Measuring Biases of Word Embeddings: What Similarity Measures and Descriptive Statistics to Use?

Hossein Azarpanah and **Mohsen Farhadloo**
John Molson School of Business
Concordia University
Montreal, QC, CA
(hossein.azarpanah, mohsen.farhadloo)@concordia.ca

Abstract

Word embeddings are widely used in Natural Language Processing (NLP) for a vast range of applications. However, it has been consistently proven that these embeddings reflect the same human biases that exist in the data used to train them. Most of the introduced bias indicators to reveal word embeddings' bias are average-based indicators based on the cosine similarity measure. In this study, we examine the impacts of different similarity measures as well as other descriptive techniques than averaging in measuring the biases of contextual and non-contextual word embeddings. We show that the extent of revealed biases in word embeddings depends on the descriptive statistics and similarity measures used to measure the bias. We found that over the ten categories of word embedding association tests, Mahalanobis distance reveals the smallest bias, and Euclidean distance reveals the largest bias in word embeddings. In addition, the contextual models reveal less severe biases than the non-contextual word embedding models with GPT showing the fewest number of WEAT biases.

1 Introduction

Word embedding models including *Word2Vec* (Mikolov et al., 2013), *GloVe* (Pennington et al., 2014), *BERT* (Devlin et al., 2018), *ELMo* (Peters et al., 2018), and *GPT* (Radford et al., 2018) have become popular components of many NLP frameworks and are vastly used for many downstream tasks. However, these word representations preserve not only statistical properties of human language but also the human-like biases that exist in the data used to train them (Bolukbasi et al., 2016; Caliskan et al., 2017; Kurita et al., 2019; Basta et al., 2019; Gonen and Goldberg, 2019). It has also been shown that such biases propagate to the downstream NLP tasks and have negative impacts on their performance (May et al., 2019; Leino et al., 2018). There are studies investigating how to mitigate biases of word embeddings (Liang et al., 2020; Ravfogel et al., 2020).

Different approaches have been used to present and quantify corpus-level biases of word embeddings. Bolukbasi et al. (2016) proposed to measure the gender bias of word representations in *Word2Vec* and *GloVe* by calculating the projections into principal components of differences of embeddings of a list of male and female pairs. Basta et al. (2019) adapted the idea of "gender direction" of (Bolukbasi et al., 2016) to be applicable to contextual word embeddings such as *ELMo*. In (Basta et al., 2019) first, the gender subspace of *ELMo* vector representations is calculated and then, the presence of gender bias in *ELMo* is identified. Gonen and Goldberg (2019) introduced a new gender bias indicator based on the percentage of socially-biased terms among the k-nearest neighbors of a target term and demonstrated its correlation with the gender direction indicator.

Caliskan et al. (2017) developed Word Embedding Association Test (WEAT) to measure bias by comparing two sets of target words with two sets of attribute words and documented that *Word2Vec* and *GloVe* contain human-like biases such as gender and racial biases. May et al. (2019) generalized the WEAT test to phrases and sentences by inserting individual words from WEAT tests into simple sentence templates and used them for contextual word embeddings.

Kurita et al. (2019) proposed a new method to quantify bias in *BERT* embeddings based on its masked language model objective using simple template sentences. For each attribute word, using a simple template sentence, the normalized probability that *BERT* assigns to that sentence for each of the target words is calculated, and the difference is considered the measure of the bias. Kurita et al. (2019) demonstrated that this probability-based method for quantifying bias in *BERT* was more effective than the cosine-based method.

Proceedings of the First Workshop on Trustworthy Natural Language Processing, pages 8–14

Motivated by these recent studies, we comprehensively investigate different methods for bias exposure in word embeddings. Particularly, we investigate the impacts of different similarity measures and descriptive statistics to demonstrate the degree of associations between the target sets and attribute sets in the WEAT. First, other than cosine similarity, we study Euclidean, Manhattan, and Mahalanobis distances to measure the degree of association between a single target word and a single attribute word. Second, other than averaging, we investigate minimum, maximum, median, and a discrete (grid-based) optimization approach to find the minimum possible association to report between a single target word and the two attribute sets in each of the WEAT tests. We consistently compare these bias measures for different types of word embeddings including non-contextual (*Word2Vec*, *GloVe*) and contextual ones (*BERT*, *ELMo*, *GPT, GPT2*).

2 Method

Implicit Association Test (IAT) was first introduced by Greenwald et al. (1998a) in psychology to demonstrate the enormous differences in response time when participants are asked to pair two concepts they deem similar, in contrast to two concepts they find less similar. For example, when subjects are encouraged to work as quickly as possible, they are much likely to label flowers as pleasant and insects as unpleasant. In IAT, being able to pair a concept to an attribute quickly indicates that the concept and attribute are linked together in the participants' minds. The IAT has widely been used to measure and quantify the strength of a range of implicit biases and other phenomena, including attitudes and stereotype threat (Karpinski and Hilton, 2001; Kiefer and Sekaquaptewa, 2007; Stanley et al., 2011).

Inspired by IAT, Caliskan et al. (2017) introduced WEAT to measure the associations between two sets of target concepts and two sets of attributes in word embeddings learned from large text corpora. A hypothesis test is conducted to demonstrate and quantify the bias. The null hypothesis states that there is no difference between the two sets of target words in terms of their relative distance/similarity to the two sets of attribute words. A permutation test is performed to measure the null hypothesis's likelihood. This test computes the probability that target words' random permutations would produce a greater difference than the observed difference. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two sets of target word embeddings and $\boldsymbol{A}$ and $\boldsymbol{B}$ be two sets of attribute embeddings. The test statistics is defined as:

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

where:

$$s(w, \boldsymbol{A}, \boldsymbol{B}) = f_{a \in \boldsymbol{A}}(s(\overrightarrow{\mathrm{w}}, \overrightarrow{\mathrm{a}})) - f_{b \in \boldsymbol{B}}(s(\overrightarrow{\mathrm{w}}, \overrightarrow{\mathrm{b}})) \quad (1)$$

In other words, $s(w, \boldsymbol{A}, \boldsymbol{B})$ quantifies the association of a single word w with the two sets of attributes, and $s(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B})$ measures the differential association of the two sets of targets with the two sets of attributes. Denoting all the partitions of $\boldsymbol{X} \cup \boldsymbol{Y}$ with $(\boldsymbol{X_i}, \boldsymbol{Y_i})_i$, the one-sided p-value of the permutation test is:

$$Pr_i(s(\boldsymbol{X_i}, \boldsymbol{Y_i}, \boldsymbol{A}, \boldsymbol{B}) > s(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}))$$

The magnitude of the association of the two target sets with the two attribute sets can be measured with the effect size as:

$$d = \frac{|\overline{s(x, \boldsymbol{A}, \boldsymbol{B})} - \overline{s(y, \boldsymbol{A}, \boldsymbol{B})}|}{\text{std-dev}_{w \in \boldsymbol{X} \cup \boldsymbol{Y}} s(w, \boldsymbol{A}, \boldsymbol{B})}$$

It is worth mentioning that d is a measure used to calculate how separated two distributions are and is basically the standardized difference of the means of the two distributions (Cohen, 2013). Controlling for the significance, a larger effect size reflects a more severe bias.

WEAT and almost all the other studies inspired by it (Garg et al., 2018; Brunet et al., 2018; Gonen and Goldberg, 2019; May et al., 2019) use the following approach to measure the association of a single target word with the two sets of attributes (equation 1). First, they use cosine similarity to measure the target word's similarity to each word in the attribute sets. Then they calculate the average of the similarities over each attribute set.

In this paper we investigate the impacts of other functions such as $min(\cdot)$, $mean(\cdot)$, $median(\cdot)$, or $max(\cdot)$ for function $f(\cdot)$ in equation (1) (originally only $mean(\cdot)$ has been used). Also, in this paper in addition to cosine similarity, we consider Euclidean and Manhattan distances as well as the following measures for the $s(\overrightarrow{\mathrm{w}}, \overrightarrow{\mathrm{a}})$ in equation (1).

Mahalanobis distance: introduced by P. C. Mahalanobis (Mahalanobis, 1936) this distance measures the distance of a point from a distribution: $s(\overrightarrow{\mathrm{w}}, \overrightarrow{\mathrm{a}}) = ((\overrightarrow{\mathrm{w}} - \overrightarrow{\mathrm{a}})^T \Sigma_A^{-1} (\overrightarrow{\mathrm{w}} - \overrightarrow{\mathrm{a}}))^{\frac{1}{2}}$. It is

worth noting that the Mahalanobis distance takes into account the distribution of the set of attributes while measuring the association of the target word w with an attribute vector.

Discrete optimization of the association measure: In equation (1), $s(w, \boldsymbol{A}, \boldsymbol{B})$ quantifies the association of a single target word w with the two sets of attributes. To quantify the minimum possible association of a target word w with the two sets of attributes, we first calculate the distance of w from all attribute words in $\boldsymbol{A}$ and $\boldsymbol{B}$, then calculate all possible differences and find the minimum difference.

$$s(w, \boldsymbol{A}, \boldsymbol{B}) = \min_{a \in \boldsymbol{A}, b \in \boldsymbol{B}} |s(\vec{\mathrm{w}}, \vec{\mathrm{a}}) - s(\vec{\mathrm{w}}, \vec{\mathrm{b}})| \quad (2)$$

3 Biases studied

We studied all ten bias categories introduced in IAT (Greenwald et al., 1998a) and replicated in WEAT to measure the biases in word embeddings. The ten WEAT categories are briefly introduced in Table 1. For more detail and example of target and attribute words, please check Appendix A. Although WEAT 3 to 5 have the same names, they have different target and attribute words.

WEAT	Association
1	Flowers vs insects with pleasant vs unpleasant
2	Instruments vs weapons with pleasant vs unpleasant
3	Eur.-American vs Afr.-American names with Pleasant vs unpleasant (Greenwald et al., 1998b)
4	Eur.-American vs Afr.-American names (Bertrand and Mullainathan, 2004) with Pleasant vs unpleasant (Greenwald et al., 1998b)
5	Eur.-American vs Afr.-American names (Bertrand and Mullainathan, 2004) with Pleasant vs unpleasant (Nosek et al., 2002)
6	Male vs female names with Career vs family
7	Math vs arts with male vs female terms
8	Science vs arts with male vs female terms
9	Mental vs physical disease with temporary vs permanent
10	Young vs old people's name with pleasant vs unpleasant

Table 1: The associations studied in the WEAT

As described in section 2, we need each attribute set's covariance matrix to compute Mahalanobis distance. To get stable covariance matrix estimation due to the high dimension of the embeddings we first created larger attribute sets by adding synonym terms. Next, we estimated the sparse covariance matrices as the number of samples in each attribute set is smaller than the number of features. To enforce sparsity, we estimated the $l1$ penalty using k-fold cross validation with k=3.

4 Results of experiments

We examined the 10 different types of biases in WEAT (Table 1) for word embedding models listed in Table 2. We used publicly available pre-trained models. For contextual word embeddings, we used single word sentences as input instead of using simple template sentences used in other studies (May et al., 2019; Kurita et al., 2019). The simple template sentences such as "this is TARGET" or "TARGET is ATTRIBUTE" used in other studies do not really provide any context to reveal the contextual capability of embeddings such as *BERT* or *ELMo*. This way, the comparisons between the contextual embeddings and non-contextual embeddings are fairer as both of them only get the target or attribute terms as input. For each model, we performed the WEAT tests using four similarity metrics mentioned in section 2: *cosine, Euclidean, Manhattan, Mahalanobis*. For each similarity metric, we also used $min(\cdot)$, $mean(\cdot)$, $median(\cdot)$, or $max(\cdot)$ as the $f(\cdot)$ in equation (1). Also, as explained in section 2, we discretely optimized the association measure and found the minimum association in equation (1). In these experiments (Table 3 and Table 4), the larger and more significant effect sizes imply more severe biases.

Model	Embedding	Dim
GloVe (840B tokens, web corpus)	-	300
Word2Vec (GoogleNews-negative)	-	300
ELMo (original)	First hidden layer	1024
BERT (base, cased)	Sum of last 4 hidden layers in [CLS]	768
GPT	Last hidden layer	768
GPT2	Last hidden layer	768

Table 2: Word embedding models, used representations, and their dimensions.

Impacts of different descriptive statistics: Our first goal was to report the changes in the measured biases when we change the descriptive statistics. The range of effect sizes was from 0.00 to 1.89 ($\mu = 0.65$, $\sigma = 0.5$). Our findings show that *mean* has a better capability to reveal biases as it provides the most cases of significant effect sizes ($\mu = 0.8$, $\sigma = 0.52$) across models and distance measures. *Median* is close to the *mean* with ($\mu = 0.74$, $\sigma = 0.48$) among all its effect sizes. The effect sizes for *minimum* ($\mu = 0.68$, $\sigma = 0.48$) and *maximum* ($\mu = 0.65$, $\sigma = 0.48$) are close to each other, but smaller than *mean* and *median*. The discretely optimized association measure (Eq. 2) provides the smallest effect sizes ($\mu = 0.39$, $\sigma = 0.3$) and reveals the least number of implicit biases. These differences as the result of applying different descriptive statistics in the association measure (Eq. (1)) show that the revealed biases depend on the applied statistics to measure the bias.

For example, in the *cosine* distance of *Word2Vec*, if we change the descriptive statistic from *mean* to *minimum*, the biases for WEAT 3 and WEAT 4 will become insignificant (no bias will be reported). In another example, in *GPT* model, while the result of *mean cosine* is not significant for WEAT 3 and WEAT 4, they become significant for *median cosine*. Moreover, almost for all models, the effect size of the discretely optimized minimum distance is not significant. Our intention for considering this statistic was to report the minimum possible association of a target word with the attribute sets. If this measure is used for reporting biases, one can misleadingly claim that there is no bias.

Impacts of different similarity measures: The effect sizes for cosine, Manhattan, and Euclidean are closer to each other and greater than the Mahalanobis distance (cosine: ($\mu = 0.72, \sigma = 0.49$), Euclidean: ($\mu = 0.67, \sigma = 0.5$), Manhattan: ($\mu = 0.63, \sigma = 0.48$), Mahalanobis: ($\mu = 0.58, \sigma = 0.45$)). Mahalanobis distance also detects the fewest number of significant bias types across all models. As an example, while *mean* and *median* effect sizes for WEAT 3 and WEAT 5 in *GloVe* and *Word2Vec* are mostly significant for *cosine, Euclidean, and Manhattan*; the same results are not significant for the *Mahalanobis* distance. That means with the Mahalanobis distance as the measure of the bias, no bias will be reported for WEAT 3 and WEAT 5 tests. This emphasizes the importance of chosen similarity measures in detecting biases of word embeddings. More importantly, as the *Mahalanobis* distance considers the distribution of attributes in measuring the distance, it may be a better choice than the other similarity measures for measuring and revealing biases with GPT showing fewer number of biases.

Biases in different word embedding models: Using any combination of descriptive statistics and similarity measures, all the contextualized models have less significant biases than *GloVe* and *Word2Vec*. In Table 3 the number of tests with significant implicit biases out of the 10 WEAT tests along with the mean and standard deviation of the effect sizes for all embedding models have been reported. The complete list of effect sizes along with their p-value are provided in Table 4.

Following our findings in the previous sections, we choose *mean* of Euclidean to reveal biases. By doing so, *GloVe* and *Word2Vec* show the most number of significant biases with 9 and 7 significant biases in 10 WEAT categories (Table 3). Using *mean of Euclidean*, our results confirm all the results by Caliskan et al. (2017), which used *mean of cosine* in all WEAT tests. The difference is that with the *mean of Euclidean* measure, the biases are revealed as being more severe. (smaller p-values). Using *mean* of Euclidean, *GPT* and *ELMo* show the fewest number of implicit biases. *GPT* model shows bias in WEAT 2, 3, and 5. *ELMo's* significant biases are in WEAT 1, 3, and 6. Using *mean Euclidean*, almost all models (except for *ELMo*) confirm the existence of a bias in WEAT 3 to 5. Moreover, all contextualized models found no bias in associating female with arts and male with science (WEAT 7), mental diseases with temporary attributes and physical diseases with permanent attributes (WEAT 9), and young people's name with pleasant attribute and old people's name with unpleasant attributes (WEAT 10).

Model	*mean cosine*	*mean Euc*	*mean Maha*	*Maha Eq.2*
GloVe	9 (1.39, 0.21)	9 (1.41, 0.2)	3 (0.79, 0.53)	0 (0.34, 0.27)
Word2Vec	7 (1.13, 0.54)	7 (1.13, 0.55)	5 (0.84, 0.52)	0 (0.32, 0.33)
ELMo	3 (0.64, 0.51)	3 (0,65, 0.52)	3 (0.61, 0.42)	0 (0.36, 0.23)
BERT	5 (0.74, 0.5)	5 (0.74, 0.48)	2 (0.47, 0.5)	2 (0.55, 0.52)
GPT	2 (0.61, 0.48)	3 (0.65, 0.42)	4 (0.59, 0.35)	0 (0.29, 0.27)
GPT2	3 (0.73, 0.46)	4 (0.71, 0.46)	3 (0.69, 0.49)	3 (0.66, 0.49)

Table 3: Number of revealed biases out of the 10 WEAT bias types for the studied word embeddings along with the (μ, σ) of their effect sizes. The larger the effect size the more severe the bias.

5 Conclusions

We studied the impacts of different descriptive statistics and similarity measures on association tests for measuring biases in contextualized and non-contextualized word embeddings. Our findings demonstrate that the detected biases depend on the choice of association measure. Based on our experiments, *mean* reveals more severe biases and the discretely optimized version reveals fewer number of severe biases. In addition, *cosine* distance reveals more severe biases and the *Mahalanobis* distance reveals less severe ones. Reporting biases with mean of Euclidean/Mahalanobis distances identifies more/less severe biases in the models. Furthermore, contextual models show less biases than the non-contextual ones across all 10 WEAT tests with GPT showing the fewest number of biases.

		Cosine					Euclidean					Manhattan					Mahalanobis				
Model	**WEAT**	**Mean**	**Median**	**Min**	**Max**	**Eq.2**	**Mean**	**Median**	**Min**	**Max**	**Eq.2**	**Mean**	**Median**	**Min**	**Max**	**Eq.2**	**Mean**	**Median**	**Min**	**Max**	**Eq.2**
GloVe	1	1.50^{****}	1.34^{****}	1.35^{***}	1.41^{****}	0.27	1.52^{****}	1.47^{****}	1.31^{****}	1.23^{****}	0.03	1.50^{****}	1.46^{****}	1.32^{****}	0.90^{**}	0.15	1.53^{****}	1.54^{****}	1.19^{****}	1.61^{****}	0.00
	2	1.53^{****}	1.37^{****}	0.83^{*}	1.57^{****}	0.08	1.53^{****}	1.42^{****}	1.42^{****}	0.03	0.13	1.51^{****}	1.43^{****}	1.44^{****}	0.27	0.24	1.61^{****}	1.63^{****}	1.49^{****}	0.98^{***}	0.28
	3	1.41^{****}	1.13^{****}	1.53^{****}	1.41^{****}	0.60^{*}	1.37^{****}	0.98^{****}	1.51^{****}	0.09	0.31	0.82^{**}	0.37	1.24^{****}	0.69^{*}	0.21	0.57	0.66^{*}	0.37	0.89^{**}	0.13
	4	1.50^{****}	1.02^{*}	1.55^{****}	1.47^{****}	0.17	1.51^{****}	0.40	1.58^{****}	0.32	0.06	0.93^{*}	0.36	1.14^{***}	0.80^{*}	0.05	0.30	0.57	0.04	0.67	0.31
	5	1.28^{***}	1.39^{***}	0.45	1.29^{***}	0.57	1.30^{****}	1.62^{****}	1.13^{**}	0.36	0.61	0.54	1.03^{*}	0.17	0.11	0.37	0.17	0.36	0.01	0.69	0.35
	6	1.81^{***}	1.83^{***}	1.70^{***}	1.67^{***}	1.01	1.80^{****}	1.75^{****}	1.75^{****}	1.56^{***}	0.17	1.78^{****}	1.78^{****}	1.71^{****}	1.46^{**}	0.86	1.17^{*}	0.83	1.27^{*}	0.60	0.43
	7	1.06	0.85	0.61	1.05	0.18	1.10	0.65	0.26	0.70	0.16	0.70	0.03	0.55	0.63	0.50	0.20	0.80	0.02	0.23	0.10
	8	1.24^{*}	0.93	1.29^{*}	1.16^{*}	0.36	1.23^{*}	1.07	1.12	0.92	0.21	1.03	0.81	0.99	0.83	0.13	0.92	0.71	0.86	0.26	0.26
	9	1.38^{*}	0.83	0.37	1.47^{*}	1.03	1.47^{*}	1.04	1.20	1.32^{*}	0.90	1.50^{*}	0.26	1.18	1.42^{*}	0.61	0.99	0.93	1.20	0.55	0.85
	10	1.21^{*}	1.05	1.01	0.75	0.99	1.26^{*}	1.42^{*}	0.84	0.64	0.41	0.70	0.90	0.34	0.46	0.25	0.47	0.83	0.45	0.60	0.71
word2vec	1	1.54^{****}	1.34^{****}	0.55	1.49^{****}	0.16	1.50^{****}	1.30^{****}	1.31^{****}	0.95^{**}	0.31	1.49^{****}	1.34^{****}	1.38^{****}	0.75^{*}	0.26	0.84^{*}	1.06^{***}	0.79^{*}	0.34	0.13
	2	1.63^{****}	1.49^{****}	1.19^{****}	1.60^{****}	0.22	1.58^{****}	1.36^{****}	1.37^{****}	0.68^{*}	0.10	1.44^{****}	1.24^{****}	1.19^{****}	0.70^{*}	0.36	1.39^{****}	0.99^{***}	0.39	0.15	0.05
	3	0.58^{*}	0.46	0.10	0.81^{**}	0.38	0.78^{**}	0.46	0.82^{**}	0.62^{*}	0.19	0.82^{**}	0.56	0.68^{*}	0.63^{*}	0.17	0.24	0.41	0.98^{****}	0.68^{*}	0.19
	4	1.31^{****}	1.21^{***}	0.44	1.27^{****}	0.09	1.49^{****}	0.80^{*}	1.66^{****}	0.60	0.35	1.44^{****}	1.13^{***}	1.37^{***}	0.55	0.86^{*}	0.55	0.16	1.30^{****}	0.49	0.28
	5	0.72	0.68	0.58	0.41	0.19	0.43	0.38	0.41	0.08	0.25	0.27	0.23	0.11	0.05	0.09	0.02	0.61	0.11	0.12	0.24
	6	1.89^{***}	1.87^{***}	1.76^{***}	1.65^{***}	0.91	1.88^{****}	1.88^{****}	1.63^{**}	1.70^{****}	0.85	1.89^{****}	1.87^{****}	1.39^{*}	1.76^{****}	0.39	1.21^{*}	0.24	1.49^{**}	0.29	0.01
	7	0.97	0.98	0.52	0.71	0.67	0.92	0.45	1.11^{*}	1.27^{*}	0.70	1.06	0.87	1.04	1.27^{*}	1.29^{*}	0.97	0.90	0.55	1.35^{*}	0.08
	8	1.24^{*}	1.23^{*}	1.18^{*}	0.99	0.59	1.25^{*}	1.09	1.21^{*}	1.49^{**}	0.60	1.47^{*}	1.36^{*}	1.33^{*}	1.67^{***}	0.00	0.40	0.30	0.48	0.52	0.88
	9	1.30^{*}	0.69	0.14	1.31	0.42	1.32^{*}	1.18	1.07	0.92	0.55	1.08	0.92	0.92	0.46	0.09	1.55^{****}	1.23	0.59	0.41	0.94
	10	0.09	0.01	0.19	0.66	0.76	0.15	0.01	0.39	0.14	0.43	0.24	0.12	0.36	0.34	0.05	1.20^{*}	1.24^{*}	1.60^{***}	0.03	0.44
ELMo	1	1.25^{****}	1.15^{***}	0.77^{*}	0.68^{*}	0.03	1.25^{****}	1.03^{***}	0.51	0.35	0.48	1.24^{****}	1.04^{***}	0.50	0.27	0.19	0.28	0.17	0.28	0.26	0.57
	2	1.46^{****}	1.37^{****}	0.87^{**}	1.37^{****}	0.08	1.46^{****}	1.28^{****}	1.14^{****}	0.71^{*}	0.51	1.50^{****}	1.22^{****}	1.25^{****}	0.75^{*}	0.27	0.67^{*}	0.11	0.79^{*}	0.11	0.15
	3	0.19	0.19	0.06	0.10	0.30	0.12	0.30	0.20	0.06	0.14	0.16	0.02	0.15	0.12	0.19	0.24	0.29	0.37	0.07	0.27
	4	0.29	0.22	0.66	0.44	0.02	0.39	0.07	0.03	0.35	0.19	0.39	0.00	0.02	0.35	0.34	0.33	0.29	0.25	0.08	0.48
	5	0.11	0.01	0.27	0.57	0.43	0.14	0.12	0.46	0.14	0.09	0.03	0.04	0.55	0.20	0.85^{*}	0.40	0.04	0.71	0.52	0.34
	6	1.24^{*}	0.95	0.61	0.10	1.00	1.27^{*}	0.50	0.02	0.44	0.53	0.30	0.59	0.51	0.20	0.49	1.34^{*}	1.10	0.06	0.50	0.22
	7	0.32	0.30	0.56	0.25	0.81	0.29	0.48	0.02	0.62	0.81	0.24	0.25	0.41	0.36	0.03	1.34^{*}	1.49^{**}	0.72	0.95	0.88
	8	0.28	0.42	0.00	0.38	0.29	0.37	0.14	0.14	0.86	0.66	0.64	0.38	0.61	0.99	0.35	0.18	1.06	0.15	0.06	0.15
	9	0.91	0.24	0.67	1.28^{*}	0.68	0.93	0.59	1.04	0.69	0.10	1.06	0.65	0.98	0.77	0.38	0.71	0.55	0.94	0.23	0.38
	10	0.37	0.81	0.53	0.56	0.23	0.33	0.93	0.36	0.13	0.62	0.28	0.74	0.49	0.06	0.62	0.61	0.73	0.26	0.48	0.20
BERT	1	0.00	0.21	0.12	0.05	0.72^{*}	0.01	0.21	0.09	0.11	0.18	0.02	0.32	0.05	0.18	0.46	0.10	0.11	0.24	0.27	0.06
	2	0.62	0.39	0.90^{**}	0.55	0.32	0.63	0.45	0.56	1.02^{**}	0.24	0.58	0.45	0.23	0.79^{*}	0.27	1.31^{****}	1.33^{****}	1.35^{****}	1.25^{****}	1.21^{****}
	3	1.04^{***}	1.02^{***}	0.75^{*}	0.83^{**}	0.58^{*}	1.05^{****}	1.07^{****}	0.77^{**}	0.72^{*}	0.21	1.04^{****}	1.09^{****}	0.82^{**}	0.76^{**}	0.01	0.24	0.30	0.23	0.29	0.17
	4	1.19^{***}	1.19^{***}	1.06^{***}	1.08^{**}	0.26	1.23^{***}	0.97^{*}	1.00^{**}	1.16^{****}	0.65	1.20^{***}	1.07^{**}	0.88^{*}	1.10^{***}	0.13	0.27	0.31	0.44	0.17	0.05
	5	0.94^{*}	0.93^{*}	0.30	0.77^{*}	0.06	0.88^{*}	0.95^{*}	0.58	0.11	0.41	0.85^{*}	0.98^{*}	0.71	0.02	0.45	0.29	0.23	0.04	0.34	0.19
	6	1.36^{*}	1.20^{*}	1.32^{*}	0.13	0.22	1.30^{*}	1.15^{*}	0.20	1.45^{**}	0.96	1.12	0.83	0.16	1.34^{*}	0.82	0.03	0.15	0.24	0.61	0.38
	7	1.14^{*}	0.85	0.75	1.01	0.07	1.18^{*}	0.75	1.03	0.95^{*}	0.40	1.20^{*}	0.90	1.09	1.02^{*}	0.85	0.29	0.09	0.49	0.18	0.77
	8	0.24	0.37	0.11	0.55	0.17	0.24	0.02	0.50	0.73	0.37	0.12	0.14	0.13	0.34	0.04	0.47	0.42	0.61	0.27	0.60
	9	0.02	0.16	0.03	1.04	0.69	0.12	0.32	0.97	0.00	0.25	0.17	0.34	0.72	0.16	0.66	1.48^{*}	1.38^{*}	1.52^{*}	1.54^{*}	1.61^{*}
	10	0.83	0.76	0.89	0.50	1.28^{*}	0.80	0.89	0.40	0.90	0.22	0.91	1.16^{*}	0.57	0.72	0.53	0.24	0.28	0.09	0.54	0.47
GPT	1	0.47	0.29	0.57	0.08	0.24	0.58	0.39	0.10	0.10	0.50	0.57	0.25	0.11	0.01	0.10	0.40	0.00	0.54	0.45	0.13
	2	1.11^{***}	0.99^{**}	0.94^{**}	0.53	0.38	1.15^{****}	0.74^{*}	0.13	0.23	0.01	1.16^{****}	0.82^{*}	0.01	0.16	0.28	0.84^{**}	0.69^{*}	0.06	1.01^{**}	0.05
	3	0.09	0.97^{***}	0.64^{*}	1.24^{****}	0.20	0.70^{*}	0.99^{***}	1.21^{****}	0.27	0.02	0.06	1.05^{****}	1.17^{****}	0.56	0.01	0.69^{*}	0.97^{***}	1.24^{****}	0.90^{***}	0.13
	4	0.33	1.54^{****}	0.88^{*}	1.48^{****}	0.30	0.51	1.51^{****}	1.37^{****}	0.28	0.10	0.31	1.52^{****}	1.33^{****}	0.50	0.29	0.91^{**}	1.36^{****}	1.26^{***}	1.07^{***}	0.42
	5	1.65^{****}	1.40^{****}	1.57^{****}	1.58^{****}	0.69	1.57^{****}	1.14^{****}	1.49^{****}	1.65^{****}	0.26	1.54^{****}	1.23^{****}	1.49^{****}	1.50^{****}	0.57	1.26^{****}	1.23^{****}	0.98^{***}	1.40^{****}	0.06
	6	0.67	0.02	0.75	0.89	0.20	0.50	0.25	0.23	0.66	0.08	0.49	0.04	0.34	0.45	0.09	0.66	0.01	1.14^{*}	0.27	0.19
	7	0.24	0.11	0.02	0.09	0.70	0.20	0.30	0.00	0.15	0.45	0.28	0.16	0.32	0.03	0.18	0.29	0.63	0.22	0.57	0.06
	8	0.10	0.16	0.35	0.13	0.40	0.08	0.10	0.32	0.03	0.93^{*}	0.12	0.14	0.37	0.08	0.47	0.19	0.13	0.22	0.58	0.43
	9	0.70	0.92	1.01	0.01	0.18	0.58	0.63	0.17	0.07	0.42	0.59	0.62	0.40	0.01	0.44	0.50	0.59	0.66	0.03	0.62
	10	0.72	0.39	0.73	0.68	0.67	0.61	0.25	0.55	1.03	0.27	0.52	0.34	0.48	0.76	0.77	0.19	0.17	0.07	0.27	0.81
GPT2	1	0.11	0.06	0.20	0.20	0.19	0.06	0.21	0.14	0.04	0.01	0.08	0.05	0.05	0.07	0.04	0.27	0.27	0.26	0.23	0.28
	2	0.64	0.28	0.50	0.24	0.04	0.51	0.63	0.21	0.61^{*}	0.02	0.47	0.69^{*}	0.18	0.53	0.56	0.44	0.45	0.45	0.41	0.34
	3	1.27^{****}	0.70^{*}	1.07^{****}	1.15^{****}	0.30	1.25^{****}	1.21^{****}	0.29	1.30^{****}	0.48	1.34^{****}	1.24^{****}	0.39	1.12^{****}	0.11	1.25^{****}	1.25^{****}	1.27^{****}	1.22^{****}	1.24^{****}
	4	1.19^{**}	0.64	1.28^{***}	0.83^{*}	0.56	1.17^{***}	1.24^{****}	0.39	1.13^{**}	0.57	1.17^{***}	1.10^{**}	0.28	1.05^{**}	0.44	1.29^{****}	1.29^{****}	1.28^{****}	1.28^{***}	1.31^{****}
	5	1.17^{**}	1.15^{**}	1.31^{***}	1.02^{**}	0.06	1.17^{***}	1.21^{****}	0.92^{*}	1.13^{**}	0.77^{*}	1.14^{**}	1.18^{***}	0.13	1.15^{**}	0.42	1.29^{***}	1.29^{****}	1.28^{****}	1.30^{****}	1.29^{****}
	6	0.79	1.06	0.94	0.11	0.20	0.86	0.90	0.28	0.94	0.54	0.66	0.74	0.19	0.11	0.63	0.94	0.55	0.52	0.89	0.71
	7	0.03	0.49	0.21	0.69	0.17	0.12	0.10	0.24	0.00	0.07	0.23	0.42	0.34	0.01	0.16	0.16	0.17	0.16	0.16	0.14
	8	0.42	1.08	0.79	0.42	0.36	0.44	0.37	0.08	0.59	0.32	0.30	0.78	0.22	0.21	0.23	0.36	0.35	0.36	0.37	0.38
	9	0.57	0.68	0.16	0.17	0.73	0.36	0.41	0.39	0.64	0.24	0.04	0.10	1.19	0.32	0.05	0.06	0.05	0.14	0.21	0.04
	10	1.14	1.12	0.77	0.63	0.03	1.13^{*}	1.11	0.09	1.12	0.14	1.12	1.12	0.24	1.13	0.71	0.80	0.83	0.82	0.65	0.86

Table 4: WEAT effect size, *: significance at 0.01, **: significance at 0.001, ***: significance at 0.0001, ****: significance at 0.00001.

References

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611*.

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

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

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

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

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998a. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998b. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Andrew Karpinski and James L Hilton. 2001. Attitudes and the implicit association test. *Journal of personality and social psychology*, 81(5):774.

Amy K Kiefer and Denise Sekaquaptewa. 2007. Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of experimental social psychology*, 43(5):825–832.

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

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. 2018. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*.

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

Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statitics. *Proceedings of the National Institute of Sciences of India*, pages 49–55.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

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

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Damian A Stanley, Peter Sokol-Hessner, Mahzarin R Banaji, and Elizabeth A Phelps. 2011. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, 108(19):7710–7715.

A The studied associations: 10 WEAT categories

WEAT	Association	N_T	N_A
1	Flowers vs insects with pleasant vs unpleasant Example: {aster, clover} vs {ant, caterpillar} with {caress, freedom} vs { abuse, crash}	25×2	25×2
2	Instruments vs weapons with pleasant vs unpleasant Example: {bagpipe, cello} vs {arrow, club} with {caress, freedom} vs { abuse, crash}	25×2	25×2
3	Eur.-American vs Afr.-American names with Pleasant vs unpleasant Example: {Adam, Harry} vs {Alonzo, Jamel} with {caress, freedom} vs {abuse, crash}	32×2	25×2
4	Eur.-American vs Afr.-American names with Pleasant vs unpleasant Example: {Brad, Brendan} vs {Darnell, Hakim} with {caress, freedom} vs {abuse, crash}	16×2	25×2
5	Eur.-American vs Afr.-American names with Pleasant vs unpleasant Example: {Brad, Brendan} vs {Darnell, Hakim} with {joy, love} vs {agony, terrible}	16×2	8×2
6	Male vs female names with Career vs family Example: {John, Paul} vs {Amy, Joan} with {executive, management} vs {home, parents}	8×2	8×2
7	Math vs arts with male vs female terms Example: {math, algebra} vs {poetry, art} with {male, man} vs {female, woman}	8×2	8×2
8	Science vs arts with male vs female terms Example: {science, technology} vs {art, Shakespeare} with {brother, father} vs {sister, mother}	8×2	8×2
9	Mental vs physical disease with temporary vs permanent Example: {sad, hopeless} vs {sick, illness} with {impermanent, unstable} vs {stable, always}	6×2	7×2
10	Young vs old people's name with pleasant vs unpleasant Example: {Tiffany, Michelle} vs {Ethel, Bernice} with {love, peace} vs {agony, terrible}	8×2	8×2

Table 1: The 10 WEAT categories.