Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation

Tianlu Wang[1*] **Xi Victoria Lin**[2] **Nazneen Fatema Rajani**[2]
Bryan McCann[2] **Vicente Ordonez**[1] **Caiming Xiong**[2]
[1]University of Virginia {tw8cb, vicente}@virginia.edu
[2]Salesforce Research {xilin, nazneen.rajani, bmccann, cxiong}@salesforce.com

Abstract

Word embeddings derived from human-generated corpora inherit strong gender bias which can be further amplified by downstream models. Some commonly adopted debiasing approaches, including the seminal Hard Debias algorithm (Bolukbasi et al., 2016), apply post-processing procedures that project pre-trained word embeddings into a subspace orthogonal to an inferred gender subspace. We discover that semantic-agnostic corpus regularities such as word frequency captured by the word embeddings negatively impact the performance of these algorithms. We propose a simple but effective technique, Double-Hard Debias, which purifies the word embeddings against such corpus regularities prior to inferring and removing the gender subspace. Experiments on three bias mitigation benchmarks show that our approach preserves the distributional semantics of the pre-trained word embeddings while reducing gender bias to a significantly larger degree than prior approaches.

1 Introduction

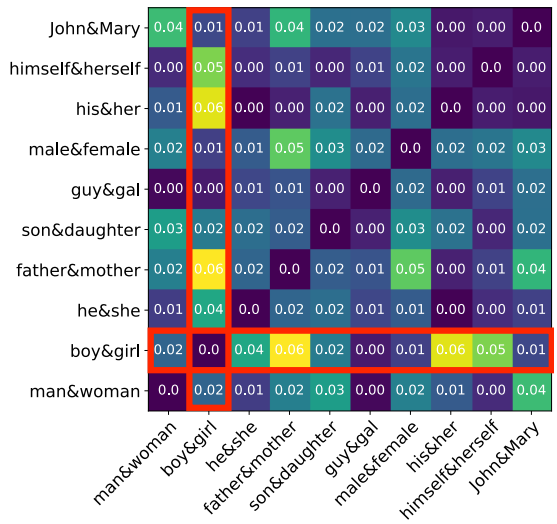
Despite widespread use in natural language processing (NLP) tasks, word embeddings have been criticized for inheriting unintended gender bias from training corpora. Bolukbasi et al. (2016) highlights that in word2vec embeddings trained on the Google News dataset (Mikolov et al., 2013a), "programmer" is more closely associated with "man" and "homemaker" is more closely associated with "woman". Such gender bias also propagates to downstream tasks. Studies have shown that coreference resolution systems exhibit gender bias in predictions due to the use of biased word embeddings (Zhao et al., 2018a; Rudinger et al., 2018). Given the fact that pre-trained word embeddings have been integrated into a vast number of NLP models, it is important to debias word embeddings to prevent discrimination in NLP systems.

To mitigate gender bias, prior work have proposed to remove the gender component from pre-trained word embeddings through post-processing (Bolukbasi et al., 2016), or to compress the gender information into a few dimensions of the embedding space using a modified training scheme (Zhao et al., 2018b; Kaneko and Bollegala, 2019). We focus on post-hoc gender bias mitigation for two reasons: 1) debiasing via a new training approach is more computationally expensive; and 2) pre-trained biased word embeddings have already been extensively adopted in downstream NLP products and post-hoc bias mitigation presumably leads to less changes in the model pipeline since it keeps the core components of the original embeddings.

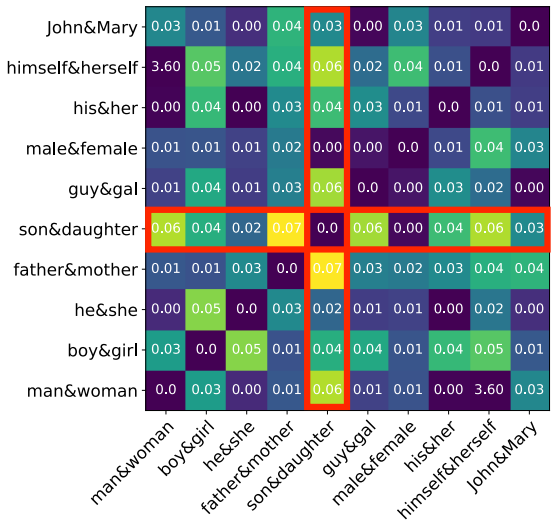
Existing post-processing algorithms, including the seminal Hard Debias (Bolukbasi et al., 2016), debias embeddings by removing the component that corresponds to a gender direction as defined by a list of gendered words. While Bolukbasi et al. (2016) demonstrates that such methods alleviate gender bias in word analogy tasks, Gonen and Goldberg (2019) argue that the effectiveness of these efforts is limited, as the gender bias can still be recovered from the geomrtry of the debiased embeddings.

We hypothesize that it is difficult to isolate the gender component of word embeddings in the manner employed by existing post-processing methods. For example, Gong et al. (2018); Mu and Viswanath (2018) show that word frequency significantly impact the geometry of word embeddings. Consequently, popular words and rare words cluster in different subregions of the embedding space, despite the fact that words in these clusters are not semantically similar. This can degrade the ability of component-based methods for debiasing gender.

[*]This research was conducted during the author's internship at Salesforce Research.



(a) Change the frequency of "boy".



(b) Change the frequency of "daughter".

Figure 1: Δ of cosine similarities between gender difference vectors before / after adjusting the frequency of word w. When the frequency of w changes, the cosine similarities between the gender difference vector ($\overrightarrow{v}$) for w and other gender difference vectors exhibits a large change. This demonstrates that frequency statistics for w have a strong influence on the the gender direction represented by $\overrightarrow{v}$.

Specifically, recall that Hard Debias seeks to remove the component of the embeddings corresponding to the gender direction. The important assumption made by Hard Debias is that we can effectively identify and isolate this gender direction. However, we posit that **word frequency in the training corpora can twist the gender direction** and limit the effectiveness of Hard Debias.

To this end, we propose a novel debiasing algorithm called *Double-Hard Debias* that builds upon the existing Hard Debias technique. It consists of two steps. First, we project word embeddings into an intermediate subspace by subtracting component(s) related to word frequency. This mitigates the impact of frequency on the gender direction. Then we apply Hard Debias to these purified embeddings to mitigate gender bias. Mu and Viswanath (2018) showed that typically more than one dominant directions in the embedding space encode frequency features. We test the effect of each dominant direction on the debiasing performance and only remove the one(s) that demonstrated the most impact.

We evaluate our proposed debiasing method using a wide range of evaluation techniques. According to both representation level evaluation (WEAT test (Caliskan et al., 2017), the neighborhood metric (Gonen and Goldberg, 2019)) and downstream task evaluation (coreference resolution (Zhao et al., 2018a)), Double-Hard Debias outperforms all previous debiasing methods. We also evaluate the functionality of debiased embeddings on several benchmark datasets to demonstrate that Double-Hard Debias effectively mitigates gender bias without sacrificing the quality of word embeddings[1].

2 Motivation

Current post-hoc debiasing methods attempt to reduce gender bias in word embeddings by subtracting the component associated with gender from them. Identifying the gender direction in the word embedding space requires a set of gender word pairs, $\mathcal{P}$, which consists of "she & he", "daughter & son", etc. For every pair, for example "boy & girl", the difference vector of the two embeddings is expected to approximately capture the gender direction:

$$\overrightarrow{v}_{boy,girl} = \overrightarrow{w}_{boy} - \overrightarrow{w}_{girl} \quad (1)$$

Bolukbasi et al. (2016) computes the first principal component of ten such difference vectors and use that to define the gender direction.[2]

Recent works (Mu and Viswanath, 2018; Gong et al., 2018) show that word frequency in a training

[1]Code and data are available at https://github.com/uvavision/Double-Hard-Debias.git

[2]The complete definition of $\mathcal{P}$ is: "woman & man", "girl & boy", "she & he", "mother & father", "daughter & son", "gal & guy", "female & male", "her & his", "herself & himself", and "Mary & John" (Bolukbasi et al., 2016).

corpus can degrade the quality of word embeddings. By carefully removing such frequency features, existing word embeddings can achieve higher performance on several benchmarks after fine-tuning. We hypothesize that such word frequency statistics also interferes with the components of the word embeddings associated with gender. In other words, frequency-based features learned by word embedding algorithms act as harmful noise in the previously proposed debiasing techniques.

To verify this, we first retrain GloVe (Pennington et al., 2014) embeddings on the one billion English word benchmark (Chelba et al., 2013) following previous work (Zhao et al., 2018b; Kaneko and Bollegala, 2019). We obtain ten difference vectors for the gendered pairs in $\mathcal{P}$ and compute pairwise cosine similarity. This gives a similarity matrix $\mathcal{S}$ in which $\mathcal{S}_{p_i,p_j}$ denotes the cosine similarity between difference vectors $\overrightarrow{v}_{pair_i}$ and $\overrightarrow{v}_{pair_j}$.

We then select a specific word pair, e.g. "boy" & "girl", and augment the corpus by sampling sentences containing the word "boy" twice. In this way, we produce a new training corpus with altered word frequency statistics for "boy". The context around the token remains the same so that changes to the other components are negligible. We retrain GloVe with this augmented corpus and get a set of new offset vectors for the gendered pairs $\mathcal{P}$. We also compute a second similarity matrix $\mathcal{S}'$ where $\mathcal{S}'_{p_i,p_j}$ denotes the cosine similarity between difference vectors $\overrightarrow{v}'_{pair_i}$ and $\overrightarrow{v}'_{pair_j}$.

By comparing these two similarity matrices, we analyze the effect of changing word frequency statistics on gender direction. Note that the offset vectors are designed for approximating the gender direction, thus we focus on the changes in offset vectors. Because statistics were altered for "boy", we focus on the difference vector $\overrightarrow{v}_{boy,girl}$ and make two observations. First, the norm of $\overrightarrow{v}_{boy,girl}$ has a 5.8% relative change while the norms of other difference vectors show much smaller changes. For example, the norm of $\overrightarrow{v}_{man,woman}$ only changes by 1.8%. Second, the cosine similarities between $\overrightarrow{v}_{boy,girl}$ and other difference vectors also show more significant change, as highlighted by the red bounding box in Figure 1a. As we can see, the frequency change of "boy" leads to deviation of the gender direction captured by $\overrightarrow{v}_{boy,girl}$. We observe similar phenomenon when we change the frequency of the word "daughter" and present these results in Figure 1b.

Based on these observations, we conclude that word frequency plays an important role in gender debiasing despite being overlooked by previous works.

3 Method

In this section, we first summarize the terminology that will be used throughout the rest of the paper, briefly review the Hard Debias method, and provide background on the neighborhood evaluation metric. Then we introduce our proposed method: Double-Hard Debias.

3.1 Preliminary Definitions

Let W be the vocabulary of the word embeddings we aim to debias. The set of word embeddings contains a vector $\overrightarrow{w} \in \mathbb{R}^n$ for each word $w \in W$. A subspace B is defined by k orthogonal unit vectors $B = \{b_1, \ldots, b_k\} \in \mathbb{R}^d$. We denote the projection of vector v on B by

$$v_B = \sum_{j=1}^{k} (v \cdot b_j) b_j. \tag{2}$$

Following (Bolukbasi et al., 2016), we assume there is a set of gender neutral words $N \subset W$, such as "doctor" and "teacher", which by definition are not specific to any gender. We also assume a pre-defined set of n male-female word pairs $D_1, D_2, \ldots, D_n \subset W$, where the main difference between each pair of words captures *gender*.

Hard Debias. The Hard Debias algorithm first identifies a subspace that captures gender bias. Let

$$\mu_i := \sum_{w \in D_i} \overrightarrow{w}/|D_i|. \tag{3}$$

The bias subspace B is the first k (≥ 1) rows of SVD($\mathbf{C}$), where

$$\mathbf{C} := \sum_{i=1}^{m} \sum_{w \in D_i} (\overrightarrow{w} - \mu_i)^T (\overrightarrow{w} - \mu_i)/|D_i| \tag{4}$$

Following the original implementation of Bolukbasi et al. (2016), we set $k = 1$. As a result the subspace B is simply a gender direction.[3]

Hard Debias then neutralizes the word embeddings by transforming each $\overrightarrow{w}$ such that every word

[3] Bolukbasi et al. (2016) normalize all embeddings. However, we found it is unnecessary in our experiments. This is also mentioned in Ethayarajh et al. (2019)

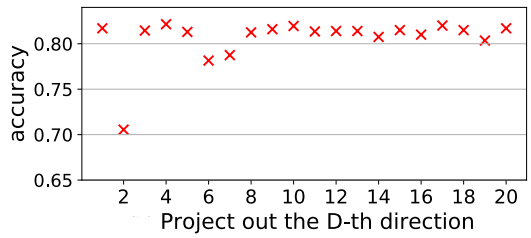


Figure 2: Clustering accuracy after projecting out D-th dominating direction and applying Hard Debias. Lower accuracy indicates less bias.

$w \in N$ has zero projection in the gender subspace. For each word $w \in N$, we re-embed $\overrightarrow{w}$:

$$\overrightarrow{w} := \overrightarrow{w} - \overrightarrow{w}_B \tag{5}$$

Neighborhood Metric. The Neighborhood Metric proposed by (Gonen and Goldberg, 2019) is a bias measurement that does not rely on any specific gender direction. To do so it looks into similarities between words. The bias of a word is the proportion of words with the same gender bias polarity among its nearest neighboring words.

We selected k of the most biased male and females words according to the cosine similarity of their embedding and the gender direction computed using the word embeddings prior to bias mitigation. We use W_m and W_f to denote the male and female biased words, respectively. For $w_i \in W_m$, we assign a ground truth gender label $g_i = 0$. For $w_i \in W_f$, $g_i = 1$. Then we run KMeans ($k = 2$) to cluster the embeddings of selected words $\hat{g}_i = KMeans(\overrightarrow{w}_i)$, and compute the alignment score a with respect to the assigned ground truth gender labels:

$$a = \frac{1}{2k} \sum_{i=1}^{2k} \mathbb{1}[\hat{g}_i == g_i] \tag{6}$$

We set $a = \max(a, 1-a)$. Thus, a value of 0.5 in this metric indicates perfectly unbiased word embeddings (i.e. the words are randomly clustered), and a value closer to 1 indicates stronger gender bias.

3.2 Double-Hard Debiasing

According to Mu and Viswanath (2018), the most statistically dominant directions of word embeddings encode word frequency to a significant extent. Mu and Viswanath (2018) removes these frequency features by centralizing and subtracting components along the top D dominant directions from the original word embeddings. These post-processed embedddings achieve better performance on several benchmark tasks, including word similarity, concept categorization, and word analogy. It is also suggested that setting D near $d/100$ provides maximum benefit, where d is the dimension of a word embedding.

We speculate that most the dominant directions also affect the geometry of the gender space. To address this, we use the aforementioned clustering experiment to identify whether a direction contains frequency features that alter the gender direction.

More specifically, we first pick the top biased words (500 male and 500 female) identified using the original GloVe embeddings. We then apply PCA to all their word embeddings and take the top principal components as candidate directions to drop. For every candidate direction $\mathbf{u}$, we project the embeddings into a space that is orthogonal to $\mathbf{u}$. In this intermediate subspace, we apply Hard Debias and get debiased embeddings. Next, we cluster the debiased embeddings of these words

Algorithm 1: Double-Hard Debias.

Input : Word embeddings: $\{\overrightarrow{w} \in \mathbb{R}^d, w \in \mathcal{W}\}$
Male biased words set: W_m
Female biased words set: W_f

$S_{debias} = []$
Decentralize $\overrightarrow{w}$: $\mu \leftarrow \frac{1}{|\mathcal{V}|} \sum_{w \in \mathcal{V}} \overrightarrow{w}$, for each $\overrightarrow{w} \in \mathcal{W}$, $\tilde{w} \leftarrow \overrightarrow{w} - \mu$;
Compute principal components by PCA: $\{\mathbf{u}_1 \dots \mathbf{u}_d\} \leftarrow \text{PCA}(\{\tilde{w}, w \in \mathcal{W}\})$;
//discover the frequency directions
for $i = 1$ *to* d **do**
  $w'_m \leftarrow \tilde{w_m} - (\mathbf{u}_i^T w_m)\mathbf{u}_i$;
  $w'_f \leftarrow \tilde{w_f} - (\mathbf{u}_i^T w_f)\mathbf{u}_i$;
  $\hat{w}_m \leftarrow HardDebias(w'_m)$;
  $\hat{w}_f \leftarrow HardDebias(w'_f)$;
  $output = KMeans([\hat{w}_m \hat{w}_f])$;
  a = eval(output, W_m, W_f);
  S_{debias}.append(a);
end
$k = \arg\min_i S_{debias}$;
// remove component on frequency direction
$w' \leftarrow \tilde{w} - (\mathbf{u}_k^T w)\mathbf{u}_k$;
// remove components on gender direction
$\hat{w} \leftarrow HardDebias(w')$;

Output : Debiased word embeddings: $\{\hat{w} \in \mathbb{R}^d, w \in \mathcal{W}\}$

and compute the gender alignment accuracy (Eq. 6). This indicates whether projecting away direction $\mathbf{u}$ improves the debiasing performance. Algorithm 1 shows the details of our method in full.

We found that for GloVe embeddings pre-trained on Wikipedia dataset, elimination of the projection along the second principal component significantly decreases the clustering accuracy. This translates to better debiasing results, as shown in Figure 2. We further demonstrate the effectiveness of our method for debaising using other evaluation metrics in Section 4.

4 Experiments

In this section, we compare our proposed method with other debiasing algorithms and test the functionality of these debiased embeddings on word analogy and concept categorization task. Experimental results demonstrate that our method effectively reduces bias to a larger extent without degrading the quality of word embeddings.

4.1 Dataset

We use 300-dimensional GloVe (Pennington et al., 2014) [4] embeddings pre-trained on the 2017 January dump of English Wikipedia[5], containing $322,636$ unique words. To identify the gender direction, we use 10 pairs of definitional gender words compiled by (Bolukbasi et al., 2016)[6].

4.2 Baselines

We compare our proposed method against the following baselines:

GloVe: the pre-trained GloVe embeddings on Wikipedia dataset described in 4.1. GloVe is widely used in various NLP applications. This is a non-debiased baseline for comparision.

GN-GloVe: We use debiased Gender-Neutral GN-GloVe embeddings released by the original authors (Zhao et al., 2018b). GN-GloVe restricts gender information in certain dimensions while neutralizing the rest dimensions.

GN-GloVe(w_a): We exclude the gender dimensions from GN-GloVe. This baseline tries to completely remove gender.

GP-GloVe: We use debiased embeddings released by the original authors (Kaneko and Bollegala, 2019). Gender-preserving Debiasing attempts to preserve non-discriminative gender information, while removing stereotypical gender bias.

GP-GN-GloVe:: This baseline applies Gender-preserving Debiasing on already debaised GN-GloVe embeddings. We also use debiased embeddings provided by authors.

Hard-GloVe: We apply Hard Debias introduced in (Bolukbasi et al., 2016) on GloVe embeddings. Following the implementation provided by original authors, we debias netural words and preserve the gender specific words.

Strong Hard-GloVe: A variant of Hard Debias where we debias all words instead of avoiding gender specific words. This seeks to entirely remove gender from GloVe embeddings.

Double-Hard GloVe: We debias the pre-trained GloVe embeddings by our proposed Double-Hard Debias method.

4.3 Evaluation of Debiasing Performance

We demonstrate the effectiveness of our debiasing method for downstream applications and according to general embedding level evaluations.

4.3.1 Debiasing in Downstream Applications

Coreference Resolution. Coreference resolution aims at identifying noun phrases referring to the same entity. Zhao et al. (2018a) identified gender bias in modern coreference systems, e.g. "doctor" is prone to be linked to "he". They also introduce a new benchmark dataset WinoBias, to study gender bias in coreference systems.

WinoBias provides sentences following two prototypical templates. Each type of sentences can be divided into a pro-stereotype (PRO) subset and a antistereotype (ANTI) subset. In the PRO subset, gender pronouns refer to professions dominated by the same gender. For example, in sentence "The physician hired the secretary because he was overwhelmed with clients.", "he" refers to "physician", which is consistent with societal stereotype. On the other hand, the ANTI subset consists of same sentences, but the opposite gender pronouns. As such, "he" is replaced by "she" in the aforementioned example. The hypothesis is that gender cues may distract a coreference model. We consider a system to be gender biased if it performs better in pro-stereotypical scenarios than in anti-stereotypical scenarios.

[4] Experiments on Word2Vec are included in the appendix.
[5] https://github.com/uclanlp/gn_glove
[6] https://github.com/tolga-b/debiaswe

Embeddings	OntoNotes	PRO-1	ANTI-1	Avg-1	|Diff-1|	PRO-2	ANTI-2	Avg-2	|Diff-2|
GloVe	**66.5**	77.7	48.2	**62.9**	29.0	82.7	67.5	75.1	15.2
GN-GloVe	66.1	68.4	56.5	62.5	12.0	78.2	71.3	74.7	6.9
GN-GloVe(w_a)	66.4	66.7	56.6	61.6	10.2	79.0	72.3	75.7	6.7
GP-GloVe	66.1	72.0	52.0	62.0	20.0	78.5	70.0	74.3	8.6
GP-GN-GloVe	66.3	70.0	54.5	62.0	15.0	79.9	70.7	75.3	9.2
Hard-GloVe	66.2	72.3	52.7	62.6	19.7	80.6	78.3	79.4	2.3
Strong Hard-GloVe	66.0	69.0	58.6	63.8	10.4	82.2	78.6	80.4	3.6
Double-Hard GloVe	66.4	66.0	58.3	62.2	**7.7**	85.4	84.5	**85.0**	**0.9**

Table 1: F1 score (%) of coreference systems on OntoNotes test set and WinoBias dataset. |Diff | represents the performance gap between pro-stereotype (PRO) subset and anti-stereotype (ANTI) subset. Coreference system trained on our Double-Hard GloVe embeddings has the smallest |Diff | values, suggesting less gender bias.

We train an end-to-end coreference resolution model (Lee et al., 2017) with different word embeddings on OntoNotes 5.0 training set and report the performance on WinoBias dataset. Results are presented in Table1. Note that absolute performance difference (Diff) between the PRO set and ANTI set connects with gender bias. A smaller Diff value indicates a less biased coreference system. We can see that on both types of sentences in WinoBias, Double-Hard GloVe achieves the smallest Diff compared to other baselines. This demonstrates the efficacy of our method. Meanwhile, Double-Hard GloVe maintains comparable performance as GloVe on OntoNotes test set, showing that our method preserves the utility of word embeddings. It is also worth noting that by reducing gender bias, Double-Hard GloVe can significantly improve the average performance on type-2 sentences, from 75.1% (GloVe) to 85.0%.

4.3.2 Debiasing at Embedding Level

The Word Embeddings Association Test (WEAT). WEAT is a permutation test used to measure the bias in word embeddins. We consider male names and females names as attribute sets and compute the differential association of two sets of target words[7] and the gender attribute sets. We report effect sizes (d) and p-values (p) in Table2. The effect size is a normalized measure of how separated the two distributions are. A higher value of effect size indicates larger bias between target words with regard to gender. p-values denote if the bias is significant. A high p-value (larger than 0.05) indicates the bias is insignificant. We refer readers to Caliskan et al. (2017) for more details.

As shown in Table 2, across different target words sets, Double-Hard GloVe consistently outperforms other debiased embeddings. For Career & Family and Science & Arts, Double-Hard GloVe reaches the lowest effect size, for the latter one, Double-Hard GloVe successfully makes the bias insignificant (p-value > 0.05). Note that in WEAT test, some debiasing methods run the risk of amplifying gender bias, e.g. for Math & Arts words, the bias is significant in GN-GloVe while it is insignificant in original GloVe embeddings. Such concern does not occur in Double-Hard GloVe.

Neighborhood Metric. (Gonen and Goldberg, 2019) introduces a neighborhood metric based on clustering. As described in Sec 3.1, We take the top k most biased words according to their cosine similarity with gender direction in the original GloVe embedding space[8]. We then run k-Means to cluster them into two clusters and compute the alignment accuracy with respect to gender, results are presented in Table 3. We recall that in this metric, a accuracy value closer to 0.5 indicates less biased word embeddings.

Using the original GloVe embeddings, k-Means can accurately cluster selected words into a male group and a female group, suggesting the presence of a strong bias. Hard Debias is able to reduce bias in some degree while other baselines appear to be less effective. Double-Hard GloVe achieves the lowest accuracy across experiments clustering top 100/500/1000 biased words, demonstrating that the proposed technique effectively reduce gender bias. We also conduct tSNE (van der Maaten and Hinton, 2008) projection for all baseline embed-

[7] All word lists are from Caliskan et al. (2017). Because GloVeembeddings are uncased, we use lower cased people names and replace "bill" with "tom" to avoid ambiguity.

[8] To be fair, we exclude all gender specific words used in debiasing, so Hard-GloVe and Strong Hard-GloVe have same acurracy performance in Table 3

Embeddings	Career & Family		Math & Arts		Science & Arts	
	d	p	d	p	d	p
GloVe	1.81	0.0	0.55	0.14	0.88	0.04
GN-GloVe	1.82	0.0	1.21	$6e^{-3}$	1.02	0.02
GN-GloVe(w_a)	1.76	0.0	1.43	$1e^{-3}$	1.02	0.02
GP-GloVe	1.81	0.0	0.87	0.04	0.91	0.03
GP-GN-GloVe	1.80	0.0	1.42	$1e^{-3}$	1.04	0.01
Hard-GloVe	1.55	$2e^{-4}$	0.07	0.44	0.16	0.62
Strong Hard-GloVe	1.55	$2e^{-4}$	0.07	0.44	0.16	0.62
Double-Hard GloVe	1.53	$2e^{-4}$	0.09	0.57	0.15	0.61

Table 2: WEAT test of embeddings before/after Debiasing. The bias is insignificant when p-value, $p > 0.05$. Lower effective size (d) indicates less gender bias. Significant gender bias related to Career & Family and Science & Arts words is effectively reduced by Double-Hard GloVe. Note for Math & Arts words, gender bias is insignificant in original GloVe.

dings. As shown in Figure 3, original non-debiased GloVe embeddings are clearly projected to different regions. Double-Hard GloVe mixes up male and female embeddings to the maximum extent compared to other baselines, showing less gender information can be captured after debiasing.

Embeddings	Top 100	Top 500	Top 1000
GloVe	100.0	100.0	100.0
GN-GloVe	100.0	100.0	99.9
GN-GloVe(w_a)	100.0	99.7	88.5
GP-GloVe	100.0	100.0	100.0
GP-GN-GloVe	100.0	100.0	99.4
(Strong) Hard GloVe	59.0	62.1	68.1
Double-Hard GloVe	**51.5**	**55.5**	**59.5**

Table 3: Clustering Accuracy (%) of top 100/500/1000 male and female words. Lower accuracy means less gender cues can be captured. Double-Hard GloVe consistently achieves the lowest accuracy.

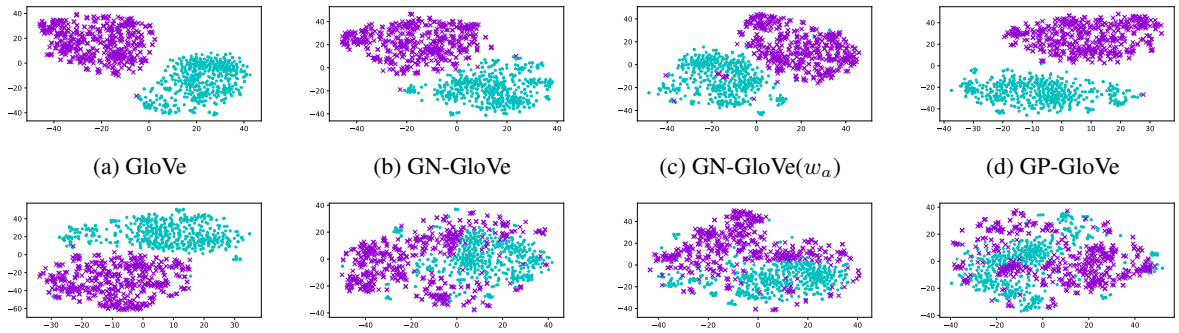
4.4 Analysis of Retaining Word Semantics

Word Analogy. Given three words A, B and C, the analogy task is to find word D such that "A is to B as C is to D". In our experiments, D is the word that maximize the cosine similarity between D and $C - A + B$. We evaluate all non-debiased and debiased embeddings on the MSR (Mikolov et al., 2013c) word analogy task, which contains 8000 syntactic questions, and on a second Google word analogy (Mikolov et al., 2013a) dataset that contains $19,544$ (**Total**) questions, including $8,869$ semantic (**Sem**) and $10,675$ syntactic (**Syn**) questions. The evaluation metric is the percentage of questions for which the correct answer is assigned the maximum score by the algorithm. Results are shown in Table4. Double-Hard GloVe achieves comparable good results as GloVe and slightly outperforms some other debiased embeddings. This proves that Double-Hard Debias is capable of preserving proximity among words.

Concept Categorization. The goal of concept categorization is to cluster a set of words into different categorical subsets. For example, "sandwich" and "hotdog" are both food and "dog" and "cat" are animals. The clustering performance is evaluated in terms of purity (Manning et al., 2008) - the fraction of the total number of the words that are correctly classified. Experiments are conducted on four benchmark datasets: the Almuhareb-Poesio (AP) dataset (Almuhareb, 2006); the ESSLLI 2008 (Baroni et al., 2008); the Battig 1969 set (Battig and Montague, 1969) and the BLESS dataset (Baroni and Lenci, 2011). We run classical Kmeans algorithm with fixed k. Across four datasets, the performance of Double-Hard GloVe is on a par with GloVe embeddings, showing that the proposed debiasing method preserves useful semantic information in word embeddings. Full results can be found in Table4.

5 Related Work

Gender Bias in Word Embeddings. Word embeddings have been criticized for carrying gender bias. Bolukbasi et al. (2016) show that word2vec (Mikolov et al., 2013b) embeddings trained on the Google News dataset exhibit occupational stereotypes, e.g. "programmer" is closer to "man" and "homemaker" is closer to "woman". More recent works (Zhao et al., 2019; Kurita et al., 2019; Basta



(a) GloVe (b) GN-GloVe (c) GN-GloVe(w_a) (d) GP-GloVe

(e) GP-GN-GloVe (f) Hard-GloVe (g) Strong Hard-GloVe (h) Double-Hard GloVe

Figure 3: tSNE visualization of top 500 most male and female embeddings. Double-Hard GloVe mixes up two groups to the maximum extent, showing less gender information is encoded.

Embeddings	**Analogy**				**Concept Categorization**			
	Sem	Syn	Total	MSR	AP	ESSLI	Battig	BLESS
GloVe	80.5	62.8	70.8	54.2	55.6	72.7	51.2	81.0
GN-GloVe	77.7	61.6	68.9	51.9	56.9	70.5	49.5	85.0
GN-GloVe(w_a)	77.7	61.6	68.9	51.9	56.9	75.0	51.3	82.5
GP-GloVe	80.6	61.7	70.3	51.3	56.1	75.0	49.0	78.5
GP-GN-GloVe	77.7	61.7	68.9	51.8	61.1	72.7	50.9	77.5
Hard-GloVe	80.3	62.5	70.6	54.0	62.3	79.5	50.0	84.5
Strong Hard-GloVe	78.6	62.4	69.8	53.9	64.1	79.5	49.2	84.5
Double-Hard GloVe	80.9	61.6	70.4	53.8	59.6	72.7	46.7	79.5

Table 4: Results of word embeddings on word analogy and concept categorization benchmark datasets. Performance (x100) is measured in accuracy and purity, respectively. On both tasks, there is no significant degradation of performance due to applying the proposed method.

et al., 2019) demonstrate that contextualized word embeddings also inherit gender bias.

Gender bias in word embeddings also propagate to downstream tasks, which substantially affects predictions. Zhao et al. (2018a) show that coreference systems tend to link occupations to their stereotypical gender, e.g. linking "doctor" to "he" and "nurse" to "she". Stanovsky et al. (2019) observe that popular industrial and academic machine translation systems are prone to gender biased translation errors.

Recently, Vig et al. (2020) proposed causal mediation analysis as a way to interpret and analyze gender bias in neural models.

Debiasing Word Embeddings. For contextualized embeddings, existing works propose task-specific debiasing methods, while in this paper we focus on more generic ones. To mitigate gender bias, Zhao et al. (2018a) propose a new training approach which explicitly restricts gender information in certain dimensions during training. While this method separates gender information from embeddings, retraining word embeddings on massive corpus requires an undesirably large amount of resources. Kaneko and Bollegala (2019) tackles this problem by adopting an encoder-decoder model to re-embed word embeddings. This can be applied to existing pre-trained embeddings, but it still requires train different encoder-decoders for different embeddings.

Bolukbasi et al. (2016) introduce a more simple and direct post-processing method which zeros out the component along the gender direction. This method reduces gender bias to some degree, however, Gonen and Goldberg (2019) present a series of experiments to show that they are far from delivering gender-neutral embeddings. Our work builds on top of Bolukbasi et al. (2016). We discover the important factor – word frequency – that limits the effectiveness of existing methods. By carefully eliminating the effect of word frequency, our method is able to significantly improve debiasing performance.

6 Conclusion

We have discovered that simple changes in word frequency statistics can have an undesirable impact on the debiasing methods used to remove gender bias from word embeddings. Though word frequency statistics have until now been neglected in previous gender bias reduction work, we propose Double-Hard Debias, which mitigates the negative effects that word frequency features can have on debiasing algorithms. We experiment on several benchmarks and demonstrate that our Double-Hard Debias is more effective on gender bias reduction than other methods while also preserving the quality of word embeddings suitable for the downstream applications and embedding-based word analogy tasks. While we have shown that this method significantly reduces gender bias while preserving quality, we hope that this work encourages further research into debiasing along other dimensions of word embeddings in the future.

References

Abdulrahman Almuhareb. 2006. *Attributes in lexical acquisition*. Ph.D. thesis, University of Essex, Colchester, UK.

Marco Baroni, Stefan Evert, and Alessandro Lenci. 2008. Bridging the gap between semantic theory and computational simulations: Proceedings of the esslli workshop on distributional lexical semantics.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783.

William F. Battig and William E. Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, 80(3p2):1.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

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

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

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. In *NAACL-HLT*.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1334–1345. Curran Associates, Inc.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *CoRR*, abs/1906.00742.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 188–197. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *EMNLP*.

A Appendices

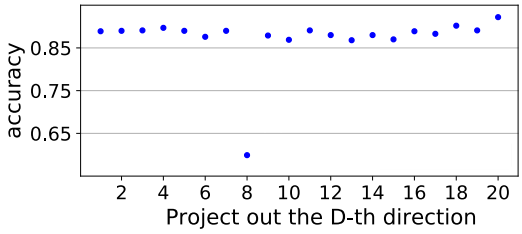


Figure 4: Clustering accuracy after projecting out D-th dominating direction and applying Hard Debias. Lower accuracy indicates less bias.

Embeddings	Top 100	Top 500	Top 1000
Word2Vec	100.0	99.3	99.3
Hard-Word2Vec	79.5	74.3	79.8
Double-Hard Word2Vec	**71.0**	**52.3**	**56.7**

Table 5: Clustering Accuracy(%) of top 100/500/1000 male and female words. Lower accuracy means less gender cues captured. Double-Hard Word2Vec consistently achieves the lowest accuracy.

We also apply Double-Hard Debias on Word2Vec embeddings (Mikolov et al., 2013b) which have been widely used by many NLP applications. As shown in Figure 4, our algorithm is able to identify that the eighth principal component significantly affects the debiasing performance.

Similarly, we first project away the identified direction $\mathbf{u}$ from the original Word2Vec embeddings and then apply Hard Debias algorithm. We compare embeddings debiased by our method with the original Word2Vec embeddings and Hard-Word2Vec embeddings.

Table 5 reports the experimental result using the neighborhood metric. Across three experiments where we cluster top 100/500/1000 male and female words, Double-Hard Word2Vec consistently achieves the lowest accuracy . Note that neighborhood metric reflects gender information that can be captured by the clustering algorithm. Experimental result validates that our method can further improve Hard Debias algorithm. This is also verified in Figure 5 where we conduct tSNE visualization of top 500 male and female embeddings. While the original Word2Vec embeddings clearly locate separately into two groups corresponding to different genders, this phenomenon becomes less obvious after applying our debiasing method.

We further evaluate the debiasing outcome with WEAT test. Similar to experiments on GloVe em-

Embeddings	Career & Family		Math & Arts		Science & Arts	
	d	p	d	p	d	p
Word2Vec	1.89	0.0	1.82	0.0	1.57	$2e^{-4}$
Hard-Word2Vec	1.80	0.0	1.57	$7e^{-5}$	0.83	0.05
Double-Hard Word2Vec	1.73	0.0	1.51	$5e^{-4}$	0.68	0.09

Table 6: WEAT test of embeddings before/after Debiasing. The bias is insignificant when p-value, $p > 0.05$. Lower effective size (d) indicates less gender bias. Across all target words sets, Double-Hard Word2Vec leads to the smallest effective size. Specifically, for Science & Arts words, Double-Hard Word2Vec successfully reaches a bias insignificant state ($p = 0.09$).

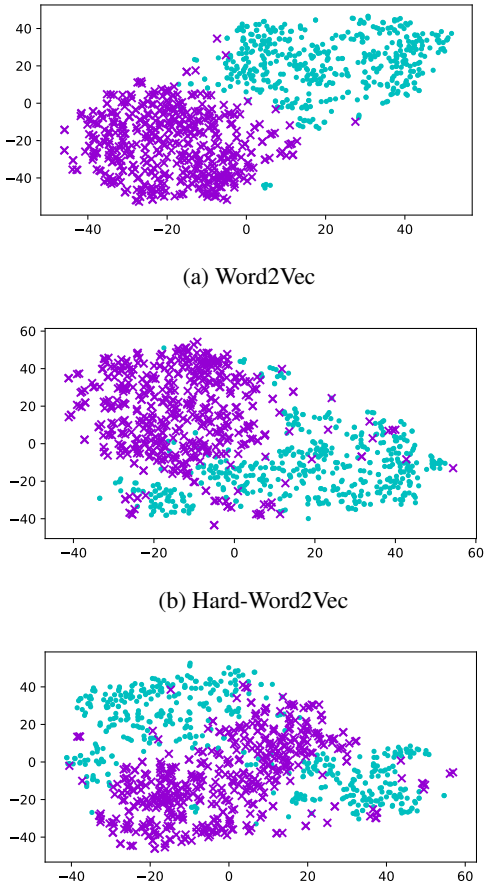
Embeddings	Analogy				Concept Categorization			
	Sem	Syn	Total	MSR	AP	ESSLI	Battig	BLESS
Word2Vec	**24.8**	**66.5**	**55.3**	73.6	**64.5**	75.0	46.3	**78.9**
Hard-Word2Vec	23.8	66.3	54.9	73.5	62.7	75.0	**47.1**	77.4
Double-Hard Word2Vec	23.5	66.3	54.9	**74.0**	63.2	75.0	46.5	77.9

Table 7: Results of word embeddings on word analogy and concept categorization benchmark datasets. Performance (x100) is measured in accuracy and purity, respectively. On both tasks, there is no significant degradation of performance due to applying the proposed method.

beddings, we use male names and female names as attribute sets and analyze the association between attribute sets and three target sets. We report effective size and p-value in Table 6. Across three target sets, Double-Hard Word2Vec is able to consistently reduce the effect size. More importantly, the bias related to Science & Arts words becomes insignificant after applying our debiasing method.

To test the functionality of debiased embeddings, we again conduct experiments on word analogy and concept categorization tasks. Results are included in Table 7. We demonstrate that our proposed debiasing method brings no significant performance degradation in these two tasks.

To summarize, experiments on Word2Vec embeddings also support our conclusion that the proposed Double-Hard Debiasing reduces gender bias to a larger degree while is able to maintain the semantic information in word embeddings.



(a) Word2Vec

(b) Hard-Word2Vec

(c) Double-Hard Word2Vec

Figure 5: tSNE visualization of top 500 most male and female embeddings. Double-Hard Word2Vec mixes up two groups to the maximum extent, showing less gender information encoded.