MDR Cluster-Debias: A Nonlinear Word Embedding Debiasing Pipeline

Yuhao Du[0000−0002−2474−8529] and Kenneth Joseph[0000−0003−2233−3976]

University at Buffalo, Buffalo NY 14260, USA
{yuhaodu,kjoseph}@buffalo.edu

Abstract. Existing methods for debiasing word embeddings often do so only superficially, in that words that are stereotypically associated with, e.g., a particular gender in the original embedding space can still be clustered together in the debiased space. However, there has yet to be a study that explores why this residual clustering exists, and how it might be addressed. The present work fills this gap. We identify two potential reasons for which residual bias exists and develop a new pipeline, MDR Cluster-Debias, to mitigate this bias. We explore the strengths and weaknesses of our method, finding that it significantly outperforms other existing debiasing approaches on a variety of upstream bias tests but achieves limited improvement on decreasing gender bias in a downstream task. This indicates that word embeddings encode gender bias in still other ways, not necessarily captured by upstream tests.

Keywords: Word Embedding · Social Bias · Debias.

1 Introduction

A literature has rapidly developed around the question of how to identify, characterize, and remove bias from ("debias") word embeddings. Attempts to do so are critical in ensuring that real-world applications of natural language processing (NLP) do not cause unexpected harm. For example, word embeddings that reflect stereotypical and/or prejudicial social norms might be used as input to other algorithms that, e.g., rank men higher than more qualified women for job searches of particular occupations [2].

However, recent work has raised questions about existing efforts to measure biases in word embeddings, and our ability to debias them. With respect to measurement, Ethayarajh et al. [4] provide both empirical and theoretical evidence that the most common method of measuring bias in word embeddings, the *Word Embedding Association Test* (WEAT), provides unreliable measures of practical and statistical significance. With respect to debiasing, Gonen and Goldberg [5] provide experimental evidence that male-stereotyped words are still easily distinguishable from female-stereotyped words after running two of the most well-known methods for debiasing, the Hard-Debias method of [1] and the Gender Neutral GloVe (GN-Glove) approach [20].

These two debiasing methods, like nearly all others, operate under the assumption that social biases in word embeddings can be defined as a specific direction (or in some cases, a subspace) of the embedding space. This direction is characterized by the difference between sets of bias-defining words. For example, the Hard-Debiasing method of Bolukbasi et al. [1] approach works roughly as follows for gender debiasing. First, a "gender direction" is identified by using differences in the embedding space between sets of gender-paired words, e.g. "man" and "woman". This direction is then essentially removed from all other words,[1] with the idea being that gender will no longer be represented by the embeddings of the remaining terms because all gender information, contained on the gender direction, is now gone.

What Gonen and Goldberg show is that while this approach removes some forms of gender information, one can still easily pick up gender stereotypes in word embedding space based on different, but equally valid, definitions of bias metrics. Inspired by their work, we propose a new debiasing procedure which combines a post-processing step, introduced in [7], to unfold manifolds in high word embedding space, followed by a simple linear debiasing approach, Cluster-Debias, that finds a better direction along which to remove bias to address these cluster-based bias measures.

We evaluate our debiasing approach for several "upstream" tasks, including bias tests and word similarity tests. In addition, we ask, what does this means for downstream performance on a standard NLP task? We compare embeddings debiased using our approach with the approaches of [20] and [1] on a coreference resolution task and a sentiment analysis task. Through these efforts, the present work makes the following contributions to the literature:[2]

- We find evidence that debiased embedding clusters are partially due to manifold structure in high dimensional word embedding space.
- We introduce a new pipeline to perform debiasing, and show it can reduce the clustering-based word embedding bias measures introduced by [5].
- However, despite significant upstream improvements, our approach does not significantly decrease bias in the downstream task of coreference resolution.

2 Related Work

2.1 Bias Definition

Critical to debiasing is how bias is actually defined. The vast majority of works use a directional definition. Under this approach, a single direction in the embedding space defines a particular bias that the authors expect to exist, e.g. the "gender direction".Detractors of the directional definition of bias, like Gonen and Goldberg, have argued that it is inappropriate, because a single gender (or

[1] Except for "gender definitional" words like "king and queen"

[2] Code and data to replicate our work are at https://github.com/yuhaodu/MDR-Cluster-Debias.git.

race, etc.) dimension may not capture all forms of bias encoded in the data. This issue has led others to define bias in terms of clustering, or word proximity. The idea is that the removal of a single dimension, or subspace [12], is not sufficient to remove bias, in that one can easily identify terms that are close to the opposing seed terms in the original embeddings in the unbiased embeddings as well. Because of this, protected information (e.g. gender) can potentially leak to machine learning algorithms in downstream tasks. The primary contribution of our work is to propose a debiasing pipeline to resolve these *cluster-based biases*.

2.2 Debiasing Word Embeddings

Several methods have been proposed to remove social biases from various kinds of NLP methods; see Sun et al. [17] for a recent review on gender specifically. Bolukbasi et al. [1] proposed two methods for word embeddings specifically, Hard Debiasing and Soft Debiasing. These methods remove gender neutral words' projection over a gender space defined by gender definitional words. Zhao et al. [20] modify the GloVe algorithm to train debiased embeddings directly from a co-occurrence matrix by adding constraints in the training objective of GloVe [14] to force gender neutral words perpendicular to some gender space.

Except these two seminal works, several others have proposed novel methods for debiasing. Most of these have been extensions of the hard-debiasing method. [12] extend the hard debias method to the multi-class setting, [4] improve the way gender-biased words are selected, and [3] propose simpler versions of the algorithm and the use of names as a means of identifying directions in the space that represent social biases. One exception is the work of Kaneko et al. [8], who propose an autoencoder based method which is able to project current word embedding into another space which preserves the word semantic information while removing the gender bias. However, their work still evaluates results using a directional approach. The present work extends current debiasing algorithms in terms of debiasing based on recently identified cluster-based bias definition.

3 Our Debiasing Pipeline

We base our debiasing pipeline on pretrained GloVe embeddings [14]; however, the approach generalizes to any other pretrained embedding. The pipeline contains two parts: the first is a post-processing procedure, which is used to re-embed original word vectors into a new space via a manifold learning algorithm. The second is the application of a direction-based debiasing method to remove gender information in the re-embedded word vectors.

3.1 Post-Processing Procedure

As a post-processing procedure, we use Manifold Dimensionality Retention (MDR) from [7]. Hasan et al. [7] are motivated by the observation that word embeddings slightly underestimate the similarity between similar words and overestimate

the similarity between distant words. This indicates that word embedding space contains non-linear manifold structure. Thus, they propose the MDR to unfold the manifold structure to improve word representation and results show that re-embedded word embeddings achieve better performance in word similarity tests. Inspired by their observation and Gonen and Goldberg's observation that gendered words are easily separated by a non-linear SVM method, we believe non-linear manifold structure in the word embedding space could potentially prevent linear directional based debias method from mitigating gender bias. Thus, we apply MDR as a post-processing procedure.

In MDR, we start from an original embedding space with vectors ordered by words frequencies. We then carry out the following steps:

1. Select a sample window of vectors that are used to learn the manifold.
2. Fit a manifold learning model to the selected sample using Locally Linear Embedding (LLE) [16].
3. The resulting fitted model is then used to transform all the word vectors in the original space to the new re-embedding space.

In Step 1, a sample window is sliding on the word vectors ordered by word frequencies. The window length L and window start S of the sample window are hyper-parameters. Additional, S will decide the computational complexity on manifold learning. As shown in prior work [6], trained word embeddings are biased toward word frequency. In order to keep learned manifold from skewing towards high frequency or low frequency words, we select S as 5000. For the choice of L, we choose 1000 following the suggestion introduced in the prior work [7]. Selections of these two parameters work well in terms of preserving semantic information in word embeddings which is shown later in Section 5.1.

3.2 Cluster-Debias

Gonen and Goldberg show that, after *debiasing*, one can still easily cluster biased words using linear K-means clustering method. We hypothesize that this observation is due to a mismatch between the direction that previous debiasing method removes and that gender bias lies along. Thus, we propose a simple approach that incorporates a cluster-based definition of bias to perform debiasing. The procedure carried out by Cluster-Debias is as follows:

1. Identify, via a particular word pair a and b (e.g. "he" and "she"), the form of bias to be addressed.
2. Identify the bias subspace by $D_{bias} = E_a - E_b$, where E_w represents the word vector of word w.
3. Calculate the bias of word vectors along D_{bias} following the methods in [1]
4. Select the top k most biased words, i.e. the k nearest neighbors to a and b
5. Apply PCA over these $2k$ word vectors and extract the first principle component D_{pc}.
6. Debias all word vectors E_ws by removing D_{pc} from them. This can be expressed as $E_w^{'} = E_w - \langle E_w \, , D_{pc} \rangle \cdot D_{pc}$.

At a high level, our approach retains much of the logic from Bolukbasi et al. [1]. However, instead of assuming that the gender direction is aligned with the word pairing(s) we identify, we instead assume that this direction can be better identified by incorporating information from the distribution of the word vectors that are proximal to a and b. We therefore assume, based on the observations of Gonen and Goldberg [5], that it is more appropriate to select a direction based on the clustered structure of the embedding space around the words of interest, rather than on those words themselves. Note that it is not guaranteed that the Cluster-Debias approach will overcome the issues with Hard-Debias. Like the Hard-Debias method, we remove only a single dimension from the embedding space. This direction is simply more informed by clustering structure than the prior work. Here, we focus on comparing to prior work, and so consider gender. Thus a and b are "he" and "she", respectively. Additionally, we set $k = 1000$ for all experiments below.

4 Evaluation Methods

There is, of course, a tradeoff between removing gender information and maintaining other forms of semantic information that are useful for downstream tasks. As such, we evaluate embeddings from bias-based evaluation measures and semantic-based evaluation measures. In addition, we are also interested in whether or not upstream evaluation results can be transferred to downstream tasks. As such, our evaluation is carried out along two dimensions – bias-based versus semantic-based and upstream versus downstream.

As an upstream measure of semantics, we focus on semantic similarity-based measures. We compute the cosine similarity between word embeddings and measure Spearman correlation between human similarity rating and cosine similarity for the same semantic relatedness datasets used to evaluate biased embeddings in prior work [18].For downstream evaluation, we identify two NLP tasks – coreference resolution and sentiment analysis. To build coreference resolution models, we use the coreference resolution system proposed in [10]. We apply the original parameter settings for the model and train each model for 100K iterations and evaluate models with respect to their performance on the standard OntoNote v5 dataset [15]. For sentiment analysis, we train an LSTM with 100 hidden units on the Stanford IMDB movie review dataset [11] and we also leverage the model [9] to train a binary classifier on the MR dataset of short movie reviews [13].

For bias-based evaluation, we use the same six cluster-based bias measures that are proposed by Gonen and Goldberg as our upstream bias-based evaluation tasks. The first one we call *Kmeans Accuracy*. We first select the top 500 nearest neighbors to the terms "he" and "she" in the original embedding space. We then check the accuracy of alignment between gendered words and clusters identified by Kmeans. The second one we call *SVM Accuracy*. We consider the 5000 most biased words (2500 from each gender) in the original embedding space. *After debiasing*, we check the accuracy of a RBF-kernel SVM trained on a random sample of 1000 of these words (500 from each gender) predicting gender bias of

the remaining 4000. The third one we call *Correlation Profession*. We extract the list of professions used in [1] and compare the correlation between the percentage of male/female socially biased words among the k nearest neighbors of the professions and their directional bias in the original embedding space. For three metrics listed above, lower scores indicate better debiasing results. The rest three are gender-related association experiments called *WEAT* introduced in [2]. Three experiments evaluate the associations between female/male and family and career words, arts and mathematics words, arts and science words respectively. For these tests, a higher p-value means lower association which indicates better debiasing results.

For downstream evaluation of bias, We again leverage the coreference model that is trained following the procedure introduced above as our model for bias-based tests. The difference between here and there is that we compare performance using the gendered coreference resolution dataset WinoBias developed by Zhao et al. [19]. The testing portion of the WinoBias dataset evaluates the extent to which a coreference resolution model exhibits gender stereotyping by assessing the degree to which it applies gender stereotypical pronouns to individuals described using a set of gender-associated occupations. They create two different datasets—"anti-stereotype", in which gender associations are reversed (e.g. "The secretary ... he"), and "pro-stereotype", in which gender associations are retained (e.g. "The secretary ... she"). Differences in performance between the two datasets are used as an indicator of gender bias in the coreference dataset and gender bias in coreference algorithm.

5 Experiments

We train GloVe [14] on a 2017 dump of English Wikipedia to obtain pre-trained 300-dimensional word embeddings for 362179 words. We then create several baselines and word embeddings debiased by our proposed methods:

GloVe: is the pretrained word embedding introduced above. This baseline denotes a non-debiased version of the word embeddings.

Hard-GloVe: We apply hard-debiasing [1] method by using released code [3] to our pretrained GloVe word embedding and obtain a hard-debiased version of the pretrained GloVE embeddings.

GN-GloVe: We apply the code [4] from original authors of GN-GloVe [20] and train our own version of GN-GloVe.

Cluster-GloVe: We apply Cluster-Debiased method to our pretrained GloVe embeddings to obtain debiased GloVe embeddings.

MDR-GloVe: We apply our Post-Processing Procedure MDR on pretrained GloVe embeddings.

MDR-Cluster: We apply the proposed Post-Processing Procedure on pretrained GloVe embedding and then use Cluster-debias method to debias it.

[3] https://github.com/tolga-b/debiaswe
[4] https://github.com/uclanlp/gn_glove

Table 1. Results for our upstream bias evaluations. The first three rows are extracted directly from prior work [5]. The last four rows are the debiased word embeddings using our proposed pipeline. Bolded results are the best-performing in each column according to the given metric.

Embedding	Kmeans Acc.	Corr. Prof.	SVM Acc.	Work/Family P-val	Math/Art 1 P-val	Math/Art 2 P-val
Original GloVe	0.999	0.820	.99			
Hard-GloVe	0.925	0.606	.89	<.0001	<.0001	.0467
GN-GloVe	0.856	0.792	.97	<.0001	<.0001	<.0001
Cluster-GloVe	**0.53**	0.74	0.80	<.0001	**0.76**	0.20
MDR-GloVe	1.000	0.88	0.99	<.0001	0.09	0.03
MDR-Cluster	0.556	**0.38**	**0.518**	0.00015	0.43	0.26
MDR-Hard	0.915	**0.38**	0.86	**0.002**	0.42	**0.51**

MDR-Hard: To test whether our Post-Processing Procedure works for other debias methods. We apply proposed Post-Processing Procedure on pretrained GloVe embedding and then use Hard-debias method to debias it.

5.1 Results

Upstream Cluster-based Bias Test Table 1 displays results from our upstream bias evaluations, and shows that our new debiasing strategies significantly improve over prior work. Results can be summarized as follows:

1. Cluster-GloVe outperforms GN-Glove and Hard-GloVe on all cluster-bias based tests because of following reasons. First, Post-debias Cluster Accuracy of Cluster-GloVe is 0.53, which means that debiased gendered words are not separable by K-means. Cluster-GloVe is also the most difficult to classify post-hoc using an SVM (it has lowest SVM Classifier Accuracy). And it shows no gender bias on two of the three WEAT tests (p-values of last two columns show no significant results). With respect to deficiencies in the Cluster-GloVe, embeddings still are highly separable by SVM, and as evidenced from the correlational professions experiment and the Work/Family WEAT, retain gender stereotypes for occupations.
2. Our Post-Processing method is able to help not only the Cluster-Debias method but also HardDebias. MDR-Cluster and MDR-Hard outperform Cluster-GloVe and Hard-GloVe respectively.
3. MDR-Cluster achieves the best overall performance and is the only method that prevents SVM from classifying gender stereotyped words, which validates the efficiency of our debias pipeline. But MDR-Cluster still struggles with work/family associations WEAT test.

These results provide two insights into the observations of Gonen and Goldberg. First, Cluster-GloVe, as a directional based debias method, out-performs

Table 2. Performance on co-reference resolution task for models trained using the given embedding. All performance scores are given as F1 scores. Bolded results are the best-performing in each column.

Model	OntoNote	Anti-Stereotype	Pro-Stereotype	WinoBias Mean	WinoBias Diff.
GloVe	72.49	60.995	81.535	71.265	20.54
Hard-Debias	71.87	63.27	77.69	70.48	14.42
GN-GloVe	**72.69**	65.47	81.415	73.4425	15.945
Cluster-debias	71.94	63.685	82.125	72.905	18.44
MDR	71.93	65.715	**83.59**	**74.6525**	17.875
MDR-Hard	71.70	66.18	79.78	72.98	**13.6**
MDR-Cluster	72.01	**66.73**	80.66	72.69	13.93

Hard-GloVe (also a directional based debias method) in terms of removing post-bias clusters identified by K-Means. This observation suggests that there is a mismatch between the direction that gendered words distribute along and the direction that prior debias methods remove. Second, the fact that Cluster-GloVe removes post-debias clusters identified by K-Means, but not non-linear SVM, and that MDR-Cluster removes both, suggests that manifold structure in the word embedding space is able to leak protected gender information to non-linear method (e.g. SVM). That validates our decision on using MDR to unfold manifold structure in the word embedding space as our post-processing step.

Upstream Semantic Similarity and Relatedness We find that, compared with others, word embeddings debiased by our proposed pipeline achieve as-good or higher performance on most benchmark datasets for our upstream semantic test. The most critical comparison is to the original embeddings, where on average, MDR-Cluster achieves 60.2 Pearson correlation with the ground truth ratings on the five benchmark tasks, while GloVe achieves 56.2. This indicates that, according to the word similarity test metric, our proposed debiased pipeline can keep or amplify the semantic information in the original word embeddings.[5]

Downstream Results - Coreference Resolution Table 2 shows the performance difference between coreference resolution algorithms based on the different debiased GloVe embeddings. Among the embeddings considered, we find that MDR-Hard and MDR-Cluster show the best performance on the WinoBias datasets on WinoBias Difference and the Anti-Stereotype metric. This suggests that in addition to showing improvements on the tasks studied by Gonen and Goldberg, MDR-Hard and MDR-Cluster methods we propose can attenuate more protected gender information in the word embeddings. We further can find that different methods' performances are similar on OntoNote dataset which indicates that our debias pipeline doesn't deprecate the semantic information that

[5] Full result tables are available at https://github.com/yuhaodu/MDR-Cluster-Debias.git

are essential for coreference resolution. However, differences between the various debiasing strategies are limited, compared to their overall difference from the unbiased, original embeddings.

Downstream Results - Sentiment Analysis Finally, we find that the models trained on MDR-cluster achieve a similar accuracy on sentiment classification on the MR dataset. However, using the MDR-Cluster embeddings, accuracy on IMDB dataset is 68.5% (95 % confidence interval [68.2%, 68.9%]), while using GloVe embeddings, accuracy is 80.9% ([80.5%,81.3%]). This drop in sentiment analysis performance is indicative that debiasing along certain dimensions of stereotyping (e.g., gender), may have important downstream effects. Although the present work focused on addressing issues raised in [5], this finding on an important downstream task suggests future work is needed on this point.

6 Conclusion

The present work addresses the fact, introduced in prior work [5], that gendered terms remain clustered in the embedding space of debiased word embeddings. We propose a two-step pipeline solution to combat this issue. Our pipeline combines a post-processing step —MDR [7] and a debiasing method—Cluster Debias. It is able to outperform state-of-art debias methods on mitigating bias on the measures proposed by Gonen and Goldberg [5]. The success of our pipeline also validates our proposed reasons behind the observations made by Gonen and Goldberg. First, that there existed a mismatch between the direction that gendered terms distributed along and the direction that debiasing methods remove. And second, that the non-linear classifier (e.g. SVM) is able to separate gendered words from manifold structure in the high dimensional word embedding space.

We also test our pipeline on downstream tasks. We find that our model outperforms existing approaches on the coreference resolution tasks in terms of mitigating gender bias. However, critically, the improvement seen is not nearly as stark as our improvement over prior methods on the upstream bias tasks we consider. This indicates that word embeddings encode gender bias in *still* other ways, not necessarily captured by the cluster-based measures from prior work. As such, in order to avoid a "whack-a-mole" approach for mitigating bias, we encourage a focus on the development of more downstream tasks, relative to further upstream analysis.

References

1. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (2016)

2. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (Apr 2017)
3. Dev, S., Phillips, J.: Attenuating bias in word vectors (2019)
4. Ethayarajh, K., Duvenaud, D., Hirst, G.: Understanding undesirable word embedding associations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
5. Gonen, H., Goldberg, Y.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: NAACL-HLT (2019)
6. Gong, C., He, D., Tan, X., Qin, T., Wang, L., Liu, T.Y.: Frage: Frequency-agnostic word representation. In: Proceedings of the 32th International Conference on Neural Information Processing Systems (2018)
7. Hasan, S., Curry, E.: Word re-embedding via manifold dimensionality retention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017)
8. Kaneko, M., Bollegala, D.: Gender-preserving Debiasing for Pre-trained Word Embeddings. arXiv:1906.00742 [cs] (Jun 2019)
9. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (2014)
10. Lee, K., He, L., Zettlemoyer, L.: Higher-order coreference resolution with coarse-to-fine inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (2018)
11. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
12. Manzini, T., Lim, Y.C., Black, A.W., Tsvetkov, Y.: Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. ArXiv **abs/1904.04047** (2019)
13. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (2005)
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543 (2014)
15. Ralph Weischedel, Martha Palmer, M.M.E.H.S.P.L.R.N.X.A.T.J.K.M.F.M.E.B.R.B.A.H.: OntoNotes v5 release. https://catalog.ldc.upenn.edu/LDC2013T19, accessed: 2019-08-22
16. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science (2000)
17. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.W., Wang, W.Y.: Mitigating Gender Bias in Natural Language Processing: Literature Review. arXiv:1906.08976 [cs] (Jun 2019)
18. Zablocki, E., Piwowarski, B., Soulier, L., Gallinari, P.: Learning multi-modal word representation grounded in visual context. In: AAAI (2017)
19. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: NAACL-HTC (2018)
20. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: EMNLP (2018)