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 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 frequently fed into NLP architectures used for more complex tasks encountered in everyday life like Machine Translation, the biased embeddings bear the risk of proliferating and strengthening existing biases and 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 a 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. Subsequent neutralizing and equalizing steps ensure that the debiased embeddings satisfy neutrality constraints without impairing desired properties of the semantic space. 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 this 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) (e.g. He, Zhang, Ren, & Sun, 2015; Huang, Liu, & Weinberger, 2016) 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 extend on the NLP knowledge gained in the course. 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 for papers proposing debiasing algorithms led us to the work described above and we decided to reimplement 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, Jeffrey Pennington & Manning, 2014). 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 contribute to the increasing body of reproducibility and replication attempts in NLP research.

Implementation

Shortly after starting the replication attempt, we noticed that the code provided on the Github repository linked in the paper by Wang et al. (2020) was poorly documented and seemed to deviate from the paper's suggestions in some aspects. This is why we sticked closely to the ideas developed in the paper when reimplementing the proposed algorithm. In the following, we will explain in detail what we did and how specific implementation choices were motivated.

Pilot Study: Impact of Word Frequency. Double-Hard Debias is built on the claim that an embedding's encoding of word frequency significantly influences the same embedding's encoding of gender which can lead to a diminished efficacy of any debiasing

algorithm (Wang et al., 2020). As interest in debiasing embeddings increased, so did awareness of the debiasing method's weaknesses. Gonen and Goldberg (2019), for example, noticed that the leading methods barely scrape the surface of removing the full bias. Wang et al. (2020) now propose that the performance of debiasing methods, specifically of post-processing methods such as Hard Debias proposed by Bolukbasi et al. (2016), can be significantly improved by mitigating frequency features before applying the debiasing step. To ground this assumption, Wang et al. (2020) perform a short pilot study. They manipulate word frequencies in the original data to investigate the impact of frequency on gender bias contained in the embedding. This is done by sampling certain sentences twice containing specific words of the set of definitional pairs introduced by Bolukbasi et al. (2016), i.e. words with a definitional gender whose difference vectors are assumed to approximate the gender direction (Wang et al., 2020). Word embeddings were trained on a GloVe (Jeffrey Pennington & Manning, 2014) model implementation that was reworked by us (original from GradySimon). The model is trained using a minibatch size of 128, a learning rate of 0.05 and the Adam optimizer over 50 epochs. The embeddings created are 300-dimensional as we work with embeddings of that size in the remaining paper as well. These hyperparameter settings depicted a good trade-off between computational needs and performance on our machines. We could not replicate the authors' approach as neatly as other parts of the paper due to hardware constraints which led us to working with a significantly smaller corpus. This left us with results that, taking into account variation over different training runs, were overall similar, but definitely less pronounced. Despite limited resources we were able to reproduce the effect that manually changing the word frequency statistics for a single word significantly influences the resulting gender direction of the embedding space. Specifically, altering frequency of smaller corpora seems to have a more overarching effect to all of the gender-depicting vectors. This supports the conclusion of Wang et al. (2020) that enhanced emphasis needs to be put on word frequency in debiasing techniques.

Datasets and Preliminaries. The pre-trained, 300-dimensional GloVe embedding used by Wang et al. (2020) is provided via link on the authors' Github repository. The original files included embeddings for 322.636 words. During preprocessing we followed common steps such as the removal of words containing digits or special characters and a restriction of the vocabularies to the 50.000 most common words which corresponds to the first 50.000 words in a GloVe embedding. These choices were motivated by computational reasons and by the steps conducted by Wang et al. (2020) in `Glove_Debias.ipynb`. For easier access to embeddings troughout coding we created vocabularies and dictionaries mapping words to embedding indices (IDs).

We made use of a number of word sets provided by Wang et al. (2020). Among the files uploaded were `definitional_pairs`, `equalize_pairs`, `gender_specific_full`, `male_word_file` and `female_word_file`. The two files `definitional_pairs` and `equalize_pairs` are directly based on the work of Bolukbasi et al. (2016), however for the remaining files it is unclear where the word sets were obtained. The gender pairs needed for the determination of the gender direction during the Hard Debias method (Bolukbasi et al., 2016), i.e. words which are associated with a gender by definition like "girl, boy" or "she, he", are provided in `definitional_pairs`. The file `equalize_pairs` includes what Bolukbasi et al. (2016) called the equality sets (word pairs such as "Women" and "Men" or "princess" and "prince"), which are needed for the equalization step. As an alternative to this, we created another set of pairs from the `female_word_file` and the `male_word_file` provided on the Github repository of Wang et al. (2020) since they also include corresponding words such as "priest" and "nun" or "spokesman" and "spokeswoman". Lastly the Hard Debias algorithm expects to be passed a set of neutral words as well as sets of female and male words (Bolukbasi et al., 2016). We will later describe how the two gendered sets are obtained using the gender direction. To obtain the neutral set, we removed all words found in the five files provided by Wang et al. (2020) from our vocabulary, since these files include words with a definitional gender.

Hard Debias. 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, which is why we sticked to the original paper from Bolukbasi et al. (2016) in order to reimplement 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. 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 (Bolukbasi et al., 2016). 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". 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 such that they are of 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 Bolukbasi et al. (2016) paper, we were successfully able to reimplement Hard Debias.

Double-Hard Debias. Our implementation of Double-Hard Debias was guided by the pseudocode presented in the paper (Wang et al., 2020). The main idea of the algorithm is to remove two different components from the embedding. The first one supposedly encodes frequency information, the second one captures the gender direction, as proposed already by Bolukbasi et al. (2016). In earlier work, Mu et al. (2018) have shown that removing dominant directions of word embeddings can increase the linguistic regularities they capture. Specifically, they first subtract the common mean vector $\mu = \frac{1}{|V|} \sum_{\vec{w} \in V} \vec{w}$, with V being the set of word embeddings $\vec{w}$, from all embeddings. Then, a few top principal components of the embedding space are identified and removed. The resulting word embeddings actually prove to serve as stronger linguistic representations than the unprocessed ones. Mu et al. (2018) noticed that some of the top PCA directions encode frequency to a significant degree. Wang et al. (2020) base their debiasing technique on those findings and apply the described steps in their algorithm as a method that identifies the directions of the embedding space which encode word frequency.

The first part of Double-Hard Debias identifies the frequency direction that is eventually ought be removed from all embeddings. This is done on a subset of 500 female and 500 male most biased words by examining different subspace projections and how

these influence Hard Debias' results. Following Mu et al. (2018), the common mean vector is removed before PCA is applied. In the pseudocode presented, Wang et al. (2020) calculate as many principal components (PCs) as denoted by the dimensionality of the embedding. In the provided code, on the other hand, they only compute the top 20 principal components. As the components do not encode much variance of the data after around 20 PCs, we also decided to compute only the first 20. For each of these PCs, the corresponding direction $\mathbf{u}$ is removed $w' = w - (\mathbf{u}^{\mathbf{T}}w)\mathbf{u}$, Hard Debias is applied, and performance of the resulting embeddings on the Neighborhood Metric (Gonen & Goldberg, 2019) is measured. This metric will be discussed in a more detailed manner in the evaluation part. Then, the PC that leads to the best performance on the Neighborhood Metric, i.e. results in least biased embeddings, is identified. This direction, which is hypothesized to encode frequency and to significantly affect gender information, is needed for the second part of the algorithm. The second part focuses on debiasing the full set of word embeddings. The choosen frequency direction is removed from all embeddings before Hard Debias is applied. By first creating the frequency-mitigated projections, removing the gender direction shows to be more successful.

Evaluation

This part deals with the reproduction of the experimental results by Wang et al. (2020) when evaluating their Double Hard debiased embedding compared to some baselines and on benchmark datasets using well-established tasks. We could reuse more of the code written by Wang et al. (2020) and provided on the authors' Github repository for the evaluations than for the implementation.

Baselines. Following Wang et al. (2020) we used GloVe embeddings that were obtained using several different debiasing methods, some applied during training and some during post-processing, as baselines for our evaluation along with the original, non-debiased embedding and the Double-Hard debiased embedding obtained by Wang et

al. (2020). The embeddings were linked on the Github repository of Wang et al. (2020).

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

Double-Hard-GloVe(Wang et al.): Double-Hard debiased GloVe embedding obtained by Wang et al. (2020) and reported to be the result of their implementation of Double-Hard Debias.

Double-Hard-GloVe(replication): Double-Hard debiased GloVe embedding obtained in this project. This is the result of applying our implementation of Double-Hard Debias to the original GloVe embedding.

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 Gender-Neutral GloVe obtained by Wang et al. (2020). Created by excluding the gender dimensions from the GN-GloVe embedding in an attempt to completely 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 which debiases all words instead of only neutral words.

Evaluation of Debiasing Performance. Wang et al. (2020) evaluated their embeddings on different tasks to test multiple of the characteristics that embeddings inherit. Even though we were unable to replicate all those tests, we focused on covering the main aspects.

Debiasing in Downstream Applications. As shortly mentioned above, a main concern regarding bias in word embeddings regards the application of pre-trained word embeddings in downstream tasks. The following method was developed to measure in how far bias contained in a pre-trained embedding influences downstream task performance.

Coreference Resolution. Wang et al. (2020) evaluated the performance of their Double-Hard debiased embedding in comparison to the baselines in coreference systems. Bias in coreference models has been shown by Zhao et al. (2018a) as models which more successfully identify the semantic reference in "The physician hired the secretary because **he** was overwhelmed with clients" due to the biased association of "the physician" with the male gender, allowing it to be more readily related to the concept *he*. In contrast, non-consistent sentences such as "The physician hired the secretary because **she** was overwhelmed with clients" showed poorer performance. The idea is that the societal bias of "the physician" is implemented within its embedding, thereby influencing the model's performance. Unbiased embeddings, on the other hand, should lead to equal model performance with both consistent and non-consistent samples. Wang et al. (2020) themselves neither provide any code for evaluations on the coreference task, nor relate to any external source of the specific results they present in their paper. Following the links to the original Zhao et al. (2018a) coreference task and WinoBias Dataset, which were the only reference points provided, implementation of those showed to be infeasible due to their huge computational demands. Ultimately, we choose to leave this evaluation out.

Debiasing at Embedding Level. Besides evaluating the performance on different downstream NLP tasks, the authors also specifically investigated how much bias can be determined in the embeddings.

The Word Embeddings Association Test (WEAT). Following along the evaluation of Wang et al. (2020) we tested the bias of all embeddings with a permutation test. This Word Embedding Association Test (Caliskan et al., 2017) takes four sets of words, two sets of so-called target words and two sets of attribute words. It calculates the relative similarity between the target words and the attribute words respectively and outputs how likely it is that this result could have been obtained from a non-biased distribution. The two values it returns are the effect size d and the p-value p. A value of $p < 0.05$ indicates a significant bias and the bias is considered more pronounced for larger effect sizes. We used a pre-implemented version of WEAT for our evaluation that can be obtained here. Instead of reusing the code of Wang et al. (2020) we opted for a different WEAT implementation that is oriented on the original paper by Caliskan et al. (2017). We decided not to use the authors' original code since it was partly incomplete and poorly understandable. Instead we decided for a more compact implementation where d and p are obtained simultanously. Wang et al. (2020) conducted WEAT three times for each embedding, once with the bias *Career & Family*, once with *Math & Arts* and lastly with *Science & Arts*. We replicated all three of these with the word lists taken directly from Caliskan et al. (2017). Following Wang et al. (2020), we made one alteration to the sets, namely exchanging "Bill" in the male names of the *Career & Family* set for "Tom". This is to avoid ambiguity due to the lower-casing of GloVe. The exact lists can be found in our implementation of the evaluation. The results we obtained can be inspected in Table 1.

With the WEAT implementation we used we were able to almost perfectly reproduce the results of Wang et al. (2020) for the *Career & Family* word set. Both d and p differ only slightly from the reported values and we can clearly see that p is larger for the Double-Hard debiased embeddings not only in comparison to the original embedding but

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	0.0001	0.6886	0.0837	1.1299	0.0116
Double-Hard-GloVe (Wang et al.)	1.531	0.0011	-0.5625	0.8702	-0.6503	0.9033
Double-Hard-GloVe (replication)	1.530	0.0012	-0.6895	0.9160	-0.9419	0.9707
GN-Glove	1.821	0.0001	-0.2564	0.6960	1.0690	0.0161
GN-Glove(a)	1.756	0.0002	0.5030	0.1586	0.8797	0.0398
GP-GloVe	1.806	0.0002	1.2085	0.0077	1.1064	0.0135
GP-GN-GloVe	1.797	0.0002	-0.0126	0.5102	0.8460	0.0447
Hard-GloVe	1.547	0.0010	-0.9826	0.9756	-0.5384	0.8586
Strong-Hard-GloVe	1.547	0.0010	-0.9857	0.9761	-0.5471	0.8620

also to the other debiasing methods. Despite the bias still being significant ($p < 0.05$), the comparably large p and small d suggest a more effective debiasing. We were also able to obtain results which are in line with Wang et al. (2020) for the Double-Hard debiased embedding provided by the original authors and for our self-debiased embedding. For both *Math & Arts* and *Science & Arts* we obtained values for d and p that differ noticably from the values reported by the original authors. However, as in Wang et al. (2020), we can see that the bias in *Math & Arts* was already insignificant in the original GloVe embedding. Additionally, the high p-values for Double-Hard, Hard and Strong-Hard debiased embeddings in *Science & Arts* still suggest that these embeddings successfully debias this concept.

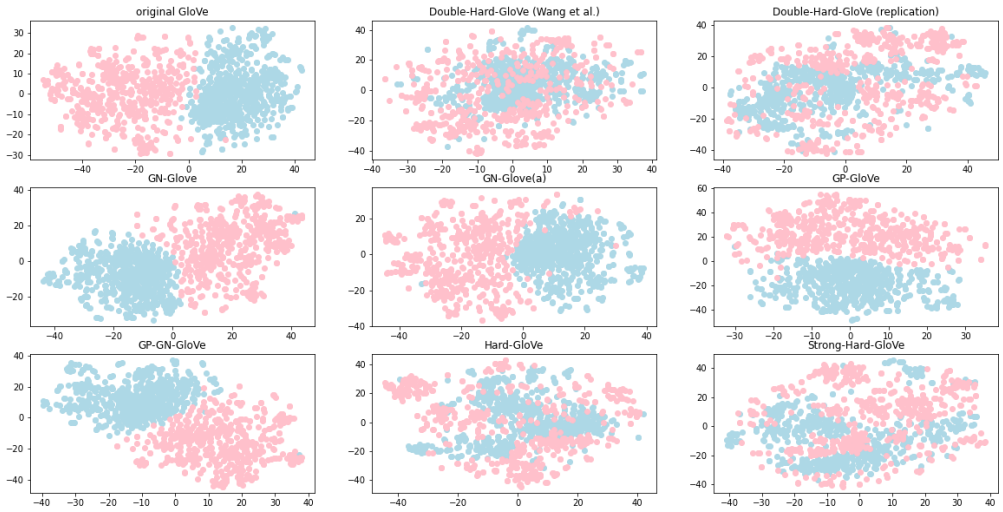
Neighborhood Metric. The Neighborhood Metric was originally introduced by Gonen and Goldberg (2019) and is based on the observation that even though debiased words no longer inherit their bias in form of a specific gender direction, it remains in the word embeddings' spatial grouping. The supposedly "removed" bias is still manifested as

one can observe those words socially-marked with the same gender situated closer to one another in the embedding space (Gonen & Goldberg, 2019). K-means clustering on a number of most biased words can often straightforwardly cluster "debiased" embeddings into the correct female and male categories, uncovering the remaining bias. Wang et al. (2020) evaluate K-means clustering performance on such most biased words 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 score value of 0.5 indicates perfectly unbiased embeddings, as the clustering algorithm failed to replicate the original gender pattern (Wang et al., 2020). Even though Wang et al. (2020) make use of this metric within their debiasing algorithm, adjusting their embeddings to optimize performance on the Neighborhood Metric, they also apply it again in the scope of evaluating their debiasing performance in comparison to other embeddings. As one would expect, the Double-Hard debiased embeddings therefore do particularly well in comparison to the other baseline embeddings. Our results can be observed in Table 2, and are visualized with the help of tSNE below. Wang et al. (2020) applied the tSNE reduction technique to the top 500 female and male biased words to be able to plot the resulting clusters in two-dimensional space. Both values in Table 2 and the tSNE visualization show that Hard-Glove, Strong-Hard-Glove and both Double-Hard-GloVe embeddings obtained by us and by Wang et al. (2020) lead to nearly bias-free clustering performances, whereby our Double-Hard debiased embeddings slightly outperform the authors' ones. The remaining baseline embeddings on the other hand can nearly perfectly be clustered, indicating remaining bias. Where our results deviate considerably to the ones from the paper is in the performance of (Strong-)Hard-Glove. In our evaluation these embeddings also show clustering accuracies alluding to basically no remaining bias, while Wang et al. (2020) present results that would indicate significantly more remaining 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	58.40
Double-Hard-GloVe (replication)	53.00	50.60	53.70
GN-Glove	100.00	99.90	99.90
GN-Glove(a)	100.00	99.70	97.90
GP-GloVe	100.00	100.00	100.00
GP-GN-GloVe	100.00	99.80	99.80
Hard-GloVe	51.50	51.50	50.15
Strong-Hard-GloVe	50.50	51.60	50.90



In the graphic above you can see a 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 high, meaning that the top 500 female (lighblue) and male (pink) biased words are still biased after applying the respective debiasing techniques.

Analysis of Retaining Word Semantics. 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. 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. For the task of Concept Categorization there are word sets provided, each of which is assigned to one of k classes. After spatial clustering of the embedding space, the accuracy of how many words of a class could correctly be assigned to the same cluster is evaluated. Higher accuracies suggest that those words with a high

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	44.10	69.28	51.01	57.80
GN-Glove(a)	43.26	69.00	50.64	57.46
GP-GloVe	42.36	73.53	51.55	59.72
GP-GN-GloVe	44.65	69.44	53.18	59.22
Hard-GloVe	54.27	73.62	62.74	66.79
Strong-Hard-GloVe	53.94	71.23	62.43	65.70

semantic relation, i.e. referencing the same concept, are also represented by spatially close embeddings as is desirable for an embedding space. Wang et al. (2020) applied this evaluation method to all embedding baselines and their Double-Hard debiased embedding. However, we found the code provided by the authors to be incomplete. This hindered us from replicating its application and due to time constraints we decided not to reimplement the missing parts.

Discussion

Reproduction of Results. The code provided by Wang et al. (2020) is incomplete and not well documented which impeded both our reimplementation and evaluation of the Double-Hard Debias algorithm. For example, the authors' notebook `GloVe_Debias` only showcases the application of Double-Hard Debias but does not match the debiased embedding uploaded by Wang et al. (2020). Several preprocessing steps such as the

restriction of the vocabulary to 50.000 words or the removal of words containing digits are applied in the exemplary notebook `GloVe_Debias`, but go unmentioned in the respective paper. It is therefore not clear at all which preprocessing steps were applied by the original authors and are demanded for a successful application of Double-Hard Debias. We also noted that judging the information provided (i.e. Github repository and paper), it was not apparent whether the original authors used the word pairs from `equalize_pairs` or those from the combination of `male_word_file` and `female_word_file` for the equalization step of the Hard Debias method. Even though we showed that this does not lead to a change in the results, this uncertainty could have been avoided by better documentation. Despite these problems, we were able to successfully reimplement the Double-Hard Debias algorithm proposed by Wang et al. (2020). This success is proven by the performance of our own Double-Hard debiased embedding which can be inspected in our evaluations.

Another issue with the code provided by Wang et al. (2020) which turned out to be a problem mainly for the evaluations is the missing `web` module. This is one of the main reasons why executing and reusing code provided on the Github repository failed. The authors did not provide this module in their repository, nor did they link to it or even mention its absence or what it used to contain. This resulted in a high reproduction effort and failure to reproduce some parts of the evaluation due to the need to self-implement functions and preprocess datasets found online such that they match the inputs required by the code of the original authors. Another failure to reproduce a part of the evaluation was due to missing code on the Coreference Resolution task. While the authors supplied two links on their Github repository, these proved not very helpful for the task at hand. Despite these challenges, we were able to obtain useful results for three parts of the evaluation, namely WEAT, the Neighborhood Metric and two Word Analogy tasks. Both our results for the Neighborhood Metric and the Word Analogy tasks closely resemble those of Wang et al. (2020). They are not only in line with the results reported by the original authors, but also show that our implementation of Double-Hard Debias turned out

to outperform all baselines including the debiased embedding provided by Wang et al. (2020). However, in the application of WEAT our results differ from those of the original authors for two of three word sets. At least for the set *Career & Family* we obtained values which closely resemble the results reported by the authors. This suggests that the method we used successfully reimplements WEAT even if we could not reproduce the original authors' results for the other two word sets where we did not only obtain differing results for all baseline embeddings, but also for the two Double-Hard debiased embeddings. Still, our results support the main conclusion drawn by Wang et al. (2020) that Double-Hard debiased embeddings (together with Hard and Strong-Hard debiased embeddings) best debias these concepts according to WEAT. .

Remarks on Double-Hard Debias. When working as closely with a paper as a reproduction study requires, this also invites to rethink and discuss main claims made by the original authors. In the case of Wang et al. (2020), the main hypothesis made by the authors appears disputable. The authors claim that the success of their algorithm can be explained by the mitigation of frequency information in embeddings before debiasing. It is argued that frequency has been shown to be a prominent characteristic of trained embeddings (Gong et al., 2018) and is represented to a significant amount in multiple of their top PCs (in line with Mu et al., 2018). Nevertheless, we would like to point out that picking out one PC and claiming it to represent any unique feature is problematic. Principal components are not only complied of a multitude of the data's characteristics, but are also features created by interaction of those characteristics. Grasping one PC's specific meaning is as complicated and convoluted as it is for the embeddings they were computed from. This is arguably also the reason why Mu et al. (2018) present their method and results without making any definite claims about general relations to word frequency information. Wang et al. (2020), on the other hand, take the observation made by Mu et al. (2018) as a universal truth and rely on it as such when making their main hypothesis. Based on these concerns we would like to argue that while the results obtained

by Double-Hard Debias might be very well evaluated in practical tests, the algorithm's theoretical foundation is not at all incontestable.

Another concern is that Wang et al. (2020) do not propose a specific PC to remove, but base their choice on the most favorable result as evaluated by the Neighborhood Metric. Specifically, they propose to remove the direction that will optimize subsequent performance of Hard Debias in every step. Ultimately, by claiming that Double-Hard Debias removes frequency information, the authors also implicitly state that these must be embedded in potentially any of the PCs, especially as their pseudocode considers all possible PCs as candidates. We believe that what the paper and algorithm truly manage to achieve is remarkable, but should also be described in a more transparent way. Double-Hard Debias extends on what Mu et al. (2018) also show: Removing dominating directions of already trained embeddings is a quite inexpensive way to considerably improve them. Not only are the resulting embeddings better linguistic representations (in line with Mu et al., 2018), but, as Wang et al. (2020) show and as we were also able to reproduce, they also lead to more effective debiasing for the post-processing technique Hard-Debias (Bolukbasi et al., 2016). Again, we want to stress that there is sufficient ground to believe that the projection mitigates frequency features, but just as other features will be mitigated. Which exactly those are will vary from embedding set to embedding set. Ultimately, claiming that Double-Hard Debias works because it removes frequency as such seems to be a simplification of the algorithm.

Our Results in Light of the Reproducibility Crisis. Preprocessing, despite being not typically reported in scientific NLP papers, is one of the main aspects that can cause experimental variation identified by Fokkens et al. (2013). In line with this, the paper by Wang et al. (2020) did not specify how the relevant datasets were preprocessed. This could be one of the reasons why our results deviate from the results obtained by the original authors. While we were able to reproduce the main conclusions and findings of the paper, we were not able to reproduce every concrete value. As Mieskes et al. (2019) point

out: “Experiments in replication fail more often than not.” Whereas the replication attempt described in this report was not a complete failure, many difficulties were identified during the process of replication. Fragmentary documentation of the original study and lack of provided code complicated our reproduction efforts. The main reasons for this appear to be common in the field: incomplete documentation, missing files and non-working links (see Mieskes, 2017; Collberg et al., 2015).

Conclusion

To sum it up, the reproduction efforts made in this study were successful. We were able to reimplement the Double-Hard Debias algorithm proposed by Wang et al. (2020) to a satisfactory degree. This was a costly and time-consuming endeavour due to incomplete code and deficient documentation provided in both paper and accompanying Github repository. But our results reward these investments with a very good performance on common embedding evaluations compared to baselines and to the Double-Hard debiased embedding provided by the original authors. Even though we were not able to replicate two out of five evaluation methods, the results obtained by the remaining methods speak for themselves. One of our main personal achievements was the ability to critically discuss the method presented by the original authors. While we agree on the algorithm’s practical success, we also pointed out where the paper makes partly extreme promises from a theoretical point of view. With this reproduction study, we contribute to the growing body of replication and reproduction attempts in the field of NLP. While we identified challenges which appear to be common for the field, we were also able to find a workaround for many of them and hope to inspire other reproduction attempts.

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

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from http://arxiv.org/abs/1512.03385

Huang, G., Liu, Z., & Weinberger, K. Q. (2016). Densely connected convolutional networks. *CoRR*, *abs/1608.06993*. Retrieved from http://arxiv.org/abs/1608.06993

Jeffrey Pennington, Richar Socher, & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1532–1543. Dohar, Qatar: Association for Computational Linguistics. Retrieved from https://nlp.stanford.edu/pubs/glove.pdf

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