Fairness in AI: Removing Bias from Word Embeddings

Kylian van Geijtenbeek
Thom Visser
Martine Toering
Iulia Ionescu

ABSTRACT

–CURRENTLY PARTLY SPECULATION, TO BE UPDATED WHEN RESULTS BECOME AVAILABLE–

In this paper we reproduce the word embedding debiasing algorithm from Bolukbasi et al. [2]. We adapt their online codebase and extend it with their soft debiasing method, integrate several popular benchmarks and investigate the effectiveness of the algorithm on the newer fastText, BERT and XLNet embeddings, besides the Word2vec embeddings used by Bolukbasi et al. [2]. We show that the removal of direct bias from all the different embeddings barely affects their effectiveness through a comparison of benchmark scores. However, we fail to reproduce the large scale soft debiasing results due to a lack of detail on the original implementation.

ACM Reference Format:

Kylian van Geijtenbeek, Thom Visser, Martine Toering, and Iulia Ionescu. . Fairness in AI: Removing Bias from Word Embeddings. In . ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

With the recent increase in automated decision making, fairness is becoming a growing concern. Group fairness is considered to be particularly problematic. Minorities can be discriminated by being treated differently which can lead to different outcomes for identical decisions as well as different error rates for automated decisions [13]. Many of these issues can be attributed to one of the 23 types of bias identified by Mehrabi et al. [13]. Most of these are a result of biases present in society or a lack of (balanced) data. These two sources coincide in the field of natural language processing, where data is created using opinions in society. This leads to underrepresentation of minority groups and propagation of social biases through the data. Massive unfairness in models that learn from this data results.

In literature, there are three common ways to treat unfairness in text classification. First, there is anti-classification, where decision making occurs without the disclosure of protected attributes as for example gender and race. Second, classification parity ensures that the error rates are similar across groups during decision making, for example it provides similar false positive and false negative rates. Lastly, calibration is used to make sure that the outcomes are independent of protected attributes [3].

Increasing concerns over gender bias in neural natural language processing have led to the development of techniques to remove this bias from word embeddings, which are key components of many neural language models. Word embeddings are semantic vector representations of words where similar vectors denote similar words. Word embeddings like Word2vec [14] and Glove [16] are learned unsupervised by word co-occurrence from a corpus. A popular post-processing technique by Bolukbasi et al. [2] aims to find the gender axis in the embedding space, which is effectively nullified for all words that should not be gendered. This approach falls within the calibration category of dealing with unfairness, helping neural language models make decisions that are independent of gender. Taking inspiration from the approach of Bolukbasi et al. [2], Zhao et al. [22] ventured to learn new embeddings with the gender axis coaxed into a single dimension of the embedding, which is easily removed or ignored. Both approaches have since had some rebuttal from Gonen and Goldberg [7], who show that the approaches only work superficially and leave the majority of the bias in the embeddings which neural models can easily recover. As transformer based neural language models such as BERT [4] and XLNet [21] increase in popularity, simple pre-trained embeddings are being replaced by full models, making aforementioned approaches unusable. Nonetheless, Bolukbasi et al. has been a simple and transparent approach. As such, we shall replicate and extend their approach in this paper.

In this paper, we follow Bolukbasi et al. [2] and attempt to investigate whether there is gender bias present in Word2vec [14] word embeddings. Analogies created by the word embedding could possibly exhibit gender bias as well as occupations in relation to gender. We replicate the method of Bolukbasi et al. [2] by first determining a general direction in vector space that captures gender and using their gender related word lists to apply debiasing algorithms. Debiasing algorithms should remove the bias while preserving the useful properties. Additionally, we analyse fastText from Bojanowski et al. [1], BERT from Devlin et al. [4], and XLNet from Yang et al. [21] for gender bias. Even though BERT and XLNet are designed to be used as full models, which are adapted in their entirety and possibly fine-tuned, we can still extract embeddings for individual words from them to use in downstream tasks where a simpler model suffices. Because it would be beneficial to debias the extracted embeddings for these tasks, we attempt to apply the approach from Bolukbasi et al. [2] to them as well.

The analysis of Word2vec word embeddings capture bias in line with the results from Bolukbasi et al. [2] ...

2 METHOD

Following Bolukbasi et al. [2], we differentiate between direct bias and indirect bias. Direct bias is the association between gender pair and gender neutral word while indirect bias manifests in associations between gender neutral words. The goal of debiasing is reducing direct and indirect bias while maintaining relationships between gender neutral words and definitional gender words. We attempt to find whether bias is present by examining analogies created by measures of distance (*man* - *woman* = *architect* - *x* corresponds to asking the model *man* is to *woman* as *architect* is to ...?).

, ,
.

The notion of gender bias used is from Bolukbasi et al. [2] where gender stereotypes were obtained via crowd-worker evaluation. Indirect bias is examined by analysing bias in occupation words as in Bolukbasi et al. [2].

Bolukbasi et al. [2] uses two steps in their debiasing algorithms. First, the gender subspace is identified to capture direction of the bias in the embeddings. This requires multiple defining sets $D_1, D_2, ..., D_n \subset W$. Although the formulae are designed to be functional for any set size, we limit each set to be a word-pair, i.e. $|D_i| = 2$ for all i. Such a word pair describes a female and male word with the same function or relation that are per definition used exclusively to indicate each respective gender, e.g. *sister-brother*. The gender subspace is then identified by calculating the means of the defining sets (Equation 1), and then calculating SVD(C) where C is calculated as in Equation 2. The k-dimensional gender subspace B is the first k rows of this result, with $k = 1$ throughout our experiments to match with Bolukbasi et al. [2].

$$\mu_i := \sum_{w \in D_i} \frac{\vec{w}}{|D_i|} \tag{1}$$

$$C := \sum_{i=1}^{n} \sum_{w \in D_i} \frac{(\vec{w} - \mu_i)^T (\vec{w} - \mu_i)}{|D_i|} \tag{2}$$

Additionally, it will be useful to denote the projection of any vector x onto subspace B as follows:

$$x_B = \sum_{j=1}^{k} (x \cdot b_j) b_j, \tag{3}$$

where b_j denotes the j^{th} component of subspace B.

Next, there are two options for debiasing: hard debiasing and soft debiasing. Both require a set of neutral word embeddings N. This set is created by subtracting a smaller set S of gender specific word embeddings from all words embeddings W, i.e. $N = W \setminus S$.

In *hard debiasing*, the gender neutral words are shifted to zero in the gender subspace (i.e. neutralized) by subtracting the projection of the neutral word embedding vector onto the gender subspace and renormalizing the resulting embedding to unit length. Next, the embedding is equalized, entailing that gender-pairs such as *princess-prince* will be adjusted in such a way that all gender neutral words are equidistant to both the female and male word in the pair. Although the notation describes a family $\mathcal{E} = \{E_1, \ldots, E_m\}$ of equality *sets*, in this paper we refer to them as equality *pairs*, i.e. $|E_i| = 2$ for all i. For each equality pair, the mean is calculated according to Equation 4, and both the female and male word in the pair is then adjusted following Equation 5[2].

$$\mu := \sum_{w \in E} \frac{\vec{w}}{|E|} \tag{4}$$

$$\vec{w} := \mu - \mu_B + \sqrt{1 - ||\mu - \mu_B||^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||} \tag{5}$$

The second option, *soft debiasing*, also decreases the difference between the sets but has a parameter that controls the similarity to the original embeddings so that valuable distinctions are not completely removed.

We extend the method of Bolukbasi et al. [2] by adding an implementation for the soft debiasing algorithm. As the details of this approach were unclear from Bolukbasi et al. [2], we adapted specifics from Manzini et al. [11]. Soft debiasing is done by solving the following optimization problem as mentioned in their papers:

$$\min_T ||(TW)^T (TW) - W^T W||_F^2 + \lambda ||(TN)^T (TB)||_F^2 \tag{6}$$

where W is the matrix of all embedding vectors, N is the matrix of the embedding vectors of the gender neutral words, B is the gender subspace, and T is the debiasing transformation that minimizes the projection of the neutral words onto the gender subspace but tries to maintain the pairwise inner products between the words. This formula can be reduced to:

$$\min_X ||\Sigma U^T (X - I) U \Sigma||_F^2 + \lambda ||N^T X B||_F^2 \qquad s.t. X \geq 0. \tag{7}$$

where $X = T^T T$, $W = U \Sigma V^T$, U and V being the orthogonal matrices and Σ being the diagonal matrix after performing singular value decomposition on W, and λ being the tuning parameter for reducing the gender bias.

3 EXPERIMENTAL SETUP

Experiments were done using various word embeddings. Following the approach in Bolukbasi et al. [2], we used a 300-dimensional Word2vec word embedding pre-trained on the Google News corpus [14]. This includes both the full embedding covering a total of 3 million words, and a smaller version consisting only of 26,423 words. This small version was created by taking the 50,000 most frequent words from the large embedding and removing any words that were longer than 19 characters, or contained any capital letters, digits, or punctuation. To extend Bolukbasi et al. [2]'s analysis, this paper also analyses a 300-dimensional *fastText* word embedding [1] consisting of 1 million words. In order to properly compare the fastText results with the Word2vec embedding and reduce any discrepancies caused by difference in vocabulary size, a smaller fastText embedding was generated using the same method that was used to create the small Word2vec embedding. This resulted in a small fastText embedding consisting of 27,014 words. - BERT and XLNet explanations coming up here!

For each embedding in this paper, the gender subspace B is defined to be a 1-dimensional subspace, i.e. a direction. This subspace is determined for all embeddings using definitional pairs. The definitional pairs used in this paper are *she-he*, *her-his*, *woman-man*, *Mary-John*, *herself-himself*, *daughter-son*, *mother-father*, *gal-guy*, *girl-boy* and *female-male*.

Both hard debiasing and soft debiasing require a set N of embeddings for neutral words. Additionally, hard debiasing requires a family of equality sets $\mathcal{E}$. We adopt the gender specific set S and equality sets $\mathcal{E}$ from Bolukbasi et al. [2].

To evaluate whether the embeddings retain some important functionality, we evaluate them on the same standard benchmarks used by Bolukbasi et al. [2]. These benchmarks consist of the similarity benchmarks RG-65 [18] and WS-353 [5], and the analogy

	RG-65	WS-353	MSR-analogy
Before	77.7	68.8	46.8
Hard-debiased	77.5	68.5	50.0
Soft-debiased	77.7	68.8	46.8

Table 1: Performance of the small set of Word2vec embeddings on benchmarks RG-65 [18] and WS-353 [5] for similarity and MSR [15] for analogies before and after debiasing.

	RG-65	WS-353	MSR-analogy
Before	83.9	74.1	55.9
Hard-debiased	83.5	74.2	56.0
Soft-debiased	84.4	73.9	55.4

Table 2: Performance of the small set of fastText embeddings on benchmarks RG-65 [18] and WS-353 [5] for similarity and MSR [15] for analogies before and after debiasing.

extreme female occupations		
homemaker	nanny	dancer
nurse	receptionist	therapist
paralegal	housekeeper	paediatrician
socialite	lifeguard	valedictorian
extreme male occupations		
cartoonist	commander	soldier
pundit	electrician	mechanic
maestro	farmer	
promoter	architect	

Table 3: FastText

extreme female occupations

extreme male occupations

Table 4: FastText after debiasing

benchmark MSR [15]. We have integrated these benchmarks in our implementation to easily verify both the quality of the embeddings before debiasing and their integrity afterwards.

In our approach to soft debiasing the word embeddings we use $\lambda = 0.2$ as parameter, as was used in the original paper by Bolukbasi et al. [2].

To solve the optimization problem we use the Adam optimization algorithm [8] with a learning rate of 0.01. For Word2vec, we run the soft debiasing for 2000 steps. The learning rate of each parameter group is decayed by 0.1 at steps 1000, 1500 and 1800. For fastText we run the algorithm for 6000 steps without learning rate decay.

4 RESULTS

Table 3 lists several examples of indirect bias in the form of gender-neutral words for professions that are close in space to either male or female for fastText embedding. These examples are similar when compared to results from Bolukbasi et al. [2] which suggests that fastText exhibit gender stereotypes. The word embeddings were evaluated on several standard benchmarks that measures the similarity between words and word embeddings and tests the embeddings on analogy tasks. Table 1 shows the results for the similarity benchmarks RG-65 and WS-353 and the analogy benchmark MSR of the original Word2vec embeddings. The results show that the performance is preserved after hard debiasing and soft debiasing. We also tested the performance after debiasing for fastText embeddings and found that there was also no significant performance loss for the small dataset as seen in table 2). Similar results were obtained in hard-debiasing the full dataset for Word2vec and fastText which can be found in table 5 and table 6 in Appendix A.

5 DISCUSSION

When comparing our results with Bolukbasi et al. [2] we see that performance on RG-65 and WS-353 benchmarks for the small Word2vec set has increased. It remains however unclear which benchmarks of multiple variants of RG and WS were originally used which makes exact reproduction difficult. Nevertheless, one claim is that the performance does not degrade after debiasing. This seems relatively accurate for the results shown here as well. When comparing the full dataset, our results match up exactly with the paper, except for missing results for soft-debiasing the full dataset. We believe that some information about the implementation of the soft debiasing algorithm is missing as we encounter memory issues when implementing the algorithm as detailed in Bolukbasi et al. [2]. Batching the process is not very straightforward as batching the SVD is hard without using all word embeddings from the set. Therefore, we question how they implemented this whilst bypassing the memory issues.

On top of repeating the debiasing of the Word2vec embeddings, we also debiased fastText embeddings to test the generalizability of the debiasing algorithms. We found that also with fastText embeddings, the performance does not drastically decrease after debiasing.

6 BROADER IMPLICATIONS

Fairness in automated decision-making can have broader implications for society over time. Liu et al. [9] argue that it is necessary to first analyse the long-term impact of “fair” machine learning as we cannot predict the outcomes for enforced fairness criteria. They show that, in general, common fairness criteria do not improve justice over time but can even cause harm [9]. The approach of Bolukbasi et al. [2] would be no exception to this. These embeddings appear to be more fair in the short term, but long term impact of the models using these embeddings could be unaffected. Louizos et al. [10] propose a different strategy to tackle fairness in machine learning. They introduce their Variational Fair Autoencoder (VFAE) and propose it to be used to develop fair classifiers that are invariant to sensitive demographic information, as they argue that it produces a better tradeoff between accuracy and invariance of the classifier [10].

- Accountability connection to be introduced here after corresponding lecture!

As noted by McMahan et al. [12], confidentiality and privacy are an ever increasing concern. Especially with recent demonstrations, like those of Shokri et al. [19], showing that the original data on which the model was trained can be extracted from solely the model's parameters. The word embeddings used by many neural language models could also contain, at least parts of, the memorised training data, though this is difficult to ascertain without thorough investigation. The approach of Bolukbasi et al. [2], though not designed for it, could reduce the amount of directly memorised information in the embeddings by debiasing them. The debiasing process should remove the desired sensitive feature, like gender, from the embeddings, thus applying noise and introducing uncertainty in any reconstructed training data involving the bias. The effectiveness of this might be limited as nearly all of the functionality of the embedding remains intact after debiasing, together with most of the bias [7].

Another growing area of research is the transparency of 'black box' models like neural networks. Gaining a better understanding of the reasoning behind the predictions of these models is important in building the trust that the end-users need to confidently use the model, as well as providing more context to the predictions. Popular methods for this are LIME [17] and Integrated Gradients [20], both of which provide external tools to analyse a model's predictions. These methods, like transparency methods in general, apply to the full model, and are thus not specifically tied to the embeddings used in language models. Methods like Integrated Gradients, which use gradient information from the network, do rely on the gradient from the embeddings, but still only as part of the whole network. The approach of Bolukbasi et al. [2] is, as a whole, not useful in providing transparency, as this is not its goal. The first step of the debiasing process, however, could be used to gain some insight in different biases in a model's embeddings. Though these insights do not simply apply to the model's predictions, they can be used to build, or break, trust in the model.

7 CONCLUSION

This paper aimed to reproduce the word embedding debiasing algorithms from Bolukbasi et al. [2]. We extended their online codebase with their soft debiasing method, integrated several popular benchmarks and investigated the effectiveness of the algorithm on the newer fastText, BERT and XLNet embeddings, besides the Word2vec embeddings originally used by them. We showed that the removal of direct bias from all the different embeddings barely affects their effectiveness through a comparison of benchmark scores. However, we failed to reproduce the large scale soft debiasing results due to a lack of detail in the original implementation.

We award the paper with two ACM badges: the Artifacts Available badge, as most of their code is easily accessible in their online repository reproducing most of their results, and the Results Replicated badge, as we were largely able to reproduce their results based on the data they provided to us [6].

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[2] 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. *arXiv preprint arXiv:1607.06520* (2016).

[3] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web*. ACM, 406–414.

[6] Association for Computing Machinery. 2018. *Artifact Review and Badging*. https://www.acm.org/publications/policies/artifact-review-badging

[7] 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. *arXiv preprint arXiv:1903.03862* (2019).

[8] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2015).

[9] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *ICML* (2018).

[10] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The variational fair autoencoder. *ICLR* (2016).

[11] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *arXiv preprint arXiv:1904.04047* (2019).

[12] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. *ICLR* (2018).

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546* (2013).

[15] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACM, 746–751.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing*. 1532–1543.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD* (2016).

[18] Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Commun. ACM* 8, 10 (1965), 627–633.

[19] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. *ICMR* (2017).

[21] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237* (2019).

[22] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *arXiv preprint arXiv:1809.01496* (2018).

A FULL WORD2VEC AND FASTTEXT EMBEDDINGS

	RG-65	WS-535	MSR-analogy
Before	76.1	70.0	47.2
Hard-debiased	76.5	69.7	47.4

Table 5: Performance of the full set of Word2vec embeddings on the benchmarks before and after debiasing.

	RG	WS	MSR analogy
Before	84.6	73.3	59.8
Hard-debiased	84.5	73.3	59.9

Table 6: Performance of the full set of fastText embeddings on the benchmarks before and after debiasing.