Fairness in AI: Removing Bias from Word Embeddings

Kylian van Geijtenbeek Thom Visser Martine Toering

Iulia Ionescu Morris Frank
Supervisor

ABSTRACT

This report attempts to reproduce the word embedding debiasing algorithm and replicate experiments from Bolukbasi et al. [3]. We adapt the publicly available implementation [2] and extend it with the soft debiasing method described in their paper. Several popular benchmarks are integrated in order to evaluate the word embeddings before and after debiasing. Besides replicating results on Word2vec [16], the effectiveness of the debiasing algorithms is investigated on GloVe [18] and fastText [1] embeddings. We show that the removal of direct bias for all the embeddings barely affects their expressiveness through a comparison of benchmark scores. However, we fail to reproduce large scale soft debiasing results as the method described by the authors faces serious computational issues.

1 INTRODUCTION

With the recent increase in automated decision making, fairness is becoming a growing concern. 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 [15]. Many of these issues can be attributed to one of the 23 types of bias identified by Mehrabi et al. [15]. 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. Key components of many neural language models are word embeddings, which are learned from word co-occurrence in a corpus. Word embeddings are semantic vector representations of words where similar vectors denote similar words.

Increasing concerns over gender bias in neural natural language processing have led to the development of techniques to remove this bias from word embeddings. This report is focused on a popular post-processing technique by Bolukbasi et al. [3] that aims to find the gender axis in the embedding space, which is effectively nullified for all words that should not be gendered. This approach increases the ability of neural language models to attain results that are independent of gender. Taking inspiration from the approach of Bolukbasi et al. [3], Zhao et al. [24] 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 [8], 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 the popularity of transformer based neural language models such as BERT [5] and XLNet [23] increases, simple pre-trained embeddings are being replaced by full transformer models. This renders the approaches listed above outdated, yet still popular. The author's method has however been a simple and transparent approach. For this reason we shall replicate and extend their approach in this report.

We aim to reproduce the findings of Bolukbasi et al. [3]. We do this by investigating whether there is gender bias present in word embeddings as well as finding whether their debiasing method effectively removes this bias. Analogies created by word embeddings could possibly exhibit gender bias and profession words could be biased in relation to gender. Debiasing algorithms should remove the bias while preserving the useful properties of the embedding. The aim is to validate results on Word2vec [16] embedding. Additionally, we chose to analyse GloVe [18] and fastText [1] to find whether the method can be generalized to other embeddings. The authors found that GloVe embedding exhibits gender stereotypes for occupations to around the same extent as Word2vec (as seen in Figure 4 from their paper). We show that fastText conveys similar gender bias for professions and both GloVe and fastText reveal gender bias in their analogies. The code used in this study is available on GitHub [1].

To validate the reported results, the main questions we will try to answer are:

- Is gender bias present in word embeddings?
- Could this bias be removed with the debiasing methods?
- Is the quality of the embeddings retained?
- Could the method be generalized to other embeddings?

2 METHOD

Following Bolukbasi et al. [3], we differentiate between direct bias and indirect bias. Direct bias is the association between a gender specific word and a gender neutral word, such as *she* and *receptionist* being close in the embedding space. Indirect bias manifests in associations between gender neutral words, such as *receptionist* being close to *nanny*. The goal of debiasing is reducing direct and indirect bias while maintaining relationships between gender neutral words not related to gender (for instance *football* and *quarterback*) and relationships between gender words (for instance *mom* and *woman*). We attempt to find direct bias by examining analogy puzzles solved by the model, for example *man* - *woman* $\approx$ *architect* - *x* corresponds to solving the analogy question: "*man* is to *woman* as *architect* is to ...". More direct bias is examined by projecting occupation words onto the gender subspace and relating it to the gender stereotypes established by the authors.

Bolukbasi et al. [3] uses two steps in their debiasing algorithms. First, the gender subspace is identified to capture direction of the bias in the embedding space. 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.

[1] https://github.com/KylianvG/Embetter

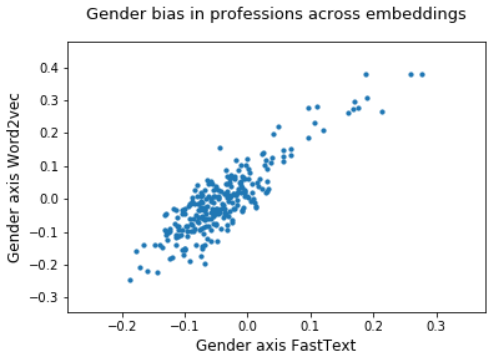

FIGURE 1: Similar bias in profession words between word embeddings Word2vec and fastText. Datapoints represent occupation words.

$|D_i| = 2$ 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*. By calculating the means of the defining sets (Equation 1) and subtracting this mean from each word embedding in that defining set, one expects to be left with mainly the gender component of the word. By summing correlation matrices of these gender components as one matrix $\boldsymbol{C}$ (Equation 2) and calculating SVD ($\boldsymbol{C}$), we can take the first k right-singular vectors to obtain a k-dimensional gender subspace B.

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

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

The projection of any vector x onto subspace B is defined 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.

The paper describes 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 equality-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 report we refer to them as equality *pairs*, i.e. $|E_i| = 2$ for all i. An equality pair is a pair of words that have essentially the same definition, except one is female and the other is male. 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[3].

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

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

The second option, *soft debiasing*, also decreases the difference between the sets but has a parameter that controls the importance of gender removal while transforming the original embeddings vectors as little as possible so that valuable distinctions are not completely removed.

We extend the implementation of the authors by adding an algorithm for soft debiasing. As the implementation details of this approach were missing from the paper, we adapted specifics from Manzini et al. [13][12], who did provide code for the soft debiasing algorithm. Their implementation relies on the same mathematical foundation, but they show that the optimization problem can be solved using stochastic gradient descent. Soft debiasing is done by solving the following optimization problem as mentioned in both 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, λ being the tuning parameter for reducing the gender bias, 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. As shown by [3], Equation 6 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 matrices containing the left-singular and right-singular vectors respectively, and Σ being the diagonal matrix after performing singular value decomposition on W.

3 EXPERIMENTAL SETUP

Experiments were done using various word embeddings. Following the authors we used the 300-dimensional Word2vec word embedding pre-trained on the Google News corpus [16]. 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 the authors by taking the 50,000 most frequent words from the large embedding. Words longer than 19 characters, or containing any capital letters, digits, or punctuation were removed. To extend the author's analysis, this report uses the pre-trained GloVe [18] embedding Common Crawl version consisting of 1.9 million words and fastText word embedding [1] trained on both Common Crawl and Wikipedia which consists of 1 million words. In order to properly compare GloVe and fastText results

with the Word2vec embedding and reduce discrepancies caused by differences in vocabulary size or vocabulary construction, smaller embeddings were generated with the aforementioned method used to create the small Word2vec embedding. This resulted in a small GloVe embedding of 42,982 words and small fastText embedding consisting of 27,014 words.

For each embedding, the gender subspace B is defined to be a 1-dimensional subspace, i.e. a direction. This subspace is determined for all embeddings using a list definitional pairs that we adapt from the authors. The definitional pairs 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 the authors.

To evaluate whether the embeddings retain some important functionality, we evaluate them on the same standard benchmarks used the authors. These benchmarks consist of the similarity benchmarks RG-65 [20] and WS-353 [6], and the analogy benchmark MSR [17]. We have integrated these benchmarks in our implementation to easily verify both the quality of the embeddings before debiasing and their integrity afterwards.

On top of these benchmarks, we implemented an objective way to evaluate the bias in the word embeddings by adopting Word Embedding Association Test (WEAT), introduced by Caliskan et al. [4]. WEAT measures the similarity between target words (male and female professions in our case) and attribute sets (female and male words, i.e. the definitional pairs) by calculating the distance between pairs of vectors and returns a total effect size. An effect size close to 2 (or -2) means there is a relatively high chance that the word embeddings contain direct bias, controlled by the p-value. A value close to zero means that if there is a bias, the bias is small.

In our approach to soft debiasing the word embeddings we use $\lambda = 0.2$ as the hyperparameter to balance between removing bias and conserving original relations, as was used in the original paper. To solve the optimization problem we use the Adam optimization algorithm [10] with a learning rate of 0.01, as in practice it proved to be a more stable and faster minimization algorithm than the basic Stochastic Gradient Descent (SGD) algorithm used by Manzini et al. [13]. For Word2vec, we run the soft debiasing for 2000 steps. The learning rate is reduced by a factor of 0.1 at steps 1000, 1500 and 1800. For GloVe we run the algorithm for 3500 steps. The learning rate is 0.01, with a reduction factor of 0.1 at 2300, 2800, and 3000 steps. For fastText we run the algorithm for 7000 steps with learning rate of 0.01 and a reduction factor of 0.1 at 5000 steps.

4 RESULTS

Table 2 lists several examples of bias in profession words that are close in space to either male or female for GloVe and fastText embedding. These examples are quite similar to reproduced results for Word2vec, which suggests that GloVe and fastText exhibit similar gender stereotypes to Word2vec. The profession words projected onto the gender subspace for both Word2vec and fastText in Figure 1 also shows results in line with Figure 4 from the author's paper. Table 1 shows the results for the similarity benchmarks RG-65 and WS-353, the analogy benchmark MSR and the bias benchmark WEAT for all embeddings. The results show that the performance is preserved after hard debiasing and soft debiasing as the scores are consistent and there is no significant performance loss after debiasing. The WEAT evaluation method resulted in relatively high effect sizes for all word embeddings before debiasing when evaluated on professions. After both hard and soft debiasing, gender bias in the word embeddings based on the professions could still exist as the WEAT remains at a value larger than zero. However, the effect size indicates that the gender bias for professions may be smaller after debiasing. For the corresponding p-values we refer to Karve et al. [9] where Word2vec, fastText, and GloVe embeddings were also used. The results show p-values close to zero, indicating significant effect sizes. This is expected since the calculation of the WEAT is relatively straightforward as described in the Method section. Similar results were obtained in hard debiasing the full dataset for Word2vec, GloVe and fastText which can be found in Table 3 in Appendix A.

5 DISCUSSION

We found that GloVe and fastText embeddings trained on different data exhibit very similar bias compared to Word2vec. After debiasing, they do not drastically decrease in performance after debiasing, as was projected in the original paper. The performance on the benchmarks is comparable to that from the original paper. In the paper it remains unclear which steps were originally used to prepare the embeddings for the benchmarks, which makes exact reproduction difficult. The benchmark results for the full dataset match up exactly with the paper, except for missing results for soft debiasing the full dataset.

We faced serious memory issues when implementing the soft debiasing algorithm as detailed by the authors. Without a more efficient implementation, soft debiasing as described in their paper is not applicable to embeddings with a larger vocabulary. Batching this process is not straightforward and generally still requires all embeddings for computation.

The addition of the WEAT effect size benchmark has provided a quantitative measure for the bias exhibited by the embeddings. This has allowed us to show an objective decrease in direct bias through the use of this method. This confirms the claim from the original paper that direct bias is indeed removed.

Several benchmarks have been used to determine that bias is removed without a decrease in performance. These benchmarks only determine the performance on analogy and similarity tasks. True preservation of the embedding performance in practice could be tested on downstream tasks such as POS-tagging, sentiment analysis or text generation.

6 BROADER IMPLICATIONS

Fairness in automated decision-making can have broader implications for society over time. Liu et al. [11] argue that it is necessary to first analyze the long-term impact of "fair" machine learning as we cannot predict the outcomes for enforced fairness criteria. They show that common fairness criteria in general do not improve justice over time but can even cause harm [11]. The approach of Bolukbasi et al. [3] would be no exception to this. These embeddings appear to be more fair in the short term. Long term impact

	Word2vec				GloVe				fastText			
	RG-65	WS-353	MSR	WEAT	RG-65	WS-353	MSR	WEAT	RG-65	WS-353	MSR	WEAT
Before	77.7	68.8	46.8	1.5	83.1	66.4	37.5	1.7	83.9	74.1	55.9	1.5
Hard-debiased	77.5	68.5	47.0	0.4	83.4	66.6	37.6	0.5	83.5	74.2	56.0	0.5
Soft-debiased	77.7	68.8	46.8	-0.1	83.1	66.4	37.4	0.8	84.3	74.1	54.7	0.4

TABLE 1: Performance of the small sets of Word2vec, GloVe and fastText embeddings on benchmarks RG-65 [20] and WS-353 [6] for similarity, MSR [17] for analogies and WEAT [4] as an objective measure before and after debiasing.

fastText		GloVe	
Professions closest to *she*		Professions closest to *she*	
nurse	librarian	socialite	librarian
socialite	dancer	nurse	receptionist
housekeeper	singer	homemaker	dancer
receptionist	vocalist	stylist	housekeeper
Professions closest to *he*		Professions closest to *he*	
inventor	commander	captain	inventor
pundit	electrician	drummer	architect
carpenter	footballer	colonel	guitarist
headmaster	architect	commander	luitenant
Analogies woman : man		Analogies woman : man	
supermodel:footballer		stunning:impressive	
beautiful:brilliant		hottest:greatest	

TABLE 2: Examples of bias present in analogies created from the word embeddings GloVe and fastText as well as gender bias in relation to occupations.

of the models using these embeddings could be unaffected if, for example, the model abuses the indirect bias.

In the area of accountability, using this approach as part of a larger project would indicate that the developer has considered existing bias and made an attempt to remove this.

As noted by McMahan et al. [14], confidentiality and privacy are an ever increasing concern. Recent demonstrations, like those of Shokri et al. [21], show 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 memorized training data. This is difficult to ascertain without thorough investigation though. The approach of Bolukbasi et al. [3], while not designed for it, could reduce the amount of directly memorized information in the embeddings by debiasing them. The debiasing process should remove desired sensitive features like gender from the embeddings. This could create noise and introduce uncertainty in any reconstructed training data involving the bias. The effectiveness might be limited as nearly all of the functionality of the embedding remains intact after debiasing, together with most of the bias [8].

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 [19] and Integrated Gradients [22], both of which provide external tools to analyze a model's predictions. These methods, such as transparency methods in general, apply to the full model. They thus are 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. [3] is not useful in providing transparency as this is not its goal. The first step of the debiasing process could however 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 report aimed to repeat the experiments from Bolukbasi et al. [3]. We extended their implementation with their soft debiasing method, repeated their experiments using several popular integrated benchmarks, quantitatively analyzed the bias with the WEAT, and investigated the effectiveness of the algorithm on GloVe and the newer fastText embedding.

The conclusions can be summarized as follows:

- Word embeddings exhibit gender bias as seen in the analogies created by the embeddings and the objective WEAT score.
- It is possible to remove the bias with the proposed debiasing methods and validate the reported results to a large extent. Hard debiasing and the benchmark results are validated. Soft debiasing remains not reproducible as the algorithm does not scale to larger embeddings or a larger vocabulary size.
- We showed that the removal of direct bias from all the different embeddings barely affects their quality through a comparison of benchmark scores.
- Similar results were obtained for GloVe and fastText, suggesting the method generalizes well to other word embeddings.

We award the paper with two ACM badges [7]: the *Artifacts Available* badge, as most of their code is easily accessible in their available 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.

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. *Debiaswe: try to make word embeddings less sexist.* https://github.com/tolga-b/debiaswe

[3] 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).

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

[5] 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).

[6] 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.

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

[8] 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).

[9] Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor Debiasing of Word Representations Evaluated on WEAT. *arXiv preprint arXiv:1906.05993* (2019).

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

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

[12] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. *Debiasing Multiclass Word Embeddings*. https://github.com/TManzini/DebiasMulticlassWordEmbedding

[13] 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).

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

[15] 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).

[16] 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).

[17] 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.

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

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

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

[21] 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.

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

[23] 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).

[24] 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 BENCHMARK RESULTS FULL EMBEDDINGS

See Table 3 for the results on the large embedding datasets. As mentioned, the soft debiasing method is not applicable to the large embedding datasets. The proposed implementation using SVD would require storage of a $N \times N$ *float-32* matrix, where N is the vocabulary size. In the case of Word2Vec with its 3 million words, this would require 36 Terabytes of memory.

B CONTRIBUTION

Equal contribution. Contributions written in no particular order. All four of us co-wrote the report. Martine spent extra time editing and rewriting. Kylian and Iulia have implemented the soft debiasing method. Thom gathered the benchmarks and implemented MSR. Martine implemented RG-65 and WS-353. Thom and Iulia implemented WEAT. Thom incorporated all benchmarks in a single object. All four have executed experiments to obtain results and test functionality. Thom ensured documentation and PEP8 style for all the code. Thom and Kylian wrote the README.md. Kylian has made sure that all available embeddings are easily downloadable from the Google drive, and embedding loading proceeds correctly. Martine sorted the data in tables and made the figure. Thom and Martine prepared the tutorial notebook. Iulia and Kylian prepared the presentation slides.

	Word2vec				GloVe				fastText			
	RG-65	WS-353	MSR	WEAT	RG-65	WS-353	MSR	WEAT	RG-65	WS-353	MSR	WEAT
Before	76.1	70.0	47.2	1.5	81.7	64.6	41.6	1.6	84.6	73.3	59.8	1.5
Hard-debiased	76.5	69.7	47.4	0.4	82.0	64.7	41.7	0.5	84.5	73.3	59.9	0.4

Table 3: Performance of the full sets of Word2vec, GloVe and fastText embeddings on benchmarks RG-65 [20] and WS-353 [6] for similarity, MSR [17] for analogies and WEAT [4] as an objective measure before and after debiasing.