Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Abstract

Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender bias in word embeddings, demonstrating convincing results. However, we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between "gender-neutralized" words in the debiased embeddings, and can be recovered from them. We present a series of experiments to support this claim, for two debiasing methods. We conclude that existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling.

1 Introduction

Word embeddings have become an important component in many NLP models and are widely used for a vast range of downstream tasks. However, these word representations have been proven to reflect social biases (e.g. race and gender) that naturally occur in the data used to train them (Caliskan et al., 2017; Garg et al., 2018).

In this paper we focus on gender bias. Gender bias was demonstrated to be consistent and pervasive across different word embeddings. Bolukbasi et al. (2016) show that using word embeddings for simple analogies surfaces many gender stereotypes. Caliskan et al. (2017) further demonstrate association between female/male names and groups of words stereotypically assigned to females/males.

Recently, some work has been done to reduce the gender bias in word embeddings, both as a post-processing step (Bolukbasi et al., 2016) and as part of the training procedure (Zhao et al., 2018). Both works substantially reduce the bias with respect to the same definition: the projection on the gender direction (i.e. $\overrightarrow{he} - \overrightarrow{she}$), introduced in the former.

We argue that current debiasing methods, which lean on the above definition for gender bias and directly target it, are mostly hiding the bias rather than removing it. We show that even when drastically reducing the gender bias according to this definition, it is still reflected in the geometry of the representation of "gender-neutral" words, and a lot of the bias information can be recovered.

2 Gender Bias in Word Embeddings

Definition Bolukbasi et al. (2016) define the gender bias of a word w by its projection on the "gender direction": $\overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$, assuming all vectors are normalized. The larger a word's projection is on $\overrightarrow{he} - \overrightarrow{she}$, the more biased it is. They also quantify the bias in word embeddings using this definition and show it aligns well with social stereotypes.

Remaining bias after using debiasing methods Both Bolukbasi et al. (2016) and Zhao et al. (2018) propose methods for debiasing word embeddings, substantially reducing the bias according to the suggested definition.

Both works provide very compelling results as evidence of reducing the bias without hurting the performance of the embeddings for standard tasks. However, both methods and their results rely on the

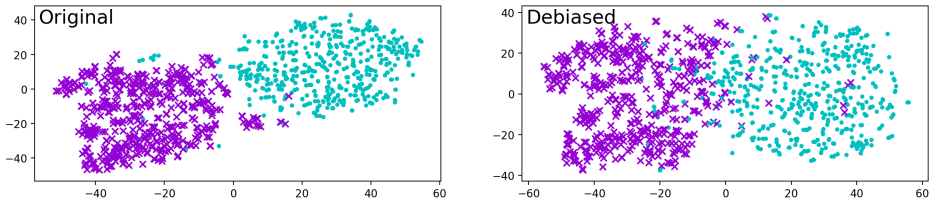

Figure 1: Clustering the 1,000 most biased words, before (left hand-side) and after (right hand-side) debiasing using HARD-DEBIASED embedding.

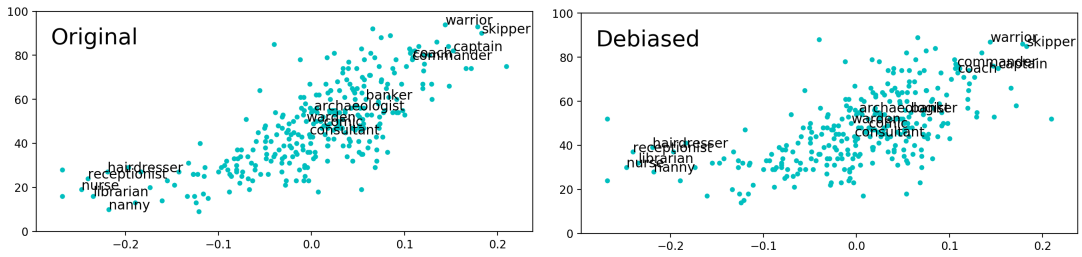

Figure 2: The number of male neighbors for each profession as a function of its original bias, before and after debiasing for HARD-DEBIASED embedding. We show only a limited number of professions on the plot to make it readable.

specific bias definition. We claim that the bias is much more profound and systematic, and that simply reducing the projection of words on a gender direction is insufficient: it merely hides the bias, which is still reflected in similarities between "gender-neutral" words.

Our key observation is that, almost by definition, most word pairs maintain their previous similarity, despite their change in relation to the gender direction. The implication of this is that most words that had a specific bias before are still grouped together, and apart from changes with respect to specific gendered words, the word embeddings' spatial geometry stays largely the same.

3 Experiments and Results

We refer to the word embeddings of the previous works as HARD-DEBIASED (Bolukbasi et al., 2016) and GN-GLOVE (gender-neutral GloVe) (Zhao et al., 2018), which are word2vec-based (Mikolov et al., 2013) and GloVe-based (Pennington et al., 2014), respectively. For each debiased word embedding we quantify the hidden bias with respect to the biased version. For HARD-DEBIASED we compare to the embeddings before applying the debiasing procedure. For GN-GLOVE we compare to embedding trained with standard GloVe on the same corpus.[1]

Male- and female-biased words cluster together We take the most biased words in the vocabulary according to the original bias (500 male-biased and 500 female-biased), and cluster them into two clusters using k-means. For the HARD-DEBIASED embedding, the clusters align with gender with an accuracy of 92.5% (according to the original bias of each word), compared to an accuracy of 99.9% with the original biased version. For the GN-GLOVE embedding, we get an accuracy of 85.6%, compared to an accuracy of 100% with the biased version. These results suggest that indeed much of the bias information is still embedded in the representation after debiasing. Figure 1 shows the tSNE (Maaten and Hinton, 2008) projection of the vectors before and after debiasing using HARD-DEBIASED. We omit the visualization of GN-GLOVE for lack of space, and refer the reader to the full version of the paper.

Bias-by-projection correlates to bias-by-neighbours This clustering of gendered words indicates that while we cannot directly "observe" the bias (i.e. the word "nurse" will no longer be closer to

[1]We use the embeddings provided by Bolukbasi et al. (2016) in `https://github.com/tolga-b/debiaswe` and by Zhao et al. (2018) in `https://github.com/uclanlp/gn_glove`.

explicitly marked feminine words) the bias is still manifested by the word being close to *socially-marked* feminine words, for example "nurse" being close to "receptionist", "caregiver" and "teacher". This suggests a new mechanism for measuring bias: the percentage of male/female socially-biased words among the k nearest neighbors of the target word.[2]

We measure the correlation of this new bias measure with the original bias measure. For the HARD-DEBIASED embedding we get a Pearson correlation of 0.686 (compared to a correlation of 0.741 when checking neighbors according to the biased version). For the GN-GLOVE embedding we get a Pearson correlation of 0.736 (compared to 0.773). All these correlations are statistically significant with p-values of 0.

Professions We consider the list of professions used in Bolukbasi et al. (2016) and Zhao et al. (2018)[3] in light of the neighbours-based bias definition. Figure 2 plots the professions, with axis X being the original bias and axis Y being the number of male neighbors, before and after debiasing using HARD-DEBIASED embedding. We omit the plot of GN-GLOVE for lack of space, and refer the reader to the full version of the paper.

We observe a Pearson correlation of 0.606 (compared to a correlation of 0.747 when checking neighbors according to the biased version) for HARD-DEBIASED and 0.792 (compared to 0.820) for GN-GLOVE. All these correlations are significant with p-values $< 1 \times 10^{-30}$.

Classifying previously female- and male-biased words Can a classifier learn to generalize from some gendered words to others based only on their representations? We consider the 5,000 most biased words according to the original bias (2,500 from each gender), train an RBF-kernel SVM classifier on a random sample of 1,000 of them (500 from each gender) to predict the gender, and evaluate its generalization on the remaining 4,000. For the HARD-DEBIASED embedding, we get an accuracy of 88.88%, compared to an accuracy of 98.25% with the non-debiased version. For the GN-GLOVE embedding, we get an accuracy of 96.53%, compared to an accuracy of 98.65% with the non-debiased version.

4 Discussion and Conclusion

Our experiments reveal a systematic bias found in the embeddings, which is independent of the gender direction. We observe that semantically related words still maintain gender bias both in their similarities, and in their representation. Concretely, we find that: a) Words with strong previous gender bias (with the same direction) are easy to cluster together; b) Words that receive implicit gender from social stereotypes (e.g. receptionist, hairdresser, captain) still tend to group with other implicit-gender words of the same gender, similar as for non-debiased word embeddings; c) The implicit gender of words with prevalent previous bias is easy to predict based on their vectors alone.

The implications are alarming: while suggested debiasing methods work well at removing the gender direction, the debiasing is mostly superficial. The bias stemming from world stereotypes and learned from the corpus is ingrained much more deeply in the embeddings space.

We note that the real concern from biased representations is not the association of a concept with words such as "he", "she", "boy", "girl". Algorithmic discrimination is more likely to happen by associating one implicitly gendered term with other implicitly gendered terms, or picking up on gender-specific regularities in the corpus by learning to condition on gender-biased words, and generalizing to other gender-biased words. Our experiments show that such classifiers would have ample opportunities to pick up on such cues also after debiasing w.r.t the gender-direction.

The crux of the issue is that the gender-direction provides a way to *measure* the gender-association of a word, but *does not determine* it. Debiasing methods which directly target the gender-direction are for the most part merely hiding the gender bias and not removing it. The popular definitions used for quantifying and removing bias are insufficient, and other aspects of the bias should be taken into consideration as well.

[2]While the social bias associated with a word cannot be observed directly in the new embeddings, we can approximate it using the gender-direction in non-debiased embeddings.

[3]https://github.com/tolga-b/debiaswe/tree/master/data/professions.json

References

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. In *Advances in Neural Information Processing Systems*.

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

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of EMNLP*, pages 4847–4853.