Replicating the Double-Hard Gender Debiasing Algorithm

Alan Ding
Princeton University
ajding@

Oleg Golev
Princeton University
ogolev@

Alik Zalmover
Princeton University
zalmover@princeton.edu

Abstract

Word embeddings computed from human-generated sources tend to encode a strong discriminative gender bias which may affect model output and performance downstream. Double-Hard Debias (DHD) is one of the most recent and effective approach for gender debiasing word embeddings (Wang et al., 2020). We reproduce the DHD algorithm on GloVe embeddings as done in the paper, and to add our own baseline, we also test the algorithm on BERTbase embeddings. Additionally, we reproduce the word embedding association test (WEAT) to measure debias performance, assess gender-based clustering performance of the most biased words on different embeddings, and evaluate the debiasing algorithm using the embedding concept categorization task to see whether it preserves the distributional semantics of the word embeddings. Using GloVe embeddings, we are able to replicate values given by the paper for WEAT and find close alignment between the values we obtain and the values reported by the paper on the categorization task. However, we are unable to reproduce as low of a clustering accuracy reported by the paper for DHD GloVe embeddings. Finally, we find that DHD does not significantly debias BERTbase but appears to improve its performance on the categorization task over GloVe.

1 Introduction

Detecting gender bias in a sentence, paragraph, or speech is a challenging endeavor. Word embeddings are a powerful tool for word representation in natural language processing. Unfortunately, word embeddings often inherit the gender bias of the source corpus as some words are used more often in male or female contexts. For instance, words used to describe a profession, such as *actor* or *nurse*, are often found to have large gender bias across not just English but other languages as well (Matthews et al., 2021). It is the case that this gender bias in word embeddings propagates to further downstream computations, often amplifying the bias from the original source (Zhao et al., 2018).

The paper we replicate builds on top of the Hard Debias algorithm (Bolukbasi et al., 2016) used to debias word embeddings against this gender bias. Wang et al. (2020) proposes a simple but effective technique, Double-Hard Debias (DHD), which purifies the word embeddings against performance-hindering information learned by the embeddings (such as word frequency) prior to inferring and removing the gender subspace.

2 Related Work

2.1 Measuring and Removing Bias in Word Representations

To understand the Double-Hard Debias (DHD) algorithm, it is important to first understand Hard Debias as it was originally proposed by Bolukbasi et al. (2016). A gender subspace (more specifically, a direction, as the subspace we take is one-dimensional) is learned from the first principal component of a set of embeddings for 10 predefined male-female word pairs. The embeddings are then projected onto the subspace orthogonal to this gender direction to debias the embeddings. In other words, each word embeddings vector is transformed such that its projection onto the bias subspace becomes zero. To evaluate a debiasing algorithm, Word Embedding Association Test (WEAT) (Caliskan et al., 2017) can be used to detect remaining bias. WEAT measurement does this by comparing two sets of target words with two sets of attribute words (e.g. *male*, *female*). The closer the difference in similarity of the target words to either of the attributes, the lower the bias. That is, we are looking for measured WEAT scores to be closer or below zero.

Mu et al. (2017) paper finds that the strongest principal components encode word frequency information which, when removed, leads to better performance on several benchmark tasks. In the paper whose experiments we replicate, Wang et al. (2020) experimentally find that applying Hard Debias to GloVe embeddings with their second principal component projected out minimizes the accuracy that the clustering algorithm achieves on the task of recovering male and female clusters in the embedding space. In this paper, we make a similar finding when projecting the first principal component of BERTbase word embeddings.

2.2 Overview of the Replicated Paper

The target paper by Wang et al. (2020) runs several experiments on the following baselines:

- GloVe (non-debiased)
- GN-GloVe (debiased Gender-Neutral)
- GN-GloVe(w_a) (gender dimension excluded)
- GP-GloVe (gender-preserving debiasing)
- GP-GN-GloVe (gender-preserving debiasing on GN-GloVe)
- Hard-GloVe (Hard Debias on neutral GloVe, preserves gender-specifi wordsc)
- Strong Hard-GloVe (Hard Debias on all GloVe)
- Double-Hard GloVe (debias GloVe using DHD).

These embeddings were then evaluated in the following ways:

- Training a coreference resolution model and computing the performance difference between the pro-stereotype and anti-stereotype subsets, where smaller difference would suggest smalles gender bias
- Evaluating the efficacy of DHD using the Word Embeddings Association Test (WEAT)
- Measuring the clustering accuracy of the most biased words based on male or female association
- Computing the accuracy of concept categorization, which indicates retention of word semantics

3 Method

The GitHub repository provided by Wang et al. (2020) contains multiple Python notebooks with code to produce the paper results. We pulled down the repository, set up the environment, and re-ran the provided experiments.

The DHD algorithm itself is the important achievement of this paper, so we also replicated it as described by the authors (Figure 1). We ran our implementation on the provided GloVe embeddings. We likewise added a baseline where we computed BERTbase embeddings on the provided datasets and applied DHD on BERTbase.

```
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:
  $\{\mathbf{u}_1 \dots \mathbf{u}_d\} \leftarrow$ 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:
       {$\hat{w} \in \mathbb{R}^d, w \in \mathcal{W}$}
```

Figure 1: The algorithm proposed by Wang et al. (2020)

3.1 Embeddings

The GitHub repository provided by Wang et al. (2020) includes GloVe embeddings, as well as several pre-debiased versions of GloVe embeddings. We examine four of the latter: GN-GloVe, GP-GloVe, Hard Glove, and Double-Hard Glove. We also examine a version of Double-Hard Glove embeddings that we reconstruct by using our own implementation of the DHD algorithm as defined in the paper. Finally, we look at BERTbase embeddings and their debiased version as they are run through our own DHD algorithm.

3.2 Evaluation

- **WEAT scores.** To evaluate the extent of debiasing that occurred, we replicate WEAT on each of the embeddings we examine. The WEAT alternative hypothesis tests for the score to be greater than zero. Scores closer to zero indicate less bias. Scores at or below zero indicate no evidence for bias.

- **Clustering accuracy.** For the Glove, Double-Hard Glove, Reconsructed Double-Hard Glove, BERTbase, and Double-Hard BERTbase embeddings, we examine the accuracy that a clustering algorithm achieves on the task of recovering male and female clusters in the embedding space.

- **Concept categorization.** To evaluate the extent to which each debiased embedding has retained proximity information (and is thus still functionally useful), we assess performance on versions of a concept categorization task.

4 Results and Discussion

4.1 PCA with BERTbase

As in Wang et al. (2020), we picked the top 1000 biased words (500 male and 500 female) using the provided GloVe embeddings. By running Principal Component Analysis, we take each of the D'th component and project all the embeddings onto the subspace orthogonal to principal component $d = 1, 2, \ldots, D$. The set of projected embeddings with the lowest clustering accuracy corresponds to the principal component we ultimately project away. On GloVe embedding, we confirm that removing the second principal component produces the lowest clustering accuracy (meaning lowest perceived bias) as found by Wang et al. (2020) (Figure 2). On BERTbase embeddings, removing the first principal component reduces the most bias (Figure 3). That is, the top principal component is the direction which most dominantly encodes the gender space in BERTbase embeddings.

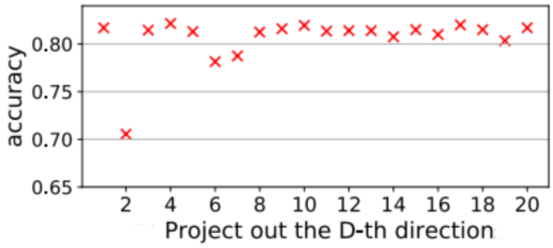


Figure 2: Clustering accuracy for top 1000 male- and female-biased words for GloVe after projecting out D-th dominating direction and applying Hard Debias. Lower accuracy indicates less bias.

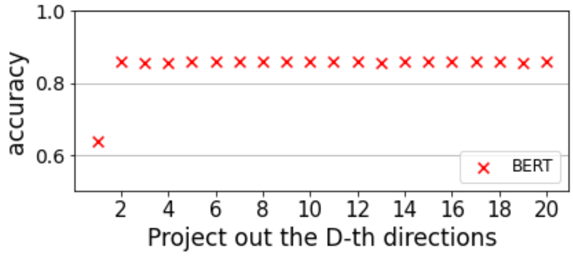


Figure 3: Clustering accuracy for top 1000 male- and female-biased words for BERTbase after projecting out D-th dominating direction and applying Hard Debias. Lower accuracy indicates less bias.

4.2 WEAT Scores

For GloVe embeddings, the numbers we obtain in Table 1 precisely match what is reported in the paper. Note that the reconstructed Double-Hard GloVe embedding performs in exactly the same way as the already-debiased Double-Hard GloVe embedding provided by Wang et al. (2020).

Observe that BERTbase is not significantly debiased by DHD (and WEAT against the *Math & Arts* and *Science & Arts* word sets fails to find bias in BERTbase itself). Generally, using BERT embeddings produces lower WEAT scores for *Career & Family* sets of words than any other embedding types. For *Math & Arts* and *Science & Arts* word sets, no bias is found in either BERTbase of Double-Hard BERTbase. This suggests BERTbase to be a better embedding to use if gender bias is of concern.

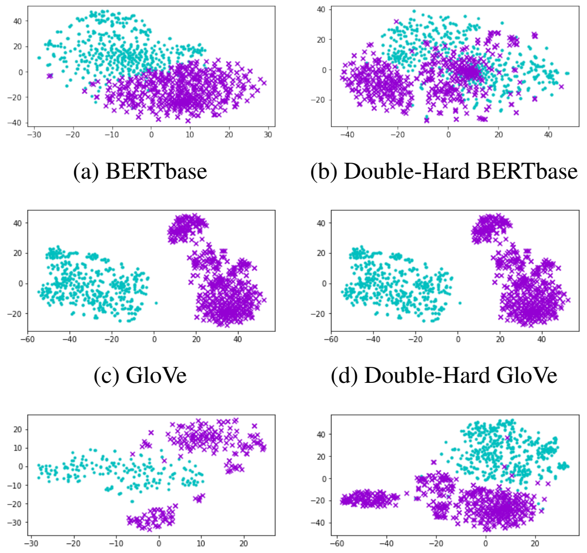
4.3 Clustering Accuracy

The Double-Hard GloVe embedding pre-debiased by (Wang et al., 2020) achieves a clustering accuracy of **62.25%** across the top 200 male-and-female-associated non-explicitly gendered words, which is higher than the reported accuracy of 51.5%

Embeddings	**Career & Family**	**Math & Arts**	**Science & Arts**
GloVe	1.8060*** (0.0000)	0.5529*** (0.1394)	0.8794* (0.0361)
GN-GloVe	1.8211*** (0.0000)	1.2069** (0.0059)	1.0244* (0.0154)
GP-GloVe	1.8059*** (0.0000)	0.8739* (0.0405)	0.9131* (0.0329)
Hard GloVe	1.5466*** (0.0002)	0.0745 (0.4425)	-0.1623 (0.6241)
Double-Hard GloVe	1.5313*** (0.0002)	-0.0944 (0.5719)	-0.1496 (0.6142)
Reconstructed Double-Hard GloVe	1.5313*** (0.0002)	-0.0944 (0.5719)	-0.1496 (0.6142)
BERTbase	1.2592*** (0.0051)	-0.6637 (0.9036)	-0.4959 (0.8335)
Double-Hard BERTbase	1.2203*** (0.0064)	-0.3016 (0.7148)	-0.2568 (0.6796)

Table 1: Computed WEAT scores on sets of target words related to different fields. A positive score indicates the presence of gender-biased associations. The p-values are provided in parentheses.

for top 100 and 55.5% for top 500. Interestingly enough, despite obtaining the same WEAT score as the pre-debiased embedding, our reconstructed Double-Hard GloVe embedding achieves an even higher clustering accuracy of **74.20%**. For BERTbase, we achieve a clustering accuracy of **90.90%** and **68.60%** before and after DHD, respectively.



(a) BERTbase

(b) Double-Hard BERTbase

(c) GloVe

(d) Double-Hard GloVe

(e) Already Debiased Double-Hard GloVe

(f) Reconstructed Double-Hard GloVe

Figure 4: tSNE visualization of top 500 most male and female embeddings.

Figure 4 presents a low-dimensional representation of the clustering algorithm run on different embeddings. The blue and purple colors indicate whether the embedding was classified as male or female respectively. The more mixed the two colors are, the less biased the embeddings are. We note that Already Debiased Double-Hard GloVe, Reconstructed Double-Hard GloVe, BERTbase, and Double-Hard BERTbase provide the lowest gender bias based on the visualizations.

4.4 Concept Categorization

For GloVe embeddings, the numbers we obtain here do not always exactly equal what Wang et al. (2020) report in their paper but are close and highly correlated (Table 2), suggesting that the different values found here are a result of stochastic variation in the evaluation tasks. We additionally find that DHD improves the performance of BERTbase on this task significantly.

5 Conclusion and Limitations

We were largely able to reproduce the numbers reported in the Wang et al. (2020) paper that we sought to replicate. We assessed the extent of debiasing with WEAT and the usefulness of the debiased embeddings with four versions of a concept categorization task that leverages these embeddings. However, we were unable to reproduce the numbers reported for the clustering accuracy of Double-Hard GloVe, nor the extent of mixing when visualizing the clusters of male and female

Embeddings	**AP**	**ESSLI**	**Battig**	**BLESS**
GloVe	59.35	72.09	49.94	81.00
GN-GloVe	56.86	68.22	48.82	85.00
GP-GloVe	56.11	68.99	49.55	78.50
Hard GloVe	63.09	74.42	51.01	84.50
Double-Hard GloVe	59.60	67.44	46.57	79.50
Reconstructed Double-Hard GloVe	59.60	67.44	46.33	79.50
BERT-Base	61.05	67.20	42.68	71.51
Double-Hard BERT-Base	73.68	70.40	48.05	76.97

Table 2: Purity of clustering performance into different categorical subsets. The evaluation was done on the Almuhareb-Poesio (AP) dataset (Almuhareb, 2006), the ESSLLI 2008 (Marco Baroni and Lenci, 2008), the Battif 1969 set (Battig and Montague, 1969-06-01), and the BLESS dataset (Baroni and Lenci, 2011), same as in the paper.

words.

WEAT fails to detect whether BERTbase is significantly debiased using DHD, but the clustering accuracy of male and female words for BERTbase decreases after DHD is applied, suggesting that debiasing has in fact occurred. BERTbase's performance on downstream concept categorization tasks improves after applying DHD as well.

Overall, we've shown that BERTbase coupled with Double-Hard Debiasing produces better performance results than any embeddings suggested by the original authors. We hope that this work encourages further research in reducing gender bias of word corpus using other dimensions of word embeddings in the future.

6 Acknowledgements

We want to thank Professor Karthik Narasimhan for his impeccable teaching, guidance and support this semester. We learned a lot, and it was a great experience.

We also want to thank our project advisor Jens Tuyls for the help and advice. Finally, many thanks to all TA's and graders in this course for making everything run smoothly and helping us on assignments.

References

Abdulrahman Almuhareb. 2006. Attributes in lexical acquisition.

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, page 1–10, USA. Association for Computational Linguistics.

William F Battig and William E Montague. 1969-06-01. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental psychology.*, 80.

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

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

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

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, Online. Association for Computational Linguistics.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.