Jasmine Zhang, Skylar Hao,
Alan Tran

Reducing Gender Bias in NLP by Modifying the Optimization Objective

University of Toronto
CSC2559, 2022

NLP models parrot biases in the data they're trained on. Our method **reduces gender bias in word2vec during training by modifying its optimization objective.**

Legend

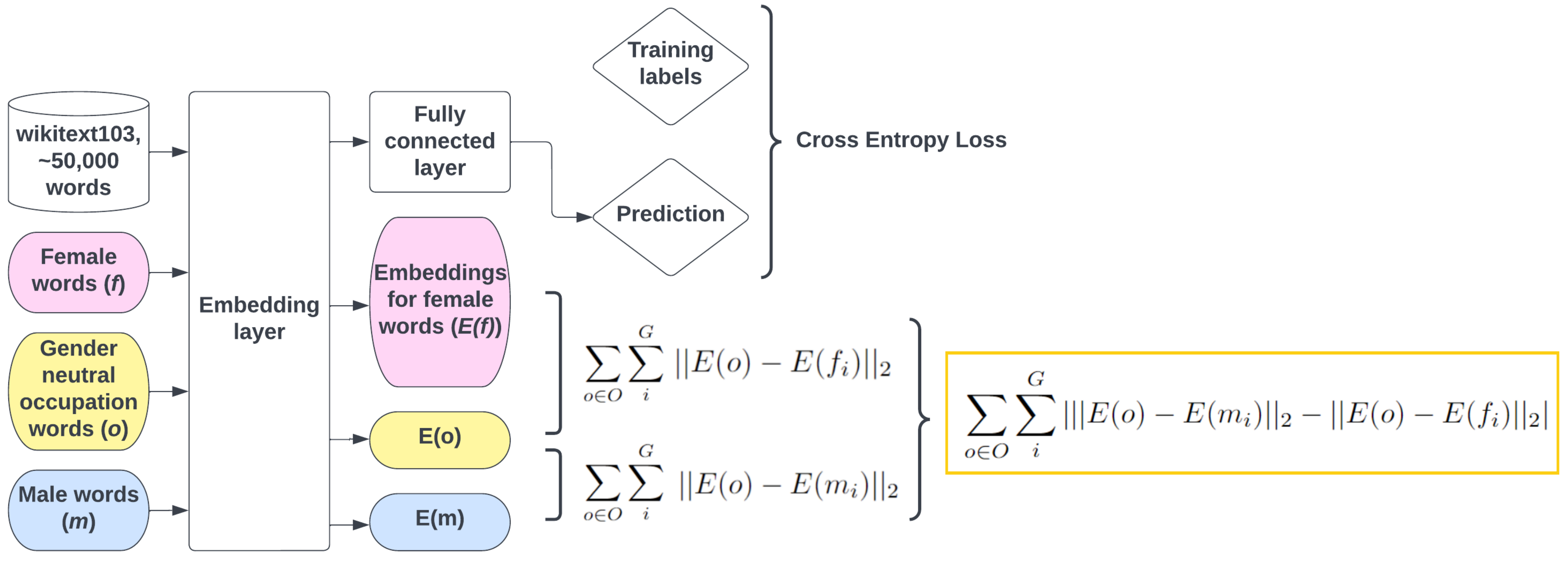
Term	Examples
Gender-neutral profession words	Doctor, engineer
Gender-pair words	He-she, man-woman
Gendered words	Father, mother

Introduction

Word embeddings in NLP represent semantically-similar words in natural human languages as real-valued vectors that occupy close spatial positions. Word embeddings can learn biases found in training data. For example, words like "engineer" may be mapped closer to male-related than female-related words [1]. This creates fairness concerns as a downstream task of NLP is resume filtering. **Our objective is to minimize gender bias** as measured by:

1. **Direct Bias:** average placement of gender-neutral profession words relative to the gender axis (where the gender information lies within the embedding). Approaches zero when there is no direct bias [1].
2. **Word Embedding Association Test (WEAT):** distance of gender-neutral profession word clusters relative to gendered word clusters. Approaches zero when the clusters are equal distance from each other [2].

Methodology



$$\sum_{o\in O}\sum_{i}^{G}||E(o)-E(f_i)||_2$$

$$\sum_{o\in O}\sum_{i}^{G}||E(o)-E(m_i)||_2$$

$$\sum_{o\in O}\sum_{i}^{G}|||E(o)-E(m_i)||_2-||E(o)-E(f_i)||_2|$$

Figure 1: We use Continuous Bag of Words **(CBOW) to create word embeddings**. The probability of each word appearing is predicted based on the distribution of words that appear close by. Intuitively, semantically-similar words are more likely to appear together. The difference in Euclidean distances between *E(o)*, *E(f)* and *E(m)* is penalized. **By adding this difference equation to the optimization objective, word embeddings are debiased during training** as each gender-neutral occupation word, *o*, is encouraged to be equidistant to each corresponding pair of male and female words, m_i and f_i .

Results

	Direct Bias	WEAT Metric
Unaltered word2vec	0.0444	0.0635
Bolukbasi (2016)	0.0014	0.0164
Zhao (2018)	0.0970	0.1058
Our Method (2022)	0.0032	0.0020

Table 1: Direct Bias and WEAT Metric values of different word embeddings. The lowest value (aka. least bias measured) is highlighted.

Table 1 compares our method against the unaltered word2vec (as a baseline) as well as:

- Bolukbasi et al.: minimizes gender projection of gender-neutral profession words along an identified gender axis [1].
- Zhao et al.: pushes male and female words further apart [3].

On WEAT, our method achieves significantly better performance than the three other word embeddings. However, on direct bias our method does not outperform Bolukbasi's, but still offers improvement over unaltered word2vec.

Figure 2 shows the original's bias and the number of male neighbors are uncorrelated after debiasing. This indicates our method reduces indirect bias as the socially-marked biased words are less clustered.

Results (Continued)

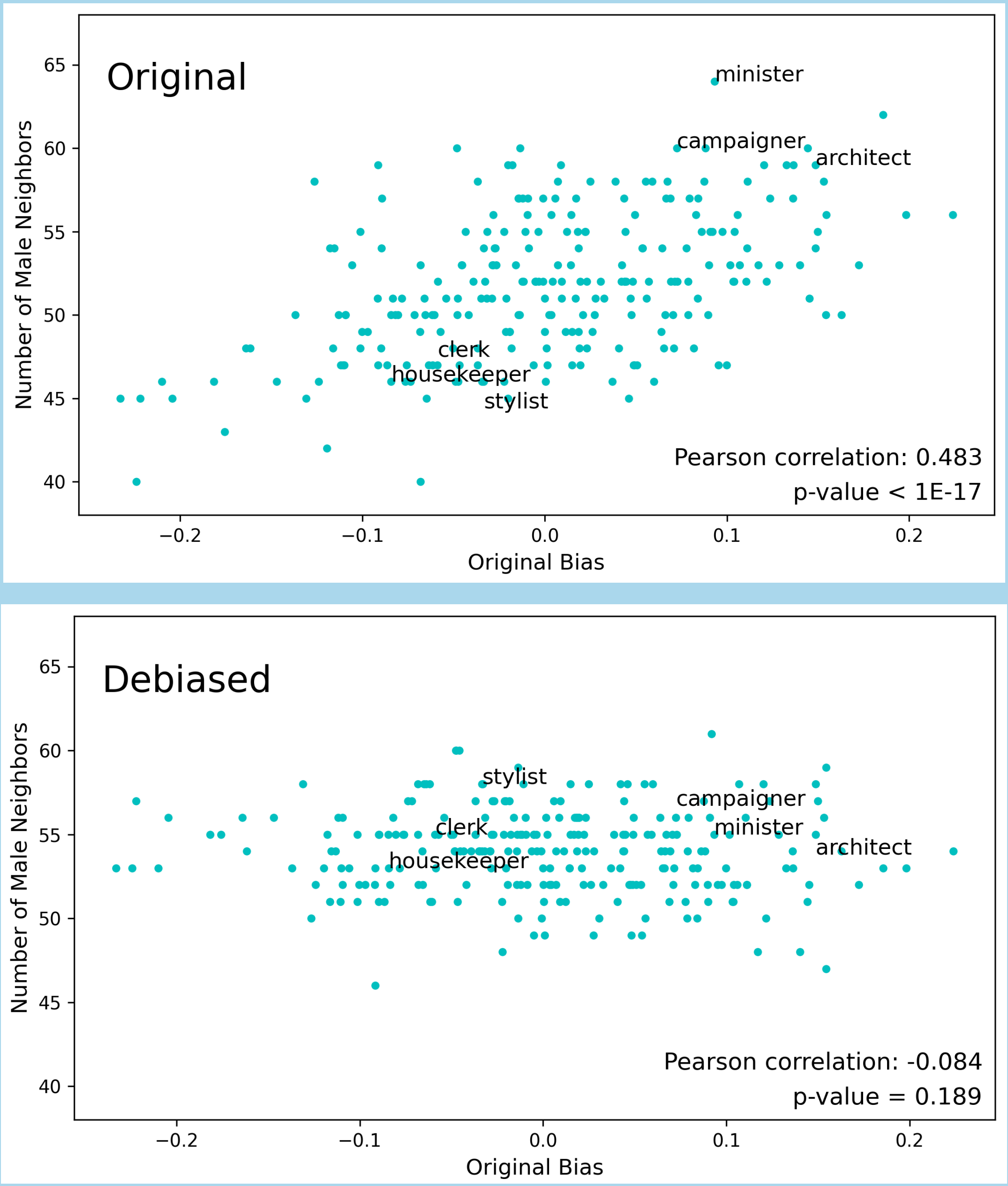


Figure 2: The number of male neighbors out of 100 nearest neighbors for each profession as a function of its original bias, before and after debiasing

Conclusion

Research on biases in NLP took off in 2016 but came to a lull in 2019 as various debiasing techniques were proposed but none were able to perform exceptionally well at both direct and indirect gender debiasing.

Our method is successful in addressing indirect gender bias (as measured by WEAT), a proven weakness of past methods [4].

Some ideas for future work include:

- Applying our generated word embeddings to a downstream task and measuring fairness
- Testing scalability of our method with different word embedding models: GLoVe, BERT, GPT
- Combining our method with other methods strong in *direct* debiasing

References

[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.
[2] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases.
[3] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings.
[4] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.