Debiasing word embeddings: replication and extension

Ivan Yovchev
University of Amsterdam
12871737
yovchev23@gmail.com

Davide Barbieri
University of Amsterdam
12871745
davidebarbieri97@gmail.com

Andras Csirik
University of Amsterdam
12521620
csirika@gmail.com

Balint Hompot
University of Amsterdam
12746452
hompot.balint@gmail.com

ABSTRACT

The word embeddings used today contain unfair biases based on gender [Zhao et al. 2019]. Some works propose algorithms which can remove said bias and present persuasive evidence based on generated analogies to support their claim [Bolukbasi et al. 2016]. Other papers argue that analogies are not a good way to measure bias and say these algorithms do not remove the bias, only hide it [Gonen and Goldberg 2019]. In our work we investigate this topic by replicating the results of a debiasing algorithm presented by Bolukbasi et al. [2016]. We conclude that while their results are replicable, the algorithm cannot be used for the intended purpose.

1 INTRODUCTION

The meaning of words are often modeled with word embeddings, which are high dimensional vector representations. Models using these embeddings perform great on a variety of problems such as sentiment analysis and analogy tasks which shows how well they capture the true meaning of words. However, as they are learned from real-world texts, they often not only encompass real-world knowledge, but also real-world sexism, racism and other biases [Bolukbasi et al. 2016], which opposes great risk in terms of fairness, as models based on e.g. sexist input representation will likely produce sexist outputs.

To tackle the problem the authors use the popular word2vec embedding [Mikolov et al. 2013a], [Mikolov et al. 2013b] trained on Google News texts. They define a gender-subspace and a gender-direction and use them to give a quantitative definition of bias. They propose an algorithm which can remove this bias while preserving the performance of the embedding for standard NLP tasks. To support their claim they test the original and debiased embeddings on analogy tasks. They also ran benchmark tests on the debiased embedding to show that the word representations still capture the meaning. They report promising results on both goals.

However, another paper [Gonen and Goldberg 2019] argues that the bias is not removed, simply hidden. The authors say the definition used by Bolukbasi et al. [2016] does not capture the essence of the bias and perform a series of experiments on the debiased word embeddings proving that the bias is still present. They conclude that "the gender-direction provides a way to *measure* the gender-association of a word, but *does not determine* it".

In our work we investigate the issue of debiasing word embeddings proposed in the previously mentioned papers [Bolukbasi et al. 2016] and [Gonen and Goldberg 2019]. Particularly we ask the following questions:

(1) Can the results reported in [Bolukbasi et al. 2016] be reproduced?
(2) Can the algorithm be applied to another embedding successfully?
(3) Is there still bias present in the debiased embeddings? If so can we identify it?

2 METHOD

In [Bolukbasi et al. 2016] the authors discuss two different types of bias: direct bias measures how supposedly neutral words coincide with gender defining words (as "he"), while indirect bias shows how neutral words coincide with words that implicitly contain gender informaion (such as "softball"). They introduce two methods to remove gender bias from word embeddings. Hard debiasing aims to completely remove any sort of bias from the embeddings, while soft debiasing aims to learn a transform that maintains pairwise comparisons of embeddings while minimizing the gender bias content. In this paper we only focus on the hard debiasing as the paper reports better results with this, and related work [Gonen and Goldberg 2019] also uses this approach. Hard debiasing consists of three steps: first we define the gender subspace, then neutralize the supposedly gender neutral words with respect to the this subspace, finally we equalize inherently gender specific word pairs (such as grandmother-grandfather) such that they are equidistant from the neutral words. Defining the gender subspace relies on defining a set of gender defining pairs D. We calculate the covariance matrix of these word pairs $\mathbf{C}$ (note that $\overrightarrow{w}$ corresponds to word embeddings). In order to have that the origin is gender neutral (for the neutral subspace), we perform this principal component analysis (PCA) on the difference between each embedding and the mean of the pair the embedding belongs to. We well call these d. Formally we thus compute:

$$\mu = \mathbf{E}\left[d\right]$$
$$C_{ij} = \mathbf{E}\left[(d_i - \mu_i)(d_j - \mu_j)\right]$$

Then we select the firs k components of the principal component analysis, which will be our gender subspace. As in the original paper, we use $k = 1$, so we can simply call it the gender axis.
For neutralization, we first need a list of gender neutral words N. Their embeddings is projected on the gender subspace, then translated and normalized such that their projection on the subspace

is 0, meaning they contain no gender information:

$$w := \frac{w - w_B}{|| w - w_B ||}$$

where w_B is w's projection on the gender space (or axis). Finally, we need a set of gender specific word pairs and equalize them with respect to the neutral terms (that are moved to the origin in the subspace). We do this by projecting them on the subspace, moving such that they are equidistant from the subspace origin. They are then projected back in the embeddings space. Formally, we equalize two word embeddings w_1 and w_2 by doing:

$$\mu = \frac{1}{2}(w_1 + w_2)$$

$$v = \mu - \mu_B$$

$$w_i := v + \sqrt{1 - || v ||^2} \frac{w_{iB} - \mu_B}{|| w_{iB} - \mu_B ||} \qquad \text{for } i \in \{1, 2\}$$

The debiasing is then assessed with two goals in mind: maintaining semantic information by proving performance on benchmark NLP tasks, and containing less gender bias by creating analogies that would reveal direct or indirect bias.
To add quentitative analysis, [Gonen and Goldberg 2019] conducted several experiments to find any remaining bias. In one of their experiments they tried to classify gender biased words after debiasing according to their original bias and found that it is still possible. We extended this experiment to get a better picture on the remaining bias.

3 EXPERIMENTAL SETUP

In our work[1] we first replicated Bolukbasi's experiments. We modeified their implementation as they made it publicly available.[2] Then we ran their algorithm on the Glove embedding [Pennington et al. 2014] and conducted the same experiments. We also checked whether the debiased embeddings perform the same on some benchmark tests. Following the [Gonen and Goldberg 2019] paper we also conducted the classification experiment and extended it using more classifiers and iterative data set extension. As we verified that the bias is still present after debiasing, we also projected the words on a larger subspace to remove the more subtle bias components. The such generated embeddings were also subjected to benchmark and quantitative analysis the same way.

3.1 Replicating debiasing

In order to fully replicate the study before the debiasing algorithm is applied the embeddings were filtered. As specified in the original paper [Bolukbasi et al. 2016] first a subset of the 50000 most frequent words was taken. Said subset was then filtered to only include words which have a length less than 20 characters and are all lower case. Words containing any upper case letters, number of any punctuation were removed. Performing said filtering on the word2vec embeddings resulted in 26,391 words. Performing filtering on the glove embeddings resulted in 23,177 words. Once the subsets had been filtered they were stored to the file system in order to compare results with the debiased embeddings.

As previously mentioned the debiasing algorithm takes as input three lists – definitional pairs, gender neutral words and equalizing pairs. In the original paper definitional and equalizing pairs were hand made lists and gender neutral words were learned with bootstrapping. We decided to directly use the lists they provided for replication and extension to the glove embedding (with cleaning instances that are not present in both).

Given the filtered word list we applied the debiasing algorithm. We stored the resulting debised embeddings to the file system for later comparison to their gender biased counterparts.

3.2 Testing Embedding Properties

The first part of evaluation is testing whether the debiased embeddings still retain a good semantic representation of words. To do so we ran a set of benchmark tests based on [Jastrzkebski et al. 2017] [3], which measure the usefulness of embeddings in four categories: word similarity, analogy, sentence properties and word properties. The embeddings were tested before and after debiasing to see the difference.

3.3 Generating he-she analogies

The second part of evaluation is checking whether the bias indeed disappeared from the embeddings. Following the original paper, we generated word pairs (z, w) such that the analogy *he* to z is as *she* to w (in notation: he : z = she : w) holds. These we will refer to as he-she analogies. First we can see that solving the aforementioned analogy is equivalent to solving he : she = z : w. This is useful because, in the embeddings' space, we solve these anaolgies by assuming that if they hold then $\overrightarrow{he} - \overrightarrow{she} \approx \overrightarrow{z} - \overrightarrow{w}$. Therefore $\overrightarrow{he} - \overrightarrow{she}$ is constant and given.

Since we could not replicate the crowd experiments of [Bolukbasi et al. 2016] we decided to take a different approach to reproduce the he-she analogies they reported. We collected all the words from their analogies corresponding to "he" then used each of them as z to solve the analogy task $he : she :: z : w$ for w. For this we computed $\overrightarrow{z} - \overrightarrow{w}$ for every w in the vocabulary. For every vector $\overrightarrow{z} - \overrightarrow{w}$ we computed the cosine distance between its direction (obtained by normilazig the vector) and that of $\overrightarrow{he} - \overrightarrow{she}$. We then select $\overrightarrow{w} = \underset{\overrightarrow{w'}}{\text{argmax}}\, d(\overrightarrow{he} - \overrightarrow{she}, \overrightarrow{z} - \overrightarrow{w'})$ as the solution for the analogy task, where $d(\overrightarrow{x}, \overrightarrow{y})$ is the cosine distance between the directions of $\overrightarrow{x}$ and $\overrightarrow{y}$ respectively. In principle, if the debiasing works properly one should not see gender biased analogies after debiasing, so for example man is to programmer would be as woman is to hacker, instead of homemaker.

3.4 Bias based on occupation words

To further check the debiasing performance we projected profession words to the gender direction. We simply took the scalar product of a profession word and the normalized gender direction to check for direct bias. Bigger negative values indicate bias towards "he" and bigger positive values towards "she". We did this for both the word2vec and Glove embeddings before and after debiasing. We also

[1] The github repository for our work can be found at https://github.com/YovchevIvan/FACT-UVA
[2] https://github.com/tolga-b/debiaswe
[3] Thei benchmark is available as Word Embeddings Benchmark (web) package: https://github.com/kudkudak/word-embeddings-benchmarks

conducted the same experiment for the softball-football direction to check for indirect bias. In principle, after debiasing, the gender specific profession analogies would disappear in both direct (he - captain) and indirect (softball - nurse) examples.

3.5 Remaining bias

[Gonen and Goldberg 2019] argued that for checking the debiasing performance more thorough quantitative assessment is needed. One of their experiments involved the 2500 most gender-biased words for males and females based on the original embeddings' projection on the gender subspace. They used a random sample of 500 for training and the remaining 2000 for testing. The authors tried to train a classifier (radial-basis SVM) before and after debiasing the embeddings, and found that the distinction can still be made after the debiasing. We extended this experiment by using 3 different classifiers (radial-basis SVM, logistic regression, MLP) and iteratively increasing amounts of training data by 10 percent points in each step to see how the classification proccess unfolds as the data size increases. We use this analysis on the original and debiased word2vec embeddings. We report the average of 10 runs, using different random seeds.

3.6 Removing remaining bias

To try to remove remaining subtle biases, instead of selecting the first principal component to define the gender subspace, we selected 18 components (the maximal subspace, as we have 18 gender-defining words), and neutralized and equalized the words using the projection on this subspace. We generated such embeddings using the word2vec embeddings and subjected them to benchmark and quantitative analysis the same way as the original ones. The comparison to the 1-component subspace is made using the average of 10 runs of the SVM-RBF model with different random seeds.

4 RESULTS

4.1 Benchmark tests

Table 1 shows the results of the web benchmark tests conducted on the original and debiased word2vec and glove embeddings. We can see that the debiased embeddings perform almost the same on all benchmark tests, which implies that the debiasing do not decrease the usefullness of word embeddings.

We can see that debiasing did not hurt the performance for neither word2vec, nor Glove embeddings. We can also observe, that removing along a larger subspace also did not hurt the performance substantially, though the results are somewhat smaller.

4.2 He-She analogies

Table 2 shows some he-she analogies generated using the debiased embeddings. We can observe, that the analogies reported in [Bolukbasi et al. 2016] could be replicated.

4.3 Occupation word experiments

Table 3-6 contains the results of the profession word projection experiments. The she-he and softball-football axes were used to illustrate the direct and indirect bias respectively. For the she-he projections those profession names which clearly indicate gender (e.g waitress, businessman) were excluded. Although the results show changes in the list of the most biased occupations, still many gender-biased profession names get in the top-5, especially along the she-he axis, which suggests remaining bias present in the debisaed embeddings.

Table 1: Benchmark tests from the Word Embeddings Benchmark package run on the original and debiased embeddings for both word2vec and glove. For word2vec we used two different debiasing methods, one where only one principal component was used for the projection (1D) and one where the full subspace (18D) The letter d stands for debiased.

	w2v	dw2v-1D	glove	dglove	dw2v-18D
AP	0.557	0.557	0.532	0.542	0.547
BLESS	0.670	0.680	0.755	0.760	0.675
Batting	0.235	0.233	0.272	0.267	0.234
ESSLI1a	0.727	0.727	0.772	0.727	0.795
ESSLI2b	0.800	0.800	0.750	0.750	0.800
ESSLI2c	0.644	0.644	0.622	0.622	0.622
MEN	0.704	0.704	0.764	0.764	0.707
MTurk	0.514	0.514	0.640	0.634	0.512
RG65	0.693	0.694	0.752	0.753	0.695
RW	0.277	0.277	0.177	0.178	0.291
SimLex999	0.435	0.438	0.398	0.401	0.443
WS353	0.649	0.645	0.698	0.697	0.641
WS353R	0.581	0.578	0.656	0.655	0.571
Google	0.334	0.332	0.386	0.384	0.322
MSR	0.570	0.570	0.550	0.552	0.565
SemEval2012	0.202	0.203	0.183	0.185	0.201

Table 2: Some he-she analogies after debiasing

sewing-carpentry	nurse-surgeon	blond-burly
petite-lanky	sassy-snappy	volleyball-football
charming-affable	giggle-chuckle	hairdresser-barber
vocalist-guitarist	diva-superstar	cupcakes-pizzas

Table 3: The top 5 most biased professions for word2vec according to the she-he axis before (up) and after (bottom) debiasing.

word2vec	she	he
1.	homemaker	maestro
2.	registered nurse	skipper
3.	nurse	protage
4.	receptionist	philosopher
5.	librarian	captain
1.	socialite	planner
2.	nurse	mechanic
3.	homemaker	gangster
4.	hairdresser	fighter pilot
5.	registered nurse	pollster

Table 4: The top 5 most biased professions for glove according to the she-he axis before (up) and after (bottom) debiasing.

glove	she	he
1.	nurse	commander
2.	homemaker	philosopher
3.	nanny	captain
4.	stylist	inventor
5.	dancer	preacher
1.	singer	officer
2.	pianist	servant
3.	warrior	technician
4.	choreographer	manager
5.	comedian	researcher

Table 5: The top 5 most biased professions for word2vec according to the softball-football axis before (up) and after (bottom) debiasing.

word2vec	softball	football
1.	bookkeper	footballer
2.	receptionist	businessman
3.	registered nurse	pundit
4.	waitress	maestro
5.	homemaker	cleric
1.	infielder	footballer
2.	major leaguer	cleric
3.	bookkeeper	vice chancellor
4.	clerk	lecturer
5.	investigator	fashion designer

Table 6: The top 5 most biased professions for glove according to the softball-football axis before (up) and after (bottom) debiasing.

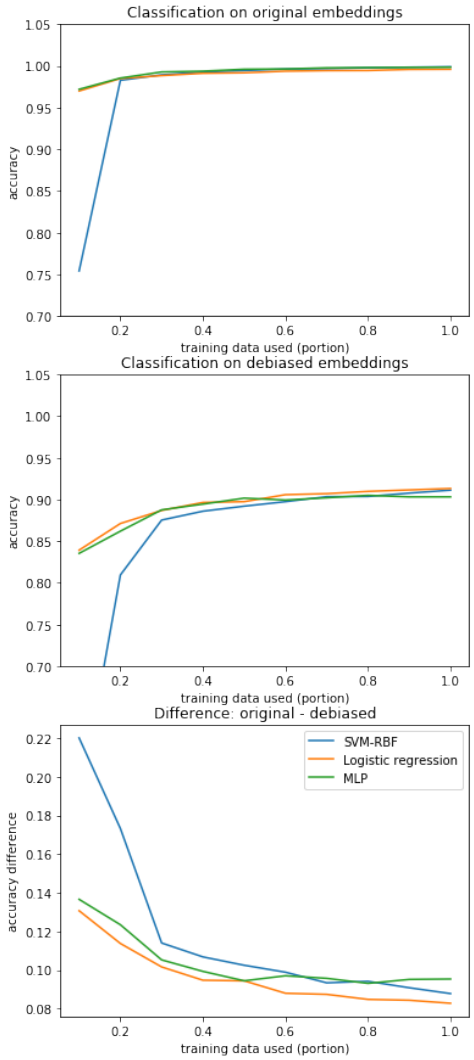
glove	softball	football
1.	counselor	footballer
2.	paralegal	boss
3.	nurse	politician
4.	realtor	coach
5.	ranger	midfielder
1.	ranger	footballer
2.	counselor	boss
3.	warden	substitute
4.	treasurer	hooker
5.	superintendent	coach

On both he-she and occupation analogies we can observe somewhat less sexist examples after debiasing, but they are still present, already hinting some remaining bias in the embeddings.

4.4 Remaining bias

Figure 1 shows the results of the classification experiment using word2vec embeddings. When we used the whole dataset the results correspond to what was reported in [Gonen and Goldberg 2019]. The iterative extension of the data set reveals that with the debiased embeddings the classifiers approach good accuracy more slowly, suggesting that, although the distinction is still present, it is better hidden in the data, thus harder to find for a classifier. We can also see on the difference chart, how the classifiers on debiased data catch up on the performance as the data set is extended. In addition the fact that the classification can be made implicates that debiasing with respect to only one principal component still leaves substantial bias within the embeddings.

Figure 1: Classification accuracy on the most biased words as incrementing the data set size

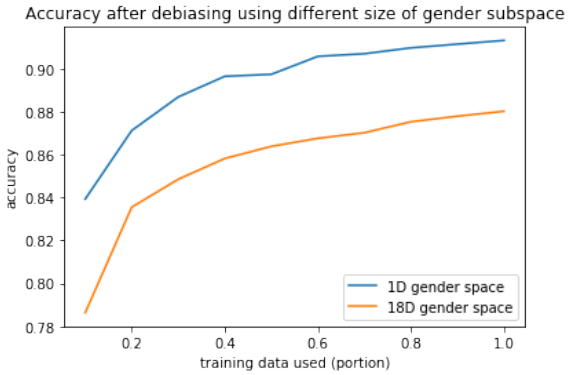


4.5 Removing remaining bias

On Figure 2 we can observe, that the classification performance on debiased data indeed drops as we project on the largest available subspace, however the difference is not large, the distinction can

still be made. This implies that the 17 other principal component also contain variance in gender, however they still do not contain all.

Figure 2: Classification accuracy after debiasing, using small and large gender space



5 DISCUSSION

Most of our results align with what was reported in [Bolukbasi et al. 2016]. The analogies could be reproduced exactly. Our occupation word projection experiment also almost replicated their results. Our lists do not exactly correspond to theirs but these are just minor differences. For example we do not have the word "pitcher" on top of our softball-lists, but that might be because we only used the most common 50000 words for generation, no the whole embedding. We consider their results replicated.

From the results we can see that the debiasing algorithm yields some positive results on the occupation words. For example "philosopher" and "captain" both vanish from the "he" list in case of the word2vec and Glove embeddings as well. Also after debiasing the word "warrior" appears as a "she" word and the word "hooker" as a "football" word in the Glove embedding.

However, some stereotypical words still remain in the top 5. The female stereotipical words "nurse", "registered nurse" and "homemaker" are not removed from the list after debiasing, which suggests there still remains bias in the embeddings after the algorithm is applied.

The main advantage of the algorithm is that it can be given any number of word pairs according to which we wish to neutralize biased words. We could put in more gender related word-pairs, but also word-pairs based on race (e.g. "white people-black people", "Europeans-Asians", "Americans-Mexicans"). Further experiments in this direction would help us find the limits of the algorithm and determine its usefulness.

The main problem of this method as argued in [Gonen and Goldberg 2019] is that the bias is not removed, only hidden. Analogy tasks and projections onto certain axes are indicators of bias, but not their whole manifestation. The bias is more deeply embedded into the geometry of the 300-dimension embedding space than it could be removed along a few axes. [Bolukbasi et al. 2016] acknowledges this problem by introducing the indirect bias, however do not offer any solutions to deal with it.

While we can see that the debiasing method makes it somewhat harder for a classifier to find gender distinctions, even simple models with few data can find the boundary with high accuracy, so it is reasonable to assume that complex methods with large amounts of data would not produce less biased models in a real-life application using this debiasing approach. We could observe, that projecting along a larger subspace slightly improves the debiasing results, which could give a method to fully mitigate the bias, but that would require defining more gender-pairs, which is hard without involving any other semantic content, and it is likely that further components capture such small variance that the debiasing performance would not be meaningfully improved. An alternative approach would be to use nonlinear dimensionality reduction, under the assumption that the gender content cannot be captured by linear methods as PCA.

This algorithm is a post-processing step to remove bias. We believe removing bias in post-processing is a very difficult problem, because the geometry of a 300-dimension space is too complicated, and as the embeddings were trained using biased data, removing the bias should mean a trade-off in performance on tasks in similar domains. The benchmark results already hinted this when extending the subspace, and we expect the drop to be even higher in case one finds more effective debiasing approaches. Another approach [Zhao et al. 2018] is to learn debiased word embeddings from scratch. This way we could have more control over the bias during the training period. Altough the [Gonen and Goldberg 2019] paper states that this method is still inefficient we believe (as do they) further research should focus in this direction.

To sum our findings we answer our questions as follows:

(1) The results reported in [Bolukbasi et al. 2016] could be reproduced by our team.
(2) The proposed algorithm is not specific to word2vec and could be successfully applied to the glove embedding as well in the sense that the performance of the debiased embedding do not drop on standard benchmark tests.
(3) Unfortunately bias still remains in the embeddings after debiasing. Some of the remaining bias is encompassed in the other 17 principal components of the definitional gender pairs, however it is still not the whole manifestation, further research is needed in this direction.

6 CONCLUSION

Our team has managed to replicate the results of the original paper [Bolukbasi et al. 2016] using, in part, artifacts provided by the authors. Therefore we choose the ACM award: Results Replicated for the paper. However, we question the usefulness of these results, because we could still find bias in the debiased embeddings. Although their work is precise and thorough we do not believe bias can be captured by a definition as simple as they used. Although ensuring fairness is undoubtedly a crucial problem for embeddings, further research is needed to tackle it.

REFERENCES

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 e-prints*, Article arXiv:1607.06520 (Jul 2016), arXiv:1607.06520 pages. arXiv:cs.CL/1607.06520

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

Stanislaw Jastrzkebski, Damian Lesniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. *CoRR* abs/1702.02170 (2017). arXiv:1702.02170 http://arxiv.org/abs/1702.02170

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. arXiv:cs.CL/1301.3781

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *CoRR* abs/1809.01496 (2018). arXiv:1809.01496 http://arxiv.org/abs/1809.01496

A CONTRIBUTIONS

The majority of the paper was written by András Csirik. He handeled the overall organization of the paper as well as the searching for relevant literature on the topic.

The coding part was done by the other three authors. Ivan Yovchev and Davide Barbieri were responsible for the reimplementation of the main debiasing code in [Bolukbasi et al. 2016] and for some of the experiments. Davide wrote the code for the analogy generation experiment and extended the debiasing algorithm to work with more than one axis. Ivan wrote the code for professions projection experiment. Both team members handled the commenting of the code and the overall environment support and documentation of the project both in the corresponding README and jupyter notebook files. Both team members also migrated over the benchmarking code to be used by the current project.

Bálint Hompot found the [Gonen and Goldberg 2019] paper and suggested to further investigate in that direction. He did the implementation of the experiments regarding that paper and also conducted all of the benchmark tests. It was also his idea to test the classification accuracies on increasing portions of the data. He also contributed to the paper by adding his respective parts to the experimental setup, the results and the conclusion sections. In addition he created the presentation.

All three of them created their respective parts of the jupyter notebook, and the explanation of their work in the readme file of the github repository.

TA

Maurice Frank