Word2Sexism: A Case Study of Gender Bias in Machine Learning and Methods for Debiasing Word Embeddings

Aditya Singh
Boston University,
College of Engineering
8 St. Mary St., Boston MA
Aditya28@bu.edu

Nidhi Tiwari
Boston University,
College of Engineering
8 St. Mary St., Boston MA
Nidhit@bu.edu

Frank Tranghese
Boston University,
College of Engineering
8 St. Mary St., Boston MA
Ftrang88@bu.edu

Abstract

Recent publications have emphasized the need to address the issue of socially biased datasets when training machine learning algorithms. This issue is especially concerning when considering methods like Word Embedding, which is used in applications such as resume parsing and web searching algorithms, where social biases can have a direct impact on people's lives. Here, we analyze several proposed methods for mitigating the effects of socially biased data on machine learning algorithms, focusing on gender biases and stereotypes. We will compare the methods' ability to mitigate direct and indirect gender bias, while maintaining semantic meaning of gender-specific terms (e.g. brother-sister), using quantification methods proposed by Bolukbasi et al. [1]

1. Introduction

Recent news articles have highlighted that several applications of machine learning algorithms learn from the negative social biases inherent in the data [11] due to the existence of said bias in commonly used datasets [9]. To this end, we aim to implement and compare four different proposed methods for reducing the bias in word embeddings which can help to reduce potential discrimination by machine learning applications.

1.1. Word Embeddings

Word embeddings are a commonly used method of representing text in Machine Learning and Natural Language Processing algorithms such as resume parsing and web searching [3][5], where gender bias can have a large impact on a person's experience. The method converts words into vectors that represents a word's semantic meaning by capturing its frequency of appearance with other words [10]. Each word is represented as a d-dimensional, normalized unit vector, where d is the size of the entire set of words W in all documents in the training set:

$$\overline{w} \in \mathbb{R}^d \text{ s.t. } ||\overline{w}|| = 1$$

where word similarity can be captured by the dot product, or vector difference, of two words, $\overline{w}_a \bullet \overline{w}_b$ [4]. In essence, words with similar semantic meaning are represented by vectors which are closer together.

1.2. Understanding Bias in Word Embeddings

The geometry of vector differences of word embeddings represents relationships between words. These relationships can also be used to quantify various biases that we observe in the data.[6]

When exploring bias with respect to gender stereotypes, there is a distinction between *gender specific* words that are gender-oriented by construction (e.g. brother–sister) and *gender neutral* words that have no implicit connection to a specific gender. It is important to maintain the meaning of gender-specific terms while mitigating the gender association of gender-neutral terms (e.g. doctor, nurse, receptionist) caused by gender bias. Additionally, it is important to distinguish between *direct bias*, when a gender-neutral term is associated directly with a gender, and *indirect bias*, which is reflected in relationships between gender-neutral words due to indirect relationships with a gender.

Quantifying gender in the embeddings: Bolukbasi *et al*. define a 'gender direction' (g) by adding the directions of gendered vector combinations, such as the $\overline{she} - \overline{he}$ and $\overline{woman} - \overline{man}$ vectors, with each other. This direction is a d-dimensional vector and will help us quantify the gender bias in word associations [1].

Quantifying Direct bias: Defining gender neutral words (N), gender direction (g) and strictness in measuring bias (c) , the direct bias from the word embedding is quantified by

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |cos(\overline{w}, g)|^c \quad [1]$$

Quantifying Indirect bias: To capture the indirect bias between two gender neutral word vectors w and v, the word w is decomposed into the gender part(w_g) and ($w_\perp$ = w - w_g) and the bias is captured as

$$\beta(w, v) = (w \bullet v - \frac{w_\perp . v_\perp}{||w_\perp||_2 \, ||v_\perp||_2}) / w \bullet v \quad [1]$$

2. Word Embedding Debiasing Methods

Outlined below are the methods of word embedding debiasing we propose to compare.

2.1. Schmidt Method

Schmidt proposes using vector rejection as a method for removing any associations a word has with gender. Vector rejection (i.e. orthogonal projection), will identify the component of word vector $\bar{x}$ that is pointing it in the direction of another vector $\bar{y}$. Removing this component, the word represented by $\bar{x}$ will no longer have any association with the word represented by vector $\bar{y}$, via the geometry of word embeddings. Schmidt proposes doing this for all gender-associated terms, thereby removing the gender subspace altogether. [8]

2.2. GloVe Method

The GloVe paper [7,12] proposes a scaling method based on the ratio of co-occurrence probabilities.. This method defines a co-occurrence matrix X, where the entries X_{ij} are the count of the times each word *j* co-occurs with word *i*. They also define $P_{ij}=P(j|i)=X_{ij}/X_i$ to encode the relationship between the words. Taking *k* as a gender neutral occupation word, we can measure its context with gender specific words. The bias is indicated when the ratio *P(k|she)/P(k|he)* is greater or less than one. It is proposed to scale the entries of the co-occurrence matrix in the following objective function by β as follows:

$$J = \sum_{i,j=1}^{V} f(\beta_{ij} X_{ij})(log(X_{ij}) - log(\beta_{ij} X_{ij}))^2$$

$$\beta_{ik} = \frac{X_{ik}+s}{X_{ik}}, \; \beta_{jk} = \frac{X_{jk}-s}{X_{jk}} \text{ here } s = \frac{X_i X_{jk} - X_j X_{ik}}{X_i + X_j}$$

s quantifies the shift in the direction of either gender.[7,12]

2.3. Hard Bias Correction

This method first requires that we identify the gender subspace. With word set W, we define subset $D_i \subset W$ as the subset of words that define the gender direction. We find the means of all D_i, with $\overline{w}$ the word embeddings in subset D_i

$$\mu_i = \sum_{w \in D_i} \overline{w} / |D_i|$$

The bias subspace is defined as the first k ≥ 1 rows of SVD(C) where

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

We *neutralize* [1] to make the gender neutral words zero in the gender subspace. For $w \varepsilon N, N \subseteq W$ are words to neutralize by:

$$\overline{w} = (\overline{w} - \overline{w}_B)^T / \| \overline{w} - \overline{w}_B \|$$

Then, for $E_i \subseteq W$ defined as the set of words outside the subspace we want to *equalize* [1] so that all neutral words are perfectly equidistant from all words in each equality set E_i

$$\mu := \sum_{w \in E} w/|E| \;, \; v := \mu - \mu_B$$

For each $w \in E$

$$\overline{w} := v + \sqrt{1 - ||v||^2} \; (\overline{w}_B - \mu_B) / \| \overline{w}_B - \mu_B \| \quad [1]$$

2.4. Soft Bias Correction

This method minimizes $T \in R^{d \times d}$ which represents the projections of gender neutral words onto the gender subspace. For W, the matrix of all embedding vectors and N, the matrix of the embedding vectors representing gender neutral words:

$$\min_T \left\| (TW)^T (TW) - W^T W \right\|_F^2 + \lambda \left\| (TN)^T (TB) \right\|_F^2$$

λ is the tuning parameter and output embedding is normalized to have a unit length [1]. Further simplifications for this equation used for solving can be found in the appendix [A.2].

3. Datasets and Testing Methods

For implementing the Schmidt Method and the Hard De-Bias method, we will be using Google's word2vec; in particular, the version pre-trained on the Google News dataset [2]. Like Bolukbasi *et al.*, we will focus on the lowercase words of 20 or less alphabetical characters, for a

total of 26,423 words. Every word vector used for the methods mentioned is 300 dimensional.

For the GloVe method, we will use the *20 newsgroups* dataset (using the top 6,000 frequency words) along with the Gensim toolbox to generate a co occurrence matrix on which the GloVe method for bias correction will be carried out. Further testing will be done using the conditional probabilities of gender neutral words with *'he'* and *'she'*. We intend to then re-embed the co occurrence matrix into word embeddings in order to carry out bias testing on the rescaled (or bias corrected) word embeddings.

We will evaluate the debiasing algorithms' effectiveness on both *direct* and *indirect bias*. Using the quantification methods described, we will be testing for the methods' abilities to mitigate direct and indirect bias on a list of gender-neutral occupational terms (e.g. doctor, nurse, receptionist, etc.) while maintaining appropriate gender association of a list of gender-specific terms (e.g. king, aunt, brother, etc.). The lists of occupations and gender-specific terms will be the same as in Bolukbasi *et al.* [1] to confirm their findings and properly compare between the different methods.

For Soft Debias method, we will use the software CVX (freely available for Matlab at http://cvxr.com/cvx/) in order to solve the minimization problem. This is a semidefinite program and mathematical details can be found in the appendix [A2]. Due to the computational complexity of solving this, we will only use the top 50 feature components associated with the top 50 singular values of the word embeddings, as opposed to using all 300 features. Additionally, due to the extensive runtime, we will use the suggested $\lambda = 0.2$ from the Bolukbasi paper [1].

We will finally compare all the methods on their ability to reduce bias while maintaining desired gender-specific associations, taking into account the algorithmic complexity.

4. Results and Discussion

4.1. Pre-debiasing Analysis of Dataset

First, we define a gender subspace. Confirming Bolukbasi *et al.* [1], we carried out principal component analysis on the difference vectors of 10 gender-defining word pairs (appendix A.1.) yielding one component (corresponding to the eigenvalue) that is much higher than the others and accounts for the majority of the variance (fig. 1). This assessment allows us to take this component as our gender subspace and use it to quantify our bias measurements.

Next, we aimed to capture the existence of bias in the dataset. Initial testing of gender-neutral occupation terms (appendix A.2, A.3), whose similarity to both genders should be equal, found a high amount of bias towards a specific gender when testing their similarity to $\overline{she}$ and $\overline{he}$ vectors (fig 2). Furthermore, quantifying both direct (table 1) and indirect bias (table 2,3) to the $\overline{football}$-$\overline{softball}$ word pairs. The comparison of word vectors with $\overline{football}$ and $\overline{softball}$ captures the indirect bias since similarity with either of the two vectors is due to gender associations with either football or softball. An example of this is seen in table 4, where the occupations $\overline{receptionist}$ and $\overline{nurse}$ have a large change in similarity with $\overline{softball}$ and $\overline{pink}$ when removing the gender direction.

	Occupations	**Gender-Specific**
Direct Bias (c=1)	0.0706	0.1737

Table 1: Measure of direct bias (average cosine similarity to gender direction) for both occupations and gender- specific words. We can see that direct bias exists since the average cosine distance for occupations is not zero.

From these values, we can see a high amount of unwanted bias (both direct and indirect) in our gender-neutral occupations. These are the basis on which we evaluate the effects of the debiasing methods.

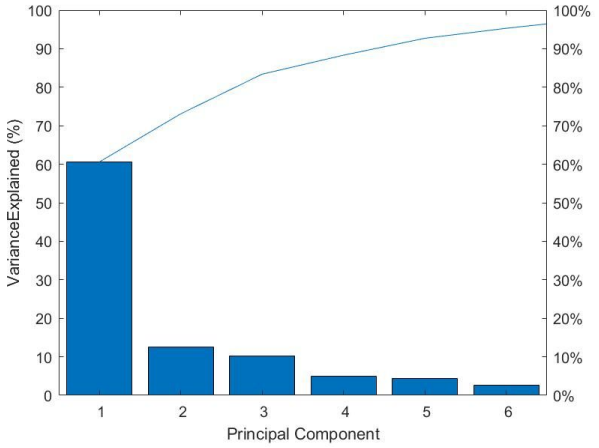

Figure 1: A bar graph of the percentage of variance explained by the principal components of the vector differences of the gender word pairs, showing the first component explaining significantly more variance than the rest (remaining four components not shown are zero-valued).

Comparison Term	Avg. Indirect Bias
Football	-0.0210
Softball	0.0624

Table 2: Average indirect bias of occupation terms when compared to $\overline{football}$ and $\overline{softball}$ vectors. Values show an overall indirect bias between occupations and football and softball due to their associations with gender.

Top Indirect Bias with Football	Top Indirect Bias with Softball
researcher	banker
interior designer	campaigner
councilor	electrician
planner	inventor
bookkeeper	doctoral-student

Table 3: List of occupations whose similarity with $\overline{football}$ and $\overline{softball}$ vectors changes the most when removing gender.

	football	softball
Nurse	-0.3976	0.3183
Scientist	0.0764	-0.2306

Table 4: Indirect bias between scientist and nurse when compared to the words softball and football. We see here the change due to removing the gender direction is significant.

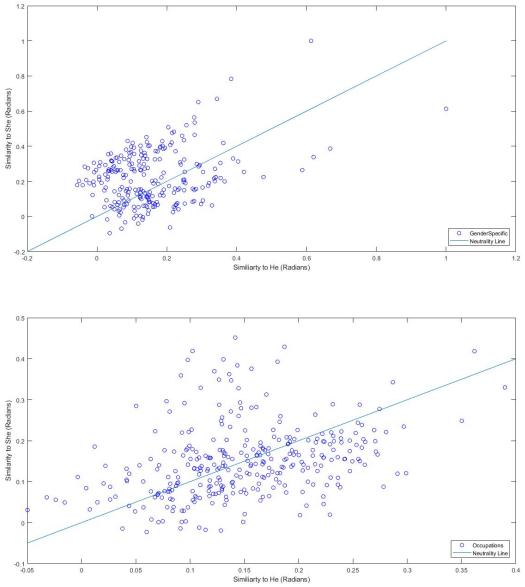

Figure 2: Graphs of occupations (bottom) and gender-specific (top) terms capturing their similarity to $\overline{she}$ and $\overline{he}$. Occupation terms should have equal similarity to both genders, however here we can see how widely the terms vary from the neutrality line.

	Occupations	Gender-Specific
Direct Bias (c=1) (before bias corr.)	0.0706	0.1552
Direct Bias (c=1) (after bias corr.)	0.0040	0.0090

Table 5: Measure of direct bias (average cosine similarity to gender direction) for both occupations and gender- specific words after Schmidt Debasing. We can see the direct bias for both Occupations and Gender-Specific are significantly decreased.

4.2. Schmidt Method Results

Since the Schmidt method involves removing the gender subspace entirely, as expected, the bias is diminished across all words (gendered and gender-neutral alike). We see a 94% decrease in direct bias for both the gender-neutral and the gender specific words (table 5). We additionally, see a marked shift in both sets towards the neutrality line on the he-she plane (Fig. 3). Additionally, the average indirect bias to football and softball have essentially become zero (Table 6).

When combined with the simplicity of the Schmidt method, we can definitely see the advantage it has for cases where gender bias should be strictly removed (i.e. resume parsing). We can see, for example, the occupations nurse and scientist are no longer affected by their indirect bias with football and softball (Table 7).. However, it does remove the association of gender with all words in the set, including the gender-specific words, where, in many cases, we would want to preserve the similarity of gendered-specific with their respective genders.

Comparison Term	Avg. Indirect Bias
Football	-1.2515 x 10^{-5}
Softball	6.5727 x 10^{-4}

Table 6: Average indirect bias of occupation terms when compared to *football* and *softball* vectors. We see a significant decrease in indirect bias for both football and softball.

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	-0.0231	0.3183	0.0057
Scientist	0.0764	-0.0029	-0.2306	-0.0344

Table 7: Indirect bias between receptionist and nurse when compared to the words softball and pink. We can not see that gender does not affect the similarity between these words now, which shows removal of indirect bias.

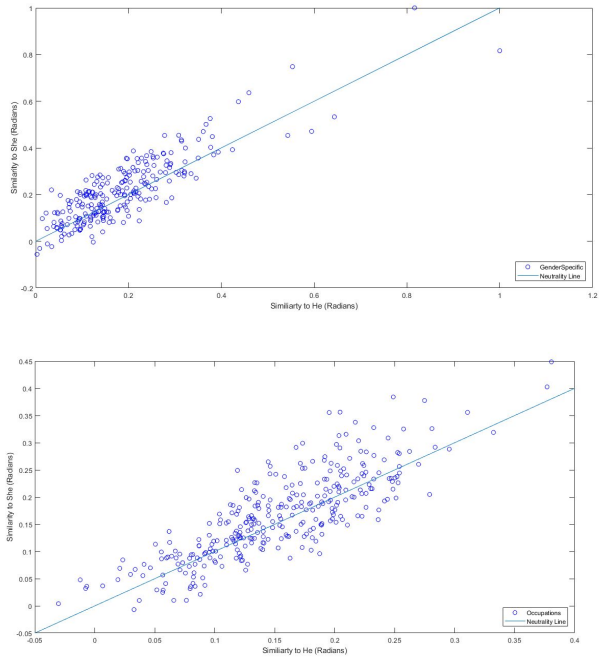

Figure 3: Graphs of occupations (bottom) and gender-specific (top) terms capturing their similarity to *she* and *he*. Here we see a shift towards neutrality line for both sets of words

4.3. Soft Bias Correction Results

Even with using only the top 50 features to solve the soft debias minimization, we see a 43% decrease in direct bias, but only see a 72% decrease in gender specific words (Table 8), meaning this method has preserved more of the semantic similarity to gendered words as desired, at the cost of not decreasing the direct bias of the occupation words as much as Schmidt. Likewise, we see the shift of indirect bias towards zero, however not nearly as much as seen in the Schmidt (Tables 9,10).

	Occupations	Gender-Specific
Direct Bias (c=1) (before bias corr.)	0.0706	0.1552
Direct Bias (c=1) (after bias corr.)	0.0396	0.0469

Table 8: Measure of direct bias (average cosine similarity to gender direction) for both occupations and gender- specific words after Soft Debasing. We can see the bias is more decreased for occupations than gender-specific, however both are significantly diminished.

Comparison Term	Avg. Indirect Bias
Football	-0.0022
Softball	-0.0269

Table 9: Average indirect bias of occupation terms when compared to *football* and *softball* vectors after soft debiasing. Again, we do see a decrease for both terms, however not a significant one.

Important to note, when looking at the plots of similarity on the he and she axis (Fig 4), we notice a drastic change in shape from both the original datasets, as well as the other methods. This may mean that that linear transform T may have affected the meanings of other words in the set with respect to each other, which was expected. Perhaps this effect would be less noticable if we were able to run on the complete set. Further testing would be needed to analyze the impact of overall change is meaning to words in the set.

Finally, we notice the examples for indirect bias are shifted appropriately in the direction towards zero (nurse's meaning shifts towards football, and away from softball, and vice-versa for scientist (table 10).

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	-0.0231	0.3183	0.0057
Scientist	0.0764	-0.0029	-0.2306	-0.0344

Table 10: Indirect bias between nurse and scientist when compared to the words softball and football. We see a shift in the meanings towards football, and away from softball.

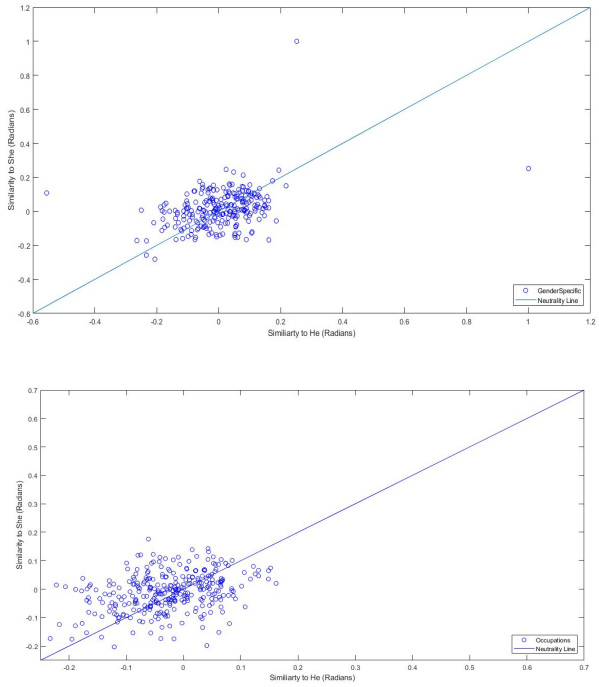

Figure 4: Graphs of occupations (bottom) and gender-specific (top) terms capturing their similarity to *she* and *he*. Here we see a shift towards neutrality line for both sets of words.

4.4. Hard De-Bias Results

The first step of the Hard De-Bias neutralises the gender neutral (occupation) words in the gender subspace. We measure the effect of the debiasing by calculating the average direct bias of all the gender neutral (occupation) words and the gender specific words. As seen in the values obtained, the direct bias for the occupation words is significantly reduced down to a fraction of its original value.

The second step for this method involves making the 'Equalise' set of words equidistant from the component of the gender neutral words in the gender subspace. The Equalise set contains word pairs like 'monastery-convent' and 'spokesman-spokeswoman'. We want to retain some gender component for these words as we want to preserve the overall utility of the word embeddings after bias correction. The direct bias for this equalise set is after Hard De-Biasing shows a stark decrease. However, it still retains significant component of bias which is what we desire.

	Occupations	**Equalise Words**
Direct Bias (c=1) (before bias corr.)	0.0706	0.1737
Direct Bias (c=1) (after bias corr.)	0.0040	0.0604

Table 11: The results for the indirect bias after hard debiasing also give us favourable results. We see that by testing for two occupation words with previously indicated male and female bias. The negative indirect bias of 'nurse' with 'football' decreases significantly and its positive indirect bias with 'softball' also reduced showing that its bias in the direction of both genders has been significantly reduced. Similar results are obtained for 'Scientist'.

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	-0.0231	0.3183	0.0057
Scientist	0.0764	-0.0029	-0.2306	-0.0344

Table 12: Indirect bias between nurse and scientist when compared to the words softball and football. We see a shift in the meanings towards football, and away from softball.

4.5 GloVe Method Results

Implementation of this method needed a co occurrence matrix which generates the count of each word occurring in context with every other word in the form of a large square matrix. We used the train set of the *20 newsgroups* dataset which we found already partitioned into a train and a test set.

Memory and computational resources constraints led us to filter the words to 6000 *most frequent* words. This led to the filtering out of a lot of useful occupation words and retention of many words redundant for the objective of our method. The resultant matrix, is very sparse, was converted into its dense form before saving into a .mat file.

We were able to locate certain useful words and test their biases using the probability measurements described in Section 2.2. Here the gender neutral word is (k). We then employed the GloVe Method to rescale the word (to

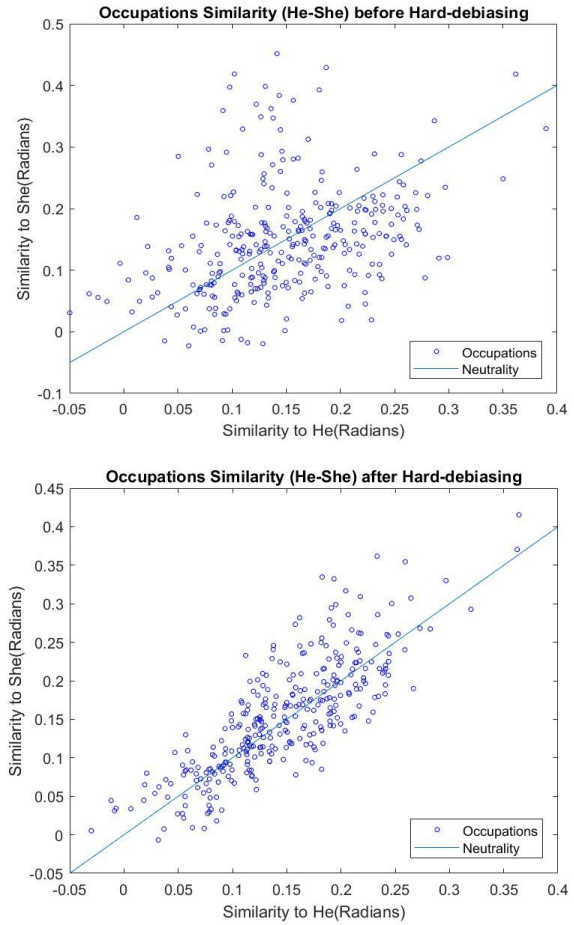

Figure 5: Graphs of occupations terms capturing their similarity to *she* and *he* before Hard De-Biasing*(top)* and after Hard De-Biasing*(bottom)*. Here we see a solid shift towards neutrality line for the occupation words. This is a desirable change.

We were able to locate certain useful words and test their biases using the probability measurements described in Section 2.2. Here the gender neutral word is (k). We then employed the GloVe Method to rescale the word (to correct the bias) and on testing the probabilities after this step we find that the values of $P(k|he)/P(k|she) \approx 1$. This shows that rescaling has resulted bias correction.

k	***P(k|he)/P(k|she)* before scaling**	***P(k|he)/P(k|she)* after scaling**
Scientist	2.757	1.000
Laboratory	4.4311	1.003

Table 13: The conditional probabilities as calculated above show bias in direction of the gender in the numerator when greater than 1. Before scaling, the probability values show considerable bias.Results after scaling with the GloVe method, the probability values are scaled down to one which we interpret as evidence of reduction of bias.

The next step was to take the co-occurrence matrix and the matrix rescaled for certain words and re-embed the matrix to get word vectors using the GloVe word2vec model. Further we intended to carry out all the bias analysis (direct & indirect) for the rescaled word vectors. To speed up the process we decided to restrict the vector dimension to 30. For more information please see Section 5 (Conclusions & Future Work).

5. Conclusions and Future Work

We can conclude from the results that the Hard Debias method offers the best mitigation of gender bias for gender neutral words while maintaining the gender associations of the gender-specific words. It also accomplishes this task efficiently [Table 14]. This all being said, the Schmidt method offers the complete removal of gender bias from all words relatively efficiently and would be useful in situations where gender bias should be completely removed.

Method	**Runtime (Billions of Cycles)**
Schmidt	2.978
Soft Debias	52,200
Hard Debias	0.0352
GloVe	1.700

Table 14: The CPU time (in billions of cycles) to run all four methods. We have excluded time to get gender direction or training the embeddings as these steps are common to all methods. We see that the Hard Debias is the most efficient.

Due to insufficient computational resources and time constraints, we were unable to properly test both the Soft Debias and GloVe methods. While we were able to train a

word2vec using smaller feature dimensions (30 as opposed to 300), we did not have the time to test the Soft Debias method to determine if performing minimization on the total feature vectors of a word set yields better performance. Though, it is important to note that Bolukbasi *et al* [1] found this method to be inferior to the Hard Debias method. Additionally, while we were able to create a co-occurrence matrix and perform the pre-processing method, we did not have the time to train a word embedding set off of this co-occurrence matrix, so we cannot confirm the GloVe method's effectiveness in removing gender bias. Due to the small size of the smaller word embedding set, many words in our original set of occupations and gender-specific terms did not exist, and many entries of the co-occurrence matrix were zero, so it is unclear whether performing the methods on the smaller set would even yield meaningful results. Overall, given more resources, both the Soft Debias and GloVe method deserve further exploration.

6. Division of Labor

Our first step is to familiarize ourselves with Google's word2vec NEWS embedding, understand it's geometry in order to use it effectively in bias quantification. We will then demonstrate and discuss the direct and indirect bias in word embeddings.

Then, we will implement the debiasing algorithms that we have listed. As a stretch goal, we aim to offer an improvement to the methods we have cited.

We will test the results of the algorithms on direct, indirect bias and their ability to maintain gender-specific meanings. We will analyse these results and discuss how the methods helped mitigate bias in word embeddings.

6.1. Pre-Debiasing Methods

- Importing and using Google News word2vec - All
- Code Direct Bias Quantification Method - Frank
- Code Indirect Bias Quantification Method - Frank
- Code to find Gender Subspace - Nidhi
- Analysis of pre-debiased data - All

6.2. Testing / Evaluating Methods

- Coding and Evaluating Schmidt Method-Frank
- Data Preprocessing - Aditya
- Training co-occurrence matrix and word2vec on 20 newsgroups dataset(30 dimensions) - Aditya
- Coding and Evaluating GloVe Method- Nidhi
- Code and Evaluating Soft Biasing Method - Frank
- Coding and Evaluating Hard Biasing Method-Nidhi

6.3. Discussion

- Each person will provide analysis and discussion on their assigned methods.
- Overall analysis will be done together.
- Stretch goal: Provide our own improved method will also be done together.

References

[1] T. Bolukbasi, K.W. Chang, J. Zhou, V. Saligrama, and A. Kalai. "Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings," *2016*. Advanced in Neural Information Processing Systems, pg 4349-4357.

[2] Google, Inc. 'word2vec', *2013*. [Online]. Available: https://code.google.com/archive/p/word2vec/ [Accessed March 25, 2018].

[3] C. Hansen *et al.* "How to get the best word vectors for resume parsing," in SNN Adaptive Intelligence / Symposium: Machine Learning 2015.

[4] T. Mikolov *et al.* "Linguistic regularities in continuous space word representations," *2013*, In HLT-NAACL, pg 746-751.

[5] E. Nalisnick *et al.* "Improving document ranking with dual word embeddings," In *www*, April 2016

[6] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 560–568. ACM, 2008.

[7] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation" Computer Science Department, Stanford University, Stanford, CA 94305

[8] B. Schmidt. "Rejecting the gender binary: a vector-space operation," October 30, 2015. [Online]. Available: http://bookworm.benschmidt.org [Accessed March 24, 2018.

[9] A. Torralba and A. Efros. "Unbiased look at dataset bias," *2012*, in CRPR.

[10] P.D. Turney and P. Pantel. "From Frequency to Meaning: Vector Space Models of Semantics," *2010*, Vol. 37, pg. 141-188

[11] D. Victor, "Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk." *New York Times*, March 24, 2016. [Online]. Available: http://newyorktimes.com [Accessed March 15, 2018].

[12] T. Chakraborty, G. Badie, and B. Rudder, "Reducing Gender Bias in Word Embeddings" [Online] Computer Science Department, Stanford University. Available: http://cs229.stanford.edu/proj2016/report/ [Accessed: March 15, 2018]

[13] T. Bolukbasi, K.W. Chang, J. Zhou, V. Saligrama, and A. Kalai. “Quantifying and Removing Stereotypes in Word Embeddings.” 2018. Poster presentation, Boston University, Boston MA.

A. Appendix

A.1. Vector Differences for Gender Direction

$\{\overline{she}-\overline{he},\ \overline{her}-\overline{his},\ \overline{woman}-\overline{man},\overline{girl}-\overline{boy},\ \overline{mother}-\overline{father},\ \overline{daughter}-\overline{son},\ \overline{gal}-\overline{guy},\ \overline{female}-\overline{male},\ \overline{Mary}-\overline{John}\}$

A.2. Solving Soft Debias

In unpublished work, Bolukbasi *et al.* [13] showed that the following equation is equivalent to solving the Soft Debiasing minimization problem presented earlier in this paper:

$$\min_X \|\Sigma V'(X-I)V\Sigma\|^2_F + \lambda \left\|NXB^T\right\|^2_F$$

where $X = TT^T$ and $W = U\Sigma V'$.

A.3. Gender-Neutral Occupations

'accountant','acquaintance','actor','actress','adjunct_professor','administrator','adventurer','advocate','aide','alderman','alter_ego','ambassador','analyst','anthropologist','archaeologist','archbishop','architect','artist','artiste','assassin','assistant_professor','associate_dean','associate_professor','astronaut','astronomer','athlete','athletic_director','attorney','author','baker','ballerina','ballplayer','banker','barber','baron','barrister','bartender','biologist','bishop','bodyguard','bookkeeper','boss','boxer','broadcaster','broker','bureaucrat','businessman','businesswoman','butcher','butler','cab_driver','cabbie','cameraman','campaigner','captain','cardiologist','caretaker','carpenter','cartoonist','cellist','chancellor','chaplain','character','chef','chemist','choreographer','cinematographer','citizen','civil_servant','cleric','clerk','coach','collector','colonel','columnist','comedian','comic','commander','commentator','commissioner','composer','conductor','confesses','congressman','constable','consultant','cop','correspondent','councilman','councilor','counselor','critic','crooner','crusader','curator','custodian','dad','dancer','dean','dentist','deputy','dermatologist','detective','diplomat','director','disc_jockey','doctor','doctoral_student','drug_addict','drummer','economics_professor','economist','editor','educator','electrician','employee','entertainer','entrepreneur','environmentalist','envoy','epidemiologist','evangelist','farmer','fashion_designer','fighter_pilot','filmmaker','financier','firebrand','firefighter','fireman','fisherman','footballer','foreman','freelance_writer','gangster','gardener','geologist','goalkeeper','graphic_designer','guidance_counselor','guitarist','hairdresser','handyman','headmaster','historian','hitman','homemaker','hooker','housekeeper','housewife','illustrator','industrialist','infielder','inspector','instructor','interior_designer','inventor','investigator','investment_banker','janitor','jeweler','journalist','judge','jurist','laborer','landlord','lawmaker','lawyer','lecturer','legislator','librarian','lieutenant','lifeguard','lyricist','maestro','magician','magistrate','maid','major_leaguer','manager','marksman','marshal','mathematician','mechanic','mediator','medic','midfielder','minister','missionary','mobster','monk','musician','nanny','narrator','naturalist','negotiator','neurologist','neurosurgeon','novelist','nun','nurse','observer','officer','organist','painter','paralegal','parishioner','parliamentarian','pastor','pathologist','patrolman','pediatrician','performer','pharmacist','philanthropist','philosopher','photographer','photojournalist','physician','physicist','pianist','planner','plastic_surgeon','playwright','plumber','poet','policeman','politician','pollster','preacher','president','priest','principal','prisoner','professor','professor_emeritus','programmer','promoter','proprietor','prosecutor','protagonist','protege','protester','provost','psychiatrist','psychologist','publicist','pundit','rabbi','radiologist','ranger','realtor','receptionist','registered_nurse','researcher','restaurateur','sailor','saint','salesman','saxophonist','scholar','scientist','screenwriter','sculptor','secretary','senator','sergeant','servant','serviceman','sheriff_deputy','shopkeeper','singer','singer_songwriter','skipper','socialite','sociologist','soft_spoken','soldier','solicitor','solicitor_general','soloist','sportsman','sportswriter','statesman','steward','stockbroker','strategist','student','stylist','substitute','superintendent','surgeon','surveyor','swimmer','taxi_driver','teacher','technician','teenager','therapist','trader','treasurer','trooper','trucker','trumpeter','tutor','tycoon','undersecretary','understudy','valedictorian','vice_chancellor','violinist','vocalist','waiter','waitress','warden','warrior','welder','worker','wrestler','writer'

A.4. Gender-Specific Words

'he','his','he','her','she','him','she','man','women','men','his','woman','spokesman','wife','himself','son','mother','father','chairman','daughter','husband','guy','girls','girl','her','boy','king','boys','brother','chairman','spokeswoman','female','sister','women','man','male','herself','lions','lady','brothers','dad','actress','mom','sons','girlfriend','kings','men','daughters','prince','queen','teenager','lady','bulls','boyfriend','sisters','colts','mothers','sir','king','businessman','boys','grandmother','grandfather','deer','cousin','woman','ladies','girls','father','uncle','pa','boy','councilman','mum','brothers','ma','males','girl','mom','guy','queens','congressman','dad','mother','grandson','twins','bull','queen','businessmen','wives','widow','nephew','bride','females','aunt','congressman','prostate_cancer','lesbian','chairwoman','fathers','son','moms','ladies','maiden','granddaughter','younger_brother','princess','guys','lads','ma','sons','lion','bachelor','gentleman','fraternity','bachelor','niece','lion','sister','bulls','husbands','prince','colt','salesman','bull','sisters','hers','dude','spokesman','beard','filly','actress','him','princess','brother','lesbians','councilman','actresses','viagra','gentlemen','stepfather','deer','monks','beard','uncle','ex_girlfriend','lad','sperm','daddy','testosterone','man','female','nephews','maid','daddy','mare','fiance','wife','fiancee','kings','dads','waitress','male','maternal','heroine','feminist','mama','nieces','girlfriends','councilwoman','sir','stud','mothers','mistress','lions','estranged_wife','womb','brotherhood','statesman','grandma','maternity','estrogen','ex_boyfriend','widows','gelding','diva','teenage_girls','nuns','daughter','czar','ovarian_cancer','he','monk','countrymen','grandma','teenage_girl','penis','bloke','nun','husband','brides','housewife','spokesmen','suitors','menopause','monastery','patriarch','beau','motherhood','brethren','stepmother','dude','prostate','moms','hostess','twin_brother','colt','schoolboy','eldest','brotherhood','godfather','fillies','stepson','congresswoman','chairwoman','daughters','uncles','witch','mommy','monk','viagra','paternity','suitor','chick','pa','soror

ity','macho','spokeswoman','businesswoman','eldest_son','gal','statesman','schoolgirl','fathered','goddess','hubby','mares','stepdaughter','blokes','dudes','socialite','strongman','witch','uterus','grandsons','bride','studs','mama','aunt','godfather','hens','hen','mommy','babe','estranged_husband','fathers','elder_brother','boyhood','baritone','diva','lesbian','grandmothers','grandpa','boyfriends','feminism','countryman','stallion','heiress','queens','grandpa','witches','aunts','semen','fella','granddaughters','chap','knight','widower','maiden','salesmen','convent','king','vagina','beau','babe','his','beards','handyman','twin_sister','maids','gals','housewives','gentlemen','horsemen','businessman','obstetrics','fatherhood','beauty_queen','councilwoman','princes','matriarch','colts','manly','ma','fraternities','spokesmen','pa','fellas','gentleman','councilmen','dowry','barbershop','monks','woman','fraternal','ballerina','dads','goddess','her','girlfriend','grandmother','restless','boyfriend','males','she','legs','grandfather','fraternity','majesty','lad','husbands','hen','handsome','boy','girl','sorority','nun','dad','nuns','girlfriends','father','couple','hubby','grandson','fella','dudes','lads','heir','brothers','whore','heiress','breasts','wife','laurels','fellas','niece','granddaughter','actress','tit','hunk','muscular','fiancee','uncles','princess','blokes','sister','daughter','gelding','grandmother','babe','nephews','grandsons','aunt','confessions','grandpa'

A.5. Equalise Words Set

'monastery'-'convent'
'spokesman'-'spokeswoman'
'councilman'-'councilwoman'
'grandpa'-'grandma'
'grandsons'-'granddaughters'
'prostate_cancer'-'ovarian_cancer'
'testosterone'-'estrogen'
'uncle'-'aunt'
'wives'-'husbands'
'boy'-'girl'
'boys'-'girls'
'brother'-'sister'
'brothers'-'sisters'
'businessman'-'businesswoman'
'chairman'-'chairwoman'
'colt'-'filly'
'congressman'-'congresswoman'
'dad'-'mom'
'dads'-'moms'
'dudes'-'gals'
'ex_girlfriend'-'ex_boyfriend'
'father'-'mother'
'fatherhood'-'motherhood'
'fathers'-'mothers'
'fraternity'-'sorority'
'gelding'-'mare'
'gentleman'-'lady'
'gentlemen'-'ladies'
'grandfather'-'grandmother'
'grandson'-'granddaughter'
'he'-'she'
'himself'-'herself'
'his'-'her'
'king'-'queen'
'kings'-'queens'
'male'-'female'
'males'-'females'
'man'-'woman'
'men'-'women'
'nephew'-'niece'
'prince'-'princess'
'schoolboy'-'schoolgirl'
'son'-'daughter'
'sons'-'daughters'
'twin_brother'-'twin_sister'