Comparing Methods for Mitigating Gender Bias in Word Embedding

Clara Biagi
Department of Statistical Sciences,
University of Bologna, Italy
clara.biagi2@studio.unibo.it

Elisabetta Ronchieri
Department of Statistical Sciences,
University of Bologna, Italy;
INFN CNAF, Bologna, Italy
elisabetta.ronchieri@cnaf.infn.it

Abstract

Word embedding captures the semantic and syntactic meaning of words into dense vectors. It contains biases learning from data that include constructs, cultural stereotypes, and inequalities of the society. Many methods for removing bias in traditional word embedding have been proposed. In this study we use the original GloVe word embedding and perform a comparison among debiasing methods built on top of GloVe in order to determine which methods perform the best removing bias. We have defined half-sibling regression, repulsion attraction neutralization GloVe method and compared it with gender-preserving, gender-neutral GloVe method and other debiased methods. According to our results, no methods outperform in all the analyses and completely remove gender information from gender neutral words. Furthermore, all the debiasing methods perform better than the original GloVe.

Keywords: Gender Bias, Natural Language Processing, Word Embedding, GloVe, Half-Sibling Regression, Repulsion Attraction Neutralization, Gender-Preserving, Gender-Neutral

1. Introduction

Word embedding, used in natural language processing (McCann et al., 2017; Zhang et al., 2015), transforms texts into numeric vectors and reproduces similar words with similar vector representations. It builds its models by using a large and structured set of texts that can contain common stereotypes (e.g. the adjective $\overrightarrow{honorable}$ would be close to the vector for man, whereas the adjective $\overrightarrow{submissive}$ would be closer to woman) and prejudices of society as reported by Garg et al. (2017). As consequences, such models can bring along human biases (Caliskan et al., 2016) causing problems on sensitive applications, such as in the machine translation e.g. from English to Italian that shows a gender bias issue in the deepL service (https://www.deepl.com/it/translator) and in the wordreference service (https://www.wordreference.com/it/), in the applicant tracking systems and in the sorting search results. Using an unfair word embedding, as outlined by Bolukbasi et al. (2016), may lead to an amplification of already present biases. For example, if a word embedding model has learnt the association between $\overrightarrow{woman}$ and $\overrightarrow{homemaker}$, and between $\overrightarrow{man}$ and $\overrightarrow{computer\ programmer}$, when looking for a $\overrightarrow{computer\ programmer}$ profile (or when filtering a curriculum vitae for finding a programmer employee) in a search engine, all web pages and resume related to women will be discarded. This mechanism may amplify the social bias instead of reducing it. Similar biases appear in word embeddings for other human stereotypes, such as race and religion (Manzini et al., 2019). The analogy *man to computer programmer and woman to homemaker*, happens because $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\ programmer} - \overrightarrow{homemaker}$. This means that in the space of word embedding, words like *homemaker* are more similar (nearer) to *woman* than to *man*. In word embedding, to evaluate word similarity it is usually used the cosine similarity between two non-zero vectors that quantifies the level of affinity between the words in the vectors (Gómez and Vázquez, 2022).

Existing literature contains methods to quantify bias (Gonen and Goldberg, 2019) and methods to remove bias, such as post-processing methods (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019; Yang and Feng, 2019; Karve et al., 2019), and word vector learning methods (Zhao et al., 2018) based on Global Vector (GloVe) that is one of the most popular word embedding technique (Pennington et al., 2014). Bolukbasi et al. (2016) proposed a distinction between a direct and an indirect bias, and two methods to reduce bias, named hard and soft debiasing methods. Kumar et al. (2020) introduced a Repulsion-Attraction-Neutralization (RAN) method based on attraction and repulsion mechanisms; words

URI: https://hdl.handle.net/10125/102720
978-0-9981331-6-4

that are clustered together for stereotypical constructs (e.g. $\overrightarrow{nurse}$ and $\overrightarrow{receptionist}$) must be disassociated or repulsed from each other. Yang and Feng (2019) detected and removed gender bias direction from gender-neutral word vectors, proposing a Half-Sibling Regression (HSR) solution. Zhao et al. (2018) proposed another debiasing method Gender-Neutral Global Vectors (GN-GloVe) by considering the SemBias analogy test and adding a new constraint to GloVe's objective function to confine gender information in the last coordinate of each vector.

In this study, we have focused on Gender-Preserving (GP) technique with GN-GloVe and HSR technique with RAN-GloVe methods. We have also compared their results with the aforementioned solutions, aiming at determining: 1. if there is a method that outperforms the others, and 2. if there is a method that truly manages to remove bias from word embedding. According to our results, no methods outperform in all the analyses and completely remove gender information from gender neutral words. However, all the debiasing methods perform better than the original GloVe.

2. Related Works

To develop our research we have searched for existing literature that include GP-GN-GloVe and HSR-RAS-GloVe methods.

With respect to the first method we have identified the following works. Kaneko and Bollegala (2019) minimised the gender bias by optimising a loss function that depends on four components: two components control the feminine/masculine element of words, one controls their neutrality and the last component controls the loss of semantic information. They introduced a Gender-Preserving Global Vector (GP-GloVe) method. Wang et al. (2020) realised that discrepancies in word frequency can influence the performance of hard-debiasing by twisting the gender direction. They introduced an improvement to the hard debiasing method. Both studies tested the new methods against a set of baselines that includes GP-GN-GloVe.

Concerning the second method at the time of this research we have just found works that talk about HSR-GloVe (Yang and Feng, 2019) and RAN-GloVe methods (Kumar et al., 2020). Yang and Feng (2019) observed that HSR-GloVe has a better performance against all other post-processing methods and outperforms the original GloVe. Kumar et al. (2020) altered the spatial distribution of word embeddings with attraction and repulsion mechanisms: repulsion is between words that are stereotyped.

In this paper, we present GP-GN-GloVe and HSR-RAN-GloVe methods. The former because it is available in literature as shown by Kaneko and Bollegala (2019), while the latter to take advantages of both methods: HSR behaves well in indirect bias as documented by Yang and Feng (2019), while RAN behaves well in SemBias Analogy test as detailed by Kumar et al. (2020)).

3. Methodology

3.1. Methods for reducing bias

To develop our research we have first identified word embedding models that have been defined to reduce gender bias: Hard-Debias (HARD) method (Bolukbasi et al., 2016), Gender-Neutral GloVe (GN-GloVe) method (Zhao et al., 2018), Gender-Preserving GloVe (GP-GloVe) method (Kaneko and Bollegala, 2019), Half-Sibling Regression (HSR) method (Yang and Feng, 2019), Double-Hard Debias (DHD) method (Wang et al., 2020), Repulsion-Attraction-Neutralisation (RAN) method (Kumar et al., 2020). Later, we have defined GP-GN-GloVe and HSR-RAN-GloVe methods.

The *Hard-Debias (HARD)* method utilizes linear projection technique for gender debiasing. The bias of a specific word is quantified comparing it to a couple of gender-specific words (e.g. *he-she*): the bias is present if the word, which is supposed to be neutral, is closer to *she* than *he* or vice-versa. The authors of this method identify two kind of biases: *direct bias*, when there is an association between gender neutral words and gender pairs (e.g. *homemaker* is closer to *woman* than *man*); *indirect bias*, when the association is between two neutral words (e.g. *receptionist* is much closer to *softball* than to *football*, and derives from the association of both *receptionist* and *softball* with female words).

The *Gender-Neutral GloVe (GN-GloVe)* method decreases gender bias while training the word embedding, instead of correcting a pre-trained one. The key idea is that each word in the embedding consists of a gendered component and a neutralized component: all the gender information is kept into one component while making the other independent of gender influence. The new vectors will have the gender component concentrated in the last coordinate, and it may be kept or not. The main limitation of this method is that it can be only applied to the GloVe embeddings, or more precisely to word embedding computed through minimisation of a loss function.

The *Gender-Preserving GloVe (GP-GloVe)* method splits the vocabulary into four mutually disjoint categories: female oriented, male oriented, neutral and stereotypical. The total vocabulary is given by the union

of these sets. Female and male oriented words are those words like *bear* or *bikini* that have an fair gender bias while the stereotypical words are those like *homemaker* that have an unfair gender bias. This method aims at protecting feminine and masculine properties when required, preserving gender neutrality when needed and removing gender bias from stereotypical words. A function that predicts the degree of femininity and a function that predicts the degree of masculinity of a given word is considered. These functions are then maximised for words belonging to female vocabulary and male vocabulary respectively, while minimized otherwise. To preserve gender neutrality for gender neutral words and to remove gender bias from stereotypical words, the strategy is to project them into a subspace orthogonal to the gender. Gender direction is found as in previous methods using a set of feminine and masculine word-pair.

The *Half-Sibling Regression (HSR)* method uses the statistical dependency between gender-definition word embeddings and gender-biased word embeddings. The key idea of this approach is to learn and then directly subtract the gender information from the non-gender-definition words. Words in vocabulary are classified in gender-definition (e.g. *she, he*) and non-gender-definition (e.g. *nurse, colonel*) words. Both sets contain underlying gender information, but semantic content is mostly present in non-gender-definition words. Debiased non-gender-definition word vectors are obtained by subtracting an approximated gender information from the original embedding. The gender information is estimated as the expected value of the word embedding conditioning on the set of gender-definition word vector.

The *Double-Hard Debias (DHD)* method is an improvement of the HD method: first it removes the influence of word frequency and then removes the gender bias. To remove the frequency features, the 500 top male- and female -biased words are considered according to the original word embedding, and principal component analysis is performed. The top principal components are then taken and word vectors are projected into the space orthogonal to each principal component. In this intermediate subspace, the standard Hard Debias method is applied.

The *Repulsion-attraction-Neutralisation (RAN)* method aims at transforming a word vector that minimise the stereotypical gender information while maintaining semantic aspects. All vocabulary is split into words for which gender carries semantic importance (e.g. *beard* and *bikini*) and all other words, which should be gender-neutral. To compute the two sets from a given dictionary, the authors of this method propose the Knowledge Based Classifier to overcome the limitation of using a classifier to select the gender neutral words as done in the Hard-Debias method. The method is based on minimising a loss function composed by three different functions. Minimizing the three functions defines three phases that give the name to the method: 1. Repulsion, minimizing the first leads to separate the debiased word from the neighbours with high indirect bias; 2. Attraction, minimizing the second function leads to the minimum loss of semantic properties of the de-biased word; 3. Neutralization, minimizing the last function leads to the minimum direct bias of the de-biased word.

3.2. Quantifying gender bias

Then we have evaluated debiasing performances by quantifying both direct and indirect bias in the original GloVe and in the debiased versions of GloVe. We have considered Word Embedding Association Test (WEAT) to detect human stereotype bias (Caliskan et al., 2016) and performed SemBias Analogy test (Zhao et al., 2018). We have also computed Gonen and Goldberg (2019)'s five tasks: clustering, correlation, profession, classification and association.

The *Word Embedding Association Test (WEAT)* is a permutation-based test that considers two sets of *target* words containing male and female stereotypes (such as *programmer, doctor, ... surgeon*, and *homemaker, nurse, ..., beautician*), and two sets of *attribute* words including gender definition words (such as *he, man, ..., male* and *she, woman, ..., female*). The null hypothesis is that the two sets of target words are related in a similar way to the sets of attribute words: if it is true, e.g. *homemaker* and *programmer* have approximately the same similarity with *she* and *he*, and also with all other words in the attribute words; on the contrary, it means that stereotypes are still relevant in the embedding. The test computes the effect size and p-value: the effect size computes differential association of the two target words; the p-value checks the significant level of bias.

With respect to Gonen and Goldberg (2019) work, we have performed five different tasks: clustering male and female biased words; correlation between bias by projection and bias by neighbours; bias by neighbours for profession words; classifying previously female and male biased words; and association.

The *SemBias Anaology* test is characterized by a set of analogy tests. It determines the word pair in best analogy to the pair *he - she* among four options: consisting of four pairs of words: a gender-definition word pair (e.g. *waiter - waitress*), a gender-stereotype word pair (e.g. *doctor - nurse*) and two other similar

bias free word pairs (e.g. *dog* - *cat*). The task is to identify the gender-definition word pair among the four pairs. The dataset contains 440 instances given by the combination of 20 gender-stereotype and 22 gender-definition pairs. A subset of 40 instances, generated by 2 gender-definition pairs, is used for testing. Accuracy is given by the number of times each type of word pair is selected. When a word embedding is free from gender bias, it should obtain high accuracy for gender-definition words and low accuracy for the other two kind of pairs.

3.3. Preservation of word semantics

To be sure that the debiasing process removes only gender bias and preserve other information, we have evaluated the word similarity and semantic accuracy.

The *Word Similarity* measure represents a quality indicator of the word embeddings. We can compute cosine similarity between word embeddings and Spearman correlation with respect to human ratings for the word pairs. To evaluate this property eight benchmark datasets are used (WB, 2018), each with its own measure of similarity: Word Similarity (WS)-353, Rubensten-Goodenough (RG)-65 dataset, Rare Words (RW) dataset, MTurk-287, MTurk-771, SimLex dataset, and SimVerb-3500. They are all composed by a list of word pairs and an associated similarity measure given by humans. The RG-65 benchmark for example contains 65 couples with an associated similarity score ranging from 1 to 5.

The *Semantic Textual Similarity (STS)* measure determines the degree of semantic similarity between two texts. To evaluate this property, the used benchmark datasets are taken from the 2012 SemEval Sentences Involving Compositional Knowledge (SICK) task and the SemEval Semantic Textual Similarity (STS) task from 2012 to 2015. They include 20 tasks: for each task, the Pearson correlation coefficient for the STS tasks between machine assigned semantic similarity scores and human judgements.

4. Experiments and Results

4.1. Evaluating debiasing performance

In this section we have detailed the performance of each method and if there are methods that work better than others by considering the GloVe word embedding and different debaised methods. We have obtained comparable results also available in the existing papers. Specifically, Gonen and Goldberg (2019) computed the five tasks for the Hard debiased Word2Vec and for the GN-GloVe; Yang and Feng (2019) reported all the following analysis for Hard-GloVe, GN-GloVe, GP-GloVe and HSR-GloVe. In our work, the results have been extended to GP-GN-GloVe, DHD-GloVe, RAN-GloVe and HSR-RAN-GloVe.

As done by Bolukbasi et al. (2016), we have selected the 50000 most frequent words from the word embedding and then excluded words with upper-case letters, digits, punctuation or words longer than 20 characters. Then, gender-specific words, such as *he* or *she*, which have a fair gender component, have been excluded as well. 47698 words have been eventually considered for this analysis. Direct bias of a word has been computed by measuring the cosine similarity between its embedding vector and the gender direction.

Table 1 shows the average direct bias. The first row shows the mean bias for the original GloVe embedding. The HARD-GloVe method gets the best results. As expected, the HARD-GloVe gets the best results. Indeed, the method is based on projecting each biased word onto the subspace that is orthogonal to the gender direction. According to Table 1, GP-GN-GloVe and GN-GloVe have a higher average bias than the original GloVe. GP-GloVe has a slightly lower bias than the original GloVe. HRS-GloVe, which does not directly minimize the projection of words onto gender direction, anyway manages to reduce direct bias. DHD-GloVe, as expected, is the second method that decreases the most the direct bias, because it computes similar steps as for the HARD-GloVe. RAN-GloVe and HSR-RAN-GloVe have a lower direct bias than the original GloVe, but it is still present in the embedding.

Embeddings	Mean Bias	Mean Male Bias	Mean Female Bias
GloVe	0.0375	0.0373	0.0378
HARD-GloVe	**0.0008**	0.0008	0.0009
GN-GloVe	0.0555	0.0618	0.0368
GP-GloVe	0.0366	0.0380	0.0343
HSR-GloVe	0.0218	0.023	0.0198
DHD-GloVe	0.0196	0.0211	0.0175
RAN-GloVe	0.0291	0.0289	0.0294
GP-GN-GloVe	0.0457	0.0431	0.0482
HSR-RAN-GloVe	0.0277	0.0281	0.0271

Table 1. Average direct bias. In the first column the best result is boldfaced.

A t-test for paired samples has been used for comparing all the average biases. As result, HSR-GloVe is not statistically different from the original bias at a 5% level and DHD-Glove and HSR-RAN-GloVe are not significantly different from GloVe at a 1% level of confidence. DHD-GloVe gets an average bias no significantly different from the RAN-GloVe at a 5% level, and HSR-GloVe and HSR-RAN-GloVe have not different average direct bias at a level of 1%. Average

	Clustering (Acc - ARI)	Correlation (Pear - Spea)	Profession (Pear - Spea)	Classification	Association
GloVe	1.0000 - 1.0000	0.7726 - 0.7486	0.8200 - 0.7882	0.9980	2
HARD-GloVe	0.8050 - 0.3715	0.6884 - 0.6801	0.7166 - 0.7026	0.9057	**1**
GN-GloVe	0.8560 - 0.5065	0.7336 - 0.7162	0.7925 - 0.7651	0.9815	3
GP-GloVe	1.0000 - 1.0000	0.7700 - 0.7457	0.8102 - 0.7407	0.9977	3
HSR-GloVe	0.9450 - 0.7919	**0.6422 -0.6430**	**0.6804 - 0.6733**	0.9055	**1**
DHD-GloVe	**0.7980 - 0.3546**	0.6645 - 0.6650	0.6975 - 0.6980	**0.8550**	**1**
RAN-GloVe	0.8240 - 0.4194	0.7130 - 0.6884	0.6873 - 0.6782	0.9183	**1**
GP-GN-GloVe	0.8920 - 0.6142	0.7676 - 0.7408	0.8127 - 0.7840	0.9813	**1**
HSR-RAN-GloVe	0.8170 - 0.4014	0.7043 - 0.6852	0.6983 - 0.6917	0.9383	**1**

Table 2. Gender bias word relation task performance. In the first four columns the best result is boldfaced. The association column with value 1 is also boldfaced.

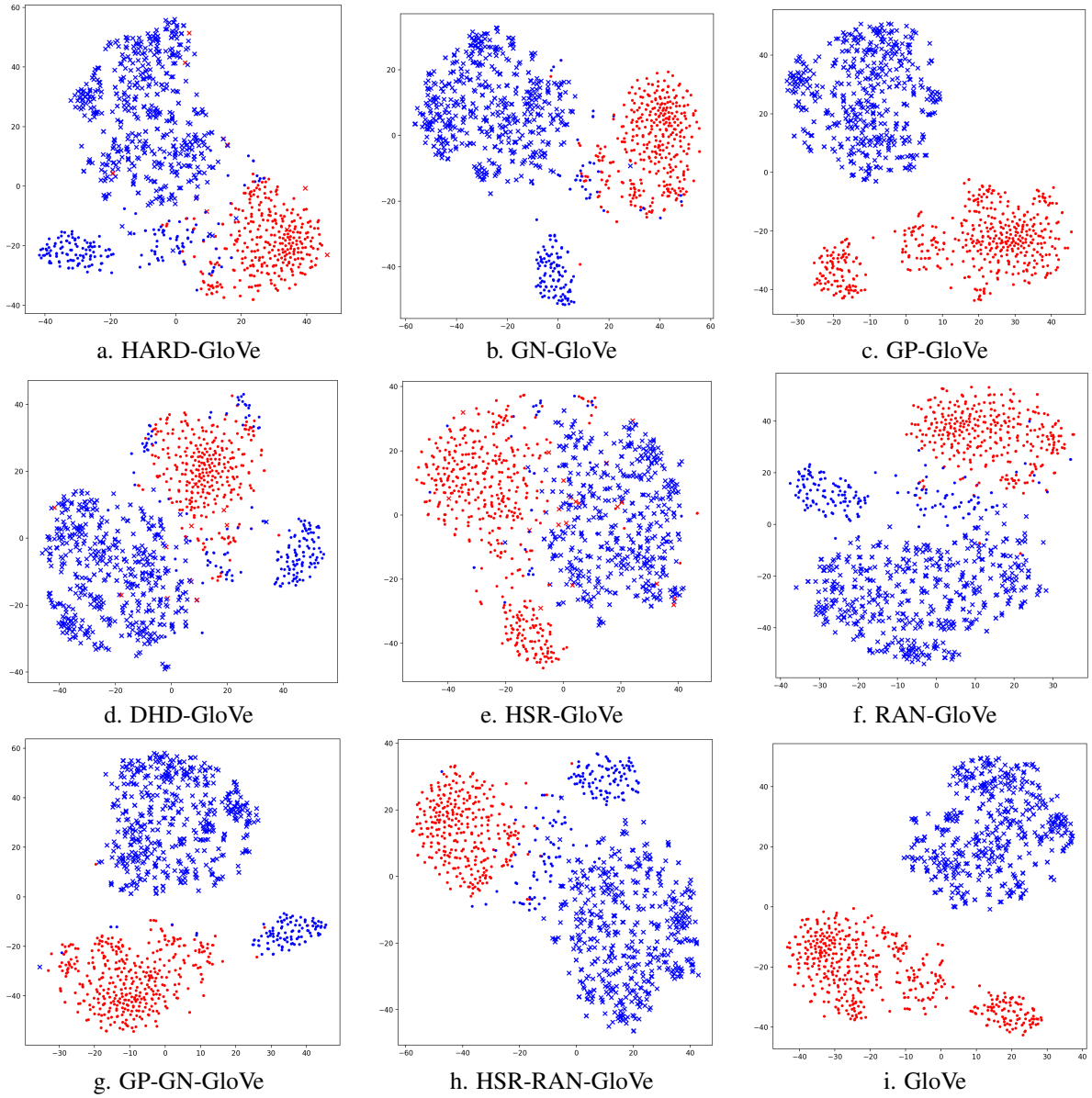
biases are obtained by averaging the absolute values of biases, so that positive and negative values do not offset each other. Male and female average biases are reported in Table 1. For all methods, gendered averages are significantly different from each other. GN-GloVe in particular has a male direct bias much higher than the female one.

Direct bias is not enough to evaluate the performance of a debiasing method, as underlined by Gonen and Goldberg (2019). Table 2 shows measures of *indirect bias* that we have computed following Gonen and Golderberg's 5 tasks: clustering, correlation, profession, association and classification. The first row shows the 5 tasks' results for the original GloVe, which are the worst possible.

The *clustering* column reports the accuracy metric as the first value and the ARI index as second value for all debiasing methods of the 2-means cluster solution: the lower the accuracy (or the ARI) value, the less the clusters align with gender and the more indirect bias is removed. Plots in Figure 1 show the t-distributed stochastic neighbor embedding (TSNE) representation of the 1000 biased words. It is clear how gender and clustering labels perfectly overlaps in the original GloVe and GP-GloVe and that genders do not splits randomly in the two clusters in any method. However, the difference with respect to the clustering solution found on the original GloVe is relevant for some methods: DHD-GloVe in particular mixes cluster and gender much more with respect to other methods (Figure 1.d). However, while the blue cluster includes mainly female biased words but also many male biased words, the red cluster contains mainly male biased words, and the same holds for HD-GloVe. In RAN-GloVe, HSR-RAN-GloVe, GP-GN-GloVe and GN-GloVe, one cluster contains only male biased words. HSR-GloVe, even if it has the second higher ARI and accuracy, is the only one in which both clusters contains some female and male words, even if in each cluster one of the two gender is certainly prevalent.

The *correlation* column contains the correlation between direct and indirect bias. The Pearson correlation is the first value, also available in Gonen and Goldberg (2019), and the Spearman correlation is the second value, added in this work. All correlations are significantly different from zero. The lower correlation is found for HSR-GloVe, meaning that an originally female/male-biased word after debiasing may have both female/male words as neighbours and some male/female words. DHD-GloVe, RAN-GloVe, HSR-RAN-GloVe and HARD-GloVe also manage to reduce the correlation with respect to the original GloVe. GN-GloVe reduce the correlation but not much. On the contrary, GP-GloVe and GP-GN-GloVe maintain more or less the same correlation as the biased GloVe, suggesting that these methods do not remove indirect bias because words with high male/female bias before debiasing have also many male/female words around them after debiasing. We can observe that even the HSR-GloVe, reducing the most the correlation, still has a correlation of 64%. This suggests that none of the proposed methods really menage to remove the gender bias from the neighbours of a given words. Even if the target word has a null direct bias, bias will still be present in its neighbour words.

The *profession* column contains the correlation measure between direct and indirect bias but considering only the profession words proposed by Bolukbasi et al. (2016). The best value again is for the HSR-GloVe. Results are similar as the previous task but correlations are all a little bit higher, suggesting that for the profession words removing bias from the neighbours word is harder, maybe because professions are one of the most stereotypical aspect of society. Graphical results are shown in Figure 2. It is clear that the two measures of bias are correlated for all methods. Profession with a relevant positive original bias also have many male neighbours and professions with negative bias are surrounded by female neighbours, before and after debiasing. Plots of debiasing methods show that the number of male neighbours increases for many female biased words but male biased words are still surrounded by many more male neighbours than female. There



a. HARD-GloVe b. GN-GloVe c. GP-GloVe

d. DHD-GloVe e. HSR-GloVe f. RAN-GloVe

g. GP-GN-GloVe h. HSR-RAN-GloVe i. GloVe

Figure 1. Clustering TSNE visualization for eight baseline debiased methods and the original GloVe. Colours are mapped to the clustering labels, while markers are mapped to gender: x are female biased words and single dot . are male biased words according to the original bias.

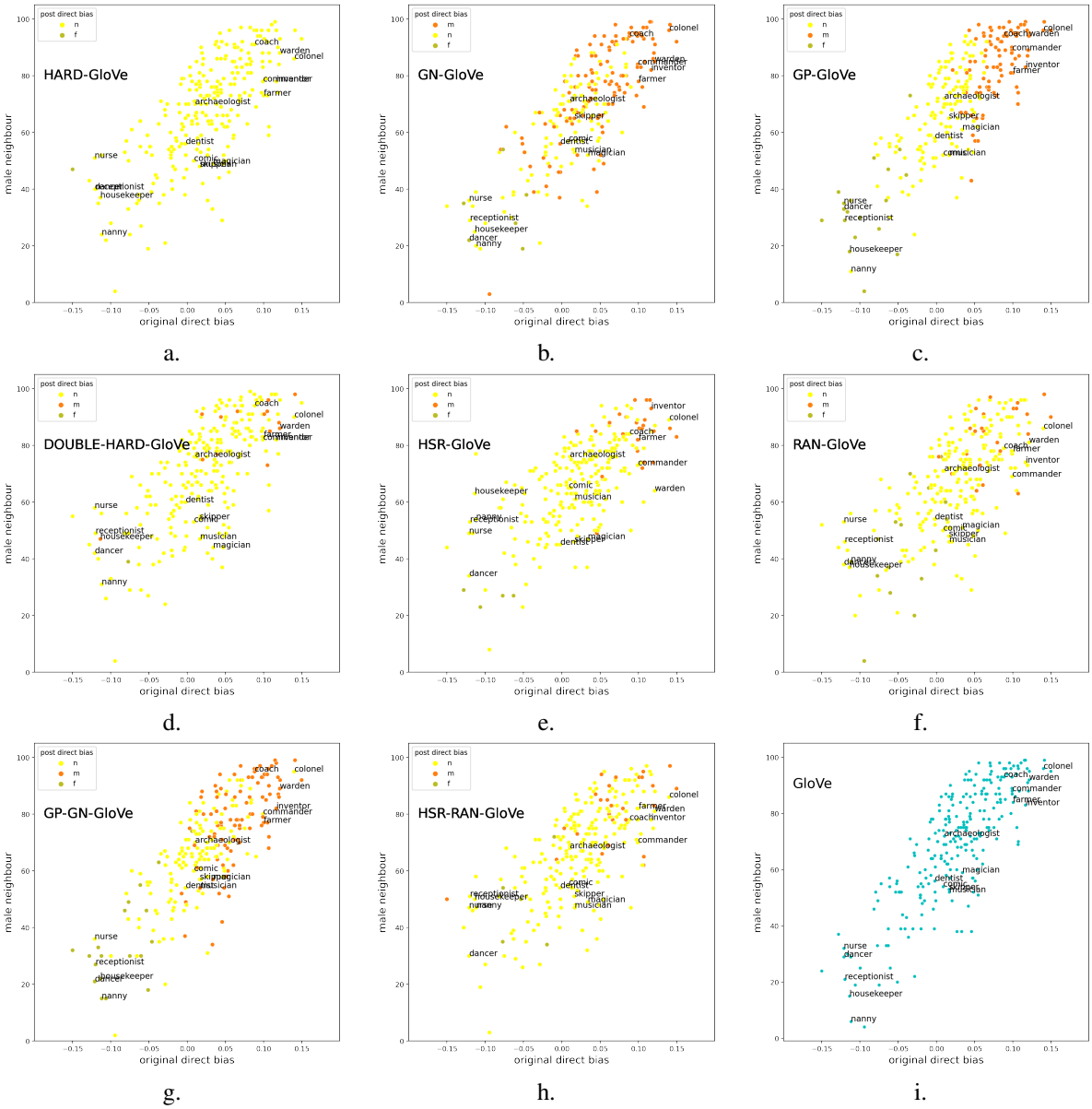


Figure 2. Profession representation for eight debiased methods and the original GloVe. Dots are yellow for neutral words, red for male words and green for female words.

is still a separation between female- and male- biased profession.

Differently from the original paper Gonen and Goldberg (2019), the colour assigned to the point reflects the direct bias of the debiased word embedding: yellow indicates words with a direct bias between -0.05 and 0.05, ocher words with a bias lower than -0.05 and orange stands for words with a bias greater than 0.05. Briefly, ocher shows female-biased, yellow neutral and orange male-biased words according to the new embedding. It is evident that a methods like the HD-GloVe reduces a lot direct bias but still maintain a separation between male and female stereotypical professions. The other methods, that reduce less the direct bias, show a clear correlation also between the direct bias computed on the new embedding and the number of male/female neighbours. As example, looking at results for HSR-GloVe (Figure 2.e) it is clear that red points (male-biased words) are all in the top right of the plot while the other points (female-biased words) are all in the left-bottom part. The new GP-GN-GloVe and HSR-RAN-GloVe embeddings in Figure 2.g and Figure 2.h still contain neutral-words that reflects both direct and indirect bias.

The *classification* column reports the accuracy of the SVM classifier. Aim of the classifier is to arrange gender after being trained on the 5000 most biased words according to the original embedding. In general, results show that accuracy metric is quite high in all methods. This again denotes that gender information is still trapped in the debiased word embeddings. DHD-GloVe gets the best result with an accuracy of only 85.5%. HARD-GloVe, HSR-GloVe, RAN-GloVe and HSR-RAN-GloVe also bring to a decreasing in the accuracy value with respect to the original GloVe, while all other methods have a very high accuracy.

The *association* column reports the number of p-value greater than 0.05 for the WEAT. A statistically significant p-value means that the null hypothesis of no difference between the two sets of target and the two sets of attribute words is rejected. There is no method that manages to get no significant p-values. This means that there is no enough evidence to prove that they are completely free from gender bias. However, some methods have better results than others. In particular, HARD-GloVe, HSR-GloVe, DHD-GloVe, RAN-GloVr, GP-GN-GloVe, and HSR-RAN-GloVe all have just one significative p-value. GN-GloVe and GP-GloVe are for sure the one with worst results, because the get all p-values greater than 0.05, even worst than the original GloVe embedding, which gets 2 low p-values.

Considering all measures of indirect bias we can observe that no methods really outperform the others, and above all no methods manage to remove completely indirect bias. Furthermore, DHD-GloVe, HSR-GloVe and RAN-GloVe are the one that decrease the most indirect bias.

Through the SemBias dataset it is possible to evaluate gender information in the word embedding. Table 3 shows how many times word embeddings recognizes gender-definition words and how many times it chooses gender-stereotype or other couples instead. Considering results on the test set, RAN-GloVe gets the best results with a 97.5% accuracy to identify gender-definition pairs. The second best result is the HSR-RAN-GloVe, with a test accuracy of 92.5%, followed by GP-GN-GloVe, that gains the higher accuracy during training and it manages to maintain a good level also in testing and. On the other side, HSR-GloVe and DHD-GloVe do not work well at all: in testing the accuracy for the DHD-GloVe is 0% and only 10% for HSR-GloVe. It is interesting noticing that DHD-GloVe and HSR-GloVe are the ones that most reduce indirect bias, while are the worst for type of gender information.

4.2. Evaluation of word embeddings

Table 4 shows the average Pearson correlation coefficients between the scores given by human and word embedding to evaluate how similar two sentences are with respect to their meaning. In this case, four methods get greater average correlations than the original GloVe and four methods get a lower average. HSR-GloVe with an average of 0.5721 gets the best result, against the GloVe that has an average of 0.5051. GP-GloVe gets the worst result, which decreases the correlation for all the five data sets and obtains a mean correlation of 0.4659. GP-GN-GloVe performs in a similar way to HSR, improving the semantic textual similarity of the word embedding. All other methods get similar results to the original GloVe, not bringing huge improvement or decreasing in the correlations.

Table 5 shows Spearman correlation coefficients between the human similarity measure given in the datasets and the similarity measure computed on the debiased word embeddings. Similarity measure is obtained through the cosine similarity. Many methods tend to improve the correlation of the original GloVe embedding: HSR-GloVe, GP-GN-GloVe and HSR-RAN-GloVe, which are the three methods with the highest correlations, obtain an average correlation of 0.5642, 0.5628 and 0.5615 respectively against the average of 0.5394 obtained by the original GloVe. Only DHD-GloVe and GP-GloVe, respectively with 0.5069 and 0.5256 of Spearman correlation coefficient, have a

	SemBias			SemBias Subset		
Embeddings	Definition	Stereotype	None	Definition	Stereotype	None
GloVe	80.2	10.9	8.9	57.5	20	22.5
HARD-GloVe	84.1	6.4	9.5	25	27.5	47.5
GN-GloVe	97.7	1.4	0.9	75	15	10
GP-GloVe	84.3	7.9	7.7	65	15	20
HSR-GloVe	85.9	3.8	10.2	10.0	30.0	60.0
DHD-GloVe	25.0	12.3	62.7	0.0	15.0	85.0
RAN-GloVe	92.7	1.1	6.1	**97.5**	**0.0**	**2.5**
GP-GN-GloVe	**98.4**	**1.1**	**0.5**	82.5	12.5	5.0
HSR-RAN-GloVe	92.3	0.9	6.8	92.5	0.0	7.5

Table 3. Percentage Results. Below the SemBias multi-columns there are values obtained during the training phase, while below the SemBias Subset multi-columns there are values obtained with the test set.

Embeddings	STS 2012	STS 2013	STS 2014	STS 2015	SICK
GloVe	0.4892	0.4690	0.5102	0.5135	0.6211
HARD-GloVe	0.4511	0.5778	0.5838	0.4620	0.4303
GN-GloVe	0.4896	**0.6175**	0.5185	0.4869	0.5331
GP-GloVe	0.4534	0.4316	0.4670	0.4729	0.5902
HSR-GloVe	0.5127	0.5245	**0.6013**	**0.6144**	0.6256
DHD-GloVe	0.4543	0.5745	0.4766	0.4426	0.4313
RAN-GloVe	0.4806	0.4703	0.5045	0.4998	0.6090
GP-GN-GloVe	**0.5333**	0.5143	0.5902	0.5831	**0.6435**
HSR-RAN-GloVe	0.4996	0.4967	0.5387	0.5342	0.6198

Table 4. Semantic textual similarity task results. In each column the best result is boldfaced.

lower average correlation than the original GloVe.

5. Conclusions

In this work we have defined HSR-RAN-GloVe method and compared it with different debiased methods with the aim at determining 1. which methods perform the best removing bias and 2. if these methods truly remove bias from word embedding.

Conclusions regarding the first aim are not straightforward. There is no method that outperforms in all the analyses: HARD-GloVe has the lower direct bias, DHD-Glove and HSR-GloVe get the better results for indirect bias measure but they both perform quite bad on the SemBias dataset, where GP-GN-GloVe and RAN-Glove perform well instead. No method leads to a decrease in performance for neither similarity or semantic tasks. GP-GN-GloVe and HSR-RAN-GloVe show similar conclusions of the singular methods. However, considering all tasks, RAN-GloVe is probably the one that in average gets the more debiased word embdedding. Especially for the SemBias dataset, the RAN-GloVe word embedding gets a remarkable accuracy in identifying the gender-definition pairs. The fact that it does not confuse the gender-definition with the gender-stereotype pairs in particular can be very useful in real application.

With respect to the second aim, it seems that none of the considered methods managed to completely remove gender information from gender neutral words. They all perform better than the original GloVe in at least one task, but the results themselves are not satisfactory and show that information related to gender is still encapsulated in the various word embeddings. In particular, following the five tasks for measuring indirect bias is evident that female-biased words in the original word embeddings are more similar to each other in the new embedding than to the originally male-biased words and vice-versa. This leads to the same conclusion of Gonen and Goldberg (2019): all these methods are only hiding gender bias but not truly removing it. Gender bias is still present in how the words are distributed in the embeddings and in their neighbours. However, the fact that these methods are not able to completely remove the bias from word embedding does not mean that they are useless. They generally have higher score than the original GloVe in the similarity and semantic tasks. And after all they reduce the original bias, both directly and indirectly. Combining one of the presented methods with a modification of the corpora and/or the machine learning algorithm may lead to a fully debiased machine learning output.

In the future, we are interested in considering other traditional word embeddings to determine what are the main difference with GloVe. Furthermore, we also intend to extend the study to non-English word embeddings to take into consideration languages with grammatical gender, like Italian. Another aspect we think it is important to investigate is removing bias not only from word embedding but also in natural language algorithm and datasets.

Embeddings	RG65	WS-353	RW	MEN	MTurk-287	MTurk-771	SimLex-999	SimVerb-3500
GloVe	0.7540	0.6199	0.3722	0.7216	0.6480	0.6486	0.3474	0.2038
HARD-GloVe	0.7648	0.6207	0.3720	0.7212	0.6468	0.6504	0.3501	0.2034
GN-GloVe	0.7457	0.6286	0.3989	0.7446	0.6617	0.6619	0.3700	0.2219
GP-GloVe	0.7546	0.6003	0.3450	0.6974	0.6418	0.6391	0.3389	0.1877
HSR-GloVe	0.7764	**0.6554**	0.3868	0.7353	0.6335	0.6652	**0.3971**	**0.2635**
DHD-GloVe	0.7478	0.5699	0.3183	0.6815	0.6284	0.6175	0.3170	0.1748
RAN-GloVe	0.7651	0.6176	0.3753	0.7205	0.6462	0.6430	0.3424	0.2061
GP-GN-GloVe	0.7248	0.6355	**0.4299**	**0.7522**	**0.6650**	**0.6791**	0.3843	0.2312
HSR-RAN-GloVe	**0.7916**	0.6445	0.3942	0.7432	0.6574	0.6630	0.3680	0.2300

Table 5. Word similarity task results. In each column the best result is boldfaced.

References

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. (2016). Semantics derived automatically from language corpora contain human-like biases. https://doi.org/10.48550/ARXIV.1608.07187

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2017). Word embeddings quantify 100 years of gender and ethnic stereotypes. https://doi.org/10.48550/ARXIV.1711.08412

Gómez, J., & Vázquez, P.-P. (2022). An empirical evaluation of document embeddings and similarity metrics for scientific articles. https://doi.org/10.3390/app12115664

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

Kaneko, M., & Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. https://doi.org/10.48550/ARXIV.1906.00742

Karve, S., Ungar, L., & Sedoc, J. (2019). Conceptor debiasing of word representations evaluated on weat. https://doi.org/10.48550/ARXIV.1906.05993

Kumar, V., Bhotia, T. S., Kumar, V., & Chakraborty, T. (2020). Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. https://doi.org/10.48550/ARXIV.2006.01938

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. https://doi.org/10.48550/ARXIV.1904.04047

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, Eds.). *30*, 6294–6305.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162

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

WB. (2018). Benchmarks for intrinsic evaluation word embeddings.

Yang, Z., & Feng, J. (2019). A causal inference method for reducing gender bias in word embedding relations. https://doi.org/10.48550/ARXIV.1911.10787

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification (C. C. adn N. Lawrance, D. Lee, M. Sugiyama, & R. Garnett, Eds.). *28*, 649–657.

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