On the Generalization of Projection-Based Gender Debiasing in Word Embedding

Elisabetta Fersini, Antonio Candelieri, Lorenzo Pastore
University of Milano-Bicocca
Milan - Italy
{elisabetta.fersini, antonio.candelieri}@unimib.it
l.pastore6@campus.unimib.it

Abstract

Gender bias estimation and mitigation techniques in word embeddings lack an understanding of their generalization capabilities. In this work, we complement prior research by comparing in a systematic way four gender bias metrics (Word Embedding Association Test, Relative Negative Sentiment Bias, Embedding Coherence Test and Bias Analogy Test), two types of projection-based gender mitigation strategies (hard- and soft-debiasing) on three well-known word embedding representations (Word2Vec, FastText and Glove). The experiments have shown that the considered word embeddings are consistent between them but the debiasing techniques are inconsistent across the different metrics, also highlighting the potential risk of unintended bias after the mitigation strategies.

1 Introduction

A recent body of work in Natural Language Processing (NLP) has focused attention on quantifying different types of bias through various approaches, spanning from psychological tests and performance differences for various tasks to the geometry of vector spaces (Sun et al., 2019). Defining the type of bias is essential to estimate and mitigate it. Several forms of biases specific to NLP application have been introduced in the literature during the last 5 years (Nozza et al., 2019; Nissim et al., 2020; Goldfarb-Tarrant et al., 2021). In (Hitti et al., 2019) the authors defined ***gender bias*** *in a text as the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender*, highlighting that gender bias can evidence itself structurally, contextually, or in both forms. Structural bias occurs when the construction of sentences shows patterns closely tied to the presence of gender bias. On the other hand, contextual bias can happen in the tone, words, or context of a sentence. Unlike structural bias, this type of bias is not evident in grammatical structure but requires contextual background information and human perception. Therefore, gender bias can be discovered using both linguistic and extra-linguistic cues and can manifest itself in subtle or explicit ways, with differing degrees of intensity (Stanczak and Augenstein, 2021; Caliskan et al., 2022; Sen et al., 2022). Furthermore, gender bias can easily propagate to models and downstream tasks, causing harm to the end-users (Bolukbasi et al., 2016). These forms of bias can emerge as representational harms and gender gaps.

The current literature about gender bias estimation and mitigation related to word embeddings lacks an understanding of their generalization capabilities. Therefore, this work complements prior research by providing the first systematic evidence on the generalization of estimating gender bias and debiasing techniques, including comprehensive quantitative and qualitative analyses. In particular, we compared in a systematic way four gender bias metrics (Word Embedding Association Test (Caliskan et al., 2017), Relative Negative Sentiment Bias (Sweeney and Najafian, 2019), Embedding Coherence Test (Dev and Phillips, 2019) and Bias Analogy Test (Dev and Phillips, 2019)), two types of projection-based gender mitigation strategies (hard- and soft-debiasing (Bolukbasi et al., 2016)) on three well-known word embedding representations (Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017) and Glove (Pennington et al., 2014)). The main findings of the systematic comparison can be summarized as follows:

- The considered word embeddings are consistent between them but the debiasing techniques are inconsistent across the different bias estimation metrics, underlying controversial generalization capabilities;
- The investigated debiasing techniques, evalu-

Proceedings of Recent Advances in Natural Language Processing, pages 336–343
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_038

ated with respect to multiple points of view, have highlighted the potential risk of unintended bias after the mitigation strategies.

The paper is organized as follows. In Section 2, the most relevant bias estimation metrics are presented. In Section 3 hard and soft debiasing strategies are reported. In Section 4, a systematic comparison is performed, detailing the main findings about the generalization capabilities in debiasing word embeddings. Finally, in Section 5 conclusions are reported and future work is discussed.

2 Measuring Gender Bias

In recent years numerous investigations have been focused on the development of measures to estimate gender bias in embedding methods. The most widely used techniques are: Word Embedding Association Test (Caliskan et al., 2017), Relative Negative Sentiment Bias (Sweeney and Najafian, 2019), Embedding Coherence Test (Dev and Phillips, 2019) and Bias Analogy Test (Dev and Phillips, 2019).

Word Embedding Association Test (WEAT). The Word Embedding Association Test (Caliskan et al., 2017) exploits the Implicit Association Test (IAT) (Greenwald et al., 1998) in order to quantify gender bias in word embeddings through the difference in the strength of association of concepts. In psychology, the Implicit Association Test (IAT) is used to assess the presence of subconscious gender bias in humans. This can be defined as "the difference in time and accuracy that humans take to categorize words related to two concepts they find similar versus two concepts they find different". In detail, WEAT compares sets of identified concepts (i.e., male and female words), denoted as X and Y (each of equal size N, with two sets of biased attributes A and B of equal size N) in order to measure bias over social attributes and roles (i.e., career/family words). The association of a single word x with the bias attribute sets A and B is computed as:

$$f(x, A, B) = \frac{1}{N} \sum_{a \in A} cos(x, a) - \frac{1}{N} \sum_{b \in B} cos(x, b) \quad (1)$$

To estimate the bias in the sets X and Y, the effect sized d is estimated as follows:

$$d(X, Y, A, B) = \frac{\mu_{x \in X} f(x,A,B) - \mu_{y \in Y}(f(y,A,B)}{std_{t \in X \cup Y} f(t,A,B)} \quad (2)$$

where $\mu_{x \in X}(f(x, A, B)$ refers to the mean of $f(x, A, B)$ with x in X and $std_{t \in X \cup Y} f(t,A,B)$ to the standard deviation over all word biases of x in X. The null hypothesis suggests that there is no difference between X and Y in terms of their relative similarity to A and B. In other words, a positive value of $d(X, Y, A, B)$ confirms the hypothesis that words in X are stereotypical for the attributes in A and words in Y stereotypical for words in B, while a negative value of $d(X, Y, A, B)$ suggest that the stereotypes would be opposite. In Caliskan et al. [2017], the null hypothesis is tested through a permutation test, i.e., the probability that there is no difference between X and Y (in relation to A and B) and, therefore, that the word category is not biased.

Relative Negative Sentiment Bias (RNSB). Relative Negative Sentiment Bias (Sweeney and Najafian, 2019) measures the fairness in word embeddings through the relative negative sentiment associated with terms from various protected groups. The idea is to use the embedding model to initialize vectors for an unbiased positive/negative word sentiment dataset. Using this dataset, a logistic classification algorithm is trained to predict the probability of any word being a negative sentiment word. After training, a selected set of neutral identity terms from a protected group (i.e., national origin) is taken to predict the probability of negative sentiment for each word in the set. Neutral identity terms that are unfairly entangled with negative sentiment in the word embeddings will be classified like their neighboring sentiment words from the sentiment dataset.

Given a gold standard of labeled positive/negative sentiment words, (x_i, y_i), where x_i is a word vector from a possibly biased word embedding model, the goal is to minimize the learned weights w of a logistic loss L:

$$min_{w \in R^d} \sum_{i=0}^{n} L(y_i, w^T x_i) + \lambda \|w\|^2, \lambda > 0 \quad (3)$$

where λ is a scalar, known as regularization rate, aimed at reducing over-fitting.

Given a set $K = k_1, ..., k_t$ identity word vectors, we define a set P containing the predicted negative sentiment probability via the minimization of the logistic loss normalized to be one probability mass:

$$P = \left\{ \frac{f^*(k_1)}{\sum_{i=1}^{t} f^*(k_i)}, ..., \frac{f^*(k_t)}{\sum_{i=1}^{t} f^*(k_i)} \right\} \quad (4)$$

The metric $RNSB(P)$ is defined as the KL divergence of P from U, where U is the uniform distribution from the t identity word elements:

$$RNSB(P) = D_{KL}(P\|U) \tag{5}$$

The RNSB metric captures the distance, via KL divergence, between the current distribution of negative sentiment and the fair uniform distribution. The fairer is the word embedding model with respect to sentiment bias, and the lower is RNSB.

Embedding Coherence Test (ECT) Embedding Coherence Test (ECT) (Dev and Phillips, 2019) measures if groups of words have stereotypical associations by computing the Spearman Coefficient of lists of attribute embeddings sorted based on their similarity to target embeddings. In particular, ECT quantifies the amount of explicit bias by comparing vectors of target sets T_1 and T_2 (averaged over the constituent terms) with vectors from a single attribute set A. ECT first computes the mean vectors for the target sets T_1 and T_2:

$$\mu_1 = \frac{1}{|T_1|} \sum_{t_1 \in T_1} t_1 \tag{6}$$

$$\mu_2 = \frac{1}{|T_2|} \sum_{t_2 \in T_2} t_2 \tag{7}$$

Next, for both μ_1 and μ_2 it computes the (cosine) similarities with vectors of all $a \in A$. Finally, the two resultant vectors of similarity scores, $s1$ (for $T1$) and $s2$ (for $T2$), are used to obtain the final ECT score. ECT corresponds to the Spearman's rank correlation between the rank orders of $s1$ and $s2$. In our specific case of gender bias, ECT quantifies the amount of explicit bias by means of the Spearman's rank correlation between the vectors of similarity scores between the attribute words set and the gender target sets. In this case, the higher the correlation and the lower the bias.

Bias Analogy Test. The Bias Analogy Test (BAT) has been introduced in (Dev and Phillips, 2019) as a set of word analogy tests. The main goal is to find the word pair in the best analogy to the pair *(he, she)*. To evaluate the extent of gender bias in word embeddings, we used the SemBias dataset, where each sample contains four-word pairs: a gender-definition word pair (Definition; e.g., *gentleman - lady*), a gender-stereotype word pair (Stereotype; e.g., *doctor - nurse*); the two other pairs consist of words similar in meaning but irrelevant to gender (None; e.g., *cat - dog*, or *flour - sugar*). To quantify the correctness of the analogy of "he-she", for each set of word pairs (Definition, Stereotype, None) the percentage of times that each class of pair is on the top based on a word embedding model is computed. The relational similarity between *(he, she)* and *(a,b)* in SemBias is computed using the cosine similarity between the *(he-she)* gender directional vector and *(a-b)* using the word embeddings under evaluation. For the four-word pairs in each instance in SemBias, we select the word pair with the highest cosine similarity with *(he-she)* as the predicted answer. If the word embedding has been properly debiased, higher values in Definition and lower values in Stereotype and None are expected.

3 Debiasing Methods

Given the potential risk of using Machine Learning algorithms that amplify gender stereotypes contained in pre-trained word embeddings, the main challenge in debiasing tasks is to strike a balance between maintaining model performance on downstream tasks while reducing the encoded gender bias (de Vassimon Manela et al., 2021). To this purpose, projection-based debiasing methods are exploited and compared to determine their generalization capabilities. In this work, we consider two main mitigation strategies, hard- and soft-debiasing.

Hard-debiasing. Hard-debiasing (Bolukbasi et al., 2016; Cheng et al., 2022), also known as *Neutralize and Equalize*, ensures that gender-neutral words are zero in the gender subspace and equalizes sets of words outside the subspace. In order to accomplish this task, hard-debiasing has the goal to satisfy the constraint that any neutral word should be equidistant to all words in each equality set (i.e., a set of words which differ only in the gender component). For instance, taking (grandmother, grandfather) and (guy, gal) as two equality sets, after equalization, *babysit* would result to be equidistant from (*grandmother, grandfather*) and (*gal, guy)*, closer to *grandparent* and further away from the *gal* and *guy*. Instead of completely removing gender information, the approach is aimed at shifting word embeddings to be equally male and female in terms of their vector direction and proposes to modify the embedding space by removing the gender component only

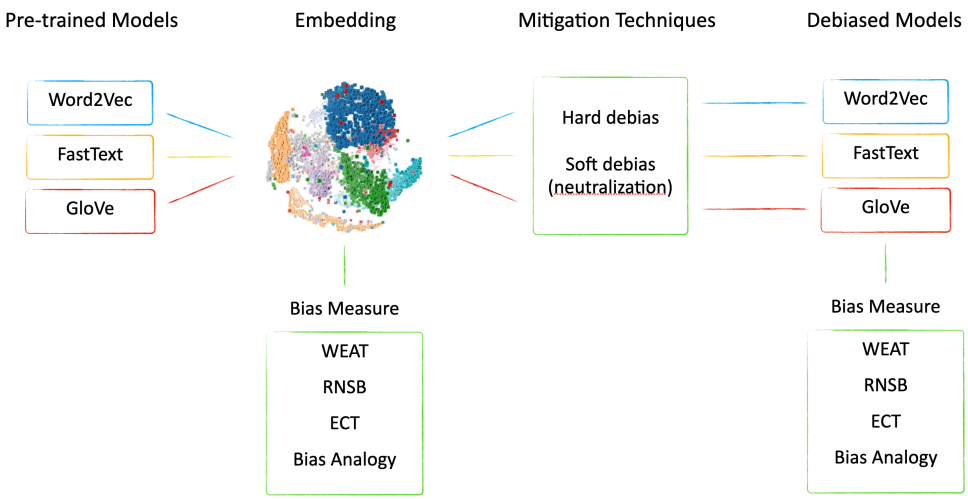


Figure 1: Proposed comparative framework

from gender-neutral words. This approach is appropriate for applications where one does not wish to display any bias in any such pair with respect to neutral words. The disadvantage of equalizing sets of words outside the subspace is that it removes certain specific distinctions that may be of value in specific applications. For instance, Bolukbasi et al. highlight that one may wish a language model to assign a higher probability to the phrase such as *grandfather a regulation* since it is an idiom, unlike *grandmother a regulation*.

Soft-debiasing. The soft-debiasing approach (Bolukbasi et al., 2016) reduces the differences between sets whilst maintaining as much similarity as possible to the original embedding, with a parameter that controls for this trade-off. More specifically, soft-debiasing applies a linear transformation that seeks to preserve pairwise inner products between all the word vectors while minimizing the projection of the gender-neutral words onto the gender subspace. In order to accomplish this task, soft-debiasing exploits a set of gender-definitional words to train a support vector machine and uses it to expand the initial set of gender-definitional words.

4 Generalization Capabilities: A Systematic Comparison

In order to perform a deep analysis of bias measures and mitigation techniques on word embeddings, we selected three of the most well-known and adopted models:

- **Word2Vec:** 300-dimensional embeddings for ca. 3M words learned from Google News corpus (Mikolov et al., 2013)
- **Glove:** 300-dimensional embeddings for ca. 2.2M words learned from the Common Crawl (Pennington et al., 2014)
- **FastText:** 300-dimensional embeddings for ca. 1M words learned from Wikipedia 2017, UMBC web base corpus, and statmt.org news (Bojanowski et al., 2017)

These three models belong to two different families. Both families learn the geometrical encoding (vectors) of words from their co-occurrence information. However, they differ because Word2Vec and FastText are *predictive* models, whereas GloVe is a *count-based* model.

In order to understand and evaluate unintentional gender bias in word embeddings from a comprehensive point of view, we adopted the framework reported in Figure 1. In particular, given the considered word embeddings, the systematic comparison for understanding the generalization capabilities of the examined gender-debiasing techniques is performed according to the following three main steps: (1) estimation of the gender-bias metrics, (2) exploiting both hard- and soft-debiasing methods and (3) evaluating the debiased embeddings using the same bias measures before and after the mitigation strategy. To evaluate the pre-trained word embeddings, we use the four metrics, comparing the results before and after the mitigation strategies.

We report in Tables 1, 2 and 3 the corresponding values according to seven sets of different target words and multiple male and female attribute words. For each metric, we computed the values obtained by the considered models according to the (**o**)original embedding, the (**s**)oft debiased, and the (**h**)ard debiased ones.

	Word2Vec			FastText			GloVe		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career-Family	0.35	-0.12	**0.03**	0.38	0.03	**0.03**	0.41	-0.10	**0.01**
Math-Arts	0.71	-0.20	**-0.09**	0.66	0.19	**0.01**	0.38	**-0.01**	-0.03
Science-Arts	0.90	-0.01	**0.00**	0.89	0.29	**0.09**	1.06	-0.07	**-0.06**
Intel.-Appearance	1.18	**-0.12**	-0.21	0.94	0.16	**-0.14**	0.96	**0.04**	-0.09
Intel.-Sensitive	0.91	0.21	**-0.07**	0.45	0.12	**-0.06**	0.69	**0.03**	-0.07
Pos-Neg words	-0.40	-0.30	**-0.18**	-0.32	-0.27	**-0.13**	-0.42	-0.23	**-0.05**
Man-Woman roles	1.83	0.97	**0.74**	1.81	1.06	**0.78**	1.78	0.87	**0.82**

Table 1: WEAT values for target word groups with respect to male and female terms.

The first measure we evaluate is the Word Embedding Association Test (WEAT) where, for each target group we computed the association with the set of male and female attribute words (pronouns). In table 1 we highlight in **bold** the best results obtained by each model. At first glance, it seems that the considered debiasing operations have affected the WEAT value for all the embeddings. Compared to the original version, all three embeddings show a significant improvement in both soft and hard debiased embeddings. Nevertheless, ***Word2Vec and FastText have a noticeable tendency to the hard debiased embedding, while Glove has very similar values for the soft and hard embeddings***.

Regarding the Relative Negative Sentiment Bias (RNSB) metric, it can be interpreted as the distance between the current distribution of negative sentiment and the fair, uniform distribution. Therefore, the fairer a word embedding model is with respect to sentiment bias, the lower the RNSB metric should be. The results in Table 2, although RNSB is not directly comparable with WEAT, seem to be coherent.

	Word2Vec			FastText			GloVe		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career-Family	.0059	**.0057**	.0065	.0026	**.0022**	.0031	.0075	.0047	**.0036**
Math-Arts	.0008	**.0006**	.0007	.0008	.0006	**.0005**	.0012	.0011	**.0010**
Science-Arts	.0005	.0006	**.0003**	.0005	.0005	**.0004**	.0006	.0006	**.0004**
Intel.-Appearance	.0069	**.0035**	.0037	.0062	**.0035**	.0042	.0100	.0059	**.0048**
Intel.-Sensitive	.0022	.0019	**.0016**	.0021	**.0014**	.0020	.0024	**.0016**	.0018
Pos-Neg words	.0204	.0165	**.0134**	.0499	.0454	**.0404**	.0339	.0324	**.0293**
Man-Woman roles	.0076	**.0011**	.0012	.0029	.0006	**.0003**	.0051	.0008	**.0005**

Table 2: RNSB values for target word groups with respect to male and female terms.

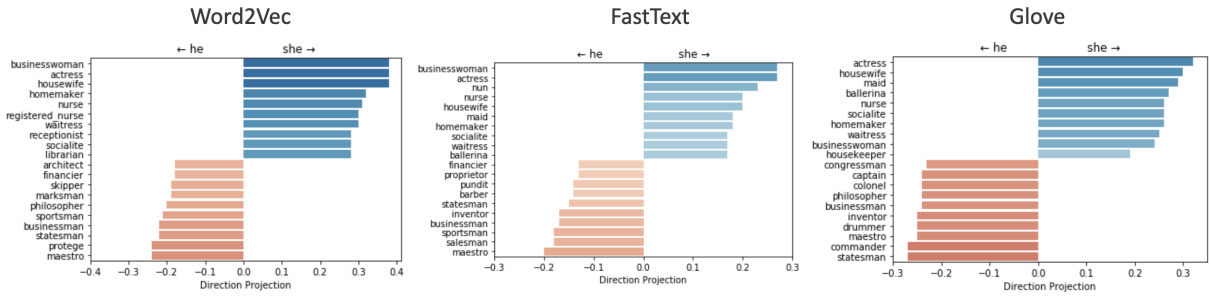
For what concerns the Relative Negative Sentiment Bias metric, it can be interpreted as the distance between the current distribution of negative sentiment and the fair, uniform distribution. Therefore, the fairer a word embedding model is with respect to sentiment bias, the lower the RNSB metric should be. The results in table 2, although RNSB is not directly comparable with WEAT, seem to be coherent. All the models seem to be improving in the debiased embedding. However, it is necessary to make a few considerations about RNSB with respect to WEAT: 1) the relative improvement from the original to the hard debiased embeddings is much more moderate in RNSB than in WEAT and 2) ***in contrast to WEAT values, GloVe's best embeddings in terms on RNSB is the hard debiased one, while Word2Vec and FastText's best model seems to swing between soft and hard***.

Regarding the Embedding Coherence Test (ECT), it quantifies the amount of explicit bias and returns the Spearman's rank correlation between the vectors of similarity scores between the attribute word set and the gender target sets. The results in Table 3 seem to confirm the considerations related to WEAT and RNSB, denoting an improved representation (less biased) with respect to the original embedding. In particular, we found out that the best debiased embedding is the one generated with the hard debiased technique. Nevertheless, we noticed that ***ECT's values are extremely high in the soft or even the original embedding for some attribute words***. In fact, the Spearman correlations are close to 1, indicating that the two variables being compared are monotonically related, even if their relationship is not linear.

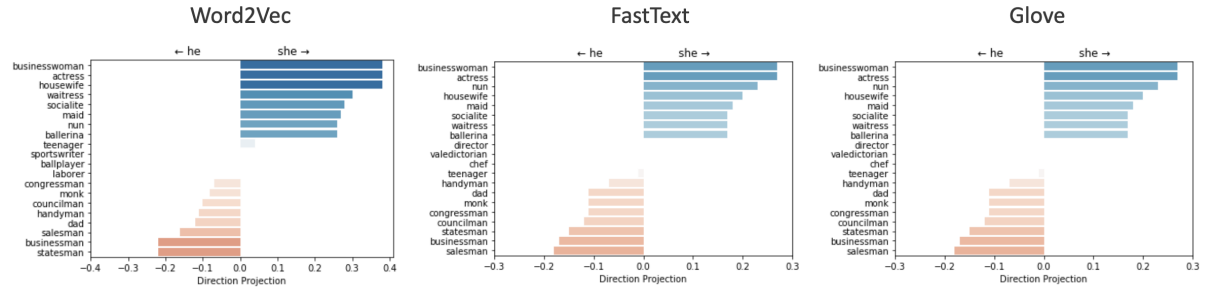
	Word2Vec			FastText			GloVe		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career	.714	**1.00**	1.00	**.952**	.929	.952	.976	.976	**1.00**
Family	.762	.833	**1.00**	.952	.976	**.976**	.905	.976	**1.00**
Science	.571	.857	**1.00**	.976	.976	**1.00**	.976	**1.00**	1.00
Arts	.810	.952	**.976**	.833	.929	**1.00**	.929	.952	**.952**
Appearance	.363	.879	**.904**	.507	.833	**.858**	.448	.952	**.965**
Intelligence	.744	.976	**.998**	.841	.943	**.991**	.916	.990	**.999**
Pleasant	.733	.978	**.983**	.943	966	**.989**	.938	.978	**.997**
Unpleasant	.800	.962	**.984**	.872	.912	**.976**	.900	.976	**.985**
Positive words	.771	.972	**.994**	.925	.982	**.997**	.936	.992	**.999**
Negative words	.791	.964	**.993**	.939	.981	**.997**	.954	.992	**.999**
Man roles	.972	.986	**.993**	.979	.972	**1.00**	.958	.958	**.993**
Woman roles	.747	**.956**	.879	.780	.885	**.901**	.511	**.923**	.736

Table 3: ECT values for target word groups with respect to male and female terms.

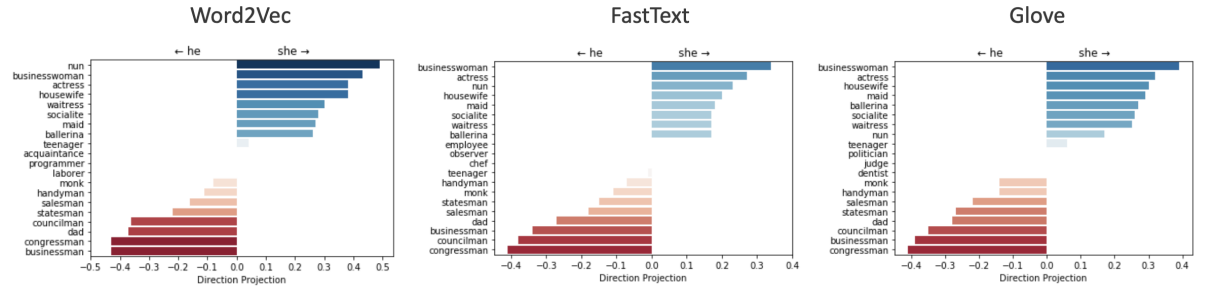
We report in Figure 2(a), 2(b) and 2(c) the *Gender Direction* for different occupations for each pre-trained model according to the original embeddings and the two debiasing techniques. Although there is an improvement for all models from the



(a) Original Embeddings



(b) Soft Debiasing



(c) Hard Debiasing

Figure 2: Gender direction for occupations in Original embeddings, Soft and Hard Debiasing.

original to the hard debiased embedding, we can observe a few potentially biased representations in the *she* direction. In particular terms such as *maid*, *waitress* and *housewife* do not constitute a form of directly observable bias, but the absence of male equivalent terms is a potential warning.

Although the analysis carried out to this point seems to confirm that the embeddings have been successfully debiased, the qualitative evaluation of the results has brought out some concerns regarding the actual presence of bias. To this purpose, we evaluated the embeddings adopting the Bias Analogy Test reporting the results in Table 4. The debiased models show lower values in **Definition** than the original embedding, suggesting the presence of bias.

In particular, for the word pairs Definition, Stereotype and None, for each pre-trained model, the only improvement from the original embedding appears to be in the **Stereotype** values of the soft embedding.

	Word2Vec			**FastText**			**GloVe**		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Definition	**.826**	.823	.795	**.911**	.777	.820	**.835**	.770	.809
Stereotype	.134	**.102**	.116	.065	**.048**	.061	.115	**.077**	.079
None	**.039**	.075	.089	.023	.175	**.119**	**.050**	.152	.111
Sub-Definition	.600	**.700**	.500	**.825**	.500	.700	**.675**	.525	.500
Sub-Steretype	.300	**.200**	.275	.125	.125	**.100**	.275	**.125**	.225
Sub-None	**.100**	**.100**	.225	**.050**	.375	.200	**.050**	.350	.275

Table 4: BAT values for pre-trained models.

Regarding the sub-metrics reported in the bottom part of the table (Sub-Definition, Sub-Stereotype and Sub-None), they spotlight a bad generalization ability for all the embeddings when compared with their corresponding original metrics. The obtained results on the BAT metric coupled with the gen-

der direction analysis, being inconsistent with the previous remarks on WEAT, RNSB and ECT, highlight the potential risk of unintended bias after the mitigation strategies.

5 Conclusions and Future Work

In this paper, a systematic comparison of different bias estimation metrics, mitigation strategies and word embeddings has been performed. The computational investigation highlighted analogies and dissimilarities among metrics, pointing out the importance of using different types of measures to have a wider overview of the generalization capabilities of the two most important debiasing techniques. The experiments have shown that the considered word embeddings are consistent between them but inconsistent across the different metrics. Although WEAT, RNSB and ECT values are coherent, the gender direction of occupations and the BAT values are signals reflecting the presence of bias in the supposed debiased models. A future research investigation relates to the evaluation of multiple bias metrics not only on word embeddings but also on transformer-based representations as contextualized word embeddings. Finally, a generalization of the proposed investigation should be pursued on generative language models.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. volume 356, pages 183–186. American Association for the Advancement of Science.

Lu Cheng, Nayoung Kim, and Huan Liu. 2022. Debiasing word embeddings with nonlinear geometry. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1286–1298.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd international conference on artificial intelligence and statistics*, pages 879–887. PMLR.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. volume 74, page 1464. American Psychological Association.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.