

Semantic Properties of Cosine Based Bias Scores for Word Embeddings

Sarah Schröder^a, Alexander Schulz^b, Fabian Hinder^c and Barbara Hammer^d

Machine Learning Group, Bielefeld University, Bielefeld, Germany

Keywords: Language Models, Word Embeddings, Social Bias.


Abstract: Plenty of works have brought social biases in language models to attention and proposed methods to detect such biases. As a result, the literature contains a great deal of different bias tests and scores, each introduced with the premise to uncover yet more biases that other scores fail to detect. What severely lacks in the literature, however, are comparative studies that analyse such bias scores and help researchers to understand the benefits or limitations of the existing methods. In this work, we aim to close this gap for cosine based bias scores. By building on a geometric definition of bias, we propose requirements for bias scores to be considered meaningful for quantifying biases. Furthermore, we formally analyze cosine based scores from the literature with regard to these requirements. We underline these findings with experiments to show that the bias scores' limitations have an impact in the application case.


1 INTRODUCTION


In the domain of Natural Language Processing (NLP), many works have investigated social biases in terms of associations in the embeddings space. Early works (Bolukbasi et al., 2016; Caliskan et al., 2017) introduced methods to measure and mitigate social biases based on cosine similarity in word embeddings. With NLP research progressing to large language models and contextualized embeddings, doubts have been raised whether these methods are still suitable for fairness evaluation (May et al., 2019) and other works criticize that for instance the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) fails to detect some kinds of biases (Gonen and Goldberg, 2019; Ethayarajh et al., 2019). Overall there exists a great deal of bias measures in the literature, which not necessarily detect the same biases (Kurita et al., 2019; Gonen and Goldberg, 2019; Ethayarajh et al., 2019). In general, researchers are questioning the usability of model intrinsic bias measures, such as cosine based methods (Steed et al., 2022; Goldfarb-Tarrant et al., 2020; Kaneko et al., 2022). There exist few papers that compare the performance of different bias scores (Delobelle et al., 2021; Schröder et al., 2023) and works that evaluate experimental setups for bias mea-


surement (Seshadri et al., 2022). However, to our knowledge, only two works investigate the properties of intrinsic bias scores on a theoretical level (Ethayarajh et al., 2019; Du et al., 2021). To further close this gap, we evaluate the semantic properties of cosine based bias scores, focusing on bias quantification as opposed to bias detection. We make the following contributions: (i) We formalize the properties of trustworthiness and comparability as requirements for cosine based bias scores. (ii) We analyze WEAT and the Direct Bias, two prominent examples from the literature. (iii) We conduct experiments to highlight the behavior of WEAT and the Direct Bias in practice.

Both our theoretical analysis and experiments show limitations of these bias scores in terms of bias quantification. It is crucial that researchers take these limitations into account when considering WEAT or the Direct Bias for their works. Furthermore, we lay the ground work to analyze other cosine based bias scores and understand how they can be useful for the fairness literature. The paper is structured as follows: In Section 2 we summarize WEAT, the Direct Bias and general terminology for cosine based bias measures from the literature. We introduce formal requirements for such bias scores in Section 3 and analyze WEAT and the Direct Bias in terms of these requirements in Section 4. In Section 5 we support our theoretical findings by experiments, before drawing our conclusions in Section 6.

^a  <https://orcid.org/0000-0002-7954-3133>

^b  <https://orcid.org/0000-0002-0739-612X>

^c  <https://orcid.org/0000-0002-1199-4085>

^d  <https://orcid.org/0000-0002-0935-5591>

2 RELATED WORK FOR BIAS IN WORLD EMBEDDINGS

2.1 WEAT

The Word Embedding Association Test, short WEAT, (Caliskan et al., 2017), is a statistical test for stereotypes in word embeddings. The test compares two sets of target words X and Y with two sets of bias attributes A and B of equal size n under the hypothesis that words in X are rather associated with words in A and words in Y rather associated with words in B . The association of a single word \mathbf{w} with the bias attribute sets A and B including n attributes each, is given by

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

To measure bias in the sets X and Y , the effect size is used, which is a normalized measure for the association difference between the target sets

$$d(X, Y, A, B) = \frac{\frac{1}{n} \sum_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \frac{1}{n} \sum_{\mathbf{y} \in Y} s(\mathbf{y}, A, B)}{\text{stddev}_{\mathbf{w} \in X \cup Y} s(\mathbf{w}, A, B)}. \quad (2)$$

A positive effect size confirms the hypothesis that words in X are rather stereotypical for the attributes in A and words in Y stereotypical for words in B , while a negative effect size indicates that the stereotypes would be counter-wise. To determine if the effect is indeed statistically significant, the permutation test

$$p = P_r[s(X_i, Y_i, A, B) > s(X, Y, A, B)]. \quad (3)$$

with subsets (X_i, Y_i) of $X \cup Y$ and the test statistic

$$s(X, Y, A, B) = \sum_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \sum_{\mathbf{y} \in Y} s(\mathbf{y}, A, B) \quad (4)$$

is done. As a statistical test WEAT is suited to confirm a hypothesis (such that a certain type of stereotype exists in a model), but it cannot prove the opposite.

2.2 Direct Bias

The Direct Bias (Bolukbasi et al., 2016) is defined as the correlation of neutral words $\mathbf{w} \in W$ with a bias direction (for example gender direction \mathbf{g}):

$$\text{DirectBias}(W) := \frac{1}{|W|} \sum_{\mathbf{w} \in W} |\cos(\mathbf{w}, \mathbf{g})|^c \quad (5)$$

with c determining the strictness of bias measurement. The gender direction is either obtained by a gender word-pair e.g. $\mathbf{g} = \mathbf{he} - \mathbf{she}$ or - to get a more robust estimate - it is obtained by computing the first principal component over a set of individual gender directions from different word-pairs.

In terms of their debiasing algorithm the authors describe how to obtain a bias subspace given defining sets D_1, \dots, D_n . A defining set D_i includes words \mathbf{w} that only differ by the bias relevant topic e.g. for gender bias $\{\mathbf{man}, \mathbf{woman}\}$ could be used as a defining set. Given these sets, the authors construct individual bias directions $\mathbf{w} - \mu_i \forall \mathbf{w} \in D_i, i \in \{1, \dots, n\}$ and $\mu_i = \sum_{\mathbf{w} \in D_i} \frac{\mathbf{w}}{|D_i|}$. To obtain a k -dimensional bias subspace B they compute the k first principal components over these samples.

2.3 Terminology

In the literature geometrical bias is measured by comparing neutral targets against sensitive attributes. By targets and attributes we refer to vector representations of words, sentences or text in a d -dimensional embedding space. However, the methodology can be applied to any kind of vector representations. While the exact notation varies between publications, we summarize and use it in the following Sections as follows:

Given a protected attribute like gender or race, we select $n \geq 2$ protected groups that might be subject to biases. Each protected group is defined by a set of attributes $\mathbf{a}_{ik} \in A_i$ with $i \in \{1, \dots, n\}$ the group's index. We summarize these attribute sets as $A = \{A_1, \dots, A_n\}$. The intuition is that the attributes define the relation of protected groups by contrasting specifically over the membership to the different groups. Therefore, it is important that any attribute $\mathbf{a}_{ik} \in A_i$ has a counterpart $\mathbf{a}_{jk} \in A_j \forall A_j \in A, j \neq i$ that only differs from \mathbf{a}_{ik} by the group membership. For instance, if we used $A_1 = \{\mathbf{she}, \mathbf{female}, \mathbf{woman}\}$ as a selection of female terms, $A_2 = \{\mathbf{he}, \mathbf{male}, \mathbf{man}\}$ would be the proper choice of male terms.

Analogously to WEAT's definition of word biases, we define the association of a target \mathbf{t} with one protected group, represented by A_i , as

$$s(\mathbf{t}, A_i) = \frac{1}{|A_i|} \sum_{\mathbf{a}_{ik} \in A_i} \cos(\mathbf{t}, \mathbf{a}_{ik}) \quad (6)$$

A similar notion is found with the Direct Bias (Bolukbasi et al., 2016). To detect bias, one would consider the difference of associations towards the different groups, i.e. is \mathbf{t} more similar to one protected group than the others. This concept is also found in most cosine based bias scores.

Whether such association differences are harmful depends on whether \mathbf{t} is theoretically neutral to the protected groups. For example, terms like "aunt" or "uncle" are associated with one or the other gender per definition, while a term like "nurse" should not be associated with gender.

3 FORMAL REQUIREMENTS FOR BIAS SCORES

3.1 Formal Bias Definition and Notations

As baseline for our bias score requirements and the following analysis of bias scores from the literature, we suggest two intuitive definitions of individual bias for target samples (e.g. one word) \mathbf{t} and aggregated biases for sets of targets T . For samples \mathbf{t} we apply the intuition of WEAT, extended to n protected groups instead of only two.

Definition 3.1 (Individual Bias). Given n protected groups represented by attribute sets A_1, \dots, A_n and a target \mathbf{t} that is theoretically neutral to these groups, we consider \mathbf{t} biased if

$$\exists A_i, A_j \in A : s(\mathbf{t}, A_i) > s(\mathbf{t}, A_j) \quad (7)$$

Definition 3.2 (Aggregated Bias). Given n protected groups represented by attribute sets A_1, \dots, A_n and a set of targets T containing only samples that are theoretically neutral to these groups, we consider T biased if at least one sample $\mathbf{t} \in T$ is biased:

$$\exists A_i, A_j \in A, \mathbf{t} \in T : s(\mathbf{t}, A_i) > s(\mathbf{t}, A_j) \quad (8)$$

The idea behind Definition 3.2 is that even when looking at aggregated biases, each individual bias is important, i.e. as long as there is one biased target in the set, we cannot call the set unbiased, even if target biases cancel out on average or the majority of targets is unbiased.

In the following we will use a notation for bias score functions in general: $b(\mathbf{t}, A)$ measuring the bias of one target and $b(T, A)$ for aggregated biases. Note that there are two different strategies in the literature: Bias scores measuring bias over all neutral words jointly (Direct Bias), which matches our notation $b(T, A)$, and bias scores measuring the bias over two groups of neutral words $X, Y \subset T$ (WEAT). In the later case, we consider the selection of subsets $X, Y \subset T$ as part of the bias score and thus treat it as a function $b(T, A)$.

Since the bias scores from the literature have different extreme values and different values indicating no bias, we use the following notations: b_{min} and b_{max} are the extreme values of $b(\cdot)$ and b_0 is the value of $b(\cdot)$ that means \mathbf{t} or T is unbiased. Note that b_{min} and b_0 are not necessarily equal.

3.2 Requirements for Bias Metrics

Based on the definitions of bias explained in Section 3.1, we formalize the properties of *trustworthiness* and *magnitude-comparability*. The goal of both

properties is to ensure that biases can be quantified in a way such that bias scores can be safely compared between different embedding models and debiasing methods can be evaluated without risking to overlook bias.

3.2.1 Comparability

The goal of *magnitude-comparability*, is to ensure that bias scores are comparable between embeddings of different models. This is necessary to make statements about embedding models being more or less biased than others, which includes comparing debiased embeddings with their original counterparts. We find a necessary condition for such comparability is the possibility to reach the extreme values b_{min} and b_{max} of $b(\cdot)$ in different embedding spaces depending only on the neutral targets and their relation to attribute vectors, as opposed to the attribute vectors themselves, which might be embedded differently given different models.

Definition 3.3 (Magnitude-Comparable). We call the bias score function $b(T, A)$ Magnitude-Comparable if, for a fixed number of target samples in set T (including the case $T = \{\mathbf{t}\}$), the maximum bias score b_{max} and the minimum bias score b_{min} are independent of the attribute sets in A :

$$\max_{T, |T|=const} b(T, A) = b_{max} \forall A, \quad (9)$$

$$\min_{T, |T|=const} b(T, A) = b_{min} \forall A. \quad (10)$$

3.2.2 Trustworthiness

The second property of *trustworthiness* defines whether we can trust a bias score to report any bias in accordance to Definitions 3.1 and 3.2, i.e. the bias score can only reach b_0 , which indicates fairness, if the observed target is equidistant to all protected groups and for target sets if all samples in the observed set of targets are unbiased. This is important, because even if a set of targets is mostly unbiased or target biases cancel out on average, individual biases can still be harmful and should thus be detected. The requirement for the consistency of the minimal bias score b_0 can be formulated in a straight forward way using the similarities to the attribute sets A_i .

Definition 3.4 (Unbiased-Trustworthy). Let b_0 be the bias score of a bias score function, that is equivalent to no bias being measured. We call the bias score function $b(\mathbf{t}, A)$ Unbiased-Trustworthy if

$$b(\mathbf{t}, A) = b_0 \iff s(\mathbf{t}, A_i) = s(\mathbf{t}, A_j) \forall A_i, A_j \in A. \quad (11)$$

Analogously for aggregated scores with a set $T = \{\mathbf{t}_1, \dots, \mathbf{t}_m\}$, we say $b(T, A)$ is Unbiased-Trustworthy if

$$b(T, A) = b_0 \quad (12)$$

$$\iff s(\mathbf{t}_k, A_i) = s(\mathbf{t}_k, A_j) \quad \forall A_i, A_j \in A, k \in \{1, \dots, m\}. \quad (13)$$

4 ANALYSIS OF BIAS SCORES

As a major contribution of this work, we formally analyze WEAT and the Direct Bias with regard to the properties defined in Section 3.2. Table 1 gives an overview over the properties. The detailed analyses follow in Section 4.1 for WEAT and Section 4.2 for the Direct Bias.

4.1 Analysis of WEAT

In the following, we detail properties of WEAT in light of the definitions stated above. First, we focus on the individual biases as reported by $s(\mathbf{t}, A, B)$.

Theorem 1. *The bias score function $s(\mathbf{t}, A, B)$ of WEAT is not Magnitude-Comparable.*

Proof. With $\hat{\mathbf{a}} = \frac{1}{|\mathbf{a}|} \sum_{\mathbf{a} \in A} \frac{\mathbf{a}}{\|\mathbf{a}\|}$ and $\hat{\mathbf{b}}$ analogously defined, we can rewrite

$$s(\mathbf{t}, A, B) = \frac{\mathbf{t} \cdot \hat{\mathbf{a}}}{\|\mathbf{t}\|} - \frac{\mathbf{t} \cdot \hat{\mathbf{b}}}{\|\mathbf{t}\|} \quad (14)$$

$$= \frac{\mathbf{t}}{\|\mathbf{t}\|} \cdot (\hat{\mathbf{a}} - \hat{\mathbf{b}}) \quad (15)$$

$$= \cos(\mathbf{t}, \hat{\mathbf{a}} - \hat{\mathbf{b}}) \|\hat{\mathbf{a}} - \hat{\mathbf{b}}\|. \quad (16)$$

Hence we can show that the extreme values depend on the attribute sets A and B :

$$\max_{\mathbf{t}} s(\mathbf{t}, A, B) = \|\hat{\mathbf{a}} - \hat{\mathbf{b}}\|, \quad (17)$$

$$\min_{\mathbf{t}} s(\mathbf{t}, A, B) = -\|\hat{\mathbf{a}} - \hat{\mathbf{b}}\| \quad (18)$$

The statement follows. \square

Theorem 2. *The bias score function $s(\mathbf{t}, A, B)$ of WEAT is Unbiased-Trustworthy.*

Table 1: Overview over the properties of bias scores.

bias score	comparable	trustworthy
$WEAT_{sample}$	x	✓
$WEAT$	✓	x
$DirectBias$	✓	x

Proof. This follows directly from the definition of $s(\mathbf{t}, A, B)$ (equation (1)):

$$s(\mathbf{t}, A, B) = s(\mathbf{t}, A) - s(\mathbf{t}, B) = 0 \quad (19)$$

$$\iff s(\mathbf{t}, A) = s(\mathbf{t}, B) \quad (20)$$

\square

Next, we focus on the properties of the effect size $d(X, Y, A, B)$, identified by WEAT in Table 1. Note that it is not specified for cases, where $s(\mathbf{t}, A, B) = s(\mathbf{t}', A, B) \quad \forall \mathbf{t}, \mathbf{t}' \in X \cup Y$ due to its denominator. This is highly problematic considering Definition 3.4, which states that a bias score should be 0 in that specific case. For Theorem 4 we need Lemma 1 from the Appendix.

Theorem 3. *The effect size $d(X, Y, A, B)$ of WEAT is not Unbiased-Trustworthy.*

Proof. For the WEAT score $b_0 = 0$. With four targets $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4$ and $s(\mathbf{t}_1, A, B) = s(\mathbf{t}_3, A, B)$ and $s(\mathbf{t}_2, A, B) = s(\mathbf{t}_4, A, B)$ the effect size

$$d(\{\mathbf{t}_1, \mathbf{t}_2\}, \{\mathbf{t}_3, \mathbf{t}_4\}, A, B) = \quad (21)$$

$$\frac{(s(\mathbf{t}_1, A, B) + s(\mathbf{t}_2, A, B)) - (s(\mathbf{t}_3, A, B) + s(\mathbf{t}_4, A, B))}{2 \cdot \text{stddev}_{\mathbf{t} \in \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4\}} s(\mathbf{t}, A, B)} \quad (22)$$

is 0, if $s(\mathbf{t}_1, A, B) \neq s(\mathbf{t}_2, A, B)$ (otherwise d is not defined). Now, for the simple case $A = \{\mathbf{a}\}, B = \{\mathbf{b}\}$ and assuming all vectors having length 1, we see

$$s(\mathbf{t}_1, A, B) = s(\mathbf{t}_3, A, B)$$

$$\iff \mathbf{a} \cdot \mathbf{t}_1 - \mathbf{b} \cdot \mathbf{t}_1 = \mathbf{a} \cdot \mathbf{t}_3 - \mathbf{b} \cdot \mathbf{t}_3$$

$$\iff \mathbf{a} \cdot (\mathbf{t}_1 - \mathbf{t}_3) - \mathbf{b} \cdot (\mathbf{t}_1 - \mathbf{t}_3) = 0$$

$$\iff (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{t}_1 - \mathbf{t}_3) = 0. \quad (23)$$

This implies that, if the two vectors $\mathbf{a} - \mathbf{b}$ and $\mathbf{t}_1 - \mathbf{t}_3$ are orthogonal (and e.g. $s(\mathbf{t}_2, A, B) = 0$), the WEAT score returns 0. In this case, there exist $\mathbf{a}, \mathbf{b}, \mathbf{t}_1, \mathbf{t}_3$ with $s(\mathbf{t}_1, A, B) = s(\mathbf{t}_3, A, B) \neq 0$ and accordingly $s(\mathbf{t}_1, A) \neq s(\mathbf{t}_1, B)$. \square

Theorem 4. *The effect size $d(X, Y, A, B)$ of WEAT with $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}, Y = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ is Magnitude-Comparable.*

Proof. With $c_i = s(\mathbf{x}_i, A, B)$, $c_{i+m} = s(\mathbf{y}_i, A, B)$, $n = 2m$, $\hat{\mu} = 1/n \sum_{i=1}^n c_i$ and $\hat{\sigma} = \sqrt{1/n \sum_{i=1}^n (c_i - \hat{\mu})^2}$, we have

$$d = \frac{1/m \sum_{i=1}^m c_i - 1/m \sum_{i=m+1}^{2m} c_i}{\hat{\sigma}} \quad (24)$$

$$= \frac{\sum_{i=1}^m c_i - \sum_{i=m+1}^{2m} c_i + \sum_{i=1}^m c_i - \sum_{i=1}^m c_i}{m \hat{\sigma}}$$

$$= \frac{2 \sum_{i=1}^m c_i - 2m \hat{\mu}}{m \hat{\sigma}}$$

$$= \frac{2}{m} \sum_{i=1}^m \frac{c_i - \hat{\mu}}{\hat{\sigma}} \in [-2, 2] \quad (25)$$

where the last statement follows from Lemma 1 (see Appendix) with $\sum_{i=1}^m \frac{c_i - \mu}{\sigma} \in [-m, m]$. The extreme value ± 2 is reached if $c_1 = \dots = c_m = -c_{m+1} = \dots = -c_{2m}$, which can be obtained by setting $\mathbf{x}_1 = \dots = \mathbf{x}_m = -\mathbf{y}_1 = \dots = -\mathbf{y}_m$, independently of A and B as long as $A \neq B$ and $\sum_{\mathbf{a}_i \in A} \mathbf{a}_i / \|\mathbf{a}_i\| \neq 0 \neq \sum_{\mathbf{b}_i \in B} \mathbf{b}_i / \|\mathbf{b}_i\|$. \square

The proof of Theorem 4 shows that the effect size reaches its extreme values only if all $x \in X$ achieve the same similarity score $s(\mathbf{x}, A, B)$ and $s(\mathbf{y}, A, B) = -s(\mathbf{x}, A, B) \forall \mathbf{y} \in Y$, i.e. the smaller the variance of $s(\mathbf{x}, A, B)$ and $s(\mathbf{y}, A, B)$ the higher the effect size. This implies that we can influence the effect size by changing the variance of $s(\mathbf{t}, A, B)$ without changing whether the groups are separable in the embedding space. Furthermore, the proof of Theorem 3 shows that WEAT can report no bias even if the embeddings contain associations with the bias attributes. This problem occurs, because WEAT is only sensitive to the stereotype "X is associated with A and Y is associated with B" and will overlook biases diverting from this hypothesis.

4.2 Analysis of the Direct Bias

For the Direct Bias, the following theorems show that it is Magnitude-Comparable, but not Unbiased-Trustworthy. The proof for Theorem 6 shows that the first principal component used by the Direct Bias does not necessarily represent individual bias directions appropriately. This can lead to both over- and underestimation of bias by the Direct Bias.

Theorem 5. *The DirectBias is Magnitude-Comparable for $c \geq 0$.*

Proof. For $c \geq 0$ the individual bias $|\cos(\mathbf{t}, \mathbf{g})|^c$ is in $[0, 1]$. Calculating the mean over all targets in T does not change this bound. The statement follows. \square

Theorem 6. *The DirectBias is not Unbiased-Trustworthy.*

Proof. For the Direct Bias $b_0 = 0$ indicates no bias. Consider a setup with two attribute sets $A = \{\mathbf{a}_1, \mathbf{a}_2\}$ and $C = \{\mathbf{c}_1, \mathbf{c}_2\}$.

Using the notation from Section 2.2 this gives us two defining sets $D_1 = \{\mathbf{a}_1, \mathbf{c}_1\}$ $D_2 = \{\mathbf{a}_2, \mathbf{c}_2\}$. Let $\mathbf{a}_1 = (-x, rx)^T = -\mathbf{c}_1$, $\mathbf{a}_2 = (-x, -rx)^T = -\mathbf{c}_2$ and $r > 1$. The bias direction is obtained by computing the first principal component over all $(\mathbf{a}_i - \mu_i)$ and $(\mathbf{c}_i - \mu_i)$ with $\mu_i = \frac{\mathbf{a}_i + \mathbf{c}_i}{2} = 0$. Due to $r > 1$, $\mathbf{b} = (0, 1)^T$ is a valid solution for the 1st principal component as it maximizes the variance

$$\mathbf{b} = \operatorname{argmax}_{\|\mathbf{v}\|=1} \sum_i (\mathbf{v} \cdot \mathbf{a}_i)^2 + (\mathbf{v} \cdot \mathbf{c}_i)^2. \quad (26)$$

According to the definition in Section 3.1, any word $\mathbf{t} = (0, w_y)^T$ would be considered neutral to groups A and C with $s(\mathbf{t}, A) = s(\mathbf{t}, C)$ and being equidistant to each word pair $\{a_i, c_i\}$.

But with the bias direction $\mathbf{b} = (0, 1)^T$ the Direct Bias would report $b_{max} = 1$ instead of $b_0 = 0$, which contradicts Definition 3.4.

On the other hand, we would consider a word $\mathbf{t} = (w_x, 0)^T$ maximally biased, but the Direct Bias would report no bias. Showing that the bias reported for single words \mathbf{t} is not Unbiased-Trustworthy, proves that the *DirectBias* is not Unbiased-Trustworthy. \square

5 EXPERIMENTS

In the experiments, we show that the limitations of WEAT and Direct Bias shown in Section 4 do occur with state-of-the-art language models. We show that the effect size of WEAT can be misleading when comparing bias in different settings. Furthermore, we highlight how attribute embeddings differ between different models, which impacts WEAT's individual bias, and that the Direct Bias can obtain a misleading bias direction by using the Principal Component Analysis (PCA). We use different pretrained language models from Huggingface (Wolf et al., 2019) and the PCA implementation from Scikit-learn (Pedregosa et al., 2011) and observe gender bias based on 25 attributes per gender, such as (**man, woman**).

5.1 Weat's Effect Size

In a first experiment, we demonstrate that the effect size does not quantify social bias in terms of the separability of stereotypical targets. We use embeddings of *distilbert-base-uncased* and *openai-gpt* to compute gender bias according to $s(\mathbf{t}, A, B)$ and the effect size $d(X, Y, A, B)$ for stereotypically male/female job titles. Figure 1 shows the distribution of $s(\mathbf{t}, A, B)$ for DistilBERT, where stereotypical male/female targets are clearly distinct based on the sample bias. Figure 2 shows the distribution of $s(\mathbf{t}, A, B)$ for GPT, where stereotypical male and female terms are similarly distributed. First, we focus on the DistilBERT model (Figure 1), which clearly is biased with regard to the tested words. We compare two cases with different targets, such that the stereotypical target groups are better separable in one case (left plot), which one may describe as more severe or more obvious bias compared to the second case, where the target groups are almost separable (right plot). However, the effect size behaves contrary to this.

Despite this, when comparing Figures 1 and 2 ,

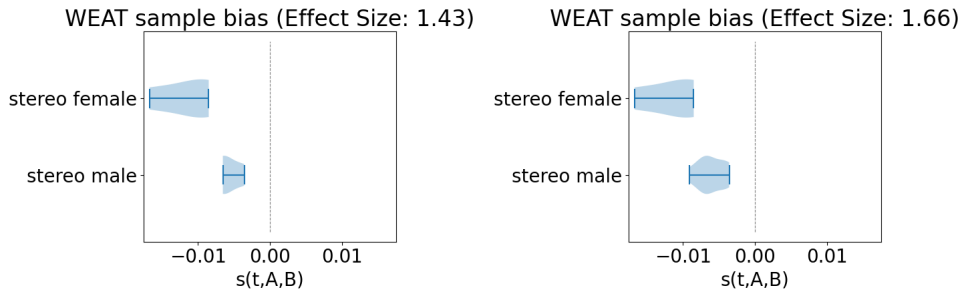


Figure 1: WEAT individual bias and effect size for distilBERT with different selections of target words. When selecting a smaller number of job titles (left), we observe that stereotypical male/female jobs are more distinct w.r.t. $s(t, A, B)$, while the effect size is lower.

one can assume that large differences in effect sizes still reveal significant differences in social bias. While high effect sizes are reported in cases where both stereotypical groups are (almost) separable, we report low effect sizes when groups achieve similar individual biases. Furthermore, one should always report the p-value jointly with the effect size to get an impression on its significance. In other terms, WEAT is useful for qualitative bias analysis (or confirming biases), but not quantitatively.

5.2 Weat’s Individual Bias

In Theorem 1 we discussed that WEAT’s individual bias depends on the mean difference of attributes $\|\hat{\mathbf{a}} - \hat{\mathbf{b}}\|$. As shown in Table 2 these vary a lot between different language models. We report the two most extreme values of 0.198 for *distilroberta-base* and 5.206 for *xlnet-base-uncased*. With such differences we cannot compare sample biases based on their magnitude between different models.

5.3 Direct Bias

Figure 3 shows the correlation of different bias directions on *bert-base-uncased* embeddings. We report bias directions of individual word pairs such as (**man**, **woman**) (left plot: 0-24, right plot: 0-22) and the resulting bias direction as obtained by PCA (last row). Overall we report very low correlations be-

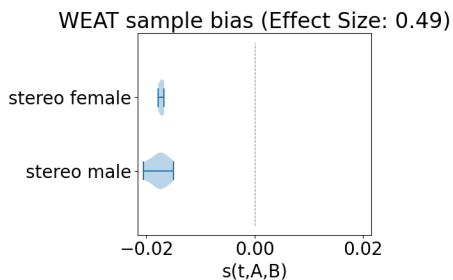


Figure 2: WEAT’s individual bias for job titles in GPT.

tween the individual bias directions. The first principal component reflects mostly individual bias direction 23 and 24 (left plot), which differ a lot from all other bias directions. On contrary, if we excluded word pairs 23 and 24 from the PCA (right plot), the first principal component would give a better estimate of bias directions 0-22. This shows that only one or few “outlier” pairs are sufficient to make the Direct Bias measure “bias” in a completely different way. From a practical point of view, by analyzing the correlation between individual bias directions we can get a good estimate whether the first principal component is a good estimate. Moreover, if we observe only weak correlations between bias directions from the selected word pairs, that is an indication that a 1-dimensional bias direction may not be sufficient to capture the relationship of sensitive groups with regard to which we want to measure bias. While Bolukbasi et al. (Bolukbasi et al., 2016) did not explicitly define that case for the Direct Bias, they proposed to use a bias subspace, defined by the k first principal components, for their Debiasing algorithm, which is related to the Direct Bias. Accordingly, this could be applied to the Direct Bias. Apart from that, one should verify how well the bias direction or bias subspace obtained by PCA

Table 2: Mean attribute difference $\|\hat{\mathbf{a}} - \hat{\mathbf{b}}\|$ for different language models given 25 attribute pairs for gender.

Model Name	Mean Attribute Diff
openai-gpt	0.728
gpt2	0.842
bert-large-uncased	1.123
bert-base-uncased	0.568
distilbert-base-uncased	0.433
roberta-base	0.235
distilroberta-base	0.198
electra-base-generator	0.518
albert-base-v2	1.123
xlnet-base-cased	5.206

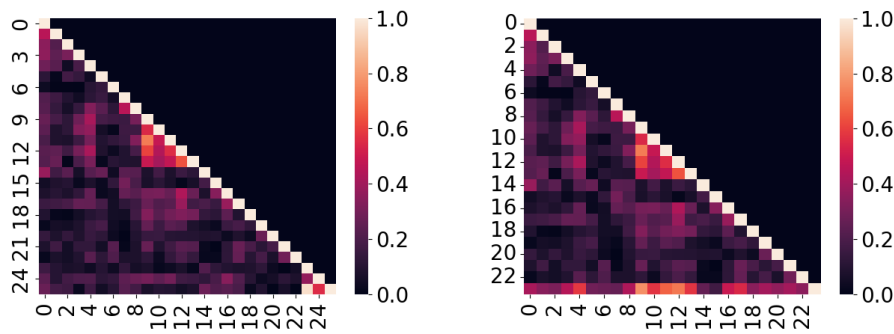


Figure 3: Correlation bias directions of individual word pairs (left: 0-24, right: 0-22) and the first principal component (last row) as selected for the Direct Bias. The lowest row in the heatmap shows the correlation of individual bias directions with the first principal component.

represents individual bias directions to make sure that they're actually measuring bias in the assumed way.

6 CONCLUSION

In this work, we introduce formal properties for cosine based bias scores, concerning their meaningfulness for quantification of social bias. We show that WEAT and the Direct Bias have theoretical flaws that limits their ability to quantify bias. Furthermore, we show that these issues have a real impact when applying these bias scores on state-of-the-art language models. These findings should be considered in the experimental design when evaluating social bias with one of these measures. Future works could build on the proposed properties to analyze other scores from the literature or propose a score that is better suited for bias quantification. The findings of our theoretical analysis open the question, whether the limitations of cosine based scores reported in the literature are due to the theoretical flaws of distinct scores, which are highlighted by our analysis, rather than limitations of geometrical properties as a sign of bias. This is an important question that should be addressed in future work. In general, we encourage other researchers to take an effort to bring the various bias measures from the literature into context and to highlight their properties and limitations, which is critical to derive best practices for bias detection and quantification.

ACKNOWLEDGEMENTS

Funded by the Ministry of Culture and Science of North-Rhine-Westphalia in the frame of the project SAIL, NW21-059A.

REFERENCES

- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.
- Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Delobelle, P., Tokpo, E. K., Calders, T., and Berendt, B. (2021). Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.
- Du, Y., Fang, Q., and Nguyen, D. (2021). Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10012–10034, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.
- Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., and Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.
- Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862.
- Kaneko, M., Bollegala, D., and Okazaki, N. (2022). Debiasing isn't enough!—on the effectiveness of debiasing mlms and their social biases in downstream tasks. *arXiv preprint arXiv:2210.02938*.
- Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.
- May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. *CoRR*, abs/1903.10561.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Schröder, S., Schulz, A., Kenneweg, P., and Hammer, B. (2023). So can we use intrinsic bias measures or not? In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods*.
- Seshadri, P., Pezeshkpour, P., and Singh, S. (2022). Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.
- Steed, R., Panda, S., Kobren, A., and Wick, M. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

APPENDIX

In order to show that the effect size of WEAT is Magnitude-Comparable (see Theorem 4), we need the following lemma.

Lemma 1. *Let $x_1, \dots, x_n \in \mathcal{R}$ be real numbers. Let $\hat{\mu}, \hat{\sigma}$ denote the empirical estimate of mean and standard deviation of the x_i . Then, for any selection of indices i_1, \dots, i_m , with $i_j \neq i_k$ for $j \neq k$, the following bound holds*

$$\left| \frac{\sum_{j=1}^m x_{i_j} - \hat{\mu}}{\hat{\sigma}} \right| \leq \sqrt{m \cdot (n-m)}.$$

Furthermore, for $0 < m < n$ the bound is obtained if and only if all selected resp. non-selected x_i have the same value, i.e. $x_{i_j} = \hat{\mu} + s\sqrt{\frac{n-m}{m}}\hat{\sigma}$ and all other $x_k = \hat{\mu} - s\sqrt{\frac{m}{n-m}}\hat{\sigma}$ with $s \in \{-1, 1\}$.

Proof. For cases $m = 0$ or $m = n$ the statement is trivial. So assume $0 < m < n$. Let $f(x) = ax + b$ be an affine function. Then the images of x_i under f have mean $a\hat{\mu} + b$ and standard deviation $|a|\hat{\sigma}$. On the other hand, we have

$$\frac{f(x_i) - (a\hat{\mu} + b)}{|a|\hat{\sigma}} = \frac{(ax_i + b) - (a\hat{\mu} + b)}{|a|\hat{\sigma}} = \text{sgn}(a) \frac{x_i - \hat{\mu}}{\hat{\sigma}}.$$

Thus, applying f does not change the bound and therefore we may reduce to case of $\hat{\mu} = 0$ and $\hat{\sigma} = 1$.

This allows us to rephrase the problem of finding the maximal bound as an quadratic optimization problem:

$$\begin{aligned} \min \quad & \mathbf{s}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^\top \mathbf{x} = n \\ & \mathbf{1}^\top \mathbf{x} = 0, \end{aligned}$$

where $\mathbf{s} = (1, \dots, 1, 0, \dots, 0)^\top$, $\mathbf{x} = (x_1, \dots, x_n)^\top$ and $\mathbf{1}$ denotes the vector consisting of ones only. Notice, that we assumed w.l.o.g. that $i_1, \dots, i_m = 1, \dots, m$. Furthermore, we made use of the symmetry properties to replace $\max |\mathbf{s}^\top \mathbf{x}|$ by the minimizing statement above, $\hat{\mu} = 0$ is expressed by the last and $\hat{\sigma} = 1$ by the first constrained (recall that $\hat{\sigma} = \sqrt{1/n\mathbf{x}^\top \mathbf{x} - \hat{\mu}^2}$). Notice, that $\nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{x} - n = 2\mathbf{x}$ and $\nabla_{\mathbf{x}} \mathbf{1}^\top \mathbf{x} = \mathbf{1}$ are linear dependent if and only if $\mathbf{x} = a\mathbf{1}$ for some $a \in \mathcal{R}$, thus, as $0 = a\mathbf{1}^\top \mathbf{1} = an$ if and only if $a = 0$ and $(\mathbf{0}\mathbf{1})^\top (\mathbf{0}\mathbf{1}) = 0$, there is no feasible \mathbf{x} for which the KKT-conditions do not hold and we may therefore use them to determine all the optimal points.

The Lagrangian of the problem above and its first two derivatives are given by

$$\begin{aligned} L(\mathbf{x}, \lambda_1, \lambda_2) &= \mathbf{s}^\top \mathbf{x} - \lambda_1 (\mathbf{x}^\top \mathbf{x} - n) - \lambda_2 \mathbf{1}^\top \mathbf{x} \\ \nabla_{\mathbf{x}} L(\mathbf{x}, \lambda_1, \lambda_2) &= \mathbf{s} - 2\lambda_1 \mathbf{x} - \lambda_2 \mathbf{1} \\ \nabla_{\lambda_1, \lambda_2}^2 L(\mathbf{x}, \lambda_1, \lambda_2) &= -2\lambda_1 I. \end{aligned}$$

We can write $\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda_1, \lambda_2) = 0$ as the following linear equation system:

$$\begin{bmatrix} 2x_1 & 1 \\ 2x_2 & 1 \\ \vdots & \vdots \\ 2x_n & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 0 \end{bmatrix}}_{=\mathbf{s}}.$$

Subtracting the first row from row 2, ..., m and row m + 1 from row m + 2, ..., n we see that $2(x_k - x_1)\lambda_1 = 0$ for $k = 1, \dots, m$ and $2(x_k - x_{m+1})\lambda_1 = 0$ for $k = m + 2, \dots, n$, which either implies $\lambda_1 = 0$ or $x_1 = x_2 = \dots = x_m$ and $x_{m+1} = x_{m+2} = \dots = x_n$. However, assuming $\lambda_1 = 0$ would imply that $\lambda_2 = 1$ from the first row and $\lambda_2 = 0$ from the m + 1th row, which is a contradiction. Thus, we have $x_1 = x_2 = \dots = x_m$ and $x_{m+1} = x_{m+2} = \dots = x_n$. But the second constraint from the optimization problem can then only be fulfilled if $mx_1 + (n - m)x_{m+1} = 0$ and this implies $x_{m+1} = -\frac{m}{n-m}x_1$. In this case the first constraint is equal to $n = mx_1^2 + (n - m)\left(-\frac{m}{n-m}x_1\right)^2$, which has the solution $x_1 = \pm\sqrt{\frac{n-m}{m}}$.

Set $\mathbf{x}^* = \left(-\sqrt{\frac{n-m}{m}}, \dots, -\sqrt{\frac{n-m}{m}}, \sqrt{\frac{m}{n-m}}, \dots, \sqrt{\frac{m}{n-m}}\right)$. Then \mathbf{x}^* and $-\mathbf{x}^*$ are the only possible KKT points

as we have just seen. Plugging \mathbf{x}^* into the equation system above and solving for $\lambda_{1/2}$ we obtain

$$\lambda_1^* = -\frac{1}{2\left(\sqrt{\frac{m}{n-m}} + \sqrt{\frac{n-m}{m}}\right)}, \quad \lambda_2^* = \frac{\sqrt{\frac{m}{n-m}}}{\sqrt{\frac{m}{n-m}} + \sqrt{\frac{n-m}{m}}}$$

Now, as $\nabla_{\mathbf{x},\mathbf{x}}^2 L(\mathbf{x}^*, \lambda_1^*, \lambda_2^*) = \left(\sqrt{\frac{m}{n-m}} + \sqrt{\frac{n-m}{m}}\right)^{-1} I$ is positive definite, we see that \mathbf{x}^* is a global optimum, indeed. The statement follows. \square

