Does Double-Hard Debias Keep its Promises? - A Replication Study

Kristina Kobrock[1], Meike Korsten[1], & Sonja Börgerding[1]

[1] University of Osnabrück

Does Double-Hard Debias Keep its Promises? - A Replication Study

Introduction

Recent research has shown that word embeddings derived from natural language corpora inherit human biases. The first seminal study on this topic was by Bolukbasi, Chang, Zou, Saligrama, and Kalai (2016) who used the word analogy task developed by Mikolov et al. (2013b) to prove that “Man is to Computer Programmer as Woman is to Homemaker” - a clearly gender biased analogical relation derived from a word2vec embedding trained on Google News. Caliskan, Bryson, and Narayanan (2017) complement this finding with a large study on human-like semantic biases in Natural Language Processing (NLP) tools. They have shown in more general that human biases as exhibited in psychological studies using, for example, the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998) are learned by NLP algorithms designed to construct meaningful word representations. According to Zhao et al. (2018a) these biases propagate to downstream tasks. As pre-trained word embeddings are often used for a lot of more complex NLP tasks and architectures of our everyday life, the biased embeddings bear the risk of proliferating and strengthening existing stereotypes in human cultures.

Debiasing Embeddings. But how can the problem of biased embeddings be solved? Several researchers have proposed post-processing techniques or algorithm modifications that promise to “debias” the word embeddings obtained by algorithms like word2vec or GloVe (e.g. Bolukbasi et al., 2016; Kaneko & Bollegala, 2019; Zhao et al., 2018b). Bolukbasi et al. (2016) have developed an algorithm called Hard Debias that is based on the idea of removing (biased) gender directions from the embedding whilst preserving desired gender relations. For example, the encoded gender of the words “king” and “queen” is given by definition. So the relation between the concept *male* and “king” and between *female* and “queen” is desired. On the other hand, words like “nurse” and “doctor” also tend to exhibit relations to a specific gender in semantic space even though the words do not have a gender-specific meaning but can be used for all genders. The proposed Hard Debias algorithm first identifies a gender subspace of the embedding that best captures the bias. Then a neutralizing step ensures that gender neutral words (like "nurse" and "doctor") are indeed neutral, i.e. zero, in the gender subspace. An equalizing step is then applied in order to ensure that useful relations apply to words from both genders and are not biased towards one gender anymore. For example, the word "babysit" should be equally distant from both "she" and "he". Wang et al. (2020) build on Hard Debias in their newly proposed Double Hard Debias algorithm. The main idea is to not only remove the gender direction of the biased embedding, but also the frequency direction as it has been shown that word frequency has a significant impact on the geometry of the embedding space (e.g. Gong et al., 2018; Mu, Bhat, & Viswanath, 2018, more on this later).

Motivation. In this project for the course "Implementing ANNs with Tensorflow", we aimed to replicate the debiasing method presented by Wang et al. (2020) and to reproduce their experimental results. We chose the paper because it is the most recent one that proposes a post-processing technique for debiasing algorithms and because it builds on the seminal paper by Bolukbasi et al. (2016). In the following, we would like to quickly motivate our choice of topic by relating it to the course and stating a motivation for replication and reproduction attempts in general. The course covered a huge breadth of topics ranging from the simplest neural network architectures to advanced Convolutional Neural Networks (CNN) and recently proposed Transformer models (Vaswani et al., 2017). One of the topics covered that stroke us the most interesting was Deep Learning for Natural Language Processing (NLP) covering word embeddings, language models as well as Seq2seq models and preprocessing techniques. So our aim was to find a final project that would combine knowledge on NLP that we learned in the course with some interesting topic that would motivate our specific choice of project. The extra session on "Ethical aspects of ML" held by Pascal and Annie was a great starting point and we decided to settle on bias in word embeddings as a topic. Our search on papers proposing debiasing algorithms led us to the work described above and we decided to do re-implement the paper by Wang et al. (2020). This gave us the opportunity to apply a variety of skills learned in the course ranging from preprocessing datasets for NLP tasks, over skillfully working with word embeddings, to understanding, fine-tuning and training a NN model that we did not know before (GloVe). The reproduction of existing results is not only good practice in science, but it is also essential for gaining a deeper understanding on the methods used and it can help to validate existing results or to shed well-grounded doubt on them where needed. Recent studies of reproducibility in the field of Computer Science (e.g. Collberg, Proebsting, & Warren, 2015) and NLP (e.g. Fokkens et al., 2013; Belz, Agarwal, Shimorina, & Reiter, 2021; Cohen et al., 2018; Mieskes, 2017) explain why reproducibility endeavours are often failing and shed light on the exact nature of the “reproducibility crisis” (term coined by e.g. Baker, 2015) in NLP research. Belz et al. (2021), for example, report that the community’s interest in topics of reproducibility have risen even though reproduction attempts still tend to fail due to problems like missing data, missing code and incomplete documentation (see also Fokkens et al., 2013; Mieskes, Fort, Névéol, Grouin, & Cohen, 2019). This project aims to make a contribution to the increasing body of reproducibility and replication attempts in NLP research.

Implementation

- task description
- explanation of model and training choices

Pilot Study: Impact of Word Frequency (Meike). Wang et al. (2020) claim, that the encoding of word frequency in embeddings also significantly influences the encoding of gender direction, thereby diminishing the effectiveness of debiasing algorithms. To ground their theory, a short pilot study is performed, in which word frequencies are artificially changed and changes in the resulting embeddings are investigated. Specifically, they look at the difference vectors of the “definitional pairs”, which should approximate the gender direction (Wang et al., 2020). We could not replicate this study as closely to the paper as the other parts, due to constraints on our hardware, therefore leading to similar, but definitely less pronounced results. Ultimately though, even based on limited ressources, we could replicate that manually changing the word frequency statistics for a single word significantly influences the resulting gender direction of the embedding space. Thus, supporting the approach to put more emphasis on word frequency in the scope of debiasing.

Datasets and Preliminaries (Sonja). As dataset we used the pre-trained, 300-dimensional GloVe embedding used and provided via link on Github by Wang et al. (2020). Further they also provided several more datasets that were obtained through different debiasing methods, some applied during training, some after. The authors] used these datasets as baselines for the evaluation to compare the perfomance of their Double-Hard Debias method to. Lastly they also provide the embedding that is the reult of applying their Double-Hard Debias alogrithm to the original, non-debiased GloVe embedding. We included all of these embeddings in our evaluation to compore to the result of our implementation of Double-Hard Debias.

The original files included 300-dimensional embeddings for 322.636 words. To avoid large downloads we decided to open the files via URL, however this was not possible for the Double-Hard Debiased embedding obtained by Wang et al. (2020), therefor this embedding needs to be downloaded and open seperatly. From all the provided files we created for each embedding a vocabulary in form of a list, a dictionary mapping all words in the vocabulary to an ID and an embedding matrix that stores the 300-dimensional embedding of each word in the row corresponding to the ID of the word. During the pre-processing we restricted the vocabularies to the 50.000 most common words, which corresponds to the first 50.000 words in a GloVe embedding. At this point we also excluded any words that contain digits or special characters from our vocabularies. Both these actions follow the implementation provided by Wang et al. (2020), however the restriction to 50.000 words is not documented and apparently only happens in the code demonstrating the Double-Hard debiasing method for computational reasons. We adopted this restriction for computational reasons as well as the hope to remove some less frequent words such as names from our embedding to obtain more general results. However unlike the authors we did not specifically remove any non lower case words as the GloVe embedding is lower-cased by default and we also refrained from excluding all words longer than 20 characters.

We made use of a number of word sets provided by Wang et al. (2020). For the application of Hard Debias as proposed by Bolukbasi et al. (2016) we need sets of male and female words as well as a set of neutral words. We will later explain how we obtained the sets of gendered words based on determining the gender direction. For this we used the supplied "definitional_pairs", a word set including pairs such as "he" and "she". To obtain the neutral set, we removed all words found in the files provided by Wang et al. (2020) from our vocabulary. These files included besides the aforementioned "definitional_pairs" also "equalize_pairs", "gender_specific_full", "female_word_file" and "male_word_file" in the assumption that all these files include definitionally gendered words that should not be included in the neutral set. However it stays unclear for some of these sets where Wang et al. (2020) obtained them, except for "definitional_pairs" and "equalize_pairs" which are the gender pairs and equality sets as suggested by Bolukbasi et al. (2016). Since the "male_word_file" and the "female_word_file" include pairs of corresponding words such as "spokesman" and "spokeswoman" or "priest" and "nun" we also created a set of pairs based on these two files to be used with the debiasing algorithm.

Hard Debias (Kristina). The authors make use of the Hard Debias algorithm proposed by Bolukbasi et al. (2016). The paper at hand does not give much information on the exact implementation of Hard Debias used. The code uploaded to the authors' Github repository is not well documented, so we were not able to find the exact parts of the algorithm that should refer to the implementation of Hard Debias. That is why we sticked to the original paper from Bolukbasi et al. (2016) in order to re-implement Hard Debias.

The paper describes two steps: First, the gender direction (or, more generally, the subspace) has to be identified. This is achieved with the help of defining sets $D_1, D_2, ..., D_n \subset W$ which consist of *gender specific* words, i.e. words which are associated with a gender by definition like "girl, boy" or "she, he". These are the words that can help to identify the gender direction by capturing the concept *female, male* in the embedding $\{\vec{w} \in \mathbb{R}^d\}_{w \in W}$. Whereas in some simple implementations of Hard Debias, only one definitional pair might be used, Bolukbasi et al. (2016) suggest to compute the gender direction B across multiple pairs to more robustly estimate the bias. In our implementation we used the 10 word pairs suggested by Bolukbasi et al. (2016) which were experimentally shown to agree with an intuitive concept of gender. For these 10 pairs the principal components (PCs) are calculated and the bias subspace B is made of the first $k \geq 1$ rows of the decomposition SVD(C). According to Bolukbasi et al. (2016), the first eigenvalue is significantly larger than the rest and so the top PC is hypothesized to capture the gender subspace. So $k = 1$ is chosen and our resulting gender subspace B is thus simply a direction. C is calculated in the following way:

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

where $\mu_i := \sum_{w \in D_i} \vec{w}/|D_i|$ are the means of the defining sets $D_1, D_2, ..., D_n \subset W$.

As a second step Hard Debias neutralizes and equalizes the word embeddings. Neutralizing means to transform each word embedding $\vec{w}$ such that every word $w \in N$ has zero projection in the gender subspace. So for each word $w \in N$ in a set of neutral words $N \subseteq W$, we re-embed $\vec{w}$:

$$\vec{w} := (\vec{w} - \vec{w}_B)/||\vec{w} - \vec{w}_B||$$

. The equalize step ensures that desired analogical properties hold for both female and male words contained in the equality sets $\mathcal{E} = \{E_1, E_2, ..., E_m\}$ where each $E_i \subseteq W$. For example, after debiasing we would like the embeddings of the pair $E = \{grandmother, grandfather\}$ contained in the equality sets to be equidistant from the embedding of "babysit"

(Bolukbasi et al., 2016). This is enforced by equating each set of words $E \in \mathcal{E}$ outside of B to their simple average $\nu := \mu - \mu_B$ where $\mu := \sum_{w \in E} w/|E|$ before adjusting vectors so that they are unit length. So for each word $w \in E$, $\vec{w}$ is re-embedded to

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

. With the help of the original paper, we were successfully able to re-implement Hard Debias.

Double-Hard Debias (Meike). As previously mentioned, the code provided by the authors' was rather unsatisfactory and superficial, which is why we were mostly guided by the Pseudocode provided in the paper (Wang et al., 2020).

The Double-Hard Debias algorithm is about removing two different directions from the embeddings. The first one being one supposedly encoding frequency information, the second one being the gender direction, as proposed by Bolukbasi et al. (2016). In earlier work, Mu and Viswanath (2018) have shown, that further dimension reduction of word embeddings actually increases the linguistic regularities they capture. To do so, they first subtracted the common mean vector and then identified a few top principal components of the embedding space. After removing both of those components, the resulting word embeddings proof to serve as stronger representations. During the application of their algorithm, Mu and Viswanath (2018) noticed, that some the top PCA directions seemed to encode frequency to a significant degree. Wang et al. (2020) base their algorithm mostly on those findings and also apply PCA to their algorithm, claiming that to be a feasible method to identify directions of the embedding space encoding word frequency. The first part of the algorithm represents picking the frequency direction that shall ultimatly be removed from all embeddings.This is done in a trial-based set up on only the most biased words (500 male and female). Following Mu and Viswanath's (2018) work, the common mean vector is removed before PCA is applied. In the Pseudocode presented, the authors depict the computation of the same number of principal components as the dimensionality that the embeddings in question have. In the provided code, on the other hand, they only compute the top 10 principal components, which is what we ultimately replicated. For each of those PCs, the corresponding direction is removed, Hard Debias is applied on those projected embeddings, and performance of the resulting embeddings on the Neighborhood Metric, originally proposed by Gonen and Goldberg (2019), is measured. This Metric and its use is discussed in a more detailed manner in the evaluation part. Then, the PC is identified that lead to the best performance on the Neighborhood Metric, here meaning resulting in least biased embeddings. This direction, supposedly encoding frequency as well as gender information, is picked out of all candidate principal components for the second part of the algorithm. This second part focuses on truly debiasing the full set of word embeddings, first by removing the recently hand-picked frequency direction and then applying Hard Debias to also remove the gender direction.

Evaluation

This part deals with the reproduction of Wang et al. (2020)"s experimental results when evaluating their Double Hard debiased embedding compared to some baselines and on benchmark datasets using well-established tasks. Due to some code provided on the authors" Github repository, the task of re-implementing the evaluation part of the paper was much more straightforward than implementing the main part.

Baselines (Sonja). Following Wang et al. (2020) we used GloVe embeddings that were obtained using several different debiasing methods as baselines for our evaluation along with the original, non-debiased embedding.

original GloVe: The original, non-debiased GloVe embedding used and provided by Wang et al. (2020). It was obtained from training on the 2017 January dump of English Wikipedia.

GN-GloVe: Gender-Neutral GloVe embedding released by Zhao et al. (2018b). This method restricts gender information in certain dimensions while neutralizing in the remaining dimensions.

GN-GloVe(a): A variant of the Gender-Neutral GloVe embedding obtained by Wang et al. (2020). It was created by excluding the gender dimensions from the GN-GloVe embedding in a try to completly remove gender.

GP-GloVe: Gender preserving GloVe embedding released by Kaneko and Bollegala (2019). This method attempts to remove stereotypical gender bias and preserve non-discriminative gender information.

GP-GN-GloVe: Gender preserving, Gender-Neutral GloVe embedding provided by Kaneko and Bollegala (2019). This is the result of applying gender preserving debiasing to an already debiased GN-GloVe embedding.

Hard-GloVe: Hard debiased GloVe embedding obtained by Wang et al. (2020) following the implementation of Bolukbasi et al. (2016). This method aims to debias neutral words while preserving the gender specific words.

Strong-Hard-GloVe: Strong-Hard debiased GloVe embedding obtained by Wang et al. (2020). A variant of Hard debias during which all words are debiased instead of only the neutral words.

Double-Hard-GloVe(Wang et al.): Double-Hard debiased GloVe embedding obtained by Wang et al. (2020). This is the result of applying their proposed Double-Hard Debias method to the original GloVe embedding.

Double-Hard-GloVe(replication): Double-Hard debiased GloVe embedding obtained by us. This is the result of applying our implementation of the Double-Hard Debias method to the original GloVe embedding.

Evaluation of Debiasing Performance.

Debiasing in Downstream Applications.

Coreference Resolution (Meike). Wang et al. (2020) also evaluated the performance of their resulting embeddings in Coreference systems. Zhao et al. (2018a) have shown that "The physician hired the secretary because he was overwhelmed with clients" is processed better, because "he" relates to "the physician", which is biased with a male gender association. In contrast, non-consistent sentences, such as "The physician hired the secretary because she was overwhelmed with clients" showed poorer performance. As the code provided by Wang et al. (2020) did not implement this evaluation and we failed to reimplement the original Zhao et al. (2018a) Coreference task due to its huge computational demands, we decided to redirect our efforts to other feasible evaluation methods.

Debiasing at Embedding Level.

Table 1

WEAT Test

Embeddings	C&F d	C&F p	M&A d	M&A p	S&A d	S&A p
original GloVe	1.806	0.0002	0.6886	0.0851	1.1299	0.0119
Double-Hard-GloVe (Wang et al.)	1.531	0.0011	-0.5625	0.8693	-0.6503	0.9028
Double-Hard-GloVe (replication)	1.530	0.0011	-0.6895	0.9160	-0.9419	0.9708
GN-Glove	1.821	0.0001	-0.2564	0.6963	1.0690	0.0162
GN-Glove(a)	1.756	0.0002	0.5030	0.1585	0.8797	0.0395
GP-GloVe	1.806	0.0001	1.2085	0.0079	1.1064	0.0132
GP-GN-GloVe	1.797	0.0002	-0.0126	0.5079	0.8460	0.0453
Hard-GloVe	1.547	0.0010	-0.9826	0.9751	-0.5384	0.8592
Strong-Hard-GloVe	1.547	0.0010	-0.9857	0.9754	-0.5471	0.8625

The Word Embeddings Association Test (WEAT) (Sonja).

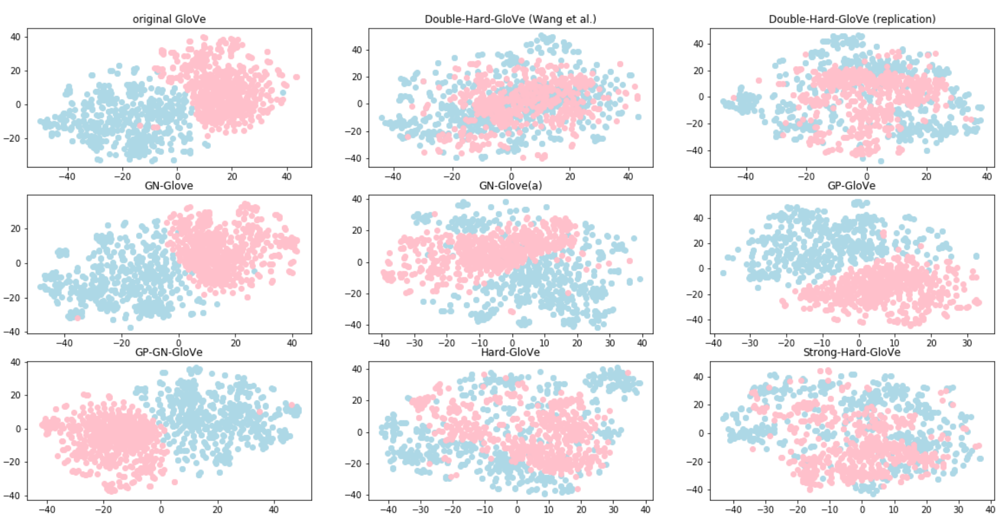
Neighborhood Metric (Meike). The Neighborhood Metric was originally introduced by Gonen and Goldberg (2019) and is based on the observation that even though debiased words no longer show their bias in form of a certain characteristic in the gender direction, the bias remains in the form of the grouping of the embeddings. The supposedly "removed" bias is still manifasted as those words socially-marked with the same gender will be situated closer to one another in the embedding space (Gonen & Goldberg, 2019). When applying K-means clustering on a number of most biased words, the clustering algorithm easily clusters "debiased" embeddings into the correct male and female categories. Wang et al. (2020) evaluate these clustering performances by simply counting the samples correctly assigned to the gender bias allegedly removed in debiasing. The alignment score a is defined as $a = \frac{1}{2k}\sum_{i=1}^{2k} 1[\hat{g}_i == g_i]$ and set to $a = max(a, 1-a)$ with k resembling the number of samples for each gender, $\hat{g}_i$ the estimated and g_i the correct label. According to this definition, an alignment value of 0.5 indicates perfectly unbiased embeddings, as the clustering algorithm failes to replicate the gender pattern (Wang et al., 2020). Even though Wang et al. (2020) make use of this metric within their proposed debiasing algorithm, thereby adjusting their embeddings to obtimize the Neighborhood Metric, they also apply it again in the scope of evaluating their debiasing performance. As one would expect, the Double-Hard debiased embeddings therefore do particularly well in comparison to the other Baseline embeddings.

Wang et al. (2020) also applied the tSNE reduction technique on the Top 500 female and male biased words to be able to plot the resulting clusters in two-dimensional space. In the graphic below you can see our plot that shows all the baseline datasets, as well as our Double-Hard debiased embedding and Wang et al. (2020)'s embedding. For the original GloVe embedding, GN-GloVe, GP-GloVe and GP-GN-GloVe the clustering accuracy is still very good meaning that the Top 500 female (lighblue) and male (pink) biased words are still biased after applying the respective debiasing techniques. Our embedding performs equally to the Double-Hard debiased embedding reported by Wang et al. (2020). Hard- and Strong-Hard-GloVe also show comparably little bias.

Table 2

Neighborhood Metric Clustering Accuracy

Embeddings	Top 100	Top 500	Top 1000
original GloVe	100.00	100.00	100.00
Double-Hard-GloVe (Wang et al.)	54.50	56.50	58.00
Double-Hard-GloVe (replication)	52.50	50.60	53.70
GN-Glove	100.00	99.90	99.90
GN-Glove(a)	100.00	99.70	98.35
GP-GloVe	100.00	100.00	100.00
GP-GN-GloVe	100.00	99.80	99.80
Hard-GloVe	51.00	51.70	50.30
Strong-Hard-GloVe	51.00	51.80	50.25



Analysis of Retaining Word Semantics (Kristina). One of the most important properties of embeddings is that they represent meaningful word semantics, for example in the form of analogies or concepts. In this section it is tested in how far the embeddings still meet this criterion after debiasing.

Word Analogy (Kristina). The word analogy task was introduced by Mikolov et al. (2013b). The task is to find the word D such that “A is to B as C is to D”. One example for an unbiased analogy is: “Man is to King as Woman is to Queen” whereas a biased analogy would be: “Man is to Computer Programmer as Woman is to Homemaker” (Bolukbasi et al., 2016). The debiased embeddings are evaluated on two word analogy test sets: the MSR (Mikolov et al., 2013b) and the Google word analogy task (Mikolov et al., 2013a) in order to find out whether they preserve desired unbiased analogies.

The MSR word analogy dataset contains 8000 syntactic questions in the form presented above. The missing word D is computed by maximizing the cosine similarity between D and C - A + B. The evaluation metric is the percentage of correctly answered questions (see Wang et al., 2020).

The Google word analogy dataset contains 19.544 (**Total**) questions, 8.869 of which are semantic (**Sem**) and 10.675 are syntactic (**Syn**) questions.

The results can be inspected in Table 3 and are in line with the results reported by Wang et al. (2020). What is interesting is that our replicated Double-Hard debiased embedding scores best in all word analogy tasks and even outperforms the original GloVe embedding.

Concept Categorization (Sonja).

Discussion

- analysis of results and evaluation of performance evaluation
- ablation studies (not applicable)
- discuss the results and what could be (partly) replicated and what not

Conclusion

Table 3

Analogy Tasks

Embeddings	MSR	Sem	Syn	Total
original GloVe	54.40	80.14	55.69	63.90
Double-Hard-GloVe (Wang et al.)	42.40	71.36	47.99	56.67
Double-Hard-GloVe (replication)	62.12	80.20	66.55	71.13
GN-Glove	51.72	75.97	54.67	61.82
GN-Glove(a)	50.72	75.89	54.26	61.52
GP-GloVe	51.62	81.92	55.42	64.32
GP-GN-GloVe	51.99	75.75	56.68	63.08
Hard-GloVe	62.57	80.00	66.10	70.76
Strong-Hard-GloVe	62.14	76.74	65.73	69.43

References

Baker, M. (2015). Reproducibility crisis: Blame it on the antibodies. *Nature*, *521*(7552), 274–276. https://doi.org/https://doi.org/10.1038/521274a

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (2021). *A systematic review of reproducibility research in natural language processing*. Retrieved from http://arxiv.org/abs/2103.07929

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR*, *abs/1607.06520*. Retrieved from http://arxiv.org/abs/1607.06520

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T., Hargraves, O., Goss, F., … Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1025

Collberg, C., Proebsting, T., & Warren, A. M. (2015). *Repeatability and benefaction in computer systems research.* studie. Retrieved from http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf

Fokkens, A., Erp, M. van, Postma, M., Pedersen, T., Vossen, P., & Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1691–1701. Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P13-1166

Gonen, H., & Goldberg, Y. (2019). *Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.* Retrieved from http://arxiv.org/abs/1903.03862

Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). FRAGE: Frequency-agnostic word representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2018/file/e555ebe0ce426f7f9b2bef0706315e0c-Paper.pdf

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

https://doi.org/10.1037/0022-3514.74.6.1464

Kaneko, M., & Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1641–1650. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1160

Mieskes, M. (2017). A quantitative study of data in the NLP community. *Proceedings of the first ACL workshop on ethics in natural language processing*, 23–29. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1603

Mieskes, M., Fort, K., Névéol, A., Grouin, C., & Cohen, K. B. (2019). NLP Community Perspectives on Replicability. *Recent Advances in Natural Language Processing*. Varna, Bulgaria. Retrieved from https://hal.archives-ouvertes.fr/hal-02282794

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N13-1090

Mu, J., Bhat, S., & Viswanath, P. (2018). *All-but-the-top: Simple and effective postprocessing for word representations*. Retrieved from http://arxiv.org/abs/1702.01417

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). *Attention is all you need*. Retrieved from http://arxiv.org/abs/1706.03762

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation. *Association for computational linguistics (acl)*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018a, January). *Gender bias in coreference resolution: Evaluation and debiasing methods*. 15–20. https://doi.org/10.18653/v1/N18-2003

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018b). Learning gender-neutral word embeddings. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1521