SADIO Electronic Journal of Informatics and Operations Research (EJS) Vol. 25 No. 2 (2026) e-ISSN 1514-6774
https://revistas.unlp.edu.ar/ejs | https://doi.org/10.24215/15146774e102

Name-Based Embedding Debiasing: Analyzing Its Impact on Gender, Religious, and Ethnic Biases

Mitigación de sesgos en embeddings basada en nombres: análisis de su impacto en sesgos de género, religiosos y étnicos

Gianmarco Cafferata[1] and
Mariano G. Beiró[1,2]

[1]Universidad de San Andrés, Victoria, Argentina
[2]CONICET, Buenos Aires, Argentina
gcafferata@udesa.edu.ar
mbeiro@udesa.edu.ar

Abstract. Word vector representations were the initial building block that started the current state-of-the-art methods for several NLP tasks. Bias metrics and debiasing methods for static embeddings have been studied with moderate success, achieving some bias reductions for specific groups and metrics. However, these methods often fail to improve multiple metrics simultaneously or to meaningfully impact extrinsic tasks. Recent research in debiasing has mainly shifted its focus towards contextual embeddings and large language models (LLMs). Here we argue that static embeddings provide a simpler and more controlled setting for testing hypotheses and techniques, which can then be extended to more complex models. We introduce a method that captures multiple demographic dimensions (gender, race, age, etc.) in static embeddings simultaneously, eliminating the need for specialized tasks or demographic-specific vocabulary.

Resumen Las representaciones vectoriales de palabras representaron el punto de inflexión técnico que dio inicio a los métodos actuales del estado del arte para diversas tareas de Procesamiento del Lenguaje Natural (PLN). Las métricas de sesgo y los métodos de mitigación para embeddings estáticos han sido objeto de estudio con un éxito moderado, logrando reducciones de sesgo para grupos y métricas específicos. No obstante, estos métodos frecuentemente no logran optimizar múltiples métricas de manera simultánea ni impactar significativamente en tareas extrínsecas. La investigación reciente en mitigación se ha reorientado principalmente hacia las representaciones contextuales y los grandes modelos de lenguaje (LLMs). En este trabajo se sostiene que las representaciones estáticas proporcionan un entorno experimental más simple y controlado para la validación de hipótesis y técnicas, las cuales pueden ser posteriormente extrapoladas a modelos de mayor complejidad. Se presenta un método que captura múltiples dimensiones demográficas (género, raza, edad, etc.) en representaciones estáticas simultáneamente, eliminando la dependencia de tareas especializadas o de vocabulario demográfico específico.

Keywords: word embeddings, embedding debiasing, language models, fairness

Received May 2026; Accepted June 2026; Published July 2026

Palabras clave: embeddings de palabras, mitigación de sesgos en embeddings, modelos de lenguaje, equidad

1 Introduction

Word embedding models are a very efficient way of encoding the meaning of a word into a vector. These embeddings exhibit a spatial property: words with similar meanings tend to be located near each other in the vector space. These vectors can solve analogies between them and find vector subspaces of meaning. Glove (Pennington et al., 2014) and Word2Vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) are common models for building these vectors, which were the state-of-the-art choice for several NLP tasks. However, they contain potentially harmful biases embedded in their distances (Bolukbasi et al., 2016) which motivated research on how to measure, reduce, and assess the impacts of these biases.

Static word embeddings have limitations in capturing polysemy and handling out-of-vocabulary words. To address this issue, contextual embeddings were introduced. Unlike static embeddings, contextual embeddings generate word vectors that depend on the surrounding context. For example, the word embedding for *"bank"* differs in *"I deposited money in the bank"* versus *"I sat on the river bank"*. ELMo (Sarzynska-Wawer et al., 2021) and BERT (Devlin et al., 2019) are two common models of contextual embeddings which have led to remarkable improvements on a variety of NLP tasks. However, they employ tokenization methods that split words into smaller subunits, which added to the fact that words' representations change according to context, makes the challenging task of studying biases in embeddings even harder.

With the advent and widespread use of LLMs (Radford et al., 2018) the quality and range of tasks that are solvable with NLP techniques is expanding (Bommasani et al., 2021; Bubeck et al., 2023). This makes the issue of fairness more urgent but also harder to study and mitigate, as the complexity of the models grows. LLMs have exhibited biases towards different demographics on open-ended generation tasks (Sheng et al., 2019).

In this work we propose an agnostic approach for debiasing embeddings based on names. Names are commonly used as proxies for diverse demographics in various bias measurement methods (Caliskan et al., 2017; Dev and Phillips, 2019; Gonen and Goldberg, 2019; Huang et al., 2020). Our method is based on the hypothesis that several demographics can be captured from distances between names. This allows our method to improve multiple metrics simultaneously and paves the way for using names' likelihoods in order to extend this method to contextualized word embeddings and LLMs.

2 Related work

As static word embeddings like GloVe and Word2Vec became widely used, one of the earliest major studies on embeddings bias focused on gender bias (Bolukbasi et al., 2016). This study showed that given an analogy puzzle *"man is to X as woman is to Y"*, using simple embedding aritmethic the vectors showed sexist analogies. For example, the word vectors produced analogies like *"man is to surgeon as woman is to nurse"*.

This indicates that stereotypes present in the training corpus were captured and encoded into the learned word representations.

The study by (Caliskan et al., 2017) introduced new metrics inspired by a psychological assessment (Greenwald et al., 1998) to detect implicit positive and negative biases in word embeddings. These metrics measure the extent to which certain demographic terms are more strongly associated with negative attributes.

(Zhao et al., 2018) introduced a modified version of Glove with a custom loss function to avoid gender biases. In their approach, the model is trained from scratch. The resulting embeddings have k additional dimensions that contain all the gender-related information, which can be kept or removed as needed for each word.

In 2019, (Gonen and Goldberg, 2019) argued that many debiasing methods failed to fully remove gender information from word representations. To support their claim, they introduced a set of metrics designed to evaluate how easily the gender of previously biased words could still be inferred after debiasing, demonstrating that gender often persists despite apparent reductions in bias.

In the same year (Manzini et al., 2019) extended the hard-debiasing method from (Bolukbasi et al., 2016) to race and religion, and proposed a new metric inspired by WEATs (Word Embedding Association Tests, (Caliskan et al., 2017)). However, this metric also requires a predefined set for each demographic.

Finaly, (Dev and Phillips, 2019) argued that bias directions can be computed using the difference between names and validated this approach through gender-related tests. They also proved that other demographics can be analyzed using name sets. However, their method requires the target demographics to be debiased, besides building a predefined set of opposite name pairs for each demographic pair.

As the use of large language models (LLMs) has grown, several methods have been proposed to measure and mitigate their biases (Dhamala et al., 2021; Huang et al., 2020; Nadeem et al., 2021; Sheng et al., 2019), often relying on prompt templates for open-ended generation and evaluating output probabilities to quantify bias.

3 Methodology

Our approach to the embeddings debiasing problem is based on the hypothesis that key directions related to differences between individuals are encoded in names' embeddings (e.g., male vs. female names, European vs. African names). Here we present a methodology to validate this hypothesis and evaluate its effectiveness.

In this section we: *(i)* introduce the data that we used; *(ii)* describe the 2 proposed models and the baseline models; *(iii)* define the metrics used for evaluation.

3.1 Data

We use the US Baby Names Dataset as a source for names. We select all cased names with a frequency larger than $1,000$ and use their lowercase counterparts if their frequency exceeds $10,000$. This results in a total of $12,447$ names that will be used for debiasing. For highly common names (e.g., *"John"*), both the cased (*"John"*) and lowercase (*"john"*) embeddings primarily represent the name itself rather than other meanings.

We depart from a cased, pretrained GloVe (Pennington et al., 2014) model trained on 840 billion tokens from the Common Crawl dataset, available on the GloVe website[1].

For the bias visualizations we used an adjective dataset studied by Condon et al., 2022 keeping the ones that were known to more than 75% of respondents, resulting in $1,906$ personality descriptors.

3.2 Debiasing model

We propose an approach based on Singular Value Decomposition (SVD) for debiasing embeddings based on names. We sample 1 million pairwise distances between names and use principal component analysis to identify the main variance directions. Then, we use the principal components that capture 35% of the variance to remove their effect from all embeddings. Denoting W for the embedding matrix and X for the matrix of the first principal components, the debiased embeddings are computed as:

$$X_{debiased} = X - XW^TW$$

Baseline models We compare our approach against two baseline models frequently used in the literature:

Gender Hard-Debias. The debias method by (Bolukbasi et al., 2016) for gender consists on using the direction of maximum variance of the distance between 10 gender pairs (he-she, her-his, etc.). This direction is defined as the *"gender component"*, $\overline{g}$, the debiasing process consists in subtracting the projection $\overline{w_b}$ of each word embedding $\overline{w}$ onto the former:

$$\overline{w}_{debiased} = \overline{w} - \overline{w_b} = \overline{w} - (\overline{w} \cdot \overline{g})\overline{g}$$

GN-Glove. These embeddings (Zhao et al., 2018) were trained to contain all the gender-related information in the last k dimensions of the embeddings. We will evaluate the metrics on these embeddings after removing those dimensions.

3.3 Experimental setup

We tested our method on multiple metrics to ensure debiasing across various demographics and to assess the quality of the embeddings. In addition to quantitative benchmarking we performed qualitative analysis of the embedding biases following the approach in Bolukbasi et al., 2016 and Gonen and Goldberg, 2019.

Embedding quality. We evaluate the quality of our debiased embeddings using five word similarity datasets: Wordsim (Finkelstein et al., 2001), Stanford RareWords (Luong et al., 2013), Cambridge Rarewords (Card660) (Pilehvar et al., 2018), Simlex-999 (Hill et al., 2015) and SimVerb-3500 (Gerz et al., 2016). These datasets consist of word

[1] https://nlp.stanford.edu/data/glove.840B.300d.zip

pairs with human-annotated similarity scores. We assess embedding quality by computing the Pearson correlation between the human similarity scores and the cosine similarity of the word embeddings. A higher Pearson correlation indicates better embedding quality.

Gender DirectBias. The Gender Hard-Debias method also introduced (Bolukbasi et al., 2016) a metric called *DirectBias* to be defined over a set of words that should be neutral N using the gender direction:

$$\mathit{DirectBias}_c = \frac{1}{|N|} \sum_{\overline{w} \in N} |cos(\overline{w}, \overline{g})|^c$$

In our tests, we are going to use c=1 as the mean cosine that each neutral word has with the gender vector and a professions word set as neutral words.

WEAT. WEAT (Caliskan et al., 2017) is a statistical test designed to measure implicit biases in word embeddings.

WEAT compares the association strength between two sets of target words' embeddings, X and Y, and two sets of attribute words' embeddings, A and B. The test computes a bias score based on the cosine similarity of word embeddings, quantifying how much closer words in X are to attributes in A compared to B, relative to words in Y.

To assess statistical significance, WEAT applies a permutation test. The test statistic is:

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

$$s(\overline{w}, A, B) = \frac{1}{|A|} \sum_{\overline{a} \in A} \frac{\overline{w} \cdot \overline{a}}{|\overline{w}||\overline{a}|} - \frac{1}{|B|} \sum_{\overline{b} \in B} \frac{\overline{w} \cdot \overline{b}}{|\overline{w}||\overline{b}|}$$

The p-value of the permutation test is taken over partitions of X and Y (X_i and Y_i):

$$Pr_i[S(X_i, Y_i, A, B) > S(X, Y, A, B)]$$

And the effect size is:

$$\frac{\sqrt{|X \cup Y|}\left(\frac{1}{|X|}\sum_{\overline{x} \in X} s(\overline{x}, A, B) - \frac{1}{|Y|}\sum_{\overline{y} \in Y} s(\overline{y}, A, B)\right)}{\sqrt{\left(\sum_{\overline{w} \in X \cup Y} s(\overline{w}, A, B) - \frac{1}{|X \cup Y|}\sum_{\overline{w} \in X \cup Y} s(\overline{w}, A, B)\right)^2}}$$

We will assess nine of the WEAT scenarios presented in the paper:

- *Flowers vs Insects:* This test compares flower names (group 1) and insect names (group 2) with positive attributes (set 1) and negative attributes (set 2). While originally used as a control to confirm that embeddings retain expected associations, we repurpose it as an additional post-debiasing validation test of embedding quality.

- *Instruments vs Weapons:* Similar to the previous test, but with musical instruments (group 1) and weapons (group 2).
- *European-American vs. African-American Names 1, 2, and 3:* These tests compare European-American names (group 1) and African-American names (group 2), always using positive attributes in set 1 and negative attributes in set 2. The three versions of the test use different sets of names or attributes, but the structure remains the same. Since our method uses names as the debiasing target, these scores are expected to be lower but will not necessarily reflect reduction in bias for our method.
- *Male vs. female names:* Measures whether male names are more strongly associated with work-related attributes than female names, which are typically associated with family-related attributes. Like the previous test, this also cannot be used as an evaluation metric because names are the debiasing target.
- *Math vs. Arts:* This test compares math-related words (group 1) with arts-related words (group 2), measuring the extent to which math is associated with male attributes and arts with female attributes. This is a relevant evaluation metric since it does not use names.
- *Science vs. Arts:* Similar to the Math vs. Arts test, but instead of math-related words, it uses a broader set of science-related terms.
- *Young vs. old people's names:* This test examines whether embeddings reflect an implicit bias associating young names more with positive attributes and old names more with negative attributes.

Name-independent WEATs. We have created a new set of WEATs to measure different demographics without relying on names. This WEATs will be part of our evaluation metrics. For each WEAT, the first attribute set contains positive words, while the second set contains negative words. We developed tests for the following six demographics:

- Western vs. Asian
- Latin American vs. Anglo-American Cultural Terms
- Heteronormative vs. Queer
- Young vs. Old
- Christian vs. Muslim
- Caucasian vs. Black

In all these tests, a higher effect size indicates that the first group is more strongly associated with positive attributes than the second. The full list of words used in these tests can be found in Appendix A.1.

Preservation of gender information. We adopt three gender information preservation tests from the literature (Gonen and Goldberg, 2019) to assess how much gender-related information remains after debiasing:

- *Clustering of male and female words:* We project word embeddings onto a gender direction and extract the 500 most gender-biased words (250 male-biased, 250 female-biased). We then apply K-means clustering to these words before and after debiasing. A lower clustering accuracy after debiasing indicates that less gender information is present in the embeddings, as they become less separable along gender lines.

- *Classifying previously female and male-biased words:* We extract the 2500 most gender-biased words for each gender and train a SVM classifier with an RBF kernel on 1000 words from the combined 5000 word dataset. We then repeat the training after debiasing. A drop in classification accuracy suggests that less gender information is available in the embeddings, making it harder for the classifier to distinguish male and female associated words.
- *Correlation between profession bias and male neighbors:* We first compute the gender projection of each profession-related word before debiasing. Then, after debiasing, we identify the top 100 nearest neighbors for each profession in the embedding space. We measure whether the pre-debiasing gender projection correlates with the number of previously male-biased words among the current debiasing-adjusted neighbors. Since professions should not retain gendered associations, a lower correlation indicates more effective debiasing.

It is important to highlight that the study aimed to demonstrate that, even after removing the gender direction, gender-related information remains embedded in word vectors. However, this does not mean that all gender information is harmful. For instance, one of the most female-biased words identified in the study was *"bra"*. An ideal debiasing method would aim to preserve associations that carry useful information.

Qualitative analysis. For our qualitative analysis, we studied the bias over a set of personality trait adjectives for three compared demographics: Male-Female, Caucasian-Black and Christian-Muslim. To measure the bias of each adjective we needed a demographic vectorial direction, like the one built with definitional pairs for gender in Bolukbasi et al., 2016. This definitional pairs require perfect analogies between pairs of each group (i.e., he-she) which we do not have for other groups than gender. We chose to build our demographic direction using our name-independent WEATs words for each group. The direction is computed averaging each word in group A substracted by each possible word in group B:

$$\overline{d}_{bias} = \frac{1}{|A| \cdot |B|} \sum_{\overline{a} \in A} \sum_{\overline{b} \in B} (\overline{a} - \overline{b})$$

Our analysis focused on the most biased adjectives in two ways:

- *Most biased adjectives:* Adjectives with the highest bias magnitude for each group, quantified by their scalar projection onto the bias direction both before and after debiasing.
- *t-SNE visualization of the most-biased adjectives:* The t-SNE (Maaten and Hinton, 2008) visualizations of the unlabeled 500 most biased word embeddings and the 50 most biased words with labels, for the vanilla and debiased version. This not only shows the most biased words but also how well-mixed these words are along the full embedding semantic space. High linear separability and high cluster density

within groups indicate the presence of strong, localized stereotyping. Conversely, increased spatial overlap and reduced cluster density suggest either successful mitigation or that the underlying stereotyping is characterized by a complex structure that cannot be visualized within a two-dimensional non-linear manifold.

4 Results

We compare the pretrained GloVe embeddings, their gender hard-debiased version, the GN-GloVe embeddings (pretrained by the original authors[2]), and our debiasing method applied to GloVe.

We first analyze the quality of the debiased embeddings. Table 1 reports the results on five standard word similarity benchmarks, which evaluate how well the embeddings capture semantic similarity between word pairs. Higher Pearson correlation scores indicate a closer alignment with human judgments.

Table 2 shows the results of two WEAT tests—Flowers vs. Insects and Weapons vs. Instruments—which are not considered biased and should ideally be preserved. These tests measure whether embeddings retain expected associations found in general semantic understanding. A preserved association is indicated by a test statistic close to the original GloVe value.

	Wordsim	**Simlex**	**Rarewords**	**Card660**	**SimVerb**
Glove (vanilla)	0.80	0.44	0.45	0.53	0.29
Gender Hard-Debias	0.80	0.44	0.45	0.53	0.29
GN-GloVe	0.72	0.38	0.39	0.44	0.22
Name-based SVD debiasing (ours)	**0.81**	**0.49**	**0.51**	**0.61**	**0.35**

Table 1: **Embedding quality**. Results of the word similarity benchmarks for our model, as compared previous ones in the literature. The values represent Pearson correlations between the human similarity scores and the cosine similarity of the word embeddings.

In Tables 3, 4 and 5 we assess the debiasing property of our model. Gender DirectBias (Table 3) measures the mean projection of the profession's embeddings over the gender component; the lower this average projection, the better. Table 4 presents the results for name-based WEATs, which assess associations between demographic name groups (e.g., African-American vs. European-American, male vs. female, young vs. old) and sets of positive and negative attributes. Table 5 reports the results of name-independent WEATs, designed to measure implicit bias along several demographic dimensions without relying on names. These include gender associations in academic domains (e.g., math vs. arts), cultural identity (e.g., Latin American vs. Anglo-American), religion, age, race, and sexual orientation. Lower effect sizes here indicate weaker, and thus less biased, associations between demographic groups and polarized attributes.

[2] https://github.com/uclanlp/gn_glove

	Flowers vs Insects	**Weapons vs Instruments**
GloVe (vanilla)	**2.24**	**2.29**
Gender Hard-Debias	**2.18**	**2.30**
GN-GloVe	1.18	1.78
Name-based SVD debiasing (ours)	**2.14**	**2.41**

Table 2: Statistics for the two unbiased (harmless) WEAT tests. High values of these test statistics ensure that the associations were preserved. We highlight in bold those cases in which the statistics are not below 90% of the values found with GloVe.

	Gender DirectBias
GloVe (vanilla)	0.106
Gender Hard-Debias	**0.019**
GN-GloVe	0.086
Name-based SVD debiasing (ours)	0.064

Table 3: **DirectBias Score**. Values represent the average the projection of the profession's embeddings into the gender component for our method as compared to previous ones in the literature.

We evaluate how much gender-related information remains in the embeddings by using three complementary metrics (Table 6): gender clustering accuracy, gender classification accuracy, and the correlation between professions and their gendered neighbors. The first two tests do not directly indicate harmful bias but rather the extent to which gender can still be inferred from the embeddings.

Finally, for both before and after debiasing, we show the most biased adjectives for different demographics in Figure 1, and a t-SNE visualization of the 500 most biased adjectives of each group in Figure 2.

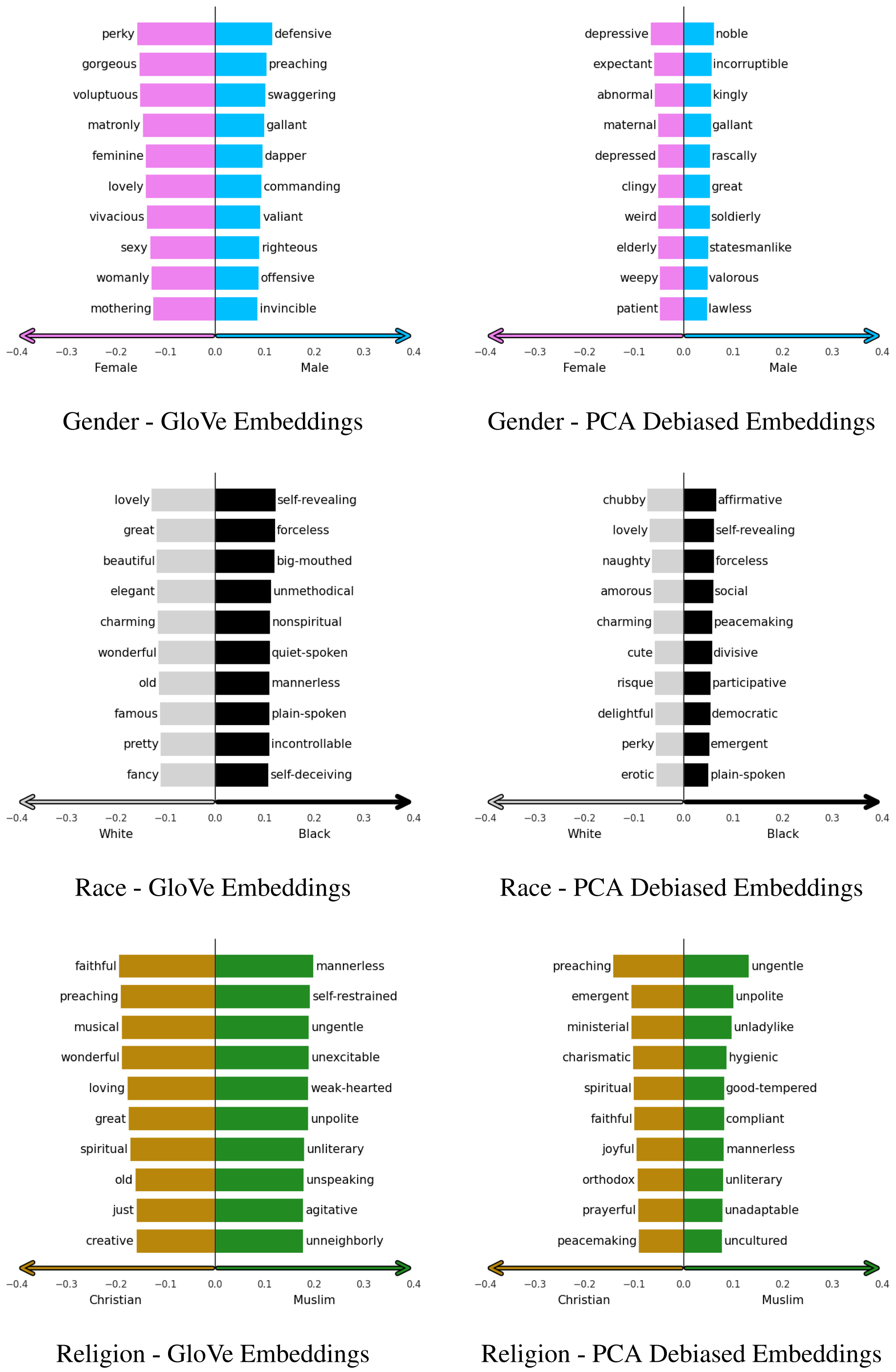
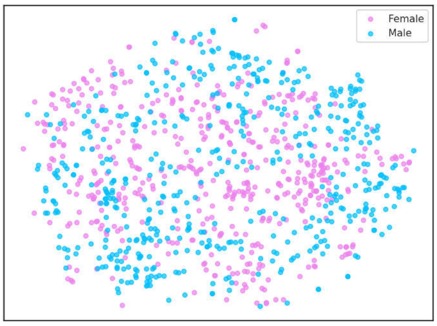
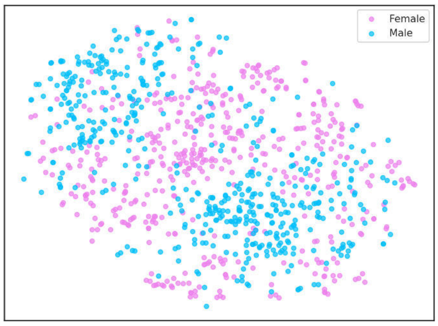


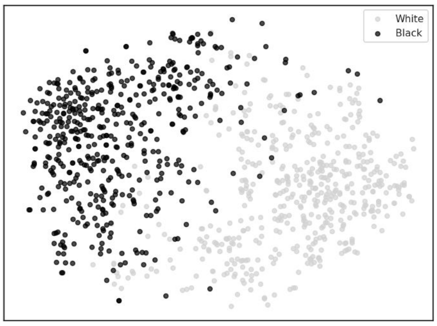
Fig. 1: Top 10 most biased adjectives for each demographic pair in the vanilla GloVe embeddings (lef) and our debiased version (right).



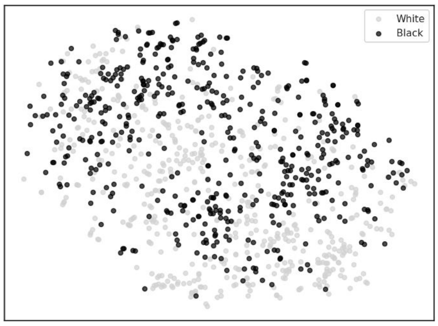
Gender - GloVe Embeddings



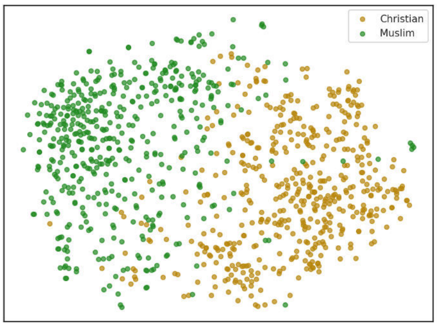
Gender - PCA Debiased Embeddings



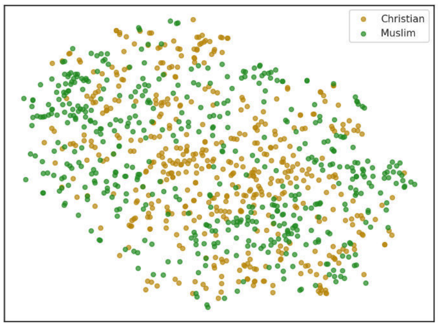
Race - GloVe Embeddings



Race - PCA Debiased Embeddings



Religion - GloVe Embeddings



Religion - PCA Debiased Embeddings

Fig. 2: t-SNE visualization of the 500 most biased adjective embeddings of each demographic group. Color encoding denotes the demographic group associated with the directional bias of each term.

	European vs. African			Male vs. Female	Young vs. Old
	1	2	3		
GloVe (vanilla)	1.73	0.73	0.92	1.27	0.38
Gender Hard-Debias	1.79	0.74	0.94	0.60	0.42
GN-GloVe	**0.44**	**0.01**	0.70	1.04	**0.05**
Name-based SVD debiasing (ours)	0.65	0.20	**-0.08**	**0.58**	-0.12

Table 4: **Name-based WEATs**. Statistics for the name-based WEAT tests. Values close to zero ensure that the biases have been removed. We highlight in bold the best method for each test.

	GloVe (vanilla)	**Gender Hard-Debias**	**GN-GloVe**	**Name-based SVD Debiasing** (ours)
Gender: Math vs. Arts	0.20	**0.02**	0.16	0.03
Gender: Science vs. Arts	0.35	-0.01	0.27	**0.00**
Western vs. Asian Associations	0.29	0.31	**0.13**	0.27
Latin American vs. Anglo-American Cultural Terms	0.38	0.41	0.53	**0.03**
Heteronormative vs. Queer Associations	0.33	0.34	**0.11**	0.26
Young vs. Old Associations	0.26	0.24	0.23	**-0.08**
Christian vs. Muslim	1.21	1.26	0.90	**0.43**
Caucasian vs. Black	0.80	0.81	**0.22**	0.46

Table 5: **Name-independent WEATs**. Statistics for the name-independent WEAT tests. Values close to zero ensure that the biases have been removed. We highlight in bold the best method for each test.

	Gender clustering accuracy	**Gender classifier accuracy**	**Gender-Professions neighbor correlation**
Glove	0.989	0.999	0.800
Gender Hard-Debias	0.914	0.975	0.713
GN-GloVe	**0.866**	0.998	0.786
Name-based SVD Debiasing	0.945	**0.973**	**0.701**

Table 6: **Preservation of gender information**. In the classification experiments (first two column) a higher accuracy indicates larger prevalence of gender information in the embeddings after debiasing. In the gender-professions neighbor correlation experiment (last column) higher values indicate larger harmful gender bias over profession embeddings. We highlight in bold the best method for each test.

5 Discussion

While pretrained word embeddings are widely used in downstream tasks such as sentiment analysis, machine translation, and question answering due to their ability to capture semantic structure, their reliance on real-world corpora makes them susceptible to encoding and amplifying social biases and stereotypes. Here we present a demographically agnostic, named-based method to reduce harmful biases in embeddings. As shown in Table 1, our method not only preserves but also improves semantic similarity scores across several benchmarks.

The proposed model outperforms existing state-of-the-art debiasing approaches, particularly GN-GloVe. While GN-GloVe achieves comparable bias reduction in some metrics, it underperforms in semantic quality benchmarks, including both word similarity tasks and WEATs designed to preserve meaningful associations (Table 2).

In the gender projection evaluation, the Hard Debiasing method achieves the lowest DirectBias score, while our method has the second lowest one (3). However, this is somewhat expected, as the Hard Debiasing method explicitly removes the embedding projection used in the metric.

In the name-dependent WEATs (Table 4), our method achieves reductions comparable to the state of the art. However, since our approach uses names for debiasing, these tests are not reliable indicators for comparison. In the name-independent WEATs (Table 5), our method achieves similar or better bias reduction than other approaches, with the exception of "Caucasian vs. Black", "Western vs. Assian associations", and "Heteronormative vs. Queer Associations" where GN-Glove reduces the size effect further by halving our result.

Regarding gender information preservation (Table 6), our method reduces both clustering and classification accuracy, which points out that it achieves a partial loss of gender signal. In contrast, GN-GloVe retains high classification performance, even though it shows a lower clustering accuracy. It is important to remark that these metrics are not necessarily desirable to become zero, as they include words with genuine, non-harmful gender associations (e.g., "*bra*"). Instead, the goal of this evaluation is to show that gender information is difficult to fully remove. Our method does not aim to eliminate all gender information, but rather to reduce unwanted harmful associations. The gender-professions neighbor correlation in the last column of this Table, instead, measures the relationship between a profession's pre-debiasing gender projection and the number of its post-debiasing neighbors previously associated with the same gender. Lower values of this metric are indicative of a more effective debiasing. Thus, in contrast with the previous metrics, an ideal debiaser should bring to zero these correlation. We find that our method achieves the weakest correlation among all the tested ones.

We hypothesize that GN-GloVe's improvements in some demographic WEATs may be more related to a loss of overall semantic meaning than to a true reduction in harmful biases. Since GN-GloVe was specifically designed and trained to address gender bias, its impact on other demographics could be incidental and just a side effect of altering the embedding space in ways that degrade overall representation quality. This is supported by its poor performance on WEATs intended to preserve meaningful associations. In contrast, our method effectively reduces undesirable biases across multiple demographics, while it preserves desirable associations without requiring demographic-specific word sets or retraining the embeddings from scratch.

The visualization in Figure 1 illustrates a systematic reduction in bias projections across all demographics following the debiasing procedure, maintaining consistency with our metric results. This analysis identifies the specific nature of the bias present in each pair of demographics. In the gender demographic, the adjectives with the highest bias magnitudes align with traditional social stereotypes. Specifically, terms such as "*sexy*" and "*lovely*" show a strong projection toward the female direction, whereas "*gallant*" and "*commanding*" project toward the male direction. For race and religion, the bias manifests as a marked valence asymmetry: White and Christian groups show high correlations with positive attributes (e.g., "*charming*", "*great*"), while Black and Muslim groups align with negative descriptors (e.g., "*mannerless*", "*forceless*"). We can visualize more detail of the biased words on Figures 3, 5 and 7.

The adjective projections over gender shown in Figure 2 illustrate the complexity of those biases. Gender-biased adjectives remain mixed together, both before and after debiasing. This overlap suggests that the bias has a complex structure that is difficult to separate, making it indistinguishable whether the debiasing had any structural impact based solely on this projection. This is also supported by the structure and adjectives for the 50 most biased words t-SNE projection before and after debiasing (Figures 3 and 4). Even after debiasing with the bias reduced in projection magnitude the words that are assigned to each gender still align to social stereotypes.

On the contrary, biases for race and religion exhibit a clear structural shift, transitioning from distinct clustering to a mixed distribution. This suggests that the stereotyping has either been rendered less explicit or effectively removed. In a more detailed visualization for race in Figures 5 and 6 we can see that although the clusters are much more mixed, the words associated with white people contain more clusters with several positive attributes (e.g. "*cute*", "*beautiful*", "*lovely*", "*charming*"). Finally, if we focus on Figures 7 and 8 we can see that not only are both groups much more mixed, but the words also do not exhibit a clear valence asymmetry within groups.

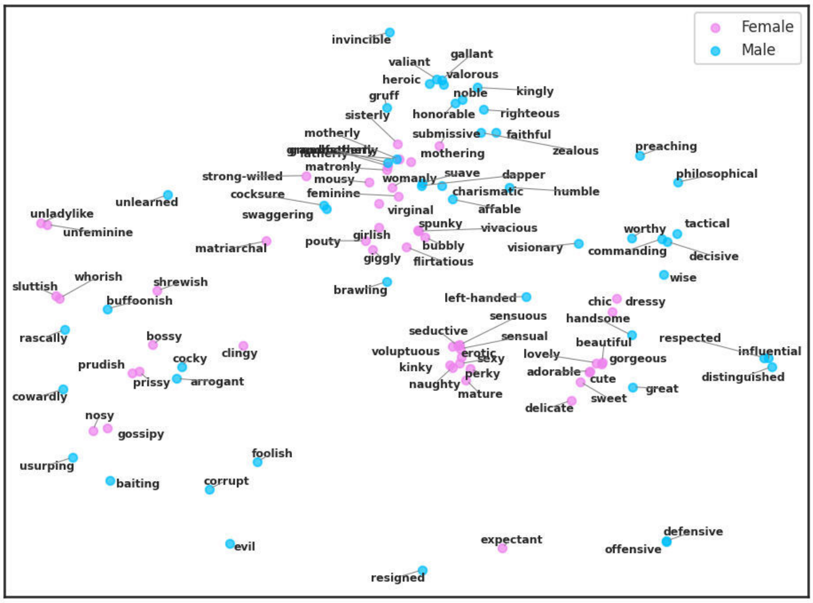


Fig. 3: t-SNE visualization of the top 50 most biased adjectives for gender before debiasing.

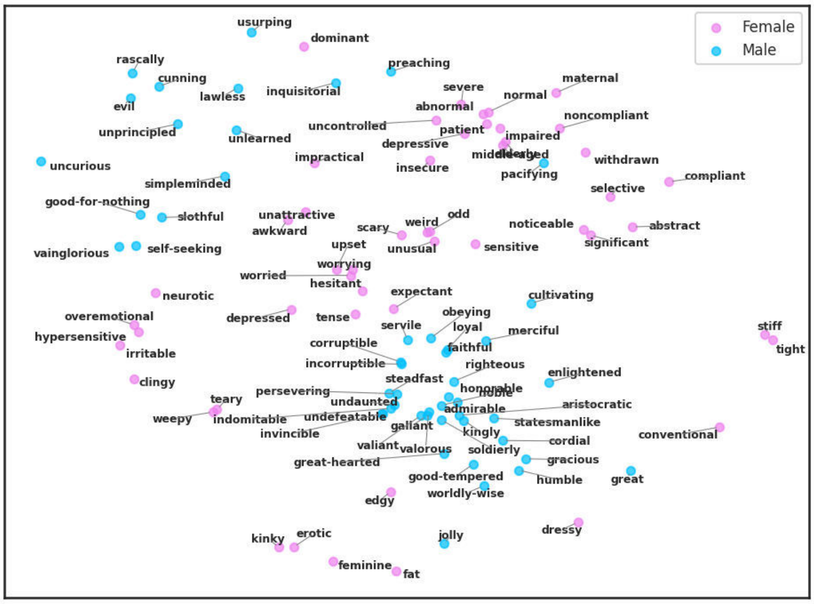


Fig. 4: t-SNE visualization of the top 50 most biased adjectives for gender after debiasing.

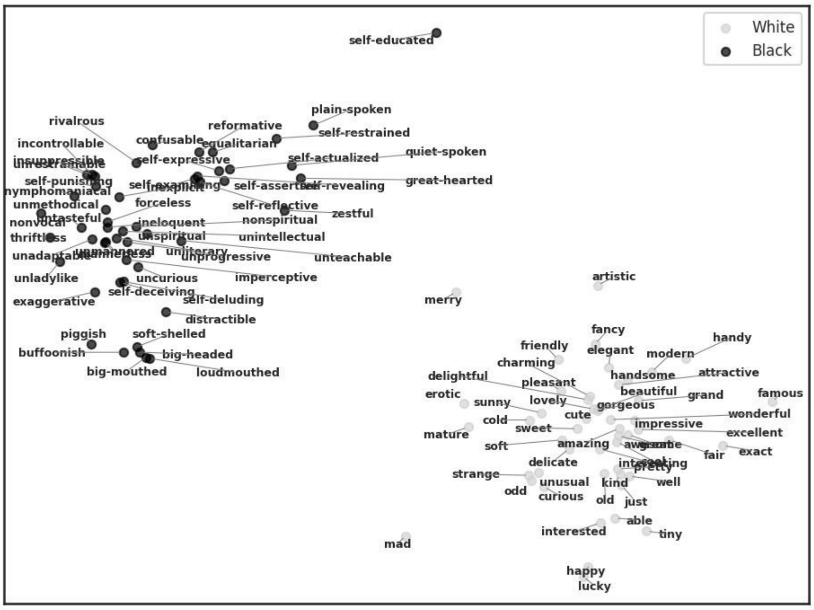


Fig. 5: t-SNE visualization of the top 50 most biased adjectives for race before debiasing.

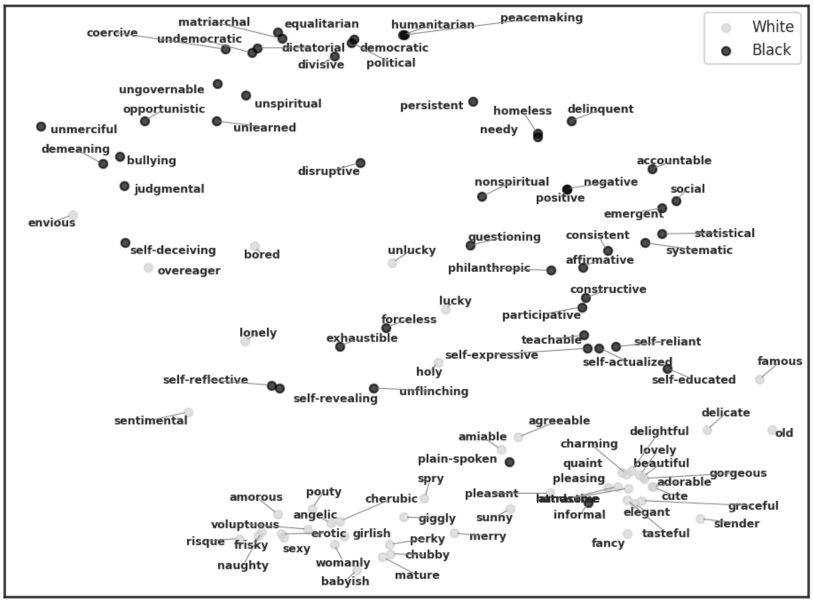


Fig. 6: t-SNE visualization of the top 50 most biased adjectives for race after debiasing.

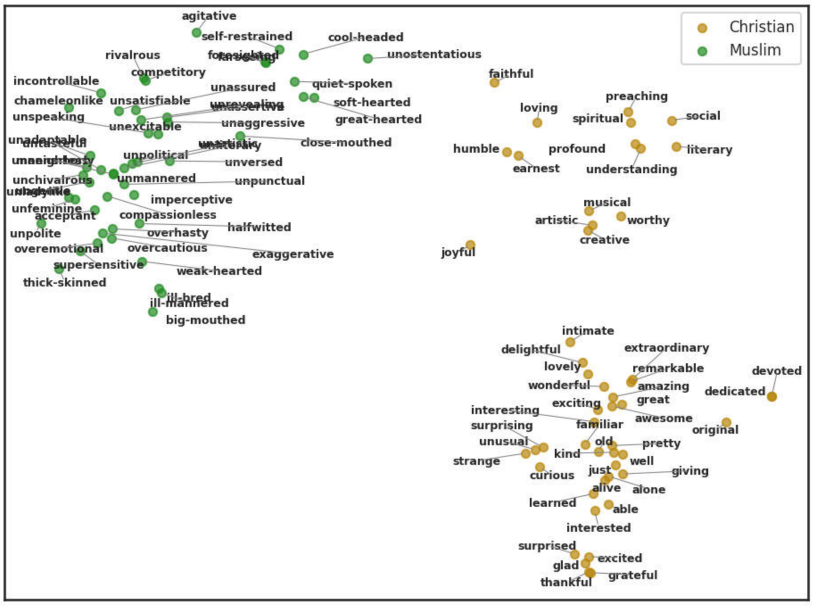


Fig. 7: t-SNE visualization of the top 50 most biased adjectives for religion before debiasing.

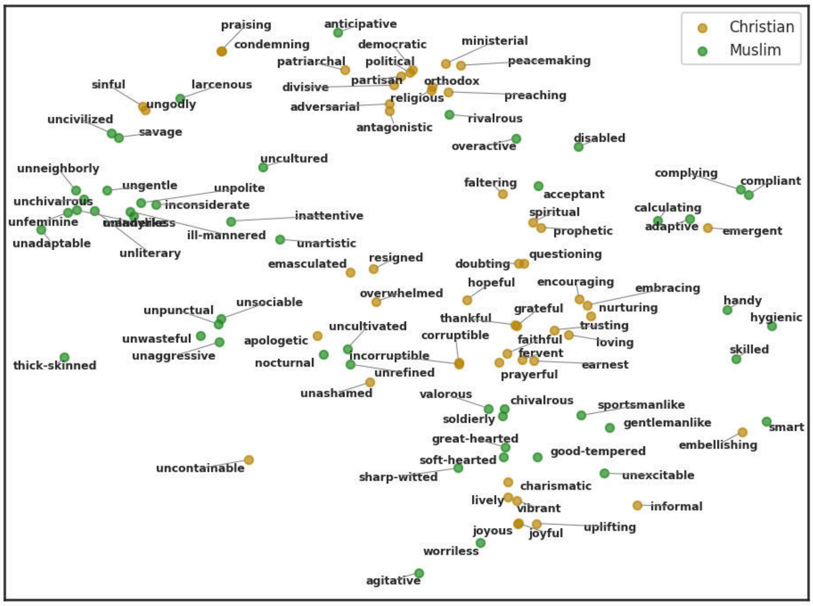


Fig. 8: t-SNE visualization of the top 50 most biased adjectives for religion after debiasing.

6 Conclusions and future work

In this work we presented a demographically agnostic model for static embedding debiasing. Our approach was effective in partially removing bias from GloVe embeddings while preserving, and in some cases improving, their semantic quality. The method is also simple to apply, as it does not require demographic-specific word sets nor retraining the embeddings from scratch.

As future work, the method should be evaluated on downstream tasks using application-specific bias benchmarks to assess whether the observed improvements translate into fairer model behavior in real-world scenarios. Also, given that most of bias over distributional semantic studies focus on gender, future research should expand to other demographic groups. This analysis could go beyond simple positive and negative attributes to explore specific semantic dimensions, verifying whether these model biases correlate with measurable societal patterns. Finally, since we achieved bias reduction using names as a general way to capture multiple demographics without specific, predefined vocabularies, we hope that our approach could be extended to Large Language Models and contextual embeddings, thus providing a generalizable way to mitigate the generation of biased content.

References

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*. https://doi.org/10.48550/arXiv.1607.06520

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://doi.org/10.48550/arXiv.2108.07258

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Condon, D. M., Coughlin, J., & Weston, S. J. (2022). Personality trait descriptors: 2,818 trait descriptive adjectives characterized by familiarity, frequency of use, and prior use in psycholexical research. *Journal of Open Psychology Data*, *10*(1). https://doi.org/10.5334/jopd.57

Dev, S., & Phillips, J. (2019). Attenuating bias in word vectors. *The 22nd international conference on artificial intelligence and statistics*, 879–887. https://proceedings.mlr.press/v89/dev19a.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web*, 406–414. https://doi.org/10.1145/503104.503110

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2173–2182. https://doi.org/10.18653/v1/D16-1235

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. https://doi.org/10.18653/v1/N19-1061

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, *74*(6), 1464. https://doi.org/10.1037/0022-3514.74.6.1464

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., & Kohli, P. (2020). Reducing sentiment bias in language models via counterfactual evaluation. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 65–83. https://doi.org/10.18653/v1/2020.findings-emnlp.7

Luong, M.-T., Socher, R., & Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. *Proceedings of the seventeenth conference on computational natural language learning*, 104–113. https://aclanthology.org/W13-3512/

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

gies, Volume 1 (Long and Short Papers), 615–621. https://doi.org/10.18653/v1/N19-1062
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*. https://doi.org/10.48550/arXiv.1310.4546
Nadeem, M., Bethke, A., & Reddy, S. (2021). Stereoset: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
Pilehvar, M. T., Kartsaklis, D., Prokhorov, V., & Collier, N. (2018). Card-660: Cambridge rare word dataset-a reliable benchmark for infrequent word representation models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1391–1401. https://doi.org/10.18653/v1/D18-1169
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135. https://doi.org/10.1016/j.psychres.2021.114135
Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412. https://doi.org/10.18653/v1/D19-1339
Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. https://doi.org/10.18653/v1/D18-1521

A Appendix

A.1 Name-independent WEATs

All tests use the same attribute sets, sourced from the original WEAT paper (Caliskan et al., 2017).

Attributes 1 (positive): caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

Attributes 2 (negative): abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit

Western vs. asian associations Group 1: western, opera, christianity, gothic, french, german, italian, spaniard, swiss, wine, christmas, cowboy, comic, shakespeare, ballet, latin, renaissance, medieval, michelangelo, vatican, Alps, scholasticism, monarchic

Group 2: asian, kabuki, buddhism, pagoda, japanese, chinese, vietnamese, korean, indonesian, sake, vesak, samurai, anime, haiku, geisha, mandarin, Edo, imperial, hokusai, shaolin, Himalayas, confucianism, dynastic

Latin American vs. Anglo-American Cultural Terms Group 1: american, festival, parade, jazz, rock, orchestra, square, brunch, cowboy, dress, tea, burger, pie, folk, hiphop, cornbread, barbecue, whiskey, bourbon, rockies, Yellowstone

Group 2: latino, fiesta, carnaval, salsa, tango, mariachi, plaza, siesta, gaucho, poncho, mate, empanada, tortilla, cumbia, reggaeton, ceviche, chimichurri, tequila, mezcal, andes, amazon

Heteronormative vs. Queer Associations Group 1: straight, binary, tradition, gala, housewife, jock, fertile, gendered, waltz, husband, monogamy, masculine, feminine, heterosexual

Group 2: queer, nonbinary, transition, ballroom, dyke, twink, asexual, genderqueer, voguing, partner, polyamory, androgynous, nonconforming, gay, lesbian

Young vs Old Associations Group 1: junior, trainee, student, novice, apprentice, juvenile, child, teenager, kid, millennial, ambition, young, intern, youth, orphan, player, cadet, backpack, skateboard, tablet, savings, bib, pacifier

Group 2: senior, manager, mentor, veteran, professor, retiree, grandparent, retiree, elderly, boomer, nostalgia, old, executive, elder, widow, coach, commander, briefcase, cane, newspaper, pension, apron, dentures

Christian vs. Muslim Group 1: Christianity, bible, church, pastor, gospel, baptism, trinity, God, Christmas, Easter, communion, prayer, worship, tithing, confession, cross, rosary, cathedral, pope, Vatican, choir, Latin, Jerusalem, Jesus, Peter, Paul, Mary, Augustine, Luther

Group 2: Islam, Quran, mosque, imam, hadith, sunnah, shahada, tawhid, Allah, hajj, Ramadan, eid, zakat, salat, sawm, wudu, halal, hijab, niqab, kufi, minaret, crescent, ummah, Arabic, Mecca, Medina, Muhammad, Umar, Omar, Ali, Fatima

Caucasian vs. Black Group 1: caucasian, opera, french, german, italian, swiss, christmas, rock, latin, Michelangelo, Alps, Madonna, zulu, celtic, bard, slavic, folk, Kennedy, Siberia, prairie, druid

Group 2: black, gospel, creole, haitian, jamaican, Gullah, kwanzaa, hip-hop, ebonics, Basquiat, Appalachians, Beyoncé, viking, igbo, maasai, Mandinka, blues, Obama, Sahara, savanna, shaman