Mitigating Gender Bias in Word Embeddings using Explicit Gender Free Corpus

David Hargrave
School of Electronic Engineering and Computer Science
Queen Mary University of London
d.r.hargrave@se19.qmul.ac.uk

***Abstract*—Words embeddings are the fundamental input to a wide and varied range of NLP applications. It has been shown that these embeddings reflect biases, such as gender and race, present in society and reflected in the text corpora from which they are generated, and that these biases propagate downstream to end use applications. Previous approaches to remove these biases have been shown to significantly reduce the direct bias, a measure of bias based on gender explicit words, but it was subsequently demonstrated that the structure of the embedding space largely retains indirect bias as evidenced by the spatial separation of words that should be gender neutral but are socially stereotyped on gender. This paper proposes a new method to debias word embeddings that replaces words in the training corpus that have explicit gender with gender neutral tokens, and creates the embeddings for these replaced words from the embedding of the gender neutral token post training utilising an added gender dimension. By design this method is able to fully mitigate direct bias and experiments demonstrate this. Experiments are also performed to investigate the effect on indirect bias, but generally are unable to achieve the reductions obtained by previous methods.**

I. Introduction

A word embedding is a compact representation of a word as a vector $\vec{w}$ in $\mathbb{R}^d$ with d usually between 50 and 300. An embedding space is thus a set of word embeddings. They are generated by algorithms such as GloVe (Pennington, Socher, and Manning 2014) and Word2Vec (Mikolov et al. 2013) that are trained on large text corpora.

Bolukbasi et al. (2016) identified that these embedding spaces contained gender bias. They defined a gender direction in the embedding space as the first principal component of the subspace spanned by a set of ten explicit gender word pairs[1]. The bias of a word embedding is then calculated as its projection, defined as the cosine similarity of normalized vectors, onto this gender direction. Using this metric they showed that gender biases defined by crowd workers were present in the embeddings, and conversely the crowd workers agreed with gender biases created from the embeddings.They coined the well known biased analogy found in the embedding space 'Man is to computer programmer as woman is to homemaker'

Caliskan, Bryson, and Narayanan (2017) demonstrated stereotyped biases, including gender bias, in word embeddings. They developed the Word-Embedding Association Test (WEAT) as analogous to the human Implicit Association Test (IAT). WEAT also used cosine similarity as a measure of correlation, and they were able to reproduce results from the IAT such as female names being more associated with family and arts as opposed to male names being more associated with career and mathematics.

They also showed that these biases propagate to downstream AI applications that use word embeddings. For example, in machine translation to English from a gender neutral language such as Turkish, "O bir doktor. O bir hemşire." translates to "He is a doctor. She is a nurse."

II. Previous Work

As well as identifying the gender bias issue, Bolukbasi et al. (2016) also implemented two different algebraic methods to debias the word embeddings after training. Neutralize and Equalize adjusts the gender neutral word embedding vectors to be orthogonal to the gender direction and equidistant to both words in a gender pair (e.g. he and she). The less rigid Soften method seeks to maintain the structure of the embedding space by preserving pairwise inner products between all the word vectors whilst minimizing the projection of the gender neutral words onto the gender subspace.

Zhao et al. (2018) take the approach of modifying the cost function to debias the word embeddings during training, with the aim of forcing the gender component into the last dimension of the embedding vectors. They too identify a gender direction from a set of predefined gender pairs, as the average of the difference between the embeddings in each pair, excluding the last dimension. They modify the standard GloVe cost function to include additional terms, one to force the gender component for male and female words apart, and the second to make gender neutral words orthogonal to the gender direction.

Lu et al. (2018) proposed a method, Counterfactual Data Augmentation (CDA), to modify the text corpus before training. They identify a list of gender pair words, and duplicate the training corpus swapping words that occur in a gender pair with the other word in the gender pair, whilst retaining semantic correctness. The aim here is to create a gender balanced corpus on which to train the word embeddings.

Gonen and Goldberg (2019) devised a set of tests to demonstrate, that whilst Bolukbasi et al. (2016) and Zhao et al.

[1]Explicit gender words are those that explicitly define a gender such as he, woman, uncle, and queen. A gender pair is an equivalent pair of gender words such as he and she, man and woman etc.

This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grant 825153 (EMBEDDIA).

(2018) did reduce the direct bias with respect to the definition as projection on to the gender direction, the embedding space still retained indirect bias, a structure between words that are not explicitly gendered but are socially stereotyped on gender, that can be used to infer gender based on the distance between vectors. Hall Maudslay et al. (2019) demonstrated a similar result for the Liu at al. (2018) method using a modification of the Gonen and Goldberg (2019) tests.

Hall Maudslay et al. (2019) also modified the CDA approach into Counterfactual Data Substitution (CDS), a technique that avoids duplication of text, by swapping gendered words pairs (from a list of 124 pairs) in situ with 50% probability. They also apply a technique called Names Intervention whereby names from the United States Social Security Administration (SSA) dataset are swapped in a manner that aims to preserve gender specificity[2] and frequency of use[3]. This approach is able to achieve a significant reduction in indirect basis using a modified method of the clustering technique defined by Gonen and Goldberg (2019). However this did not fully resolve the gender bias problem, and issues associated with it still remain.

In this paper an alternative pre-processing approach is proposed, whereby explicit gender is removed from the corpus before training. By design, this approach will yield equivalent results to the Neutralize and Equalize method of Bolukbasi et al. (2016) and will retain gender appropriate analogies. This paper aims to demonstrate these results and investigate the effect on indirect bias.

III. Methodology

Word embedding models such as GloVe (Pennington, Socher, and Manning (2014)) and Word2Vec (Mikolov et al. (2013)) use co-occurrence of words within a small window, to generate the word embedding vectors. Words such as he and she, king and queen, and man and woman, and given names have explicit gender, so that words that co-occur with these can inherit gender associations, and thus be biased, towards a particular gender. So an approach that can remove such associations from the training corpus may be able to mitigate this gender bias.

Word pairs such as he and she, king and queen, and man and woman essentially represent the same thing with the sole difference being gender. This suggests that the embedding vectors for these pairs of words should vary only in a gender direction. Given names also have gender (although to varying degrees) but are just labels assigned to people and really should not carry any other meaning, which suggests that they should cluster around some fixed vector and again vary only in a gender direction.

So the approach taken here is to substitute these gender pair words and names in the training corpus with gender neutral tokens before training. After the word embeddings have been trained, word embeddings for the individual gender pair words and names are created from the substituted token embedding vectors by the addition of an extra gender dimension. The values in this extra dimension are set to small non-zero values for the substituted words as described later, and to zero for all other words.

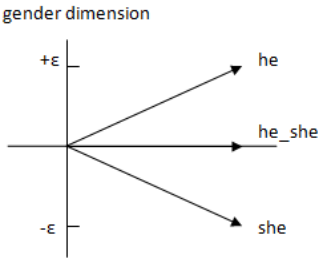

Fig. 1. Creation of word embedding for he and she from he_she. Here the he_she vector is shown as 1-d whereas in actuality it will be higher dimensional. The gender dimension is always 1-d.

For the gender pair words, a new token is created by combining the male word and the female word, separated by an underscore, e.g. man or woman are combined into man_woman. This new gender neutral token then replaces all occurrences of man or woman in the corpus. The list of gender pairs to be substituted is compiled as a subset of the definitional and equalize pairs used by Bolukbasi et al. (2016), and the list of male and female words used by Zhao et al. (2018).

Given names are replaced by a new token _name_ which is gender neutral. The identification of names uses the same source as used in the Names Intervention of Hall Maudslay et al. (2019), namely the United States Social Security Administration (SSA) dataset.

After the word embeddings have been created, embeddings for the substituted words are created from the embeddings of the substituting tokens. Firstly an extra dimension to represent gender is added to the embedding space.

For the substituted gender pair words, the male and female embedding vectors are created from the embedding of the combined token vector with the value in this extra dimension set to $+\epsilon$ for the male word in the pair and to $-\epsilon$ for the female word, for some small value of ϵ. An example is shown in Fig 1.

Embedding vectors for all substituted names are created as the embedding vector for the substituting token _name_ with a non-zero value in the added gender dimension. The value in the extra dimension is set to a value between $-\epsilon$ and $+\epsilon$, depending on the gender specificity of the name. The gender specificity is determined using the same method as Hall Maudslay et al. (2019). The SSA dataset, in addition to listing all names given in the US since 1880, also lists gender and frequency of use. It is thus possible to calculate a weighting between -1 (100% female) and +1 (100% male) and the value in the gender dimension is set to ϵ multiplied by this weighting. All other words will be considered gender neutral and have the gender dimension value set to 0.

[2]The degree to which a name is more male or female

[3]Names are swapped with other names that have similar frequency of use

This results in the difference between words in a gender pair being solely in the gender dimension e.g. $\vec{he}$-$\vec{she}$ is a vector of all zeros except for a value of 2ϵ in the gender dimension, whereas gender neutral words have a value of zero in the gender dimension. Thus this approach leads to the removal of direct bias;

- Gender neutral words are orthogonal to the gender pair directions such as $\vec{he}$-$\vec{she}$, which has been used as a definition of direct bias.
- Gender neutral words will be equidistant to both words in a gender pair e.g. man and woman.
- Correct gender associations for words in gender pairs, such as he is to she as man is to woman, will be present.
- And by controlling the size of ϵ it will be possible to ensure that the analogy test 'man is to surgeon as woman is to w?' will result in w=surgeon.

Experiments are performed to create the word embeddings using this prescribed methodology, and to demonstrate the removal of direct bias and investigate the effect on indirect bias.

IV. Experimental Setup

Wiki data is downloaded and preprocessed to create a training corpus. Original (biased) word embeddings are then created using the GloVe model. Gender pairs and names are then replaced in the training corpus with the substitution gender neutral token. The debiased embeddings are then created, after which the gender dimension is introduced. Word embeddings for the substituted gender pair words and names are then created from the embedding of the substitution token.

A. Data source

Wikipedia dumps are downloaded to create 3 separate corpora (500A, 500B and 500C) to train the word embeddings. Each corpus has approximately 500 million tokens. Details of the dumps used for these corpora are in Appendix A.

B. Data preparation

1) Preprocessing: Basic preprocessing is performed to prepare the corpus for training;

- split words separated by hyphen or forward slash
- remove words containing non-latin characters
- remove punctuation and special characters
- remove words that are not either all alphabetic or all numeric
- convert word to lowercase

Both sets of embeddings are created from the corpora that has had preprocessing applied.

2) Gender pairs: Gender pairs were collated from the definitional and equalize pairs of Bolukbasi et al. (2016) and the seed words of Zhao et al. (2018). Not all words were included.

- Pairs of words that were not considered to be exact matches that differ only in gender were removed e.g. fella and lady, beard and toque, and Catholic_priest and nun.
- Animal pairs were also removed because the selection is somewhat limited and arbitrary, and the language used to describe animals is considered to be less dependent on gender.
- Pairs where one of both words occur with low frequency were removed. However such pairs where the words could co-occur in a unique context were retained.

This resulted in a list of 77 unique gender pairs, which can be found in Appendix B.

3) Names: Names are retrieved from the same source as used by Hall Maudslay et al. (2019), namely the United States Social Security Administration (SSA) dataset. The dataset contains annual lists of all given names in the U.S. since 1880, including gender and count, e.g. the 2020 file has two entries for the name Taylor:

Taylor,F,1729
Taylor,M,456

From this, a gender specificity percentage (gsp) is calculated as:

$$gsp = (tmuc - tfuc)/(tmuc + tfuc) \quad (1)$$

where tmuc is the total male usage count and tfuc is the total female usage count.

4) Substitution: To prepare the corpus for the creation of the debiased word embedding vectors, individual male and female words in the corpus that occur in a gender pair, and names, are replaced by the appropriate substitution token.

The male and female words are replaced by the appropriate gender pair token e.g. 'he' or 'she' is replaced by 'he_she'.

Names are replaced with the '_name_' token. There are over 100,000 unique names in the SSA. Many of these names have a very low frequency usage count and thus name substitution was restricted to those that had been given at least 2,000 times in total, across both genders. This resulted in 7,082 names that were replaced. The usage counts in the SSA dataset show that these accounted for over 95% of given name usage. Seven names were removed from the list as they were also appeared as words in gender pairs.[4].

C. Word embedding creation

Word embeddings are created using the GloVe model. The parameters for the GloVe model are left unchanged from those in the demo.sh script downloaded from the Stanford GloVe website (Pennington (2014)), with the one exception being the vector size.

A set of original word embeddings are created from the original corpus with the vector size set to 200.

Substitution is then applied to the corpus, and a set of debiased word embeddings created, with a vector size of 199. The extra gender dimension is then added by increasing the vector size to 200, and setting the value in this gender dimension to zero.

[4]The removed names are Duke, Prince, Baron, Guy,King, Queen and Princess

Embedding vectors in the debiased space are then created for the substituted words, from the embedding vector of the substitution token. For words substituted by a gender pair, the individual word vectors are created from the gender pair vector with the value in the gender dimension set to $+\epsilon$ for the male word and to $-\epsilon$ for the female word. vectors for substituted names are created from the vector for the _name_ token. The value in the gender dimension is set to a value of ϵ multiplied by the gender specificity percentage calculated in equation 1.

The original and debiased embedding vectors are now ready to be used in the experiments.

V. Experimentation

In all experiments the gender dimension value ϵ is to 0.1. Both the original and debiased vectors are normalised for all experiments.

A. Direct bias

These 4 experiments are performed with the debiased embeddings and demonstrate the assertions that direct has been removed.

1) This test checks that all gender neutral words are unbiased. This is calculated as the sum of the absolute projection of all the gender neutral words (V_{GN}) onto the $\vec{he}$-$\vec{she}$ direction,
$$\sum_{w \in V_{GN}} \mid \vec{w}.(\vec{he} - \vec{she}) \mid$$
and is expected to be 0.
2) This test checks that gender neutral words are equidistant to both words in each gender pair. It calculates the difference between the distance to the male word and the distance to the female word, and sums this value for all gender neutral words (V_{GN}) and gender pairs (GP),
$$\sum_{w \in V_G N} \sum_{p \in GP} \|\vec{w} - \vec{p_m}\|_2 - \|\vec{w} - \vec{p_f}\|_2$$
where p_m is the male word in p and p_f is the female word in p. Again this is expected to be 0.
3) This test checks that appropriate analogies are present for the words in gender pairs. It uses the gensim (Řehůřek 2009) (version 3.8.3) 'most similar' function[5] to check that for each gender pair combination, p_1 and p_2, the question 'p_{1m} is to p_{1f} as p_{2m} is to w?' returns w=p_{2f} where p_{1m} and p_{2m} are the male words in the pairs and p_{1f} and p_{2f} are the female words in the pairs, e.g. for the pairs man_woman and king_queen, the question 'man is to king as woman is to w?' returns w=queen.
4) This test checks that the specific analogy 'man is to surgeon as woman is to w?' returns w=surgeon. Gensim is used to find the most similar word, and it's cosine similarity. Since gensim will not return any of the 3 input words, the cosine similarity of the these is also calculated. The word with the overall highest cosine similarity is deemed to be the best answer to the question. i.e. this test determines $\arg\max_{w \in V}(\vec{w}.(\vec{surgeon} - \vec{man} + \vec{woman}))$ where V is all words in the embedding space.

[5]Gensim uses cosine similarity to determine similarity, the word with the highest cosine similarity being the most similar. For normalised vectors this is equivalent to finding the nearest vector in Euclidean space.

B. Indirect bias

Gonen and Goldberg (2019) proposed 5 experiments to measure indirect bias, for which the code is supplied by the authors. They firstly reduce the embedding space to the 50,000 most common words, and from that remove the gender specific words used by Bolukbasi et al. (2016) and Zhao et al. (2018). The approach taken here is similar, but modified to remove the gender explicit words that have been substituted in the corpus prior to running the GloVe model, i.e. the gender pairs (which is a subset of their list), and the given names that have been substituted. This leaves only those words created by the GloVe model without intervention, and that are considered to be gender neutral, to be used in the experiments.

In all experiments the male/female gender bias of a word is defined as the projection of the word onto the $\vec{he}$-$\vec{she}$ direction in the original embedding space only (the same projection in the debiased space is now 0 for all gender neutral words).

1) Correlation between bias-by-projection and bias-by-neighbours: Implicitly gendered words, such as nurse or warrior, will no longer show direct bias in the debiased embedding space. This experiment suggests a measure of the indirect bias of a word as the correlation between the male/female bias of a word and the number of similarly biased words amongst it's nearest neighbours. Lower correlation will indicate less indirect bias.
2) Clustering: This experiment looks at how biased male and female words cluster together. It takes the 500 most biased male words and the 500 most biased female words, and performs k-means clustering (k=2), and then calculates the prediction accuracy of the clusters. The lower the accuracy[6], the more merged the male and female words have become, reducing indirect bias.
3) Professions: This experiment calculates the correlation between the gender bias of gender stereotypical professions, and the gender bias of the nearest 100 neighbours of the profession. The list of professions is taken from Bolukbasi et al. (2016). Less correlation indicates less indirect bias.
4) Classification: This experiment determines how well biased male and female words can be separated by an RBF-kernel SVM. It uses 5,000 words made up from the 2,500 most biased male words and the 2,500 most biases female word. It randomly takes 1,000 words to train an the classifier (500 from each gender) and then calculates the gender prediction accuracy on the remaining 4,000 words. Lower accuracy indicates less separation of the words and thus less indirect bias.

[6]This cannot be below 50% for 2 clusters

5) Association: This experiment replicates the gender related association experiments from Caliskan et al. (2017), but uses names as the gender identifier rather than gendered words (e.g. girl, her, brother). The experiments evaluate the association between male and female names, and 3 pairs of concepts that are considered to be gender biased, namely family and career, arts and maths, and arts and science. The experiments calculate the p-value. They do this by calculating the bias as the average absolute cosine similarity of the female names and female concepts, and male names and male concepts, and then calculating the same value for all combinations of names, and reports the percentage of times the combination of names has a higher bias than the original bias. The larger the p-value, the less likely there is an association between the names and concepts. Gonen and Goldberg (2019) use the terms 'art' and 'symphony' in experiments 2 and 3 as female concepts. These are also respectively a male and female name manipulated in the names processing, and so have been changed to 'theatre' and 'music' in experiment 2 and 'painting' and 'classics' in experiment 3.

   In addition Hall Maudslay et al. (2019) proposed adaptations to two of the Gonen and Goldberg (2019) experiments to measure indirect bias. The code for these experiments had to be modified to fit into the experimental framework provided by the Gonen and Goldberg (2019).
6) V-measure: This experiment reproduces the clustering experiment of Gonen and Goldberg (2019) with two variations.
   Firstly a different gender dimension is used to calculate bias in the original embedding space. It is defined as the first principal component of the subspace spanned by the difference between the word embeddings and the pair mean for each of the 23 pairs of words in the Google Analogy family test subset (GAF).
   $$\{p_m - \tfrac{p_m+p_f}{2}, p_f - \tfrac{p_m+p_f}{2} | p_m, p_f \epsilon p, p \epsilon GAF\}$$
   where p_m, p_f are the male and female words in the gender pair. And secondly tSNE (van der Maaten and Hinton (2008)) is done prior to the clustering.
7) Classification: This experiment reproduces the classification experiment of Gonen and Goldberg (2018) but uses the same definition of gender direction and bias as used in the V-measure experiment.

VI. Results

The set of experiments were run for all three datasets 500A, 500B and 500C. In addition experiments were run for two combinations of these datasets, 1000AB and 1000AC, both consisting of approx. 1 billion tokens (500A and 500B combined, and 500A and 500C combined respectively).

Results for the Gonen and Goldberg (2019) experiments are given along with those from their paper. Using their convention the results for Bolukbasi et al. (2016) are referred to as HARD-DEBIASED and for Zhao et al. (2018) as GN-GLOVE. Their results are based on word embeddings obtained from different datasets and thus the results are not directly comparable.

Results for the two Hall Maudslay et al. (2019) experiments are given along with the results for Counterfactual Data Substitution with Names Interventions (nCDS), and Counterfactual Data Augmentation (Lu et al. 2018) with Names Intervention (nCDA) from their paper as these were the two best performing techniques. Again these results are obtained from different datasets, albeit from Wikipedia.

A. Direct bias

These results apply to all datasets.

1) The projection of the gender neutral words onto $\vec{he} - \vec{she}$ is shown to be 0 as expected.
2) The gender neutral words are shown to be equidistant to the male and female words in the gender pairs as expected. The total difference is 0.
3) All combinations of gender pairs are shown to exhibit correct gender associations.
4) The test returns the word surgeon as the answer. As an example, gensim gives businesswoman[7] as the most similar word with a cosine similarity of 0.705, and the three input words have similarities of; woman 0.283, man 0.283, surgeon 1.000

Thus all four criteria are met demonstrating that direct bias has been removed.

B. Indirect bias

1) Correlation: The results show the Pearson correlation between the male bias of a word and the number of male biased words in it's 100 nearest neighbours.

	Before	After	Change
HARD-DEBIASED	0.741	0.686	-7.4%
GN-GLOVE	0.773	0.736	-4.8%
500A	0.729	0.680	-6.7%
500B	0.707	0.658	-6.9%
500C	0.714	0.664	-7.0%
1000AB	0.711	0.679	-4.5%
1000AC	0.703	0.672	-4.4%

   The results are consistent with those for HARD-DEBIASED and GN-GLOVE and show that this form of indirect bias still remains.

2) Clustering: The results show the cluster prediction accuracy.[8] 500A, 500C and 1000AC show a significant improvement over HARD-DEBIASED and GN-GLOVE, showing that there has been a siginifcant reduction of indirect bias. However, 1000AC does not show any improvement over the two smaller datasets, suggesting a limit may have been reached.

[7]Since gensim cannot return surgeon, this may be considered a more acceptable answer than nurse

[8]As there are two clusters, the accuracy cannot be below 50%

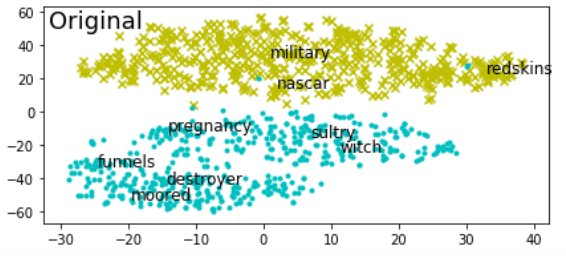

Fig. 2. Original clustering for the 500C dataset. Yellow represents the male words and cyan the female words.

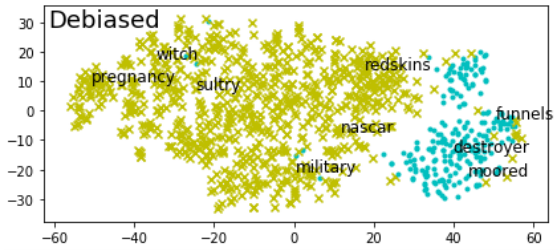

Fig. 3. Debiased clustering for the 500C dataset. The nautical words have formed a separate cluster (cyan) and the remaining female words have been incorporated into a single cluster with the male words.

	Before	After	Change
HARD-DEBIASED	0.999	0.925	-7.4%
GN-GLOVE	1.0	0.856	-14.4%
500A	1.0	0.695	-30.5%
500B	1.0	0.988	-1.2%
500C	1.0	0.727	-27.3%
1000AB	1.0	0.991	-0.9%
1000AC	1.0	0.705	-29.5%

500B and 1000AB perform poorly. Analysis shows that the most biased female words in 500A, 500C and 1000AC include a wide variety of nautical terms (e.g. destroyer, funnels, sank, torpedoed and drydock). This is presumably since ships are referred to with female pronouns. By substituting these female pronouns with gender neutral pairs, this association would be broken and it is observed that the nautical terms separate sufficiently from the other most biased female words to allow one of these groups to be incorporated into the male cluster. 500B and 1000AB do not exhibit this, so this could simply be an anomaly of the data. Fig 2 and Fig 3 shows this effect for 500C.

3) Professions: The results show the Pearson correlation between the male bias of a profession and the number of male biased words in it's 100 nearest neighbours.

	Before	After	Change
HARD-DEBIASED	0.747	0.606	-18.9%
GN-GLOVE	0.820	0.792	-3.4%
500A	0.817	0.788	-3.5%
500B	0.783	0.753	-3.8%
500C	0.794	0.745	-6.2%
1000AB	0.796	0.766	-3.8%
1000AC	0.766	0.720	-6.0%

500A and 500B perform similarly to GN-GLOVE, whereas 500C and 1000AC show some improvement but still well short of HARD-BIASED. This bias still remains.

4) Classification: The results show the prediction accuracy of the SVM classifier.

	Before	After	Change
HARD-DEBIASED	0.983	0.889	-9.6%
GN-GLOVE	0.987	0.965	-2.2%
500A	1.0	0.986	-1.4%
500B	1.0	0.979	-2.1%
500C	1.0	0.979	-2.1%
1000AB	1.0	0.990	-1.0%
1000AC	1.0	0.988	-1.2%

All datasets perform similarly to GN-GLOVE, but still short of HARD-BIASED. The bias still remains.

5) Association: The results show the p-values using the list of names supplied by Gonen and Goldberg (2019).

	Family-Career	Arts-Maths	Arts-Science
HARD-DEBIASED	0.0	0.0	0.047
GN-GLOVE	0.0	0.0	0.006
500A	0.524	0.476	0.476
500B	0.524	0.476	0.476
500C	0.524	0.476	0.476
1000AB	0.524	0.476	0.476
1000AC	0.524	0.476	0.476

The identical results are initally surprising. The word embeddings for the names have been created from the embedding of the _name_ token, based on gender specificity. This will result in the name embeddings having the same relative position in the gender dimension, and thus the projection of a word onto names will give the same relative (not absolute) values, in all experiments. But the experiment is also dependent on the size of the differences between the projections. The method in which names are created means that when they are normalised the values in all other dimensions change very slightly, and so the differences in the projection of a word onto names are very small ($<$1e-06). This makes this experiment very sensitive to the names used.
For example changing one of the female names from Joan to Karen has little effect on GN-GLOVE (0.0, 0.0

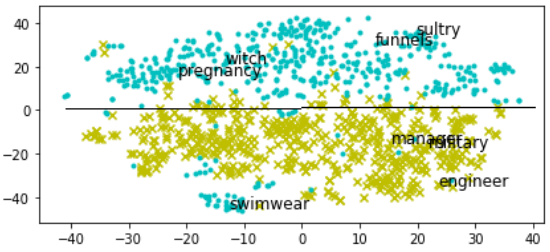

Fig. 4. tSNE mapping of the 500 most biased male and female words from which the V-measure is calculated for dataset 500B. The line is where the clusters will separate and thus shows where misclassification will occur

and 0.014)[9], but markedly (and consistently) changes the other results to (0.039, 0.960, 0.960). So rather than using a fixed set of names, a better approach is to use random samples of male and female names and average the results, and on dataset 500A for 500 iterations this gave results of 0.352, 0.671 and 0.671. Very similar results were seen for the other datasets. This suggests that this form of indirect bias has been removed.

6) V-measure:

	Before	After	Change
nCDA	1.0	0.594	-40.6%
nCDS	1.0	0.609	-39.1%
500A	0.963	0.262	-72.8%
500B	0.928	0.567	-38.9%
500C	0.940	0.354	-62.3%
1000AB	0.955	0.446	-53.3%
1000AC	0.903	0.683	-24.4%

There is an excellent reduction in cluster purity, with some datasets performing better than nCDA and nCDS[10]. In this experiment tSNE is performed before the kMeans clustering[11] However tSNE is highly sensitive to the order of input data such that if the order of the embeddings input to tSNE is switched, the V-measures change to: 500A-0.786, 500B-0.283, 500C-0.568, 1000AB-0.719, 1000AC-0.714. If tSNE is to be done first, it would be better to average over a large number of runs, but based on the results above the variance may be high.

Alternatively it may be preferable to perform kMeans without the prior tSNE, in which case the V-measures are: 500A-0.827, 500B-0.827, 500C-0.299, 1000AB-0.871, 1000AC-0.829.

7) Classification:

[9]It was not possible to rerun the HARD-DEBIASED test due to system constraints

[10]The results shown here are from a dataset taken from Wikipedia as this is considered more appropriate. nCDS did achieve a higher reduction of 58% on a dataset from Gigaword

[11]Hall Maudslay et al. (2019) state that 'For each biased embedding we then project these words into 2D space with tSNE (van der Maaten and Hinton (2008)), compute clusters with k-means, and calculate the clusters' V-measure (Rosenberg and Hirschberg, 2007).' The supplied code is consistent with this.

	Before	After	Change
nCDA	1.0	0.944	-5.6%
nCDS	1.0	0.889	-11.1%
500A	1.0	0.959	-4.1%
500B	1.0	0.957	-4.3%
500C	1.0	0.959	-4.1%
1000AB	1.0	0.963	-3.7%
1000AC	1.0	0.971	-2.9%

The results are slightly better than those from the previous classification experiment but not as good as nCDA and nCDS.

VII. Discussion

A. Results summary

This new approach to debiasing has, by design, removed direct bias, given a definition of the gender direction as $\vec{he} - \vec{she}$, Since, for all gender pairs, $\vec{p_1} - \vec{p_2}$ only has a value in the gender dimension, this result extends to all gender pairs, and any combination thereof.

It has also reduced clustering purity in both experiments. In the first experiment the results are mixed. Three datasets perform better than HARD-DEBIASED and GN-GLOVE, whereas the other two have hardly any effect. This could be related to the number of nautical terms present in the datasets and would be better assessed using a much larger and generalised dataset. With all five datasets it was observed that the cluster centers do move closer together, by up to 20%, showing that there is an degree of convergence of the most biased words.

The results of the second experiment are difficult to interpret given the sensitivity of tSNE. The results obtained without running tSNE first seem more reliable, and apart from the 500C figure, seem very consistent, although not performing as well as some of the datasets in the first experiment. In this experiment the definition of the gender direction is now more general than just $\vec{he} - \vec{she}$ and there are far fewer nautical terms in the most biased female words suggesting a more general degree of merging of the male and female words has occurred.

The method used to create name embeddings has also removed the connection between names and male/female concepts in the association experiments.

The results of the correlation experiment are similar to HARD-DEBIASED and GN-GLOVE. There is a small improvement over GN-GLOVE in the professions experiment, but well short of HARD-DEBIASED, and in the classification experiments the results are below that of HARD-DEBIASED, and nCDA and nCDS. It must be remembered that the results for HARD-DEBIASED, GN-GLOVE, nCDA and nCDS were produced on different datasets and so are not necessarily directly comparable.

So whilst this technique has performed reasonably well in removing indirect bias it generally has not been able to achieve superior performance to other techniques and much indirect bias still remains.

B. Methodology issues

All words that are not treated as gender pairs or names will be orthogonal to the gender direction and appear as gender neutral in the embedding space whilst clearly not all being so. This may have an effect on downstream applications.

Hall Maudslay et al. (2019) raise problems and limitations associated with an approach based on gender pairs. These include different spelling (mum v mom) and one-to-many associations (her v his and him, ladies v gentlemen and lords). In this paper, dad has been paired with mom, her with his, and ladies with lords since mom, his and lords are more frequent in the corpus but this approach is not wholly satisfactory.

There is also the issue of anatomical differences and related words. These can all imply gender (e.g. that person is pregnant) but do not pair off exactly (ovarian cancer is not the exact female equivalent of prostate cancer). Due to co-occurrence, these terms may well play a part in the gender separation of the embedding space.

C. Names polysemy

Many names are also words. Of the 7082 names substituted, 516 can be found in the US dictionary in the python enchant package (version 2.0.0). If substitution is extended to all 100,000+ names then 2,643 can be found in the dictionary. Substituting all 100,000+ names with the _name_ token reduces the quality of the embeddings as shown by the lower accuracy in the evaluation tests that follow the the creation of the embeddings by GloVe. Due to this, and since most names are used infrequently, only names with a total usage count count over 2,000 were selected, and it was seen that there was minimal impact on the evaluation accuracy. However this still leaves 516 words that will have an debiased embedding that is solely representative of the name, and other meanings will be lost, and this may reduce the performance of the embeddings in downstream applications. In the original embedding space these words will have an embedding that is a combination of the word and the name.

This limitation on the number of names that can be replaced without degrading the word embeddings means that there will be untreated names that can contribute to the gender separation of the embedding space, and additionally exhibit gender bias e.g. in the association experiment.

D. Experimental issues

The data preprocessing used here, together with the minimum vocabulary count in GloVe set to the default 5, created vocabularies with over 600K types (unique tokens) for datasets 500A and 500C, over 700K for 500B and over 1.0M for 1000AB and 1000AC. These vocabularies are too large. A separate experiment test run was done on dataset 500C with the minimum vocabulary count set to 10, which reduced the vocabulary size to 414K but performance was similar to the previous experiments except for the professions experiment which had a slightly larger reduction of -8.06%.

E. Observations

In the approach taken here, gender explicit terms are replaced with gender neutral terms, with the intention that this will give biased words a common reference in the cooccurrence matrix of Glove, i.e. whereas a female biased word would cooccur with 'she' and a male biased word with 'he', they would instead both cooccur with 'he_she', and the convergence of the clusters suggests that this is having the desired, although insufficient, effect. There are two points to make here. Firstly there are only 77 gender pairs and one name token that have been created to give this common cooccurrence, and a significant number of the gender pairs may not occur frequently in the training corpus. And secondly, the higher counts of cooccurrence of the gender pairs and especially the _name_ token will be penalised by the GloVe model as it both takes the log of the cooccurrence count and has a weighting function to limit the effect of frequent cooccurrences set at a value of 100 (Pennington, Socher, and Manning 2014). This may limit the effectiveness of this common cooccurrence to reduce indirect bias.

The approach taken here, and in the work of Bolukbasi et al. (2016), Zhao et al. (2018), Lu et al. (2018) and Hall Maudslay et al. (2019), is to focus on gender explicit terms and the removal of direct bias. However it may be that this approach is simply insufficient to address indirect bias.

It is interesting to note that the HARD-DEBIASED embeddings used by Gonen and Goldberg (2018), and which perform best in 3 of the experiments, are created using the Word2Vec model, whereas all other embeddings are created using GloVe. It may be worth creating HARD-DEBIASED embeddings from GloVe embeddings for a better comparison.

VIII. Further Work

The datasets used here are relatively small, so it is necessary to obtain results from a much larger dataset, and to produce results for the other methodologies on that dataset so that direct comparisons can be made. It would also be necessary to reduce the size of the vocabularies used, either through improved preprocessing or use of the minimum vocabulary count setting in GloVe.

Evaluation of the debiased embeddings in gender sensitive downstream applications should be investigated. This could include applications such as sentiment analysis on named individuals, coreference resolution, or CV and application form processing.

Modification of the GloVe cost function to allow for a greater contribution from the gender pair and _name_ tokens may be justified in that these terms represent more than a single word/name, and could lead to reduced indirect bias.

A solution to the issue of name polysemy needs to be found.

IX. Acknowledgements

I would like to thank Rowan Hall Maudslay for kindly providing the code for the experiments in Hall Maudslay et al. (2019).

References

Bolukbasi, Tolga et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186. ISSN: 0036-8075. DOI: 10.1126/science.aal4230. eprint: https://science.sciencemag.org/content/356/6334/183.full.pdf. URL: https://science.sciencemag.org/content/356/6334/183.

Gonen, Hila and Yoav Goldberg (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614. DOI: 10.18653/v1/N19-1061. URL: https://aclanthology.org/N19-1061.

Hall Maudslay, Rowan et al. (Nov. 2019). "It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5267–5275. DOI: 10.18653/v1/D19-1530. URL: https://aclanthology.org/D19-1530.

Lu, Kaiji et al. (2018). "Gender Bias in Neural Natural Language Processing". In: *CoRR* abs/1807.11714. arXiv: 1807.11714. URL: http://arxiv.org/abs/1807.11714.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., pp. 3111–3119.

Pennington, Jeffrey (2014). *GloVe: Global Vectors for Word Representation*. URL: https://nlp.stanford.edu/projects/glove/ (visited on 04/30/2021).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Řehůřek, Radim (2009). *Topic Modelling for Humans*. URL: https://radimrehurek.com/gensim/ (visited on 03/15/2021).

van der Maaten, L.J.P. and G.E. Hinton (2008). "Visualizing High-Dimensional Data Using t-SNE". English. In: *Journal of Machine Learning Research* 9.nov. Pagination: 27, pp. 2579–2605. ISSN: 1532-4435.

Zhao, Jieyu et al. (Oct. 2018). "Learning Gender-Neutral Word Embeddings". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4847–4853. DOI: 10.18653/v1/D18-1521. URL: https://aclanthology.org/D18-1521.

Appendix A: Data Sets

Dataset 500A:

Downloaded from https://dumps.wikimedia.org/enwiki/20210601/ [12]

enwiki-20210601-pages-articles1.xml-p1p41242.bz2
enwiki-20210601-pages-articles10.xml-p4045403p5399366.bz2
enwiki-20210601-pages-articles12.xml-p7054860p8554859.bz2
enwiki-20210601-pages-articles13.xml-p10672789p11659682.bz2
enwiki-20210601-pages-articles14.xml-p11659683p13159682.bz2
enwiki-20210601-pages-articles15.xml-p14324603p15824602.bz2
enwiki-20210601-pages-articles16.xml-p17460153p18960152.bz2
enwiki-20210601-pages-articles17.xml-p20570393p22070392.bz2
enwiki-20210601-pages-articles18.xml-p23716198p25216197.bz2
enwiki-20210601-pages-articles19.xml-p27121851p28621850.bz2

Dataset 500B:

Downloaded from https://dumps.wikimedia.org/enwiki/20210701/ and https://dumps.wikimedia.org/enwiki/20210720/ [12]

enwiki-20210701-pages-articles20.xml-p32808443p34308442.bz2
enwiki-20210701-pages-articles20.xml-p34308443p35522432.bz2
enwiki-20210701-pages-articles21.xml-p35522433p37022432.bz2
enwiki-20210701-pages-articles21.xml-p37022433p38522432.bz2
enwiki-20210701-pages-articles21.xml-p38522433p39996245.bz2
enwiki-20210701-pages-articles22.xml-p39996246p41496245.bz2
enwiki-20210701-pages-articles22.xml-p41496246p42996245.bz2
enwiki-20210701-pages-articles22.xml-p42996246p44496245.bz2
enwiki-20210701-pages-articles22.xml-p44496246p44788941.bz2
enwiki-20210701-pages-articles23.xml-p44788942p46288941.bz2
enwiki-20210701-pages-articles23.xml-p46288942p47788941.bz2
enwiki-20210720-pages-articles23.xml-p47788942p49288941.bz2
enwiki-20210720-pages-articles23.xml-p49288942p50564553.bz2
enwiki-20210720-pages-articles24.xml-p50564554p52064553.bz2
enwiki-20210720-pages-articles24.xml-p52064554p53564553.bz2
enwiki-20210720-pages-articles24.xml-p53564554p55064553.bz2
enwiki-20210720-pages-articles24.xml-p55064554p56564553.bz2

Dataset 500C:

Downloaded from https://dumps.wikimedia.org/enwiki/20210701/ and https://dumps.wikimedia.org/enwiki/20210720/ [12]

enwiki-20210701-pages-articles27.xml-p66975910p68108549.bz2
enwiki-20210720-pages-articles26.xml-p62585851p63975909.bz2
enwiki-20210720-pages-articles27.xml-p63975910p65475909.bz2
enwiki-20210720-pages-articles27.xml-p65475910p66975909.bz2
enwiki-20210720-pages-articles27.xml-p66975910p68286200.bz2
enwiki-20210720-pages-articles3.xml-p151574p311329.bz2
enwiki-20210720-pages-articles4.xml-p311330p558391.bz2
enwiki-20210720-pages-articles5.xml-p558392p958045.bz2
enwiki-20210720-pages-articles6.xml-p958046p1483661.bz2
enwiki-20210720-pages-articles7.xml-p1483662p2134111.bz2

[12] All links correct as of 31st July 2021

Appendix B: Gender pairs

From Bolukbasi et al. (2016b) definitional pairs

man	woman
boy	girl
he	she
father	mother
son	daughter
guy	gal
male	female
his	her
himself	herself

From Bolukbasi et al. (2016b) equalisation pairs

monastery	convent
spokesman	spokeswoman
monk	nun
dad	mom
men	women
councilman	councilwoman
grandpa	grandma
grandsons	granddaughters
uncle	aunt
husbands	wives
husband	wife
boys	girls
brother	sister
brothers	sisters
businessman	businesswoman
chairman	chairwoman
congressman	congresswoman
dads	mums
boyfriend	girlfriend
fatherhood	motherhood
fathers	mothers
fraternity	sorority
lord	lady
lords	ladies
grandfather	grandmother
grandson	granddaughter
king	queen
males	females
nephew	niece
prince	princess
schoolboy	schoolgirl
sons	daughters

Additional from Zhao et al. (2018)

countryman	countrywoman
actor	actress
bachelor	bachelorette
papa	mama
governor	governess
sir	madam
househusband	housewife
god	godess
groom	bride
emperor	emperess
landlord	landlady
duke	duchess
fiance	fiancee
stepfather	stepmother
policeman	policewoman
paternity	maternity
masseur	masseuse
mr	mrs
headmaster	headmistress
czar	czarina
stepson	stepdaughter
homosexual	lesbian
waiter	waitress
heir	heiress
monks	nuns
hero	heroine
abbot	abbess
widower	widow
baron	baroness
host	hostess
godfather	godmother
priest	priestess
patriarch	matriarch
actors	actresses
paternal	maternal
kings	queens