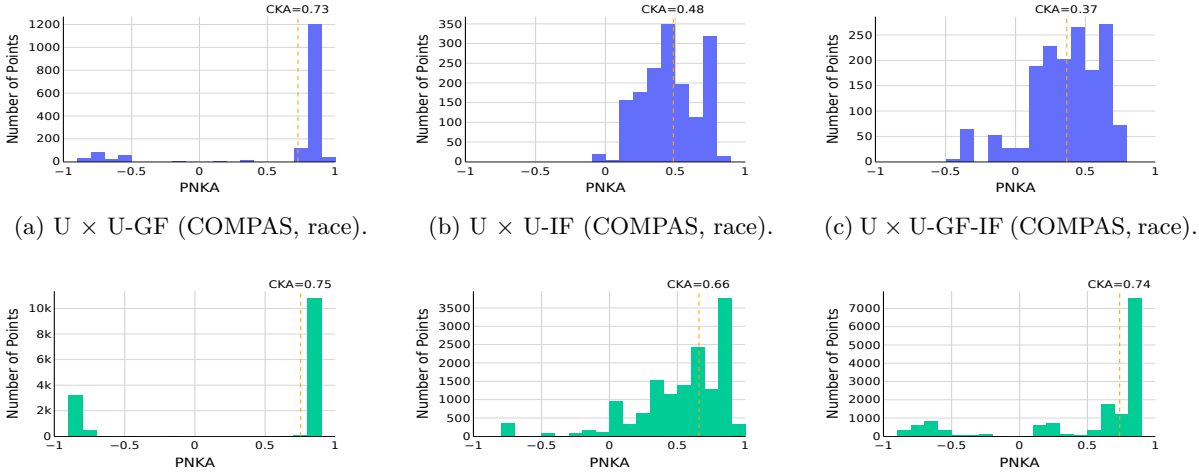
Relationship of PNKA with aggregate measures of representational similarity. PNKA is inspired by CKA, and in Appendix A.2 we empirically show that the aggregate version of our measure ($\overline{\text{PNKA}}$) produces results close to CKA.

Invariance properties. Previous work (37) has identified invariance to orthogonal transformations and isotropic scaling as desirable properties for representational similarity measures. We provide a mathematical proof that our measure possesses both of these invariance properties in Appendix A.3.

3 Investigating Debiased Tabular Data Representations

Prior work (92; 14; 51; 86) has proposed learning fair data representations which retain minimal information about sensitive features. Here we conduct a case study to show how PNKA can be used to audit the effect of such debiasing approaches, by comparing the original (baseline) and the debiased representations.

One example of these debiasing techniques is the approach proposed by Zemel et al. (92), which learns representations of tabular data by optimizing a loss function that maintains as much utility as possible, while removing information about protected attributes. The learning algorithm modifies the data representation based on three distinct objectives: classification accuracy (denoted by us as utility (U)), statistical parity (to achieve group fairness (GF)), and data loss (as a proxy for achieving individual fairness (IF)). By applying



(a) U × U-GF (COMPAS, race). (b) U × U-IF (COMPAS, race). (c) U × U-GF-IF (COMPAS, race).

(d) U × U-GF (Adult, gender). (e) U × U-IF (Adult, gender). (f) U × U-GF-IF (Adult, gender).

Figure 1: Distribution of PNKA similarity scores. The first row (blue plots) shows results for the COMPAS dataset, debiased with respect to race, while the second row (green plots) displays results for the Adult dataset, debiased based on gender. The vertical dotted line shows the overall similarity scores provided by CKA (37). We compare the representations obtained from the baseline (U, only utility) with each of the three other debiased options: (1) utility, group and individual fairness (U-GF-IF), utility and group fairness (U-GF), and utility and individual fairness (U-IF).

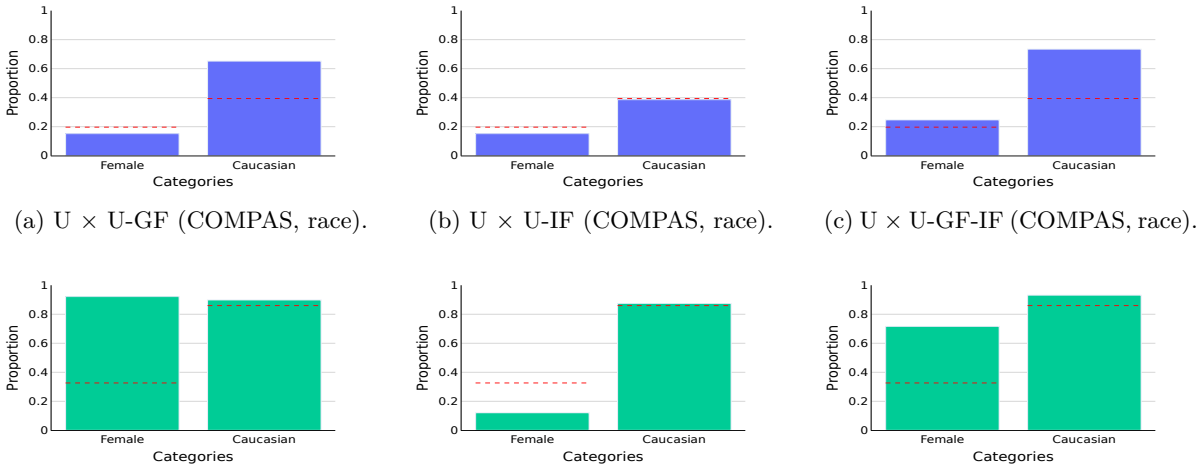
combinations of these objectives, we obtain four types of representations, one based on utility (U), serving as the baseline, and three debiased versions: utility and group fairness (U-GF), utility and individual fairness (U-IF), and a combination of all three objectives (U-GF-IF).

We use the COMPAS (45) dataset, debiased for race, and the Adult (4) dataset, debiased for gender [3], to analyze with PNKA the effect of the different debiasing objectives. We first investigate the overall effect of debiasing in Section 3.1, followed by a more detailed analysis of the individuals whose representations change the most in Section 3.2. Finally, we show that the predictions made based on PNKA scores match the predictions observed on downstream tasks in Section 3.3.

3.1 How Do Individual Representations Change?

The distribution of PNKA similarity scores, obtained by comparing the representations of the baseline with each of the debiasing methods, is visualized in Figure 1, with COMPAS (45) results depicted in the first row (Figures 1a– 1c), and Adult (4) shown in the second row (Figures 1d– 1f). As shown in Figures 1a and 1d, for both datasets, under group fairness, the majority of individuals maintain high representational similarity, suggesting that the majority of individuals' representations remain similar to the baseline, with significant changes primarily occurring for a small subset of the population. In contrast, as depicted in Figures 1b and 1e, individual fairness constraints lead to more uniform representational change across the entire population, slightly impacting nearly every data point's representation. Finally, when combining group fairness and individual fairness, as observed in Figures 1c and 1f, the impact on the data points' representations varies depending on the dataset. For the COMPAS dataset, the resulting pattern closely resembles that observed under the influence of only group fairness. In contrast, the pattern observed for the Adult dataset aligns more closely with the effects seen when applying only individual fairness.

[3]We pre-process each dataset as done by Zemel et al. (92). More information can be obtained in `https://github.com/Trusted-AI/AIF360`.



(a) U × U-GF (COMPAS, race). (b) U × U-IF (COMPAS, race). (c) U × U-GF-IF (COMPAS, race).

(d) U × U-GF (Adult, gender). (e) U × U-IF (Adult, gender). (f) U × U-GF-IF (Adult, gender).

Figure 2: Distribution of the binary attributes for the 10% most affected individuals (i.e., lowest PNKA score). The first row (blue plots) shows results for the COMPAS dataset, debiased with respect to race, while the second row (green plots) displays results for the Adult dataset, debiased based on gender. We compare the data representations of the baseline (U, only utility) with each of the three other debiased options: (1) utility, group and individual fairness (U-GF-IF), utility and group fairness (U-GF), and utility and individual fairness (U-IF).

3.2 Whose Representations Change the Most?

Next, we analyze the 10% of the population with the lowest PNKA similarity scores, representing those whose representations have undergone the most significant changes due to the applied fairness constraints. We focus on the distribution of the protected attributes in this group, specifically on race and gender. Figure 2 illustrates the distribution of the binary sensitive attributes for the subset of the most affected people in both COMPAS, in the first row, and Adult, in the second row. The horizontal red dotted line shows the population average per attribute [4].

In the COMPAS dataset (Figures 2a– 2c), where data representations are debiased according to race, we observe that the gender distribution remains relatively stable, aligning closely with the population baseline distribution, while the race distribution exhibits notable shifts. More specifically, for COMPAS, under the group fairness objective, the individuals whose representations are altered the most are primarily Caucasians. This happens regardless of whether the group fairness constraint has been used alone or in combination with individual fairness. However, when only individual fairness is applied, the distribution resembles that of the baseline one, suggesting a minimal impact of individual fairness on attribute distribution compared to group fairness.

In contrast, as shown in Figures 2d–2f, for the Adult dataset, whose representations are debiased with respect to gender, under group fairness conditions, the female category is the most drastically affected group. This happens regardless of whether individual fairness is applied in combination with group fairness. Interestingly, however, when individual fairness constraints are applied, the male representations are more affected than female ones. The race attribute largely retains its baseline distribution.

3.3 Do PNKA's Predictions Match Downstream Outcomes?

The previous results suggest that under some debiasing objectives, the data representations of specific demographics are disproportionally affected. However, do the changes in the data representations also translate

[4]We show the distribution of the remaining attributes in Appendix B.1.

Dataset	Constraint Type	Accuracy	Statistical Parity	Consistency
COMPAS	Utility	0.6780	0.2308	0.9811
	Utility + Group Fairness	0.6585	0.0376	0.9220
Adult	Utility	0.8015	-0.1784	0.9385
	Utility + Group Fairness	0.7878	-0.0426	0.9968

Table 1: Overall accuracy, statistical parity, and consistency results for linear regression models trained with baseline (Utility) data representation and with the group fairness debiased representations. For statistical parity, in the context of the COMPAS dataset, where the model predicts the risk of criminal recidivism, a positive score indicates a higher predicted risk for non-Caucasians. On the other hand, in the Adult dataset, where the model is trained to predict whether income exceeds 50K per year, a negative score means fewer females are likely to receive a high annual income. Ideally, statistical parity should be 0, while a score of 1 is optimal for both accuracy and consistency.

to changes in the behavior of models that use them? In other words, can the analysis at the representation level provide insights at the output level?

To test the utility of PNKA in predicting downstream behavior, we use the insights obtained from PNKA in Sections 3.1 and 3.2, together with prior knowledge about the design of the fairness constraints, to predict what outcomes we can expect from models that use the modified data representations. We then train models on the representations to test whether the predictions made using PNKA match the actual outcomes. In our analysis, we focus on the effect of the GF-constraints, because (i) we observe in Section 3.1 that they strongly affect a small part of the population that we can easily characterize in Section 3.2 (as opposed to the more uniform changes across the board for the other constraints, which make it harder to characterize the demographics of the affected individuals), and (ii) it is easy to understand their effect, namely minimizing disparities between groups that differ in the protected attribute.

Hypothesis: For the COMPAS dataset, we observe in Figure 2a that the people whose representations are changed the most by the group fairness constraints are predominantly Caucasians. Given that the COMPAS dataset is known to exhibit a bias in favor of Caucasians, *i.e.* that they are less likely to recidivate than non-Caucasians, a change in the representations of Caucasians therefore suggests that a model using the group fairness debiased representations might assign Caucasians less favorable outcomes than one that uses the baseline representations. Conversely, for the Adult dataset we observe in Figure 2d that females are the predominant demographic in the 10% of most affected representations. Given that in this dataset males tend to be more likely to have the positive label (income above 50K per year), we thus expect that a model using group fairness debiased representations will assign more favorable outcomes to females compared to a model using baseline representations.

To test whether these hypotheses, based on PNKA's similarity scores, are correct, we train logistic regression models for each of the representation types. On the COMPAS dataset the goal is to predict recidivism, whereas on the Adult datasets the goal is to predict whether an individual's income is above 50K dollars per year. The overall accuracy and debiasing results of these models are presented in Table 1. Following the approach used by Zemel et al. (92), we measure group fairness through the statistical parity difference, which is the ratio of favorable outcomes received by unprivileged versus privileged classes. In the context of the COMPAS dataset, where the model predicts the risk of criminal recidivism, a positive (negative) score indicates a higher predicted risk for non-Caucasians (Caucasians). For the Adult dataset, where the model is trained to predict whether income exceeds 50K per year, a positive (negative) score means females (males) are less likely to receive a high annual income. Individual fairness (92) is assessed using the consistency score, which evaluates how similar the predicted labels are for neighboring instances. Table 1 shows that in both the COMPAS and Adult datasets, the inclusion of group fairness leads to models exhibiting a lower statistical parity difference compared to the baseline. This suggests a reduction in bias against non-Caucasians for COMPAS, and females for the Adult dataset.

To better understand whether the reduced statistical parity on both datasets is indeed due to the hypothesized changes for each of the groups, we investigate how the benefits that each of the groups receives on the

Measure	Dataset	Protected Attribute	Data	
			Utility	Utility + Group Fairness
Precision	COMPAS	Caucasian	0.6797	0.5465
		Non-Caucasian	0.6980	0.7090
Recall	Adult	Male	0.4340	0.3731
		Female	0.0752	0.4146

Table 2: Precision and recall scores for COMPAS and Adult datasets, respectively, of linear regression models trained with baseline (U) data representation and the group fairness debiased data representations. In the COMPAS dataset, where there is a known bias against non-Caucasians, we found that precision for Caucasians is reduced due to group fairness interventions, indicating an adverse effect. In the Adult dataset, characterized by a bias against females, females are positively affected by group fairness.

two datasets changes. In the case of COMPAS, false positives are deemed to be far more costly – the widely accepted Blackstone's ratio posits that "it is better that ten guilty persons escape than that one innocent suffer" (6). On the other hand, in the case of Adult, false negatives are deemed to be more costly, as it suggests that the individual has a lower income, which can be associated with limited financial resources, reduced access to opportunities, and potentially lower socio-economic status. Thus, for COMPAS, we measure precision and for the Adult dataset, recall, since these measures capture the change in beneficial outcomes.

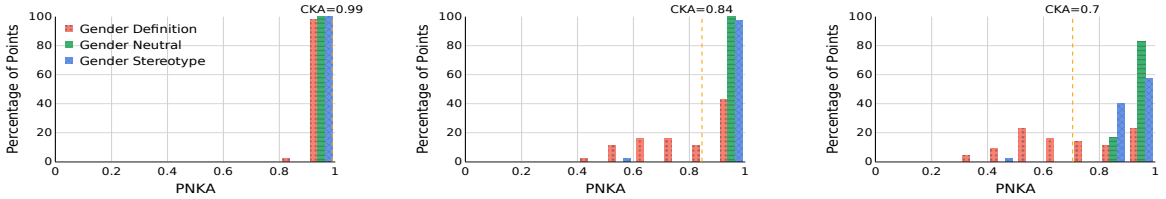
The results for each protected group are shown in Table 2. The results corroborate the earlier predictions made by PNKA. For the COMPAS dataset, the use of group fairness significantly impacts Caucasians, leading to a notable decrease in their precision score. In particular, the precision score changes more strongly for Caucasians ($\sim$ 13%) than for non-Caucasians ($\sim$ 1%). On the Adult dataset, the use of group fairness results in a substantial increase of recall scores for females. Again, females experience a much larger change in recall ($\sim$ 34%) than males ($\sim$ 6%). Both of these findings match the predictions made using PNKA, *i.e.* the groups achieving the largest change in their outcomes are those that are most prevalent among the individuals whose representations changed the most due to the group fairness constraints.

Takeaways. The results in this section show that PNKA can be a valuable auditing tool, as it enables a comprehensive and detailed analysis of the changes that occur when modifying representations. Further, we demonstrate that the insights provided by PNKA are reliable, exhibiting predictive power for downstream applications and aligning with findings from prior outcome-based studies. Importantly, unlike prior work, which requires studying each downstream application of the representations separately and depends on prior knowledge to identify relevant groups, PNKA directly quantifies the extent of change in representations at the individual level, offering a more targeted and efficient approach.

4 Investigating Debiased Language Embeddings

In our next case study, we analyze PNKA as a tool for investigating debiased word embeddings. Previous work (7; 25; 52) has identified the presence of stereotypical biases in word embeddings, *i.e.* vector representations of words, and since has developed several debiasing methods (7; 94; 33). Among them are: (1) Gender Neutral (GN-)GloVe (94), which focuses on disentangling and isolating all the gender information into certain specific dimension(s) of the word vectors; and (2) Gender Preserving (GP-)GloVe (33), which targets preserving non-discriminative gender-related information while removing stereotypical discriminative gender biases from pre-trained word embeddings. The latter method is also used to finetune GN-GloVe embeddings, creating a third, combined method called GP-GN-GloVe.

We start, in Section 4.1, by briefly explaining how these debiasing approaches are traditionally evaluated. In Section 4.2, we show how PNKA can be leveraged to investigate whether these approaches modify the group of words (*i.e.*, stereotypical words) as originally intended. Using the insights from PNKA, we formulate hypothesis about the effects of the debiasing approaches, and test them in Section 4.3.



(a) GloVe × GP-GloVe. (b) GloVe × GN-GloVe. (c) GloVe × GP-GN-GloVe.

Figure 3: Distribution of PNKA scores per group of words for SemBias dataset (94). We compare the baseline (GloVe) model and its debiased versions. Words with the lowest similarity scores are the ones that change the most from the baseline to its debiased version. Across all debiased embeddings, the words whose embeddings change the most are the gender-definition words.

4.1 Traditional Measures for Debiased Word Embeddings

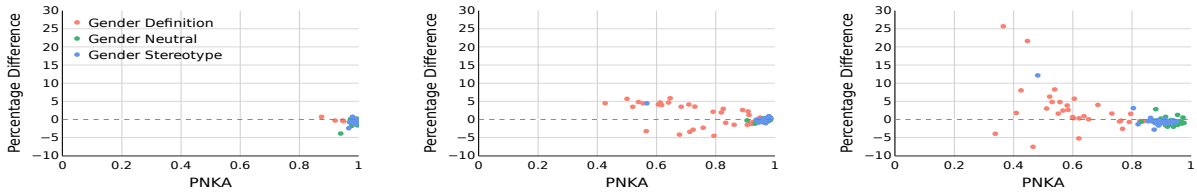
The aforementioned debiasing models for word embeddings have been originally evaluated using SemBias (94), a dataset designed to assess whether the debiasing methods have successfully removed stereotypical gender information from the word embeddings. Each instance in SemBias consists of four word pairs: a *gender-definition* word pair (e.g. "waiter - waitress"), a *gender-stereotype* word pair (e.g. "doctor - nurse"), and two other word-pairs that have similar meanings but no gender relation, named *gender-neutral* (e.g. "dog - cat"). To assess the bias in word embeddings, the evaluation scheme measures the cosine similarity with the canonical gender vector, i.e., $cos(\overrightarrow{a} - \overrightarrow{b}, \overrightarrow{he}$ - $\overrightarrow{she})$ for each of the four word pairs (a, b) in a SemBias instance. The word pair with the highest cosine similarity is selected as the "predicted" answer. If the word embeddings are correctly debiased, then the cosine similarity of the $\overrightarrow{he}$ - $\overrightarrow{she}$ vector with the gender-definition words should be high, and the similarity with the gender-stereotype words should be low, i.e., the frequency of predictions for these categories should be high for the gender-defining word pairs and low for gender stereotypical word pairs.

Thus, GP- and GN-GloVe evaluate how (de)biased embeddings are based on whether they predict stereotypical or definitional word pairs in each instance of the SemBias dataset. We show results for evaluating GP- and GN-GloVe on SemBias in Appendix C.1. The evaluation shows that GP-Golve embeddings offer only a marginal improvement over the baseline embeddings, while GN-GloVe and GP-GN-GloVe embeddings show substantial reductions in bias in the prediction task. These findings suggest that the latter models are more effective in mitigating gender bias in word embeddings.

4.2 How Do Individual Representations Change?

We next employ PNKA to better understand which word embeddings have changed the most due to the debiasing procedure of GP- and GN-GloVe methods. As before, we use PNKA to measure similarity between the original GloVe (baseline) and the debiased versions of GloVe embeddings, i.e., (GN-)GloVe (94), Gender Preserving (GP-)GloVe and GP-GN-GloVe (33). Figure 3 shows the distribution of PNKA similarity scores for words in the SemBias dataset grouped by their category (i.e., gender defining, gender neutral, and gender stereotype).

We first observe, in Figure 3a, that GP-GloVe representations exhibit a high degree of similarity to the original GloVe embeddings for almost all words. This suggests that the GP-GloVe method may not significantly alter the word representations, maintaining a close resemblance to the original embeddings for most words. However, Figures 3b and 3c show that GN-GloVe and GP-GN-GloVe, respectively, considerably change the representations for a subset of the words. This observation aligns well with the results of the prior evaluation, detailed in Table 4 of Appendix C.1, in which GP-GloVe was shown to yield results similar to GloVe (suggesting similar representations), while GN-GloVe and GP-GN-GloVe achieve better debiasing results.



(a) GloVe × GP-GloVe. (b) GloVe × GN-GloVe. (c) GloVe × GP-GN-GloVe.

Figure 4: Relationship between PNKA scores (x-axis) and percentage difference (y-axis) in magnitude of the projection on the gender direction $\overrightarrow{he}$ - $\overrightarrow{she}$. A positive or negative percentage difference value indicates a shift in magnitude along the gender direction. Word embeddings that change their gender information are the ones that obtain low PNKA scores.

Moreover, as illustrated in Figure 3, an intriguing pattern emerges across all three debiasing methods: the words with lower PNKA scores, i.e. the words whose representations change the most, belong predominantly to the gender-definition category. This finding stands in stark contrast to the expected behavior: As discussed in Zhao et al. (94) and Kaneko & Bollegala (33), the primary goal of these debiasing techniques for word embeddings is to retain gender-specific information in feminine and masculine words (i.e., gender-defining words), maintain neutrality in gender-neutral words, and eliminate biases in stereotypical words. Thus, the expectation is that debiasing should primarily alter gender-stereotypical word embeddings, while mostly preserving gender-definitional ones.

4.3 Do PNKA's Predictions Match the Projection Analysis?

Our analysis using PNKA reveals that, contrary to the expected effect of debiasing, the most profound changes occur in the gender-defining words, challenging the conventional understanding of how these debiasing methods function and prompting a more careful analysis of their effects. We formulate a new hypothesis about the impact of debiasing on word embeddings: *instead of removing the gender information in gender-stereotypical words as initially intended, debiasing methods inadvertently amplify gender information in the gender-definition words.* Such an effect could be missed by the conventional evaluation procedure discussed previously, which only assesses the *relative* cosine-similarity. Increasing the gender-related information in gender-defining words would make gender-defining words more gender-aligned, and thus increase their prediction frequency, relative to other word pairs in SemBias. This effect could be achieved, however, without removing gender-related information from the gender stereotypical words, as intended by the debiasing methods.

To test this hypothesis, we measure for each word how much its embedding changed in terms of gender information, when compared to the original GloVe embedding, by projecting it onto the canonical gender vector $\overrightarrow{he}$ - $\overrightarrow{she}$. More specifically, for each embedding approach e and word i, we project the corresponding word embeddings $\phi_i^{(e)} = e(i)$ onto the gender vector direction $g^{(e)} = \overrightarrow{\phi_{he}^{(e)}}$ - $\overrightarrow{\phi_{she}^{(e)}}$ and compute the projection magnitudes $p_i^{(e)} = \phi_i^{(e)} \cdot \hat{g}^{(e)\top}$, where $\cdot$ represents a dot product, and $\hat{g}$ is the normalized gender direction. The higher $p_i^{(e)}$ is, the more gender information is contained in the word embedding vector $\phi_i^{(e)}$. To understand how much each of the debiased embedding methods change the amount of gender information, relative to the original *glove* embeddings, we analyze the percentage difference in magnitude, defined as $\omega_i^{(e)} = \frac{p_i^{(e)} - p_i^{(glove)}}{|p_i^{(glove)}|}$. A $\omega_i^{(e)} = 0$ indicates that the gender information in the debiased embedding has not changed relative to (baseline) GloVe, while $\omega_i^{(e)} > 0$ ($\omega_i^{(e)} < 0$) indicates an increase (decrease) in the male (female) gender information associated with word i.

Figure 4 depicts the relation between PNKA scores (x-axis) and the percentage difference in magnitude (y-axis) for each word in the SemBias dataset. We can see that the GN-GloVe and GP-GN-GloVe debiasing methods primarily amplify the gender information in gender-definition words (red dots), rather than reduce it

for gender-stereotype words (blue dots), *i.e.*, $\omega_i^{(e)} \neq 0$ for gender-defining and $\omega_i^{(e)} \approx 0$ for gender-stereotype words. The words exhibiting the most significant change in gender information, identified as the ones with lower PNKA similarity scores, predominantly fall within the gender-defining category. This observation supports our alternative explanation that these debiasing methods might be enhancing gender information in gender-defining words, rather than diminishing it in gender-stereotypical words.

Finally, we applied a similar analysis using PNKA to investigate bias mitigation efforts in contextualized embeddings of transformer-based language models. Unlike static word embeddings, these models generate word representations that depend on surrounding context, which makes identifying and mitigating biases challenging yet crucial, given that biases in the pre-trained models can propagate to numerous downstream tasks. We focused on a debiasing method that aims to remove gender bias from contextual representations of stereotypical words (e.g., associating "math" with a male bias or "poetry" with a female bias) while retaining gender-specific information in gender-defining words. Our analysis using PNKAreveals that, similar to findings with static embeddings, the debiased contextualized embeddings showed only minimal shifts along the gender direction, failing to reduce bias in the intended stereotypical contexts. Detailed results from this investigation are available in Appendix D.

Takeaway. The findings in this section shows that PNKA is a powerful tool for auditing word embedding debiasing techniques, as it allows for a detailed, representation-level analysis of how gender information is altered by these methods. Through PNKA, we gain reliable insights that not only complement traditional outcome-based evaluations but also reveal underlying representational shifts that might otherwise go unnoticed. Crucially, unlike prior approaches that rely on predictive task assessments or predefined gender categories, PNKA quantifies representational changes at the individual word level, enabling a more precise and efficient evaluation of bias mitigation efforts. This ability to identify specific shifts in gender-related representations positions PNKA as an essential tool in refining and improving debiasing strategies.

5 Related Work

In this section, we first review prior work on developing and assessing debiasing methods in the fairness domain. We then review work on representational similarity measures.

5.1 Fairness in ML Systems

ML algorithms have become fundamental to several aspects of daily life, yet these models often exhibit discriminatory behavior that negatively impacts protected groups. To address these biases, several debiasing methods have been developed, intervening at different stages of the ML pipeline (15; 57). Pre-processing debiasing methods, including reweighting, resampling, data augmentation, or even learning fair data representations, modify the training data to remove the underlying discrimination before it is used by the model (10; 32). In-processing methods modify the learning algorithms, by, for example, adding constraints to the optimization objective, in order to remove discrimination during the model training process (90; 93). Finally, post-processing methods change the decision thresholds or apply recalibration methods to adjust model outputs to achieve fairness after the model has been trained (27; 7). PNKA can be used to analyze any debiasing method that operates at the pre- or in-processing stages of the ML pipeline.

With the increased adoption of debiasing methods, auditing these approaches has received considerable attention. Most fairness measures fall under two main categories (64): (1) group fairness, which requires parity of some statistical measure across protected groups (18; 27). (2) individual fairness, which requires that similar individuals be treated similarly (31; 29). Both categories are output-based, *i.e.*, they solely look at the predictions/decisions made by the model.

Some work (25; 12; 7; 20; 22) has analyzed representations, mostly in the realm of word embeddings. For instance, Bolukbasi et al. (7) evaluates gender bias in word embeddings by identifying a gender subspace using explicitly gendered word pairs and analyze the proximity of gender-neutral words to gender-specific terms, uncovering societal stereotypes. Caliskan et al. (12) uses the WEAT test to systematically measure biases in word embeddings by comparing the association strength between pairs of target words (*e.g.*,"male" and "female" names) and attribute words (*e.g.*, "career" and "family").

The closest work to ours is the one by Gonen & Goldberg (25), which investigates word embeddings and shows that debiasing methods for word embeddings mostly hide the bias, instead of removing it. More specifically, using a clustering algorithm, they show words that receive implicit gender from social stereotypes (e.g., receptionist, captain) still tend to group with other implicit-gender words of the same gender, similar as for non-debiased word embeddings. Using PNKA, we reach a similar conclusion: these debiasing models primarily mask, rather than mitigate, the bias. However, as we showed in Section 4, our analysis offers a fresh perspective and instead reveals that these methods are augmenting the gender-related information of the gender-defining words, instead of removing this information from the gender-stereotypical words. None of the previous work provide a general measure that can be applied to representations in different contexts. To the best of our knowledge, our measure is the first one that can be broadly applied to any debiasing method that alters the representations used by models.

5.2 Representational Similarity Measures

Several measures have been proposed and used as tools to better understand the internal representations of machine learning (ML) models. Recently, approaches that compare the representational spaces of two models by measuring representational similarity have gained popularity (42; 48; 80; 68; 58; 37). At their core, representational similarity measures (RSMs) quantify how a set of points are positioned relative to each other within the representation spaces of two different models. Among the RSMs proposed in the literature, CKA (37) has gained popularity and has now been extensively used to study representations (60; 70; 67; 69). CKA is based on the idea of first choosing a kernel and then measuring similarity as the alignment between these two kernel matrices. We take inspiration from this insight to propose PNKA.

Most widely used RSMs, however, yield only an aggregate estimate (i.e., a single score) of similarity across an entire set of points. Being limited to aggregate measurements makes these measures unsuitable to study nuances of representational similarity at a more granular, local level. Therefore, approaches of representational similarity for individual points have been proposed, such as for words (26) and nodes in graphs (13). However, the applicability of these measures is constrained due to their task-specific nature. Work by Shah et al. (73) proposes a method to estimate the contribution of individual points to learning algorithms, but mainly focuses on understanding what *features* of the inputs are encoded in the representations, and do not evaluate the *similarity* of representations directly. Finally, a pointwise RSM proposed by Moschella et al. (59) resembles our proposed measure PNKA. However, the goal of Moschella et al. (59) differs significantly from ours, as they use their measure to enable model stitching, whereas we employ PNKA to audit representations. To the best of our knowledge, this is the first time representational similarity has been employed to study the effects of debiasing methods at a fine-grained level.

6 Discussion

We introduce a framework that leverages representational similarity to assess how debiasing methods affect individual data representations. Crucial to this framework is PNKA, a pointwise representational similarity measure that quantifies changes in an individual's representation due to fairness interventions.

For datasets like COMPAS and Adult, PNKA corroborates established findings and reveals new insights, showing notably different effects on individuals when learning group-level versus individual-level fair representations. PNKA also enables anticipation of biases and fairness constraints' impacts before model training and deployment. Applying PNKAto (non-)contextual embeddings, we find that debiasing methods do not consistently alter targeted groups, like gender-stereotypical terms, as intended. Our findings also indicate that current fairness evaluation measures are limited, unable to fully assess debiasing methods – a gap that PNKA addresses.

The impact of this work has broader implications for the community of fair ML. Our results show that pointwise representational similarity can be a valuable tool to understand the effects of debiasing methods on individuals in a nuanced way, and we propose PNKA as an effective way to measure it. By employing PNKA, we can perform more comprehensive audits of fairness interventions, ensuring they achieve the desired outcomes not just at the group level but also at the individual level.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, volume 80, pp. 60–69, 2018. URL `http://proceedings.mlr.press/v80/agarwal18a.html`.

[2] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93, 2016.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2016.

[4] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[6] William Blackstone. *Commentaries on the Laws of England*. Oxford University Press, 2016.

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

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pp. 803–811. PMLR, 2019.

[11] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

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

[13] Zuohui Chen, Yao Lu, Jinxuan Hu, Wen Yang, Qi Xuan, Zhen Wang, and Xiaoniu Yang. Graph-based similarity of neural network representations. *arXiv preprint arXiv:2111.11165*, 2021.

[14] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

[15] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[16] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

[17] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *NIPS*, pp. 2796–2806, 2018.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, pp. 214–226, 2012. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

[19] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *ICLR*, 2016. URL http://arxiv.org/abs/1511.05897.

[20] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1166. URL https://aclanthology.org/P19-1166.

[21] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pp. 259–268, 2015.

[22] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[23] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

[24] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. *AAAI*, 32(1), 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11662.

[25] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pp. 609—-614, 2019.

[26] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, pp. 2116. NIH Public Access, 2016.

[27] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, volume 29, pp. 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *IJCAI*, pp. 2248–2254, 2018.

[30] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[31] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

[32] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

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

[34] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*, 2021.

[35] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kearns18a.html`.

[36] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pp. 5264–5274. PMLR, 2020.

[37] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

[38] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, pp. 4, 2008.

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[40] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[41] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[42] Aarre Laakso and Garrison Cottrell. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000.

[43] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pp. 1334–1345. IEEE, 2019.

[44] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

[45] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Data and analysis for 'How we analyzed the COMPAS recidivism algorithm'. https://github.com/propublica/compas-analysis, 2016.

[46] Yann LeCun, Ido Kanter, and Sara Solla. Second order properties of error surfaces: Learning time and generalization. *Advances in neural information processing systems*, 3, 1990.

[47] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.

[48] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

[49] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with emrs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 556–559. IEEE, 2014.

[50] Weiyang Liu, Rongmei Lin, Zhen Liu, James M Rehg, Liam Paull, Li Xiong, Le Song, and Adrian Weller. Orthogonal over-parameterized training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7251–7260, 2021.

[51] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[52] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pp. 189–202, 2020.

[53] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, volume 80, pp. 3381–3390, 2018. URL http://proceedings.mlr.press/v80/madras18a.html.

[54] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *ICML*, volume 119, pp. 6755–6764, 2020. URL http://proceedings.mlr.press/v119/martinez20a.html.

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

[56] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*, 2021.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[58] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

[59] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodola. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

[60] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021.

[61] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[62] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *IJCAI-20*, pp. 2277–2283, 7 2020. doi: 10.24963/ijcai.2020/315. URL https://doi.org/10.24963/ijcai.2020/315.

[63] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

[64] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[65] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, pp. 8227–8236. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00842. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Quadrianto_Discovering_Fair_Representations_in_the_Data_Domain_CVPR_2019_paper.html.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[67] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[68] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[69] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[70] Vinay V Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*, 2020.

[71] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2018.

[72] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Post-hoc methods for debiasing neural networks. *arXiv preprint arXiv:2006.08564*, 2020.

[73] Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms. In *International Conference on Machine Learning*, pp. 30646–30688. PMLR, 2023.

[74] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 international conference on electronics and sustainable communication systems (ICESC)*, pp. 490–494. IEEE, 2020.

[75] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2383–2389, 2021.

[76] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

[77] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[78] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32, 2019.

[79] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, 2020. URL `https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf`.

[80] Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31, 2018.

[81] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, pp. 9322–9331, 2020.

[82] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, pp. 5310–5319, 2019.

[83] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation, 2020.

[84] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

[85] Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6176–6185, 2017.

[86] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

[87] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[88] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *WWW*, Apr 2017. doi: 10.1145/3038912.3052660. URL `http://dx.doi.org/10.1145/3038912.3052660`.

[89] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, volume 54, pp. 962–970. PMLR, 2017. URL `http://proceedings.mlr.press/v54/zafar17a.html`.

[90] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

[91] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *JMLR*, 20(75):1–42, 2019. URL `http://jmlr.org/papers/v20/18-262.html`.

[92] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, number 3, pp. 325–333, 17–19 Jun 2013. URL `http://proceedings.mlr.press/v28/zemel13.html`.

[93] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

[94] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, 2018.

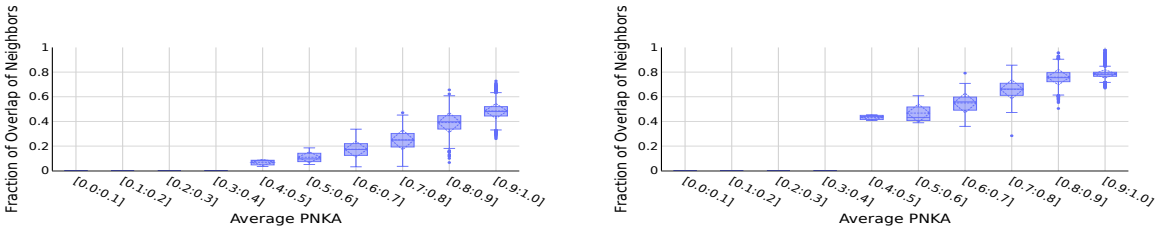
[95] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.

A Properties of PNKA

A.1 Overlap of neighbors

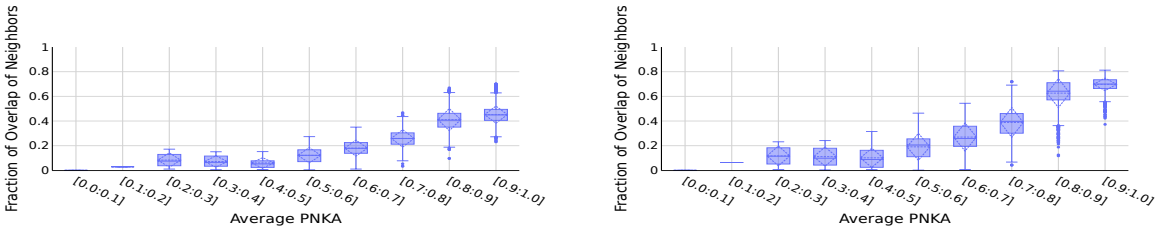
In Section 2.1, we argue that for an input example to be similarly represented in two representations, its neighborhood should be similar across both of them. Here, we empirically show that this intuition applies to PNKA, and that if the PNKA score of point i is higher than that of j, then i's nearest neighbors in representations Z and Z' overlap more than those of j. To show this, we train two models that only differ in their random initialization and compute their representations on the test set (10K instances). We use the penultimate layer (*i.e.*, the layer before logits) for the analysis. For each model, we determine a point's k nearest neighbors by ranking a point's representation distance (via either *cosine similarity* or *L2 distance*) to every other point in that representation. We then compute the fraction of those two sets of k neighbors that intersect.

In the following plots we depicts the relationship between PNKA similarity scores (x-axis) and the fraction of overlapping k nearest neighbors of each point (y-axis), *i.e.* 1 means all k nearest neighbors are shared between both representations. We report the analysis on CIFAR-10 and CIFAR-100 (39), for ResNet-18 (28), VGG-16 and Inception-V3, for different k values, up to $k = 20\%$ of the dataset size. All the results are reported over 3 runs. In all cases of Section A.1 we see a clear relationship between high PNKA scores and high overlap of nearest neighbors across representations, indicating that PNKA captures how similar the neighborhoods of the points are. We further show in Figures 7 and 8 that the direct comparison of representations through cosine similarity does not exhibit the same trend, *i.e.* there is not a positive correlation between cosine similarity scores and the fraction of overlap of the k nearest neighbors.



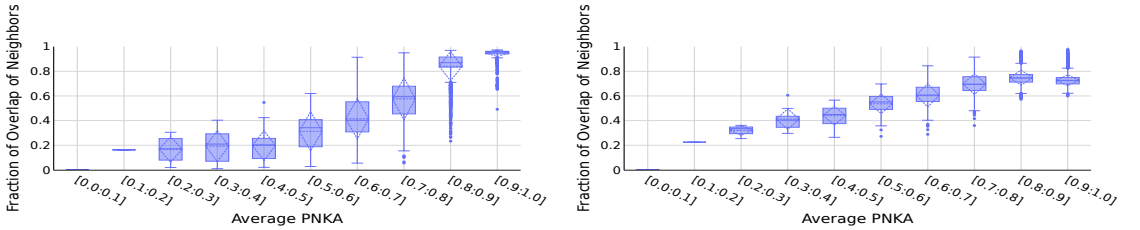
(a) ResNet-18, k=250 (2.5% total points)

(b) ResNet-18, k=2000 (20% total points)



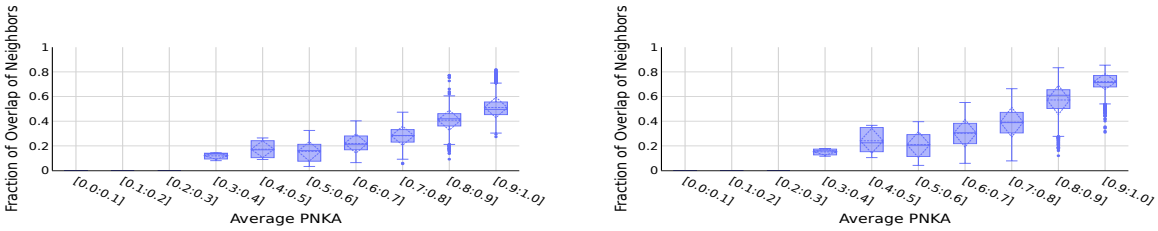
(c) VGG-16, k=250 (2.5% total points)

(d) VGG-16, k=500 (5% total points)



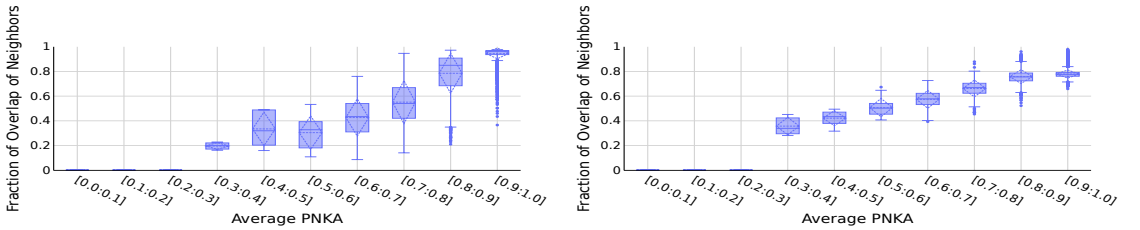
(e) VGG-16, k=1000 (10% total points)

(f) VGG-16, k=2000 (20% total points)



(g) Inception-V3, k=250 (2.5% total points)

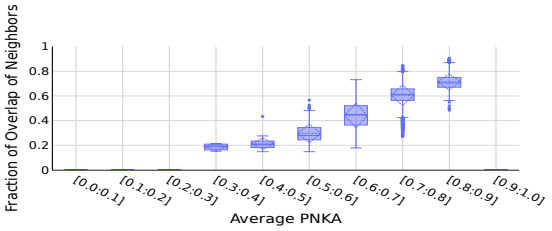
(h) Inception-V3, k=500 (5% total points)



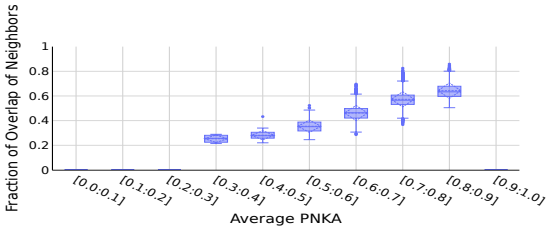
(i) Inception-V3, k=1000 (10% total points)

(j) Inception-V3, k=2000 (20% total points)

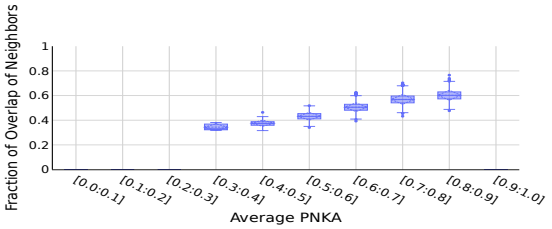
Figure 5: **PNKA** captures the overlap of k nearest neighbors between two representations, *i.e.*, the higher PNKA scores, the higher the fraction of overlapping neighbors. Results are an average over 3 runs, each one containing two models trained on **CIFAR-10** (39) dataset with the same architecture but different random initialization.



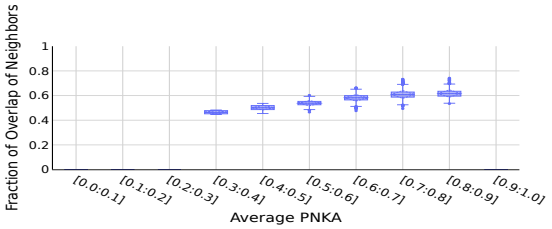
(a) ResNet-18, k=250 (2.5% total points)



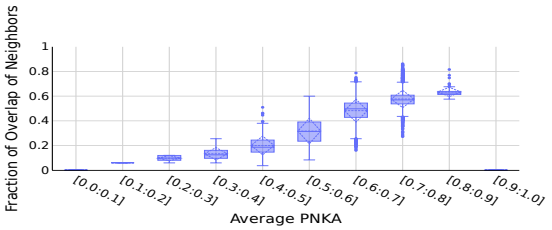
(b) ResNet-18, k=500 (5% total points)



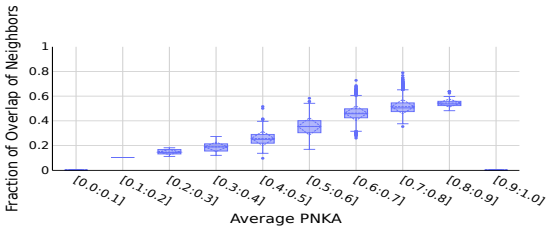
(c) ResNet-18, k=1000 (10% total points)



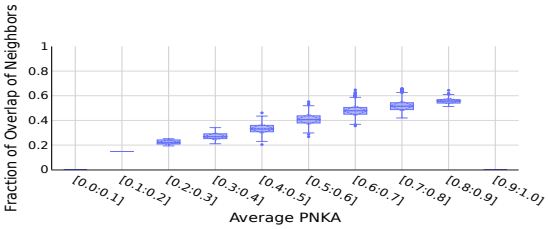
(d) ResNet-18, k=2000 (20% total points)



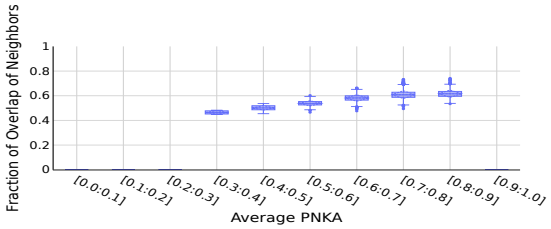
(e) VGG-16, k=250 (2.5% total points)



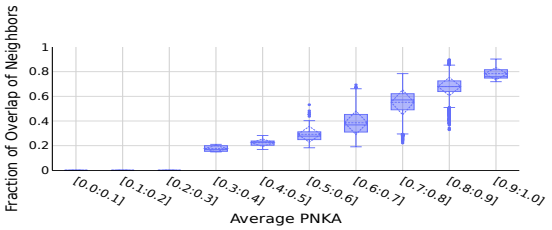
(f) VGG-16, k=500 (5% total points)



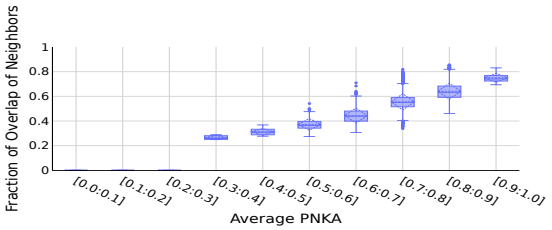
(g) VGG-16, k=1000 (10% total points)



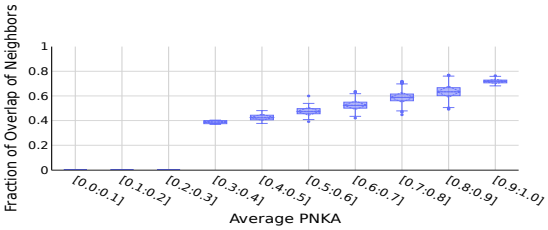
(h) VGG-16, k=2000 (20% total points)



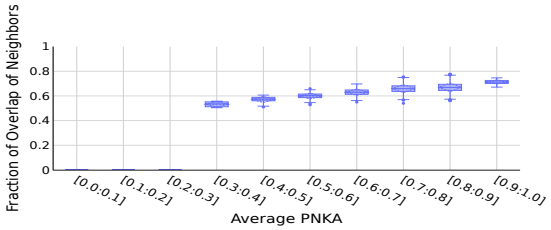
(i) Inception-V3, k=250 (2.5% total points)



(j) Inception-V3, k=500 (5% total points)

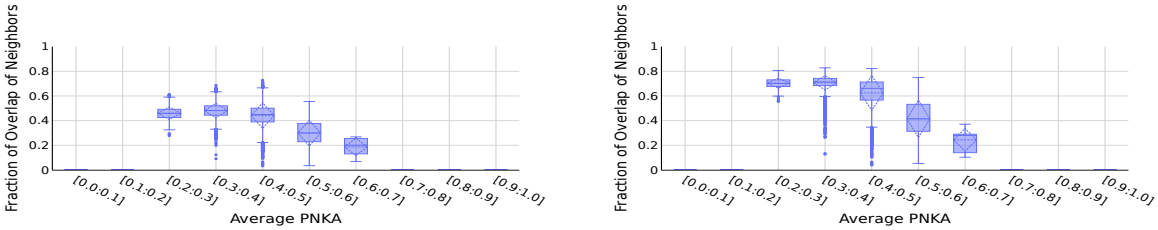


(k) Inception-V3, k=1000 (10% total points)



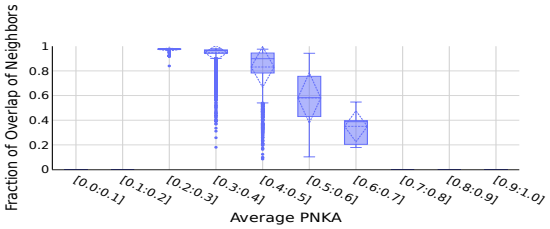
(l) Inception-V3, k=2000 (20% total points)

Figure 6: **PNKA** captures the overlap of k nearest neighbors between two representations, *i.e.*, the higher PNKA scores, the higher the fraction of overlapping neighbors. Results are an average over 3 runs, each one containing two models trained on **CIFAR-100** (39) dataset with the same architecture but different random initialization.

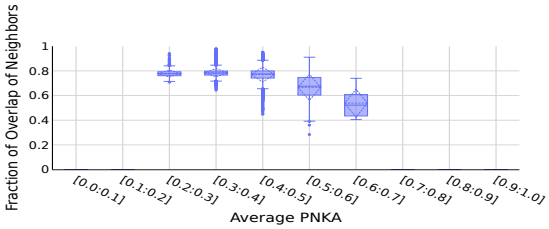


(a) ResNet-18, k=250 (2.5% total points)

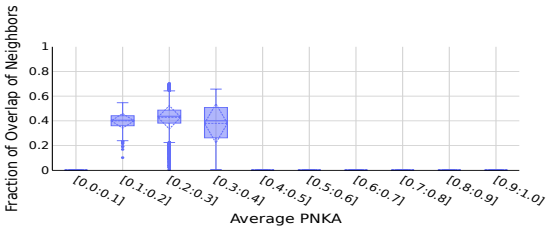
(b) ResNet-18, k=500 (5% total points)



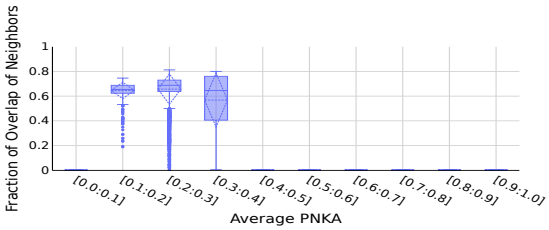
(c) ResNet-18, k=1000 (10% total points)



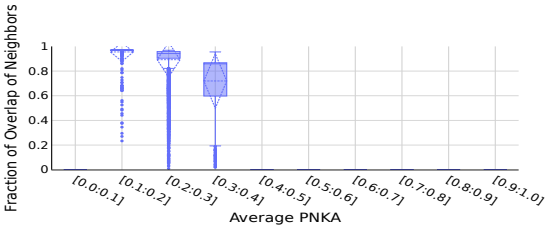
(d) ResNet-18, k=2000 (20% total points)



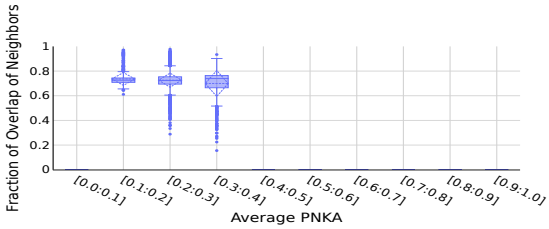
(e) VGG-16, k=100 (2.5% total points)



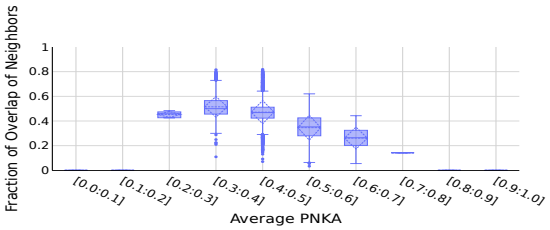
(f) VGG-16, k=500 (5% total points)



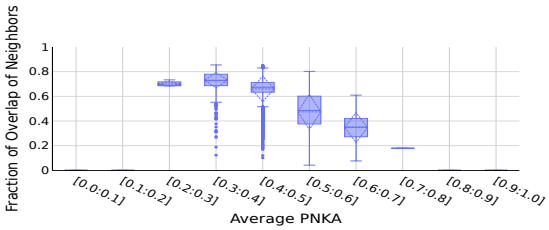
(g) VGG-16, k=1000 (10% total points)



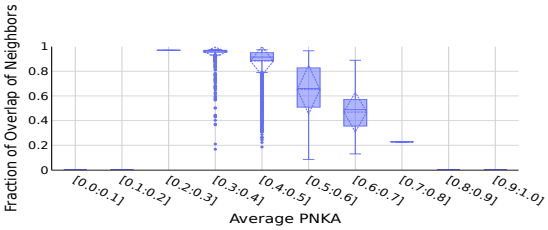
(h) VGG-16, k=2000 (20% total points)



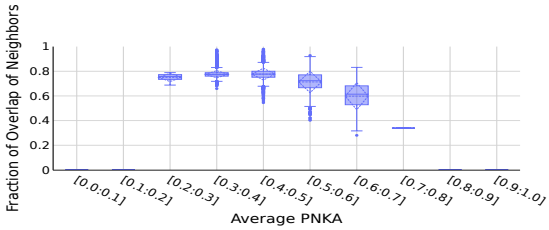
(i) Inception-V3, k=250 (2.5% total points)



(j) Inception-V3, k=500 (5% total points)

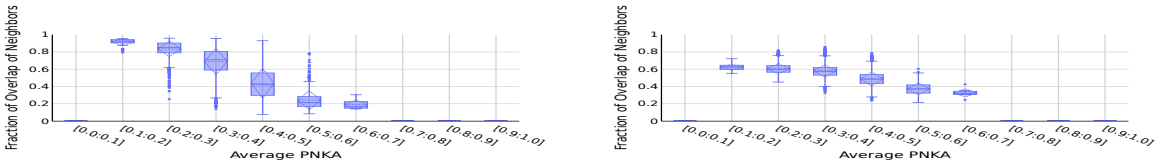


(k) Inception-V3, k=1000 (10% total points)



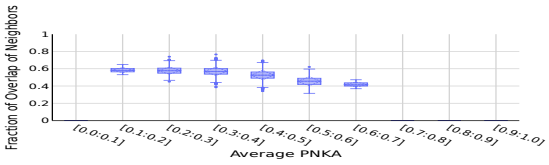
(l) Inception-V3, k=2000 (20% total points)

Figure 7: **Cosine similarity** is not able to captures the overlap of k nearest neighbors between two representations *i.e.* there is not a positive correlation between cosine similarity scores and the fraction of overlap of the k nearest neighbors. Results are an average over 3 runs, each one containing two models trained on **CIFAR-10** (39) dataset with the same architecture but different random initialization.

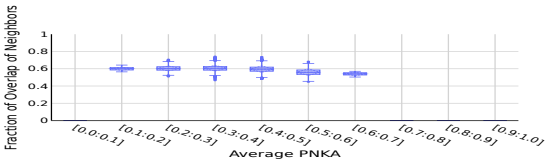


(a) ResNet-18, k=100 (1% total points)

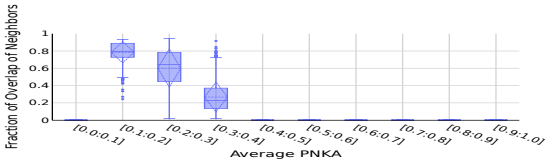
(b) ResNet-18, k=500 (5% total points)



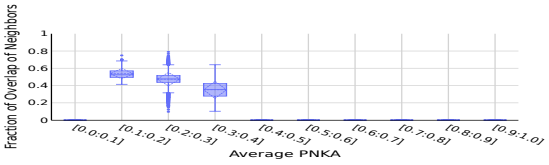
(c) ResNet-18, k=1000 (10% total points)



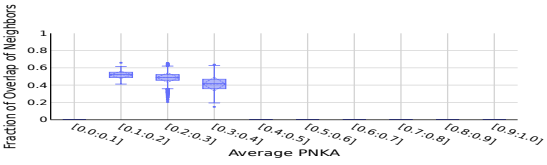
(d) ResNet-18, k=2000 (20% total points)



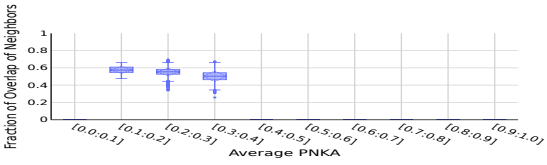
(e) VGG-16, k=100 (1% total points)



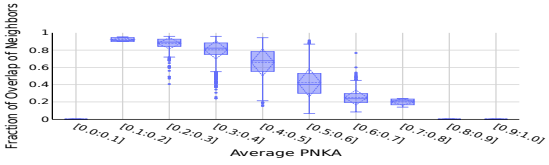
(f) VGG-16, k=500 (5% total points)



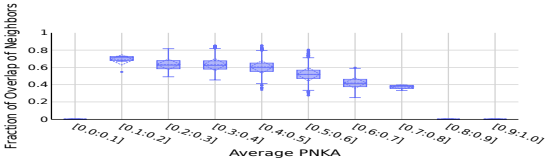
(g) VGG-16, k=1000 (10% total points)



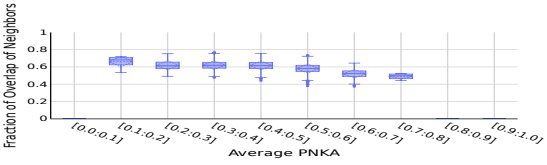
(h) VGG-16, k=2000 (20% total points)



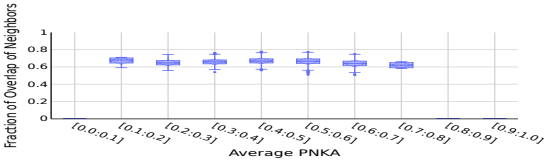
(i) Inception-V3, k=100 (1% total points)



(j) Inception-V3, k=500 (5% total points)



(k) Inception-V3, k=1000 (10% total points)



(l) Inception-V3, k=2000 (20% total points)

Figure 8: **Cosine similarity** is not able to captures the overlap of k nearest neighbors between two representations *i.e.* there is not a positive correlation between cosine similarity scores and the fraction of overlap of the k nearest neighbors. Results are an average over 3 runs, each one containing two models trained on **CIFAR-100** (39) dataset with the same architecture but different random initialization.

A.2 Relationship of PNKA with aggregate measures of representation similarity

Dataset	Model	CKA	$\overline{\text{PNKA}}$
CIFAR-10	ResNet-18	0.925 ($\pm$0.005)	0.925 ($\pm$0.022)
	VGG-16	0.895 ($\pm$0.013)	0.893 ($\pm$0.039)
	Inception-v3	0.916 ($\pm$0.001)	0.915 ($\pm$0.023)
CIFAR-100	ResNet-18	0.741 ($\pm$0.00)	0.733 ($\pm$0.033)
	VGG-16	0.658 ($\pm$0.010)	0.668 ($\pm$0.049)
	Inception-v3	0.798 ($\pm$0.009)	0.792 ($\pm$0.032)

Table 3: Comparison between CKA (37) and the aggregate version of PNKA ($\overline{\text{PNKA}}$). Results are an average over 3 runs, each one with two models that only differ in their random initialization. We capture the representations of the penultimate layer (*i.e.*, the layer before logits) for the analysis. We show that both measures produce similar overall scores.

A.3 Proof of invariances

A.3.1 Invariance to orthogonal transformations

Proof. Given an orthogonal matrix Q, it suffices to show that

$$\begin{aligned} K(ZQ) &= ZQ(ZQ)^\top \\ &= ZQQ^\top Z^\top \\ &= ZQQ^{-1}Z^\top \\ &= ZZ^\top \\ &= K(Z) \end{aligned}$$

Here we have used that for an orthogonal matrix Q, $Q^\top = Q^{-1}$. By substituting $K(ZQ)$ and $K(Z'R)$ in $\text{PNKA}(ZQ, Z'R, i) = \cos(K(ZQ)_i, K(Z'R)_i)$ with $K(Z)$ and $K(Z')$, respectively, we obtain $\text{PNKA}(ZQ, Z'R, i) = \text{PNKA}(Z, Z', i)$. Thus, PNKA is invariant to orthogonal transformations. □

A.3.2 Invariance to isotropic scaling

Proof. Note that because of the bilinearity of the dot-product, we have $K(\alpha Z)_i = \left[(\alpha Z)(\alpha Z)^\top\right]_i = \alpha^2 K(Z)_i$. By substituting into PNKA, we get

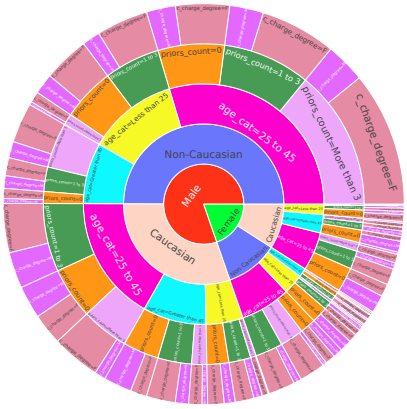
$$\begin{aligned} \text{PNKA}(\alpha Z, \beta Z', i) &= \frac{K(\alpha Z)_i^\top K(\beta Z')_i}{||K(\alpha Z)_i||_2 ||K(\beta Z')_i||_2} \\ &= \frac{\alpha^2 K(Z)_i^\top \beta^2 K(Z')}{||\alpha^2 K(Z)_i||_2 ||\beta^2 K(Z')_i||_2} \\ &= \frac{\alpha^2 K(Z)_i^\top \beta^2 K(Z')}{\alpha^2 ||K(Z)_i||_2 \beta^2 ||K(Z')_i||_2} \\ &= \text{PNKA}(Z, Z', i). \end{aligned}$$

Thus, PNKA is invariant to isotropic scaling. □

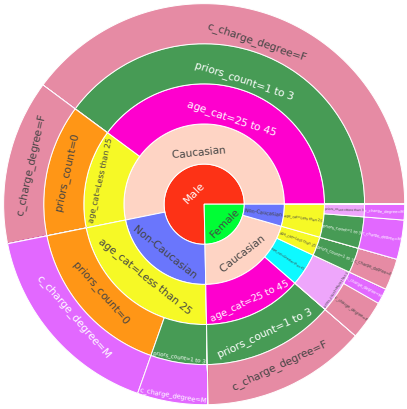
B Investigating Debiased Tabular Data Representations

B.1 Distribution of attributes for the data points whose representation changed the most

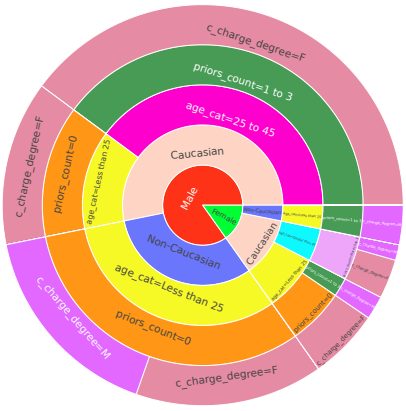
In this section, we present graphical illustrations of the distribution of all the attributes within the COMPAS and Adult datasets using a series of concentric rings. Each ring functions as a pie chart, depicting the distribution of a specific attribute. For instance, the innermost two rings represent the gender (male/female) and racial (Caucasian/non-Caucasian) distributions in both datasets. The subsequent rings are tailored to each dataset. For the COMPAS dataset, the third, fourth, and fifth outer rings display the distributions of age categories (under 25, 25-45, over 45 years old), prior counts (0, 1-3, more than 3), and type of charge (felony or misdemeanor), respectively. For the Adult dataset, the third and fourth outer rings showcase the age categories (grouped by decades) and education levels. Additionally, the first plot of Figure 9 and Figure 10, located in the top left, illustrates the overall distribution of these attributes across the entire dataset. The subsequent plots provide a detailed view of the attribute distributions for the bottom 10% of data points, which exhibited the most significant changes in their representation.



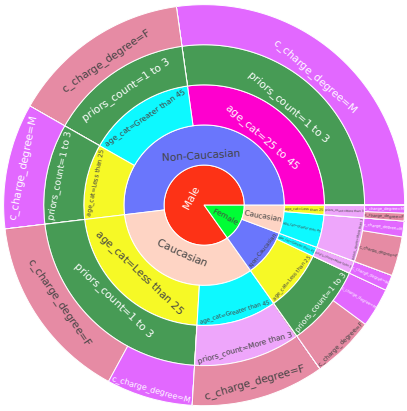
(a) Distribution of dataset (baseline).



(b) U × U-GF-IF.

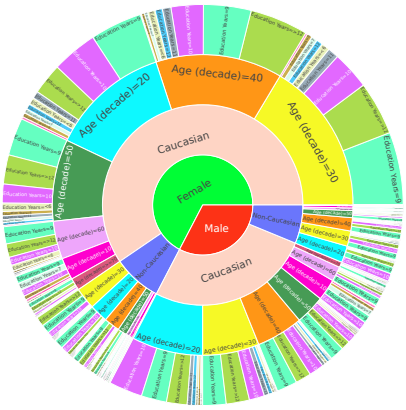


(c) U × U-GF.

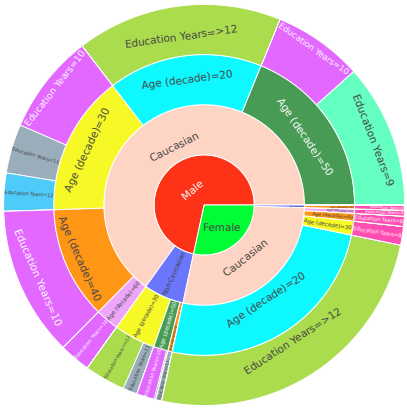


(d) U × U-IF.

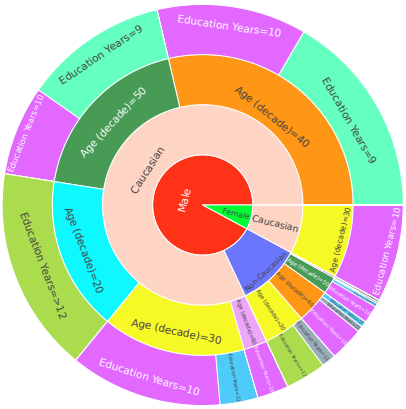
Figure 9: Distribution of all the attributes of the 10% of instances with the lowest PNKA scores for COMPAS dataset with the protected attribute as race.



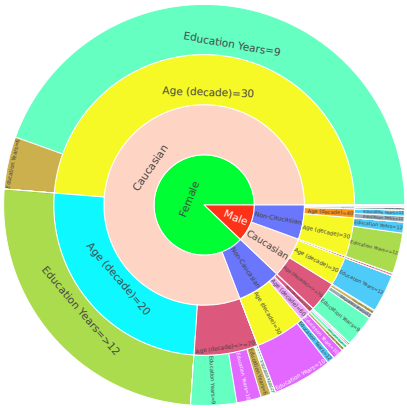
(a) Distribution of dataset (baseline).



(b) U × U-GF-IF.



(c) U × U-GF.



(d) U × U-IF.

Figure 10: Distribution of all the attributes of the 10% of instances with the lowest PNKA scores for Adult dataset with the protected attribute as gender.

C Investigating Debiased Word Embeddings

C.1 Results on SemBias dataset

For each of the four word pairs (a, b) in a SemBias instance, GP- and GN-Glove measure its cosine similarity with the canonical gender vector, *i.e.* $cos(\overrightarrow{a} - \overrightarrow{b}, \overrightarrow{he}$ - $\overrightarrow{she})$. The word pair with the highest cosine similarity is selected as the "predicted" answer. If the word embeddings are correctly debiased, then the cosine similarity of the $\overrightarrow{he}$ - $\overrightarrow{she}$ vector with the gender-definition words should be high, and the similarity with the gender-stereotype words should be low, *i.e.* the frequency of predictions for these categories should be high and low, respectively. Table 4 depicts the results for the GN- and GP-Glove (94; 33) methods.

Embeddings	SemBias		
	Definition ↑	Stereotype ↓	Neutral ↓
GloVe	80.2	10.9	8.9
GN-GloVe	97.7	1.4	0.9
GP-GloVe	84.3	8.0	7.7
GP-GN-GloVe	98.4	1.1	0.5

Table 4: Frequency of predictions for gender relational analogies (33). Each column shows the frequency with which the respective word-pair category (gender-definitional, gender-stereotype, gender-neutral) is predicted as having the highest cosine similarity with the canonical gender vector $\overrightarrow{he}$ - $\overrightarrow{she}$. The more often gender-definition words are predicted as being most gender-aligned, as opposed to gender-stereotype words, the less biased an embedding approach can be considered.

D Investigating Debiased Contextual Embeddings

Transformer-based language models are currently the leading approach for many NLP tasks. In contrast to the previously discussed word embeddings, these models use contextualized embeddings where the representation of a word or token also depends on the preceding and possibly succeeding tokens. However, just like their non-contextualized predecessors, these models have also been found to embed several unfair biases (95; 78; 40). The predominant approach to using transformer-based (large) language models is to adapt pre-trained models to downstream tasks, either with or without fine-tuning. Any biases that the pre-trained base-models exhibit might thus be inherited by the downstream tasks. Since a small number of base-models are adapted for many different downstream tasks, it is infeasible to audit models on each of those tasks separately. Rather, analyzing biases in the representations that base models use provides a scalable way to study their fairness.

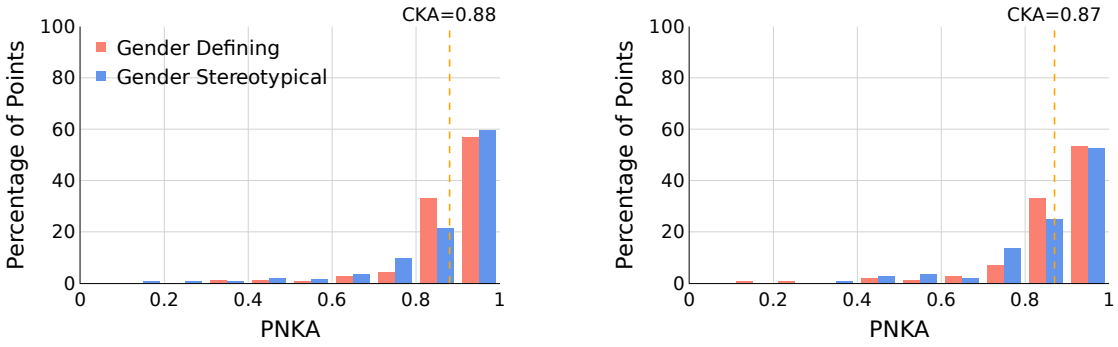
Given the importance of addressing these biases for transformer-based language models, researchers have started proposing methods aimed to debias the contextualized embeddings of these models. One method by Kaneko & Bollegala (34) proposes to remove a potential gender bias by fine-tuning a contextual baseline model to "preserve semantic information with respect to sentences with sensitive attributes (*e.g.* 'she', 'he'), while removing any discriminatory biases with respect to sentences with stereotypical words (*e.g.* 'poetry', 'math')". In other words, the goal of this method is to remove gender bias from the representations of sentences containing potentially stereotypical words, such as ones related to poetry (with a presumed female bias) and ones related to math (with a presumed male bias).

We leverage PNKA to investigate whether this debiasing method changes the representations of sentences in the intended manner. For our analysis, we use the Albert ('albert-base-v2') model as baseline, and evaluate it and its debiased versions on the same dataset used to study the gender bias in the original work of Kaneko & Bollegala (34), specifically the SEAT-7 and SEAT-8 datasets (55). These datasets are composed of simple sentence templates such as "This is a [BLANK]", and create gender defining and gender stereotypical sentences by substituting "[BLANK]" with gender defining (*e.g.*, 'she', 'he') and target stereotypical words (*e.g.*, 'poetry', 'math'), respectively. The specific words were chosen based on the WEAT measure (12), which measures the associations between concepts like math/art and science/art with male/female attributes in SEAT-7 and SEAT-8, respectively. A more detailed explanation on WEAT and SEAT is provided in Appendix D.1 and D.2.

Figure 11 shows the distribution of PNKA similarity scores for the two categories of sentences in SEAT-7 (Figure 11a) and SEAT-8 (Figure 11b), respectively. In both cases, we observe an overall high PNKA score, with more than 80% of points obtaining PNKA scores higher or equal to 0.8, and CKA of 0.88. This suggests that most of the instances of both gender defining as well as gender stereotypical sentence categories have not drastically changed from the baseline to the debiased model. More surprisingly, there is no clear distinction between the two groups of sentences, *i.e.*, both gender defining (*e.g.*, "This is a woman." or "This is a man.") and gender stereotypical (*e.g.*, "This is poetry." or "This is math.") sentences change in a similar proportion.

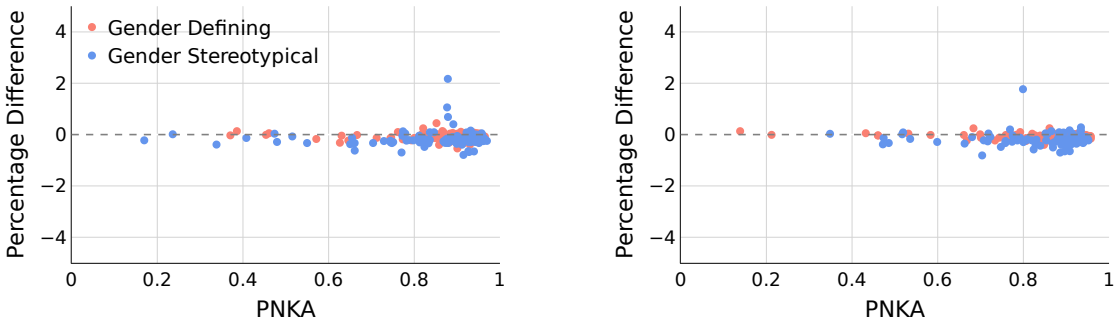
To investigate whether the high representation similarity indicates a lack of change in the gender properties of the sentence, we follow the procedure used for the non-contextual embeddings described in Section 4. Again, we project the sentence representations onto the gender vector $\overrightarrow{he}$ - $\overrightarrow{she}$ [5] and measure the change in projection magnitude from the baseline to the debiased version. We follow Kaneko & Bollegala (34) and obtain representations of the [CLS] token at the last layer. Figure 12 displays the relationship between PNKA scores and the percentage difference $\omega_i^{(baseline)}$ for the baseline ('albert-base-v2') as well as $\omega_i^{(debiased)}$ for debiased model. In accordance with the high PNKA scores, we observe that all the contextual embeddings exhibit a minimal shift along the gender direction. Once again, PNKA is a reliable indicator that the debiasing technique is not effective in the intended way, and that the bias of gender stereotypical sentences is not reduced as expected. Our insights complement previous work, which also identified that using SEAT alone is insufficient to evaluate debiasing methods (56; 75; 55). Future work could explore the effects of this method further to understand why it does not significantly alter the representations in the intended way.

[5] In Appendix D.3 we also explore using the average contextual gender vector and observe a similar pattern.



(a) SEAT-7. (b) SEAT-8.

Figure 11: Distribution of PNKA scores per group of sentences in SEAT-7 and SEAT-8 dataset (55). We compare the baseline ('albert-base-v2') model and its debiased version (34). Sentences with the lowest similarity scores are the ones that change the most from the baseline to the debiased version. Across all the datasets, we observe that most of the sentences maintain high PNKA scores, which indicates that they have not substantially changed their representations. Moreover, there is no clear difference between the groups of gender defining and gender stereotypical sentences in how they change their representations.



(a) SEAT-7. (b) SEAT-8.

Figure 12: Relationship between PNKA scores (x-axis) and percentage difference (y-axis) in magnitude of the projection on the gender direction $\overrightarrow{he}$ - $\overrightarrow{she}$. A positive or negative percentage difference value indicates a shift in magnitude along the gender direction. Overall, the contextual embeddings exhibit a low shift along the gender direction, with no clear distinction between groups of sentences.

D.1 Word Embedding Association Test (WEAT)

The description below is provided by Li et al. (47). Word Embedding Association Test (WEAT) (12) measures the association between two sets of attributes words (e.g., male and female) and two sets of targets words (e.g., family and career). Formally, the sets of attribute words are indicated by A and B, and the sets of target words are denoted by X and Y. Then, the WEAT test is as follows:

$$s(A, B, X, Y) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(x, A, B), \tag{2}$$

where $s(w, A, B)$ represents the difference between the average of the cosine similarity of word w with all words in A and the average of the cosine similarity of word w to all words in B, and it is defined as follows:

$$s(w, X, Y) = \frac{1}{|A|} \sum_{a \in A} cosine(w, a) - \frac{1}{|B|} \sum_{b \in B} cosine(w, b), \tag{3}$$

where $w \in X$ or Y, and $cosine(\cdot, \cdot)$ represents the cosine similarity. The normalized effect size is as follows:

$$d = \frac{\mu(\{s(x, A, B)_{x \in X}\}) - \mu(\{s(y, A, B)_{y \in Y}\})}{\sigma(\{s(t, X, Y)_{t \in A \cup B}\})}, \tag{4}$$

where $\mu(\cdot)$ is the mean function and $\sigma(\cdot)$ is the standard deviation.

D.2 Sentence Embedding Association Test (SEAT)

Sentence Embedding Association Test (SEAT) (55) adapts WEAT to contextual embeddings, which uses simple sentence templates such as "This is a [BLANK]" to substitute attribute words and target words to obtain context-independent embeddings. Then the SEAT test statistic between the two sets of embeddings (represented by the '[CLS]' of the last layer) is calculated similarly to Equation 4.

D.3 Results of Projection with Average Contextual Gender Direction

To investigate whether the high representation similarity indicates a lack of change in the gender properties of the sentence, we follow the procedure used for the non-contextual embeddings described in Sections 4. However, in this case, instead of just using the gender directions of the words *he* and *she*, we use a gender direction from an aggregated embeddings of male and female sentences. In other words, we project the sentence representations of the SEAT dataset onto the average contextual gender vector originally used by Kaneko & Bollegala (34) to debias the model. This average contextual gender vector is obtained as follows. First, we obtain an average representation for the sentences $S(w)$ where the gendered-word w appears. This gendered-word w belongs to an attribute $A \in \mathcal{A}$, where $\mathcal{A} = \{A_f, A_m\}$, and A denotes a set of words associated with the corresponding gender, where f stands for female and m stands for male.

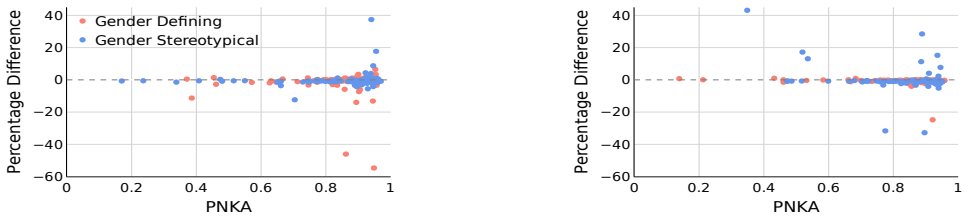
$$e_A(w) = \frac{1}{|S(w)|} \sum_{s \in S(w)} e(s). \tag{5}$$

Next, we take the final average contextual gender vector for attribute A, which can be A_f for female and A_m for male, as follows:

$$\overline{e}_A = \frac{1}{|A|} \sum_{w \in A} e_A(w). \tag{6}$$

Thus, the projection will be computed using the average contextual gender direction $\overline{e}_{A_m} - \overline{e}_{A_f}$.

Figure 13 displays the relationship between PNKA scores and the percentage difference for the baseline ('albert-base-v2') as well as for debiased model according to the average contextual gendered direction. As before, and in accordance with the high PNKA scores, we observe that most of the contextual embeddings exhibit a minimal shift along the gender direction.



(a) SEAT-7.

(b) SEAT-8.

Figure 13: Relationship between PNKA scores (x-axis) and percentage difference (y-axis) in magnitude of the projection on the average gendered direction. A positive or negative percentage difference value indicates a shift in magnitude along the gender direction. Overall, the contextual embeddings exhibit a low shift along the gender direction, with no clear distinction between groups of sentences.