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

Tanushree Banerjee Jessica-Ann Ereyi Kevin Castro
Princeton University {tb21, jereyi, kcapupp}@princeton.edu

Abstract

Double-Hard Debias is a technique proposed in "Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation" (Wang et al., 2020) to reduce the gender bias present in pre-trained word embeddings. In this study, we first reproduce the results of the original paper, comparing the Double-Hard debiased word embeddings with five baselines (GloVe, GP-GloVe, GN-GloVe, GP-GN-GloVe, and Hard-GloVe) using WEAT, two benchmark tasks (word analogy and concept categorization), and the neighborhood metric test. Additionally, we evaluate the proposed technique and the aforementioned baselines on Spanish GloVe embeddings to assess the extent to which these debiasing methods generalize to non-English languages. We also evaluate the debiased embeddings on an additional, more robust bias metric, RIPA. We receive similar results as the original paper on the English word embeddings. However, we find that Double-Hard Debias does not outperform Hard-Debias on the neighborhood metric test for the Spanish word embeddings. Moreover, all the debiasing methods are found to perform significantly worse on the Spanish word embeddings, suggesting that existing debiasing methods do not generalize well to languages other than English.

1 Introduction

Word embeddings pre-trained on large human-generated corpora such as the Wikipedia dump dataset are widely used in NLP systems. However, since pre-trained embeddings are derived from human-generate corpora, they often encode human gender bias, significantly affecting the reliability of NLP systems using them. For example, Bolukbasi et al., 2016, find that word2vec embeddings (Mikolov et al., 2013) trained on the Google News dataset associate the word 'programmer' more closely with 'man' and 'homemaker' with 'woman'.

Gender bias encoded into word embeddings propagates to downstream tasks such as coreference resolution models (Wang et al., 2020). Thus, it is crucial to debias these embeddings to ensure NLP systems deployed in the real world do not perpetuate human gender-based discrimination.

Past efforts to mitigate bias encoded in word embeddings include post-processing techniques such as Hard-Debias (Bolukbasi et al., 2016) as well as modified training schemes that compress gender information into a few dimensions such as Gender-Neutral Debias (Zhao, Zhou, et al., 2018) (Wang et al., 2020). Post-processing methods like Hard-Debias are more practical to implement since they are less computationally expensive and lead to fewer changes in model pipelines that already use biased pre-trained word embeddings (Bolukbasi et al., 2016).

Although Hard-Debias reduces gender bias to some extent in word analogy tasks (Bolukbasi et al., 2016), Gonen and Goldberg, 2019, demonstrate that Hard-Debias fails to completely debias embeddings. To improve the debiasing algorithm, Wang et al., 2020 modify Hard-Debias to more effectively isolate the gender dimension in the embeddings. Next, they evaluate their proposed method on the following evaluation metrics: representation level evaluation using the WEAT test (Caliskan et al., 2017), the neighborhood metric test (Gonen and Goldberg, 2019), and downstream task evaluation on coreference resolution (Zhao, Wang, et al., 2018). The Double-Hard debiased embeddings are compared against embeddings debiased using other baseline methods. Wang et al., 2020, find that their proposed method outperforms all evaluated baselines. Additionally, the semantic and syntactic information lost due to the Double-Hard debiasing method, evaluated using benchmark datasets, is found to be comparable to the information lost using other baselines.

Yet, Wang et al., 2020, do not analyse whether their method generalizes well to languages other than English. Such an analysis is crucial to ensure that models deployed in the real world are unbiased for non-English languages, especially for languages whose structure differs significantly from English. For example, Spanish has grammatical gender, causing gender to be more deeply rooted into the language and potentially making the task of debiasing embeddings harder.

In this paper, we first summarize the original paper. Next, we describe the work related to our novel contributions: the reproduction of the baselines and Double-Hard Debias methods on English and Spanish corpora as well as the evaluation on an additional, more robust bias metric that is not used by the original paper – RIPA (described in detail in the "Analysis and Discussion of Results" section). Then, we describe the method we use to reproduce the experiments and evaluations of the original paper and an analysis of our findings. Finally, we offer potential directions for future work. Our results correspond with those of Wang et al., 2020, for the English embeddings. However, our results indicate that the Double-Hard method does not outperform Hard-Debias for Spanish embeddings on one bias evaluation metric. Moreover, all debiasing methods are significantly worse at debiasing Spanish word embeddings. This indicates that existing methods do not generalize well to non-English languages.

2 Related Work

2.1 Hard-Debias (Bolukbasi et al., 2016)

The key idea behind the approach in this paper is to remove the component of the pre-trained embedding associated with gender. First, for each pair in a set of ten gendered word pairs, the difference vector is calculated as shown below.

$$v_{boy,girl} = w_{boy} - w_{girl} \tag{1}$$

Next, the first principal component of the ten difference vectors is calculated. This is defined as the 'gender direction' g in the embedding space. Finally, the biased word embeddings w are projected onto the subspace orthogonal to the computed gender direction g to get the debiased word embeddings w'. Thus, the projection of the debiased embeddings w' onto the gender direction g is 0.

This paper demonstrates that the Hard-Debias method reduces the bias found in the embeddings. However, as Gonen and Goldberg, 2019, later demonstrate, this bias removal is superficial as the gender direction can still be recovered from the debiased embeddings.

2.2 Double-Hard Debias (Wang et al., 2020)

Wang et al. hypothesize that the true gender direction is difficult to identify in the original Hard-Debias algorithm. Moreover, the work of Mu et al., 2017, and Gong et al., 2018, shows that word frequency significantly impacts the geometry of word embeddings, which in turn can impact the identification of the gender direction in the original Hard-Debias algorithm, thereby reducing the ability of the original algorithm to debias embeddings.

Inspired by these hypotheses, Wang et al., 2020, propose Double-Hard Debias – a modification to Hard-Debias that projects embeddings into an intermediate subspace independent of word frequency before applying the standard algorithm, thereby computing a more accurate gender direction. It does this by finding the dimension that encodes frequency information for the word, which distracts from the gender direction computation.

The Double-Hard Debias Algorithm: First, the principal components of all the word embeddings are computed and considered as candidates for the frequency dimension as shown below.

$$u_1 ... u_d \leftarrow PCA(\{\tilde{w}, w \in W\}) \tag{2}$$

Next, the set of most biased male and female words are selected. For each possible frequency dimension $u_i, 1 \leq i \leq d$, repeat the following three steps:

1. Project each word embedding into an intermediate space orthogonal to u_i to get the revised embeddings.

$$w'_m \leftarrow \tilde{w}_m - (u^T w_m) u_i \tag{3}$$

$$w'_f \leftarrow \tilde{w}_f - (u^T w_f) u_i \tag{4}$$

Note: (w_m,w_f) represents a pair of male and female words

2. Apply the Hard-Debias algorithm on these revised embeddings.

$$\hat{w}_m \leftarrow HardDebias(w'_m) \tag{5}$$

$$\hat{w}_f \leftarrow HardDebias(w'_f) \tag{6}$$

3. Cluster the top biased words using the embeddings from the previous step and compute their clustering accuracy as follows.

$$output = KMeans([\hat{w}_m, \hat{w}_f]) \quad (7)$$

$$a = eval(output, W_m, W_f) \quad (8)$$

The better the K-means algorithm clusters the top biased words into two gender-aligned groups, the worse the chosen u_i does in improving the degree to which the embeddings are debiased. Thus, the u_i resulting in the worst clustering accuracy a is chosen as the frequency dimension u_k, and the components along this dimension are removed from the word embeddings as follows.

$$w' \leftarrow \tilde{w} - (u_k^T w)u_k \quad (9)$$

Finally, the components of the embeddings along the gender direction are removed using the regular Hard-Debias algorithm.

$$\hat{w} \leftarrow HardDebias(w') \quad (10)$$

Wang et al., 2020, find that for GloVe embeddings pre-trained on the Wikipedia dataset (Pennington et al., 2014), removing components along the second principal component significantly decreases clustering accuracy, leading to the best debiasing results. In addition, they demonstrate the effectiveness of their technique by comparing Double-Hard debiased embeddings against other baseline debiased embeddings: GloVe, Gender-Neutral GloVe (GN-GloVe), GN-GloVe(w_a), Gender-Preserving GloVe (GP-GloVe), GP-GN-GloVe, Hard-GloVe and Strong Hard-GloVe. Each of these baseline approaches are described in detail under "Setup and Experiments". The downstream tasks on which the debiasing methods are evaluated and the evaluation metrics used to make comparisons across debiasing methods are described below.

2.2.1 Downstream tasks used for evaluations

Word analogy. Given words A, B and C, the analogy task involves finding a fourth word D such that "A is to B as C is to D", i.e. the D maximizes the cosine similarity between D and C – A + B (Wang et al., 2020). The Microsoft Research(MSR) and Google word analogy datasets(Aekula et al., 2021) are used containing both semantic and syntactic questions. The evaluation metric is the percentage of questions for which the correct answer is assigned the maximum score by the algorithm (Wang et al., 2020).

Concept categorization. This task clusters a set of words into different sub-categories. Clustering performance is evaluated on purity, i.e. the fraction of the total words correctly classified (Wang et al., 2020). Four benchmark datasets are used for evaluation: 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).

The word analogy and concept categorization tasks are used to measure the degree to which the debiased embeddings retain word semantics, allowing us to evaluate the quality of the debiased embeddings.

Coreference resolution. This task identifies noun phrases referring to the same entity. The WinoBias dataset is used as a benchmark to evaluate gender bias in coreference resolution (Zhao, Wang, et al., 2018). Performance on coreference resolution is used to evaluate the quality and usability of debiased embeddings in downstream tasks.

2.2.2 Bias evaluation metrics

The Word Embeddings Association Test (WEAT). This is a permutation test measuring the degree of significance of bias in word embeddings (Caliskan et al., 2017).

Neighborhood metric test. Gonen and Goldberg, 2019, introduces this metric to measure the degree of bias in word embeddings based on the k-means clustering accuracy with two gender-aligned clusters. An accuracy value of around 0.5 indicates gender-neutral word embeddings.

Through their analysis, Wang et al., 2020, demonstrate that their proposed method mitigates the impact of word frequency on embeddings thereby producing better debiased embeddings. In addition, their method preserves the quality of word embeddings, making them suitable for use in downstream tasks.

2.3 Work related to our contributions

Previous attempts at reproduction. There have been other attempts to reproduce the results of the original Double-Hard Debias paper such as that of Aekula et al., 2021. Aekula et al., 2021 are unable to reproduce the evaluation on the coreference resolution task of the original paper due to the poor readability of the code base for Wang et al., 2020. Moreover, Aekula et al., 2021 determine that the neighbourhood metric test is not reproducible with the information provided by Wang et al., 2020. Nevertheless, they attempt to reproduce the results of Wang et al., 2020 by filling in the missing information with their own approximations. This approach produced different results from the original paper, particularly for the t-SNE visualisations for the embeddings.

On the other hand, the benchmark tasks, word analogy and concept categorization were found to be reproducible within 0.5 percent of the values reported in Wang et al., 2020.

Given the findings of Aekula et al., 2021, we only reproduce the results of the neighborhood metric test, WEAT, and the benchmarking tasks as implementing the coreference resolution task is beyond the scope of this project and its time constraints.

Debiasing Spanish Word Embeddings. In their paper, Shin et al., 2020, investigate the efficacy of existing debiasing algorithms such as GP-Debias and Hard-Debias on Spanish and Korean fastText word embeddings. Additionally, to evaluate the non-English embeddings using the Sembias gender analogy test, Shin et al. translate the English analogy questions into the other languages using machine translation with human corrections. Shin et al. (2020) are unable to reproduce the GN-Debias and GP-GN-Debias baselines for the Spanish and Korean fastText word embeddings due to the close ties between the GloVe vectors and the implementation of the GN debiasing algorithm. To avoid this restriction, we decide to use Spanish Glove embeddings rather than the fastText embeddings proposed in Shin et al., 2020.

3 Statement of Purpose

In this paper, we aim to reproduce English embeddings debiased using the Double-Hard Debias algorithm and compare them with five additional baseline embeddings: GloVe, GP-GloVe, GN-GloVe, GP-GN-GloVe, and Hard-GloVe. We choose these five baselines since they are popularly used in NLP tasks, and hence their implementations are well documented. In addition, we investigate the effectiveness of Double-Hard Debias and the five baseline debiasing techniques on Spanish embeddings to determine whether these methods generalize well to languages other than English, particularly since Spanish is a language with grammatical gender. Finally, we use the Relational Inner Product Association (RIPA) test to evaluate our word embeddings for gender bias as RIPA is a more robust alternative to WEAT.

4 Setup and Experiments

GloVe. As in Wang et al., 2020, we use 300-dimensional GloVe embeddings pre-trained on the English Wikipedia corpus (Pennington et al., 2014). Due to Google Colab's memory constraints, we use a subset of the GloVe embeddings trained on the Spanish Billion Words Corpus (Cardellino, 2019) to derive the Spanish word embeddings.

GN-GloVe. GN-GloVe restricts the gender information in certain dimensions while removing it in the other dimensions (Wang et al., 2020). Unlike the other baselines, GN-GloVe uses a modified training scheme to produce its debiased word embeddings from human-generated corpora rather than modifying pre-trained GloVe embeddings. Since the dataset that the original English GloVe embeddings were trained on is not open-access, we instead derive our GN-Glove embeddings using the open-access 2022 Wikipedia dump. Due to Google Colab's memory constraints, we use a 1GB sample of this corpus containing 123,991 unique words to derive our embeddings. It takes about 6 hours to derive the embeddings for this subset of the full corpus. Similarly, we use a sample of 38,826 unique words from the Spanish Billion Word Corpus (Cardellino, 2019) to derive the GN-GloVe Spanish word embeddings, which takes about 3 hours to run.

Since the GN-GloVe algorithm is written in the C language, we write a shell script to run the algorithm in Google Colab.

We use the list of male-female word pairs provided in the original GN-GloVe paper (Zhao, Zhou, et al., 2018) to reproduce results for the English corpus, and use machine translation with human correction to generate the male-female word pairs for the Spanish corpus (See Appendix B).

GP-GloVe, GP-GN-GloVe. GP-GloVe preserves non-discriminative gender information, while removing stereotypical gender bias, while GP-GN-GloVe applies the gender-preserving debiasing algorithm on the debiased GN-GloVe embeddings.

To reproduce the results for these baselines, we use the code base for the original GP-GloVe paper (Kaneko and Bollegala, 2019), containing the original python code and files required (list of male-female word pairs, gender-neutral words and gender-stereotyped words). We also write a bash script to run the code on Google Colab. We run the English version of GP-GloVe and GP-GN-GloVe on the same word embeddings as Wang et al., 2020: the 300-dimensional GloVe and GN-GloVe embeddings trained on the English Wikipedia corpus (Zhao, Zhou, et al., 2018). This algorithm took approximately 2 hours to run.

For the Spanish GP-GloVe embeddings, we translate the necessary files into Spanish using machine translation. Human correction was limited due to the large size of the dataset and the time constraints of the project. Additionally, while we run the gender-preserving debiasing algorithm on the pre-trained word embeddings generated from the Spanish Billion Word Corpus, we create the Spanish GP-GN-GloVe embeddings by running the same algorithm on the Spanish GN-GloVe word embeddings that we previously derived. This algorithm also took approximately 2 hour to run.

Hard-GloVe. Hard-GloVe attempts to debias the neutral words and preserve the gender specific words.

Similar to how we reproduce GP-GloVe, we use the python and data files from the Hard-Debias GitHub repository (Bolukbasi et al., 2016), which includes two python scripts: one that learns a larger list of gender-specific words from a seed set and another that outputs the debiased word embedding given the original word embedding, a set of ten pairs of words used to define the gender direction, a list of gender-specific words, and around 50 crowd-sourced female-male word pairs that represent gender direction. We use the pre-trained GloVe vectors used in Wang et al., 2020, and other data required is in the repository. Finally, we convert the python files to a Colab notebook, which took around 10 minutes to run.

For the Spanish version of Hard-GloVe, we follow the same procedure as the English version, but use Spanish GloVe vectors and translate the data files from English to Spanish using machine translation with human correction. Running Hard-Debias on the Spanish embeddings also took 10 minutes.

Double-Hard GloVe. To reproduce the word embeddings produced by the Double-Hard Debiasing algorithm, we use the code base provided in the reproduction of the original study by Aekula et al., 2021. We use code from the reproduction of the original paper rather than the original paper since it is more readable and stores the debiased embeddings into a file unlike the code written by Wang et al., 2020. We use the same pre-trained GloVe embeddings and data files (original and translated) as Hard-Debias. Executing the code takes around 10 minutes.

4.1 Evaluations

To evaluate the five English baseline embeddings, we use the scripts provided in the GitHub of the reproduction of the Double-Hard Debias paper (Aekula et al., 2021). This repository also includes links to the MSR analogy dataset and the Google analogy dataset. Since it is challenging to run multi-file programs on Colab, we combine all files related to evaluations and their dependencies into a single Colab notebook. The execution of all the English evaluations took approximately 30 minutes.

For the evaluation of the Spanish baseline embeddings, we adapt this procedure as follows. First, we translate all of the WEAT words into Spanish, changing the most common names in the English-speaking world to the most common names in the Spanish-speaking world. Additionally, we replace the Google analogy dataset with a human translation of this dataset into Spanish (Rukua95, 2020). The same could not be done for the MSR analogy dataset due to the fact that the majority of the dataset consists of superlatives, which cannot be translated effectively into Spanish as the translations typically consist of more than one word (e.g. *rough, rougher* → *áspero, más áspero*). Thus, using the CATS analogy dataset (Rukua95, 2020), we create a Spanish version of the MSR analogy dataset by writing a python script that generates word analogy questions from pairs of words in different tenses (e.g. *aproximar aproximación autorizar autorización*). Finally, we replaced all the necessary data files with their Spanish translations.

Due to Google Colab's memory limits, we limit the size of the embedding vector files to 1.5GB for the evaluations on the Spanish embeddings. Additionally, since the datasets used for the English concept categorization task are not open-access, we are unable to reproduce the results for this task on the Spanish word embeddings.

Moreover, since the English and Spanish GN-GloVe word embeddings are trained on a subset of the data the original English and Spanish GloVe vectors were trained on, we add checks to the t-SNE visualization algorithm such that out of the 500 most gendered male and female words displayed in the visualization, the ones that do not appear in the GN-GloVe vocabulary are excluded.

We also evaluate the debiased embeddings on an additional bias metric - the Relational Inner Product Association (RIPA) test. This test is a more robust measure of the degree to which the embeddings are gender biased. Since the original paper introducing the test Ethayarajh et al., 2019, is not accessible, we are left to recreate RIPA on our own. We create a python function to find the first principal component for a set of gendered pairs similarly to Bolukbasi et al., 2016. This is defined as the relation vector that is used for an inner product of word embeddings in the same embedding space and the relation vector. The RIPA score for the debiased GloVe embeddings of the baselines are calculated and defined as the genderedness. This is compared to the starting genderedness in the corpus (Ethayarajh et al., 2019).

5 Analysis and discussion of results

The debiased embeddings are evaluated on the following characteristics:

1. the extent to which they exhibit gender bias using the neighborhood metric test, WEAT, and RIPA.

2. the degree to which semantic information is retained based on the performance on benchmarking tasks (word analogy and concept categorization)

Since we cannot access the necessary Spanish datasets, we do not reproduce the analysis on concept categorization for the Spanish embeddings.

5.1 Associations

Word Embeddings Association Test (WEAT). WEAT measures bias in word embeddings using two metrics: (1) the effect size between a target set of words and a gender attribute set and (2) the p-value giving the significance of the effect size on the word embeddings. Table 1 displays the effect size and p-value for each baseline English embedding on the target word sets of "Career & Family", "Math & Arts", and "Science & Arts". A lower effect size and a p-value > 0.05 indicate less bias. We find that the embeddings debiased by the Double-Hard algorithm outperform all of the other embeddings on the "Career & Family" word set, achieving an effect size of 1.5313. For the "Science & Arts" and the "Math & Arts" word sets, the Double-Hard Debias embeddings are beat marginally by the Hard-Debias ones, producing the second lowest effect size of 0.1496 and 0.0943 respectively. These results are near identical to those reported in the original paper with the effect sizes differing by approximately ± 0.001. Additionally, the increase in the p-value from the original GloVe embeddings to the Double-Hard Debias embeddings (0.14 to 0.57 for the "Math & Arts" set and 0.04 to 0.61 in the "Science & Arts" set) indicate that the Double-Hard algorithm was able to make the gender bias insignificant in the word embeddings. However, while Double-Hard Debias produces a low effect size for the "Career & Family" word set, it also produces a small p-value of 0.0001, suggesting that some bias remained significant even after debiasing. The original paper achieves similar p-values on the "Career & Family" word set. Intuitively, this makes sense as the discourse surrounding career and family tends to be highly gendered, more so than the other categories.

Table 1: WEAT test results of English embeddings before/after Debiasing.

WEAT scores for English Embeddings						
Embeddings	Career & Family		Math & Arts		Science & Arts	
	d	p	d	p	d	p
GloVe	1.8059	0.0	0.5528	0.14	0.8793	0.04
GN-GloVe	-	-	-	-	-	-
GP-GloVe	1.8042	0.0	0.8519	0.04	0.8441	0.04
GP-GN GloVe	1.7987	0.0	1.4145	0.001	1.0511	0.01
Hard-GloVe	1.5466	0.0001	0.0745	0.44	0.1622	0.62
DH-GloVe	1.5313	0.0001	0.0943	0.57	0.1496	0.61

The results for the Spanish embeddings are similar to those of the English embeddings. Those debiased by the Double-Hard algorithm slightly outperform the embeddings debiased by the other baselines producing the lowest effect size for the "Math & Arts" and the "Science & Arts" sets, 0.0616 and 0.0833 respectively, and the third lowest effect size for the Career & Family set (1.0316). The p-value of the Double-Hard Debias embeddings increases from 0.09 to 0.45 for the "Math & Arts" set and from 0.24 to 0.56 for the "Science & Arts" set, showing that the effect size and hence bias becomes insignificant. We see the "Career & Family" set have a p-value that is significant (0.01), implying words related to this topic are still significantly biased. Again, it is intuitive that the the effect size and its significance for the Career and Family set is high as the Spanish-speaking world has relatively strong gender norms around family and careers.

Table 2: WEAT test results of Spanish embeddings before/after Debiasing.

WEAT scores for Spanish Embeddings			
Embeddings	Career & Family	Math & Arts	Science & Arts
GloVe	1.4033 0.0002	0.6698 0.09	0.3604 0.24
GN-GloVe	- -	- -	- -
GP-GloVe	0.8713 0.03	0.31 0.30	0.3088 0.27
GP-GN GloVe	0.8713 0.03	0.30 0.30	0.3088 0.27
Hard-GloVe	1.3433 0.0	0.4965 0.17	0.2070 0.17
DH-GloVe	1.0316 0.01	0.0616 0.45	0.0833 0.56

The Relational Inner Product Association (RIPA) is a subspace projection method. Ethayarajh et al., 2019 criticize that the WEAT method's use of a cosine similarity based measurement allows for the attribute word sets used to change the gender direction the embedding can take and overestimate the association. To solve this issue, RIPA generalizes (Bolukbasi et al., 2016)'s idea of measuring bias by projecting onto $\vec{he} - \vec{she}$ by replacing the difference vector with a relation vector $\vec{b}$, where $\vec{b}$ is the first principal component of the difference vectors of a set of gender word pairs (Ethayarajh et al., 2019). This allows RIPA to adapt more effectively to the choice of word pairs that define the association than WEAT does to its attribute word sets (Ethayarajh et al., 2019). RIPA evaluates gender bias by comparing the 'genderedness' in embedding space with the genderedness in the corpus to figure out the absolute change in genderedness induced by the embedding model (Ethayarajh et al., 2019). We find that the GloVe embeddings before debiasing recieve an average RIPA score of -0.189, meaning that the model significantly increases the genderness of the words in the training corpus. However, the Double-Hard Debias provided an average RIPA score of 0.009, which indicates a drastic decrease in the genderedness induced by the model. This means that Double-Hard debiasing algorithm essentially makes the words no more gendered than they are in the training corpus. The other baseline embeddings, GN-GloVe, GP-Glove, GP-GN-GloVe and Hard-GloVe, receive RIPA scores of -0.09, -0.03, -0.024, and 0.010 respectively, indicating that Double-Hard Debias induces the least amount of genderedness in the embedding space.

The Spanish embeddings received different results. The GloVe embeddings before debiasing got an average RIPA score of 0.006 while the Double-Hard Debias embeddings achieved an average RIPA score of -0.007, which presents a marginal change in the gender induced into the embedding space. GN-GloVe, GP-Glove, GP-GN-GloVe and Hard-GloVe produce RIPA scores of 0.0067, -0.007, 0.17, and -0.001, indicating that debiasing does little to change the genderness induced in the embedding space. Moreover, the small RIPA score for the original Spanish GloVe embeddings suggest that the embedding model does not make the words any more gendered than they are in the training corpus, which is distinct from our findings for the English embeddings.

Table 3: Clustering Accuracy (%) of top 100/500/1000 male and female English words.

Neighborhood Metric for English Embeddings			
Embeddings	Top 100	Top 500	Top 1000
GloVe	100	100	100
GN-GloVe	94.11	61.1	61.01
GP-GloVe	100	100	100
GP-GN GloVe	100	100	99.95
Hard-GloVe	76	80.5	80.25
DH-GloVe	66.5	74.2	70.4

Table 4: Clustering Accuracy (%) of top 100/500/1000 male and female Spanish words.

Neighborhood Metric for Spanish Embeddings			
Embeddings	Top 100	Top 500	Top 1000
GloVe	100	100	100
GN-GloVe	96.8	69.47	70.74
GP-GloVe	100	100	99.85
GP-GN GloVe	100	100	99.95
Hard-GloVe	97.5	94.3	91.7
DH-GloVe	100	95.9	93.3

Neighborhood Metric Test A percentage score that is closer to 0.5 indicates that the embedding is less biased. All baselines produced percentage scores greater than 0.5, with some producing scores of 1, such as GloVe, GP-GloVe, and GP-GN GloVe. Embeddings debiased on the Double-Hard method had more difficulty clustering words into a male or female gender group as indicated by the lowest scores of 66.5%, 74.2%, and 70.4% for the top 100, top 500, and top 1000 most biased words, respectively. This shows that the method allowed for the gender of a word to be taken out of the most biased words to the point where the gender of the word could not be recognized, and embeddings could not be clustered well.

Although we see the Spanish embeddings debiased by Double-Hard produce effect sizes that indicate that the bias in the embedding is low, the Neighborhood Metric suggests that bias is still highly present within the embeddings as it was able to cluster the gender of the words. For the top 100 most biased words, it had a 100% score similar to the GloVe embeddings that did not go through a debiasing process. The Hard-GloVe baseline outperformed it for all three top k words . DH-GloVe had a 95.9% score for the top 500 words and 93.3% for the top 1000 words, while Hard-GloVe had a 94.9% score for the top 500 words and 91.7% for the top 1000 words. The ability to be able to cluster the words into a certain gender space indicates the bias still being present within the embeddings.

Visualization The original GloVe embeddings for English and Spanish differ greatly in the initial bias that we see present. The English embeddings are presented to have a more clear separation of gender in different regions, whereas the Spanish embeddings have the gender projected into a space in which they begin to overlap. This was surprising since Spanish is commonly known to be a more gendered language. Once debiased, the embeddings projections become more intermixed, indicating that the embeddings are encoding less gender information and bias. (See Figure 1 in Appendix A) The Spanish embeddings present similar findings. In the visualizations, there are various points of overlap for the gender spaces for many baselines. It is important to note that the non-debiased Spanish embeddings also presented the overlap of the gender regions. Therefore the reduction of bias was not so clear in this metric. GloVe, GP-Glove, Hard-GloVe, and DH-GloVe all present these results.

5.2 Semantics

Word Analogy. The biased English embeddings produced an 80.48% semantic accuracy score, 62.76% syntactic accuracy score, 70.80% total accuracy and 54.24% MSR accuracy score. The embeddings debiased by double-hard produced an 80.94% semantic accuracy score, 61.64% syntactic accuracy score, 70.40% total accuracy and 53.21% MSR accuracy score. These comparable accuracy scores reflect the fact that the Double-Hard Debias embeddings were able to retain the semantic information encoded in them.

The Spanish GloVe embedding produced a 30.75% semantic accuracy score, 43.67% syntactic accuracy score, 41.73% total accuracy and 26.14 MSR accuracy score. The embeddings debiased by Double-Hard produced an 53.65% semantic accuracy score, 45.68 syntactic accuracy score, 46.41 total accuracy and 32.06% MSR accuracy score. These high accuracy scores suggest that, similar to the English embeddings, the Spanish debiased embeddings perserv the semantic makeup of the words

Concept Categorization. Through clustering the set of words into categorical subsets, we were able to get the performance scores of the baselines. GloVe embeddings achieved a 55.6% accuracy on AP, 72.7% accuracy on ESSLI, 48% on Battig, and 81% accuracy on BLESS. The Double-Hard GloVe embeddings achieved similar accuracy scores of 58.9% on AP, 72.7% on ESSLI, 37.6% on Battig, and 79.5% on BLESS. The numbers of the accuracy scores for the Double-Hard embeddings compared to the GloVe embeddings present an insignificant difference that indicates the semantic information being preserved for the embeddings after debiasing. Due to the unavailability of the AP, ESSLI,Battig, and BLESS datasets for Spanish words, we were not able to conduct the concept categorization for the Spanish embeddings as we did on the English embeddings. Although this was the case, we see through the word analogy that the embeddings after debiasing altered the semantic information of the word embeddings as before the debiasing.

Table 5: Results of English word embeddings on word analogy and concept categorization benchmark datasets

Benchmark Tasks for English Embeddings								
Embeddings	Word Analogy				Concept Categorization			
	Sem	Syn	Total	MSR	AP	ESSLI	Battig	BLESS
GloVe	80.48	62.76	70.80	54.24	55.61	72.72	48.00	81.00
GN-GloVe	00.48	00.31	00.32	00.44	19.14	46.34	09.27	25.88
GP-GloVe	80.55	61.78	70.30	51.48	55.86	72.72	50.03	78.50
GP-GN GloVe	77.57	77.57	68.88	51.75	60.59	70.45	52.24	77.50
Hard-GloVe	80.28	62.74	70.70	54.27	61.59	79.54	50.55	84.50
DH-GloVe	80.94	61.64	70.40	53.81	58.95	72.73	37.62	79.50

6 Limitations

A major limitation of reproducing the results of the baseline debiasing algorithms was the RAM allowance of Google Colab pro, which forced us to train our embeddings on a smaller subset of data.

Table 6: Results of Spanish word embeddings on word analogy benchmark dataset

Word Analogy for Spanish Embeddings				
Embeddings	Sem	Syn	Total	MSR
GloVe	30.75	43.67	41.73	26.14
GN-GloVe	40.83	6.75	10.87	8.87
GP-GloVE	56.93	44.13	45.30	31.96
GP-GN GloVe	71.74	7.67	11.04	15.85
Hard-GloVe	36.91	43.82	43.04	26.46
DH-GloVe	53.65	45.68	46.41	32.06

This was particularly consequential when reproducing GN-GloVe, which uses a modified training scheme instead of a post-processing technique to debias the word embeddings. This means that GN-GloVe had to be trained on a small subset of the full corpus, unlike the pre-trained GloVe embeddings trained on the full corpus, consisting of hundreds of millions of words. Thus, a large number of words were missing when executing the t-SNE-visualizations and the WEAT test. Hence, the t-SNE visualizations (See figure 1 in Appendix A) and the WEAT scores are not informative for the GN-GloVe embeddings and for the GP-GN-GloVe embeddings. GP-GN-GloVE English is a notable exception as it was trained on the pre-trained debiased word embeddings generated from the original GN-GloVe paper instead of the ones we reproduced in this study.

7 Conclusion

The baselines are good methods to use on the English language, however these methods are not as effective on the Spanish language. Double-hard produced debiasing results that outperformed any other debiasing method used as a baseline for the English embeddings. The WEAT, Neighborhood metric and RIPA metric show the degree to which Double-Hard embeddings are debiased. Thus, the analysis shows that the Double-Hard Debias algorithm produces embeddings that are semantically similar with less gender bias present. The same method, however, was not able to reproduce these results for Spanish embeddings. The Neighborhood Metric and RIPA show that the debiasing method allowed for a significant bias to be present in the word embeddings after debiasing although the semantic integrity of the embeddings are retained. This suggest that the debiasing methods do not effectively target the gender direction for non-English languages, causing bias to be perserved post-debiasing.

8 Future Work

Current debiasing techniques work relatively well on English embeddings while preserving semantic and syntactic information, rendering the embeddings suitable in downstream NLP tasks. However, these techniques may not work as well on other languages. Future work on developing or modifying existing debiasing techniques to generalize well to other languages (especially ones like Spanish that are inherently more gendered and containing grammatical gender) is crucial to ensure NLP models are un-biased for languages other than English as well.

Zhou et al., 2019 propose a revised definition of gender bias in languages with grammatical gender such as Spanish (Zhou et al., 2019). Future work can extend the analysis in this paper by evaluating debiased Spanish embeddings using this revised definition of word embeddings. Another potential avenue for future work involves extending the analysis of the performance of the Double Hard debias method by comparing its performance with the use of ”bilingual word embeddings to analyse and mitigate gender bias” as proposed by Zhou et al., 2019.

References

Aekula, H., Garg, S., & Gupta, A. (2021). [re] double-hard debias: Tailoring word embeddings for gender bias mitigation. https://doi.org/10.48550/ARXIV.2104.06973

Almuhareb, A. (2006). Attributes in lexical acquisition /.

Baroni, M., Evert, S., & Lenci, A. (2008). Esslli workshop on distributional lexical semantics bridging the gap between semantic theory and computational simulations.

Baroni, M., & Lenci, A. (2011). How we blessed distributional semantic evaluation. *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 1–10.

Battig, W. F., & Montague, W. E. (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.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. https://doi.org/10.48550/ARXIV.1607.06520

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

Cardellino, C. (2019). Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705. https://doi.org/10.18653/v1/P19-1166

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. https://doi.org/10.48550/ARXIV.1903.03862

Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). Frage: Frequency-agnostic word representation. https://doi.org/10.48550/ARXIV.1809.06858

Kaneko, M., & Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. https://doi.org/10.48550/ARXIV.1310.4546

Mu, J., Bhat, S., & Viswanath, P. (2017). All-but-the-top: Simple and effective postprocessing for word representations. https://doi.org/10.48550/ARXIV.1702.01417

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162

Rukua95. (2020). Spanish_word_embedding_ evaluations. https%5C%5C://github.com/Rukua95/Spanish%5C_Word%5C_Embedding%5C_Evaluations#word%5C-analogy

Shin, S., Song, K., Jang, J., Kim, H., Joo, W., & Moon, I.-C. (2020). Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation. https://doi.org/10.48550/ARXIV.2004.03133

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation. https://doi.org/10.48550/ARXIV.2005.00965

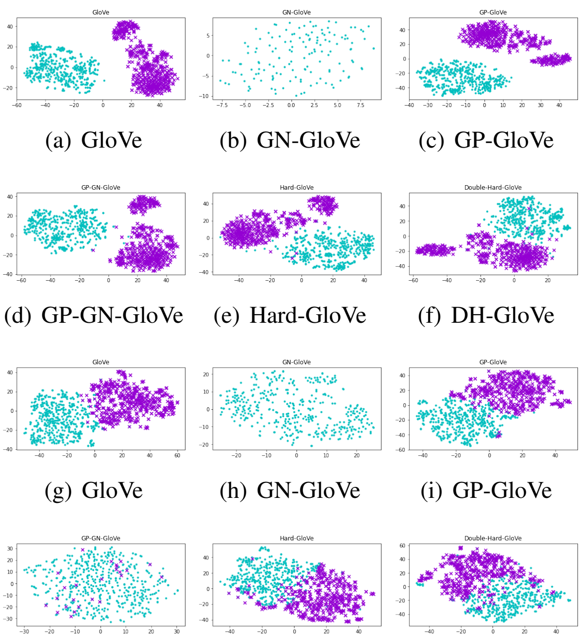
Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. https://doi.org/10.48550/ARXIV.1804.06876

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning gender-neutral word embeddings. https://doi.org/10.48550/ARXIV.1809.01496

Zhou, P., Shi, W., Zhao, J., Huang, K., Chen, M., Cotterell, R., & Chang, K. (2019). Examining gender bias in languages with grammatical gender. *CoRR*, *abs/1909.02224*. http://arxiv.org/abs/1909.02224

Appendix

A t-SNE Visualizations for English and Spanish GloVe embeddings



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

(d) GP-GN-GloVe (e) Hard-GloVe (f) DH-GloVe

(g) GloVe (h) GN-GloVe (i) GP-GloVe

(j) GP-GN-GloVe (k) Hard-GloVe (l) DH-GloVe

Figure 1: t-SNE Visualizations for English GloVe embeddings (a-f) and Spanish GloVe embeddings (g-l)

B Machine-Translated Spanish Gender Word Pairs

1. paisana paisano
2. sororal fraternal
3. brujas magos
4. criada criado
5. madres padres
6. diva divo
7. actriz actor
8. solterona soltero
9. mamá papá
10. duquesas duques
11. camarera camarero
12. paisanas paisanos
13. dote dote
14. anfitrionas anfitriones
15. aviadora aviadora
16. menopausia andropausia
17. clítoris pene
18. princesa príncipe
19. institutrices gobernadores
20. abadesa abad
21. mujeres hombres
22. viuda viudo
23. señoras caballeros
24. hechiceras hechiceros
25. señora señor
26. novias novios
27. baronesa barón
28. amas de casa amos de casa
29. diosas dioses
30. sobrina sobrino
31. viudas viudos
32. dama señor
33. hermana hermano
34. novias novios
35. monja sacerdote
36. adúlteras adúlteros
37. obstetricia andrología
38. camperas botones
39. ella él
40. marquesa marqués
41. princesas príncipes
42. emperatrices emperadores
43. yegua semental
44. presidenta presidente
45. convento monasterio
46. sacerdotisas sacerdotes
47. niñez niñez
48. señoras muchachos
49. reina rey
50. chicas tipos
51. mamis papis
52. criada sirviente
53. eyaculación femenina semen
54. portavoz portavoz
55. costurera sastre
56. vaqueras vaqueros
57. chica amigo
58. solteronas solteros
59. peluquería barbería
60. emperatriz emperador
61. mamá papi
62. feminismo masculinismo

63. chicas tipos
64. encantadora encantador
65. chica chico
66. maternidad paternidad
67. estrógeno andrógino
68. camarógrafas camarógrafos
69. madrina padrino
70. mujer fuerte hombre fuerte
71. diosa dios
72. matriarca patriarca
73. tía tío
74. presidentas presidentes
75. señora señor
76. hermandad fraternidad
77. anfitriona anfitrión
78. estradiol testosterona
79. esposa esposo
80. mamá padre
81. azafata azafato
82. hembras varones
83. viagra cialis
84. portavoces portavoces
85. mamá papá
86. belleza galán
87. descarada semental
88. doncella soltero
89. bruja mago
90. señorita señor
91. sobrinas sobrinos
92. dar a luz engendrado
93. vaca toro
94. bellas galán
95. concejales concejales
96. caseras caseras
97. nieta nieto
98. prometidas prometidos
99. madrastras padrastros
100. amazonas amazonas
101. abuelas abuelos
102. adúltera adúltero
103. colegiala colegial
104. gallina gallo
105. nietas nietos
106. soltera soltero
107. camarógrafa camarógrafo
108. mamás papás
109. ella él
110. amante maestro
111. muchacha muchacho
112. mujer policía policía
113. monja monje
114. actrices actores
115. vendedoras vendedores
116. novia novio
117. concejala concejal
118. dama amigo
119. estadista estadista
120. materno paternal
121. muchacha tío
122. dueña propietario
123. hermanas hermanos
124. señoras señores
125. mozas chicos
126. hermandad femenina fraternidad

127. botones botones
128. duquesa duque
129. bailarina Bailarín
130. chicas tipos
131. novia prometido
132. potrancas potros
133. esposas maridos
134. pretendiente pretendiente
135. maternidad maternidad
136. ella él
137. mujer de negocios empresario
138. masajistas masajistas
139. heroína héroe
140. gama ciervo
141. meseras meseros
142. novias novios
143. reinas reyes
144. hermanas hermanos
145. amantes amantes
146. maestras maestros
147. madrastra padrastro
148. novias novios
149. hija hijos
150. vaquera vaquero
151. dama caballero
152. hijas hijos
153. mezzo barítono
154. vendedora vendedor
155. amante amante
156. anfitriona anfitrión
157. monjas monjes
158. sirvientas sirvientes
159. señora señor
160. directoras directores
161. muchachas muchachos
162. congresista congresista
163. aviadora aviador
164. ama de casa amo de casa
165. sacerdotisa sacerdote
166. camareras camareros
167. baronesas barones
168. abadesas abades
169. toque barba
170. hermandades femeninas fraternidades
171. azafatas mayordomos
172. potra potro
173. czarina czar
174. hijastras hijastros
175. ella misma él mismo
176. muchachas niños
177. leonas leones
178. dama caballero
179. vagina pene
180. masajista masajista
181. vacas toros
182. tias tíos
183. esposa marido
184. leona león
185. hechicera hechicero
186. afeminado macho
187. madre padre
188. lesbianas homosexuales
189. femenino masculino
190. camareras camareros

191. óvulo próstata esperma
192. glándulas de skene utrículo prostático
193. hijastra hijastro
194. empresarias empresarios
195. heredera heredero
196. camarera camarero
197. directora de escuela director de escuela
198. mujer hombre
199. institutriz gobernador
200. diosa dios
201. novia novio
202. abuela abuelo
203. novia novio
204. chica amigo
205. lesbiana gay
206. señoras caballeros
207. muchacha chico
208. abuela abuelo
209. yegua caballo castrado
210. gallinas gallos
211. útero utrículo prostático
212. monjas sacerdotes
213. sirvientas sirvientes
214. costurera costurero
215. mesera mesero
216. heroínas héroes