Final Project - Trade-off Between Debiasing and Word Embedding Quality

Farhana Faruqe*, Luming Wang*, Peiyu Wang*, Wei Guo*
*George Washington University, faruqe,lmwang,oliviawpy,weiguo@gwu.edu

***Abstract*—Word embedding is a representation for words according to word co-occurrence statistics. Embeddings are used for various types of down-stream tasks for the information that it carries which could be used to enrich NLP models. There is on going research on the topic of debiasing word embeddings, because with the bias embedded, the use of word embedding will very likely amplify the bias in downstream tasks. The focus of this project is on identifying the tradeoffs between debiasing performance and linguistic properties preservation. To test the trade off and to quantify the word embedding quality, we use intrinsic and extrinsic evaluation methods to measure the the preservation of linguistic and semantic meaning, and the extent of gender bias left after debiasing.**

***Keywords*—Word embedding, word embedding evaluation, machine learning, debiasing.**

I. Introduction

As word embeddings are attracting much attention from many computer scientists and researchers for the role it plays in Natural Language Processing tasks, the potential amplification of bias embedded in word embedding has also become a major concern in this field. Therefore, many researchers have been focusing on finding systematic approaches of debiasing word embeddings while maintaining the linguistic properties.

Some previous research showed promise by showing having bias removed from the word embeddings while sustaining quality, later research overturned the conclusions and showed that the bias removed was shallow, which could be easily recovered. We were inspired to conduct research incorporating more principal components to see if the debiasing performance would increase while the linguistic properties are properly preserved.

The specific contributions of this paper are as follows: we debiased word embeddings with top 1 to 5 principal components and tested our embeddings with intrinsic and extrinsic tasks and reported key statistic. We tried quantifying the bias level and linguistic property preservation ability of word embeddings with multi-rater scoring system in analogy generation tasks, and reported the level of agreement among raters. We also applied the methodologies proposed by [6] and [4] to our embeddings in order to intuitively visualize debiasing performance and to test embedding linguistic property preservation ability in downstream tasks.

In the end, our work showed that incorporating more principal components may not necessarily improve debiasing performance, while it could decrease the word embedding's ability to preserve useful semantic meanings.

II. Problem Statement

The objective of our project is to find out whether incorporating more principal components into the debiasing process is a better way of debiasing. However, debiasing the word embedding often comes with the deterioration of quality of word embedding when used for downstream tasks. We also want to quantify the trade-offs between incorporating more principle components in the debiasing process and the linguistic property preservation ability of the word embedding after debiasing. To do that, we plan to use multiple intrinsic and extrinsic evaluation methods.

III. Related Work

Bolukbasi et al.[3] first proposed the "Hard Debiasing" and "Soft Debiasing" approaches to debias word embedding, and showed that after debiasing, the gender bias has been significantly mitigated while the semantic meanings embedded in the word embeddings were preserved. However, Gonen and Goldberg[6]'s work used k-means clustering and showed that the debiasing methods proposed by Bolukbasi et al.[3], which leaned on the above definition for gender bias and directly targeted it, were mostly hiding the bias rather than removing it. Their research also showed that the gender bias was still reflected in the geometrical representation of the words that were supposed to be gender-neutral after debiasing. Moreover, the bias can even be recovered from the debiased word embedding. In Karve et al.[8]'s work, the idea of "conceptors debiasing" was proposed and they used "conceptors debiasing to post-process both traditional and contextualized word embeddings.". They claimed that their debiasing conceptor did a successful soft damping of the relevant principal components. However, like Gonen and Goldberg [6], they found that embedding debiasing may leave bias undetected by some measures like WEAT, and they proposed that all debiasing methods should be tested on end-tasks such as emotion classification and co-reference solution.

The above research findings motivated us to define our problem statement.

IV. Dataset

Considering our goal to explore the trade off after debiasing, we would want to use a dataset that holds a large number of biases in its corpus. In Corpus of Historical American English (COHA)[5] embeddings, the average percentage of difference of women in the same occupations increases constantly from 1910 to 1990. This makes it a good source for PC discovering because both data corpus and pre-trained word embeddings in a historical sense can be accessed.

As mentioned above, COHA is an open-source dataset with a 400-million-word data corpus from 1810s to 2000s, which includes a balanced number of fiction, popular magazine, newspaper, and non-fiction books in American English, and there are also pre-trained word2vec historical word vectors based on COHA, which can be convenient in applying our methodology on. For its balanced cover range of time of decade and genres, it can be a good source for this project. By adapting historical dataset to how Bolukbasi et al. [3] found the first PC, we will be able to acknowledge the existence of bias in other PCs. For this project, we specifically focus on data corpus from the 1990s. The reason why we select the dataset from this time is that it contains the most recent data corpus compare to other eras. It misses the least amount of words and holds word embeddings that have the closest meaning to modern semantics.

During our evaluation process, for analogy generation tasks, we applied the analogy dataset used by previous research [3][6]. These datasets [3][6] also have been used for concept categorization evaluation process. We have used Caliskan et al. [4] datset for Word-Embedding Association Test (WEAT). For sentiment classification of external evaluation, we applied data corpus from Mass et. al.[11] for testing if the debiased word embeddings can result a better performance of classification.

V. Approach

We focus on the trade-off of debiasing results between the reduction of biases from original word embedding and the preservation of desirable features.

There are primarily two steps in our method. First, we use the debiasing methods Bolukbasi et al. proposed[3] to generate multiple debiased embedding from COHA word embeddings with different number of principal components incorporated in the debiasing process. In this step, we obtain the 'purified' representations of the words. After that, we will test the representations that are free of gender bias as mostly discussed by Bolukbasi et al[3]. We are planning to follow the same processes to generate the representations by debiasing the race bias from the original word embedding.

To debias the static word embeddings, we use the hard debiasing method from Bolukbasi et al. bolukbasi2016man. Firstly, the algorithm finds the gender subspace and identify the gender direction. Secondly, the algorithm hard-debiasing the gender bias by neutralizing the gender neutral words and equalizing the set of words outside the gender subspace.

In addition to that, we will also generate the third kind of new word embedding by debiasing the gender and race bias simultaneously. Second, we will use intrinsic evaluation and extrinsic downstream tasks[13] to show the trade off between debiasing performance and word embedding linguistic property preservation ability. From the literature review, we have noticed that these methods are more appropriate for our project. We will use intrinsic methods to evaluate the debiasing performance and linguistic property preservation ability. We select the word groups related to gender for intrinsic evaluation tasks. To evaluate whether the preferred features are preserved in new representation, we adopted downstream machine learning applications. There are 12 extrinsic and 16 intrinsic methods [1]. For our project, we have selected some specific methods as following:

Intrinsic Evaluation:

- **Word Embedding Association Test**: we use Word Embedding Association Test to find the stereotypical associations between different groups.
- **Relatedness**: we use the cosine similarity of the word embedding to compare with the ground truth relatedness score of word pairs.
- **Analogy**: We follow Mikolov et al.[12], Bolukbasi et al.[3] and Gonen and Goldberg[6]. The target is to find a word x for a given word y so that (x, y) and sample word pairs (a, b) share the same relationship.
- **Concept Categorization**: We follow the approach proposed by Gonen and Goldberg [6] and cluster a group of debiased word embedding and measure the purity of returned clusters.

Extrinsic Evaluation:

- **Sentiment Classification**: We classify the sentence with a binary polarized label using a movie reviews dataset. To represent the whole sentence, we use the TF-IDF feature for each word as the weight and calculate the weighted sum of word embedding vectors of the sentence.

VI. Experiments

A. Word-Embedding Association Test

We use Word-Embedding Association Test (WEAT) from Caliskan et al.[4] to measure the biases from COHA embeddings and the debiased embeddings.

Let X and Y be two sets of target words of equal size, and A, B be two sets of attribute words. Let $cos(\vec{a_i}, \vec{b_j})$ stands for the cosine value between embeddings vector of word a and b. Here the vector $\vec{a}$ is a contextual embedding vector for word a. The test statistics is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

and

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b})$$

and the **Effect Size** is

$$ES = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std\_dev_{w \in X \bigcup Y} s(w, A, B)}$$

Let $\{X_i, Y_i\}$ denote all the partitions of $X \bigcup Y$ into two sets of equal size. The one-sided P value of the permutation test is

$$Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

Due to the small size of COHA, some words such as names from Caliskan et al.[4] are missing. We only replicate the experiments that less than 5 words miss because to build the experiment, we need to delete 2

words because the missing of 1 word. So We replicate test 2(Weapon/Instruments,Pleasant/Unpleasant), test 7(math/arts,male/female), test 8(science/arts,male/female) and test 9(mental/physical, temporary/permanent). The results are reported in Table I.

B. Concept categorization

As we have mentioned earlier, concept categorization is an intrinsic evaluator, which is not the same as word analogy and word similarity concept. The idea is, for a given set of words, split it into different categories where the subset of words belong. As words are categorized based on the concepts, the semantic relation can be tested [2] and to do that we have done several experiments. We have followed the approach proposed by Gonen and Goldberg [6] for concept categorization (k-means clustering) and also conducted three more gender-biased related experiments which are also proposed by Gonen and Goldberg [6] to find out whether gender exists in the word vector relation. We have used COHA embedding as we referring to it as original or raw embedding throughout the paper. We also have used five debiased word embeddings (with PCs) to understand the trade-off between the reduction of biases from original word embedding and the preservation of desirable features. Our original vocabulary size is 71, 097. We have used three datasets, gender_specific_full dataset, definitional_pairs dataset and equalize_pairs dataset [6], [3] to create vocabularies and embeddings without gender specific words. Also, we have used 320 professions to conduct profession related experiment. We have removed punctuation, digits, word with upper-case letters and words with length greater than 20 characters.

C. Analogy generation

As we previously have seen and mentioned, the approach of using analogy pairs generated by word embedding to measure the quality and stereotypes has been utilized by many researchers [3][12]. For consistency and comparison purposes, we used the same data set as Bolukbasi et al. [3] and Gonen and Goldberg [6], with 320 profession words and 222 female words in total. We used all of the profession words and the first 50 of the female words as our data set for analogy generation task. We also utilized the analogy generation metrics proposed by Bolukbasi et al. [3], where the formula is

$$S_{(a,b)}(x,y) = \begin{cases} \cos(\vec{a}-\vec{b}, \vec{x}-\vec{y}) & \text{if} ||\vec{x}-\vec{y}|| \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

In this formula, $\vec{a}-\vec{b}$ stands for female-male direction, $\vec{x}$ stands for the word input for analogy generation tasks and $\vec{y}$ stands for the analogy generated by the model. We also extended Bolukbasi et al.'s proposal and set the δ to be 1 so that the analogy pairs would be semantically similar.

We used the original COHA word embedding plus the five debiased word embeddings to generate analogies for the data set. In order to quantify the debiasing performance and semantic rationality of analogies generated, we had 3 raters rate for stereotypes and semantic rationality respectively from 0-5 (with 0 standing for no semantic meaning, 1 as having little semantic rationality and 5 as having the most semantic rationality), and we averaged the ratings given by all 3 coders in order to mitigate some of the potential cognitive errors. Then we conducted analysis and comparison regarding the ratings across the 6 word embeddings. We also conducted a inter-coder reliability analysis with Fleiss' κ to check for the consistency of the ratings given by the 3 raters to ensure the quality of the ratings[9].

D. Sentiment Classification

In the external evaluation of sentiment classification, we used data source from Maas et al. contains movie reviews with binary labels from Internet Movie Database (IMDB)[11], and evaluate whether there's a difference of applying COHA embeddings and the debiased embeddings.

After mapping each word from data corpus to the corresponding vector, we selected two classic classifiers which are Linear Regression and Supporter Vector Machine with RBF kennel to compare the precision, recall rate, and accuracy to reveal if there's any improvement after debiasing for sentiment classification.

VII. Results

A. Word-Embedding Association Test

We report the results in Table I. For 'math/arts' experiment, the effect size is significant for raw embeddings and insignificant for all debiased embeddings. We find that the Hard debiasing methods decrease the intensities of gender bias in math and arts.

For 'science/arts' experiment, all effect sizes are negative whhich means in COHA embedding, science is associated more with females than males. It is contradictory with the general case in word embeddings that males are more related to science and females are more related to arts[4], [3]. However, we notice that the absolute values become smaller with more Principle Components in debias method, which means the difference in association becomes smaller and debias method works.

Besides the two cases about gender bias, we also report Weapon/instrument and Mental/physical illness cases. Since we only debias the gender bias, the effect sizes of these two biases do not decrease.

B. Concept categorization

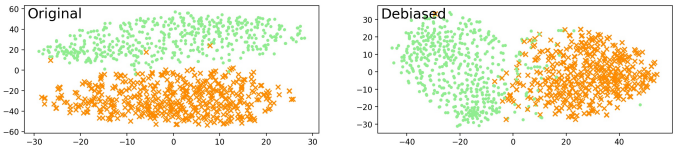

Fig. 1: Before and after debiasing (PC=1),1000 most biased words represented into two clusters

1) Clustering female-biased words and male-biased words (KMeans) : For concept categorization, clustering is popular in conducting experiments. The idea is to calculate the vector for each word and then apply the clustering algorithm to separate the vectors into n categories. In our case, n will be 2. A total of 1000 most biased words according to the original

TABLE I: **Summary of WEAT tests.** Here we report the effect sizes and P values for four different tests. The results for 6 different embeddings are reported.

	Weapon/Instrument, Pleasant/Unpleasant		Math/Arts, Male/Female		Science/Arts, Male/Female		Mental/Physical ill, Temporary/Permanent	
	Effect Size	P value	Effect Size	P value	Effect Size	P value	Effect Size	P value
Raw	**1.53**	**6.00E-08**	**1.29**	**4.67E-3**	-1.57	1.00	**1.17**	**0.02**
Debiased(PC=1)	**1.53**	**1.88E-07**	-0.43	0.79	-1.08	0.97	**1.16**	**0.02**
Debiased(PC=2)	**1.54**	**1.39E-07**	-0.31	0.74	-1.06	0.96	**1.15**	**0.02**
Debiased(PC=3)	**1.54**	**4.91E-07**	-0.28	0.71	-1.05	0.96	**1.16**	**0.02**
Debiased(PC=4)	**1.53**	**1.34E-07**	-0.29	0.72	-0.99	0.95	**1.17**	**0.02**
Debiased(PC=5)	**1.53**	**9.89E-08**	-0.23	0.66	-0.96	0.95	**1.17**	**0.02**

embedding have been selected (male-biased: 500 and female-biased: 500) and then we used k-means clustering algorithm to cluster them. As a performance metric, we have tested whether individual cluster contains the concept of the same category or different categories. The accuracy of the debiased embedding (PC=1) is 95.1% that means with 95.1% accuracy the cluster aligns with gender in comparison to an accuracy of 99.6% of the original biased embedding. We have seen the same trend for debiased embedding (PC=5) in which accuracy decreased compared to the original embedding. However, the rest of debiased embeddings don't show this trend in comparison with the original biased embedding. The outcome also indicates that bias information is still present in the representation after debiasing. Figure 1 shows the existence of gender bias in word vector relation by using the t-SNE (t-Distributed Stochastic Neighbor Embedding) technique [10].

2) Bias-by-projection and bias-by-neighbors correlation: Previous clustering experiment represented the bias-by-projection. Gonen and Golberg [6] suggested a new mechanism for measuring bias. The idea of this mechanism is to get an approximation of female/male socially based word by using gender-direction in the non-debiased embedding among the KNN (k nearest neighbors, k=100) of the target word. This is referred to as bias-by-neighbors. In this experiment, Pearson correlation has been calculated with the original and debiased embeddings and all correlations are identified as statistically significant. We have reported the result in Table II. For debiased embedding (PC=1) correlation is 0.66 with p-values <0.01 , which is less than the biased embedding where the correlation is 0.69. We have seen the same trend for debiased embedding (PC=5) in which correlation decreased compared to the original embedding; which means the difference in association becomes smaller and debias method works. However, the correlation values for the remaining debiased embeddings don't show the same trends.

3) Professions: For this experiment, a list of profession words is considered based on the definition of the neighbors-based bias, which was adopted by Bolukbasi et al. [3] and Zhao et al. (2018)[7]. In this experiment, Pearson correlation has been calculated with the original and debiased embeddings and all correlations are identified as statistically significant. We have reported the result in Table II. For debiased embedding (PC=1) correlation is 0.71 with p-values $<10^{-37}$, which is less than the original biased embedding where the correlation is 0.74. We have seen the same trend for debiased embedding (PC=5) in which correlation decreased compared to the original embedding. It shows that debias method works since the difference in association becomes smaller. However, the rest of debiased embeddings don't show this trends. Figure 2 plots the number of male neighbors as Y and the original bias as X. From the plot we can clearly see the correlation between X and Y.

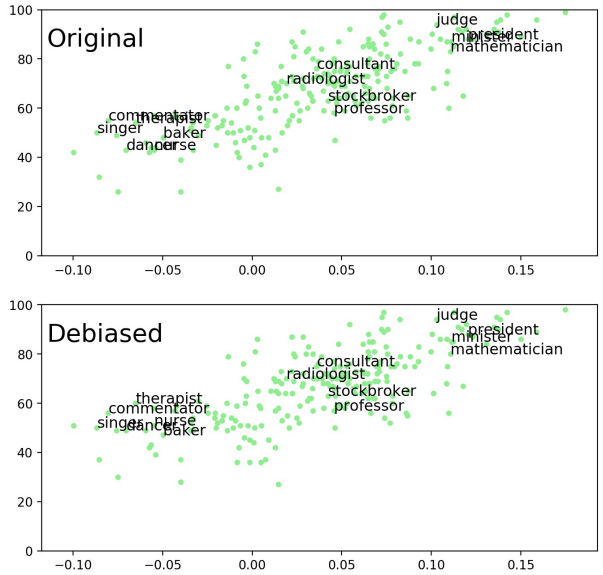

Fig. 2: The number of male neighbors for each profession as a function of its original bias. Only a limited number of professions are on the plot to make it readable (debiased (PC=1).

4) Classifying previously female-biased and male-biased words: In this experiment, the top 5000 female-biased and male-biased words (2500 for each gender) are sampled. To predict the gender, an RBF-kernel SVM classifier is trained on a subset of 1000 (female:500, male: 500) words, and the remaining 4000 words were allocated for test and evaluation. The classification accuracy for debiased (PC=1) is 92.5% compared to the original embedding of 94%. The debiased accuracy is less than the original embedding. We have observed the similar trend for biased embedding (PC=5) but not for the other embeddings. So it seems like based on the representation, a classifier can learn to generalize from some gendered words to others.

The above experiments for concept categorization evaluation illustrate that debias method works with some scenarios. However, debiased embeddings continue to maintain gender bias in their representations and in their similarities.

TABLE II: **Summary of concept categorization [correlation].** Here we report the Pearson correlation coefficients and P values for two tests. The results for different embeddings are reported.

	Bias-by-projection and bias-by-neighbors		Profession (bias-by-neighbors)	
	Correlation	P value	Correlation	P value
Raw	**0.69**	**0.00**	**0.74**	**6.88E-41**
Debiased(PC=1)	**0.66**	**0.00**	**0.71**	**9.81E-37**
Debiased(PC=3)	0.69	**0.00**	0.74	**4.31E-41**
Debiased(PC=4)	0.69	**0.00**	0.74	**4.31E-41**
Debiased(PC=5)	**0.68**	**0.00**	**0.73**	**2.06E-39**

C. Analogy generation

1) Gender specific analogy pairs: Similar to Bolukbasi et al.[3] and Gonen and Goldberg[6], we use a set of gender specific words for analogy generation task. We extract the Top 50 gender specific words for our analogy task and we generate analogy pairs on the "female-male" direction with all 6 word embeddings. Then we have 3 raters rate for the semantic meaning of the analogy pairs to measure the ability of the word embedding generating rational analogies.

We report the percentage of the analogies generated out of all word inputs, the percentage of analogies with no semantic meaning for generated analogies, the average rating of the analogies generating and the average rating of all analogies, as shown in Table III.

Besides the ratings, we conduct an inter-coder reliability analysis, which is to test for the rating consistency among multiple raters[9], and the results have been reported in Table III as well.

The data in the table show that as the number of Principal Components involved in the debiasing process increases, the percentage of successfully generated analogies decreases, as well as the percentage of analogies that have rational semantic meaning.

However, for the average scores for analogies generated, we first see an increase from the original COHA embedding to the debiased embedding (with PC=2), and then a decrease.

As for inter-coder reliability score, we see the observed agreement among 3 raters drops from substantial agreement ($0.61 \leq k \leq 0.80$) to moderate agreement ($0.41 \leq k \leq 0.60$)[9].

2) Gender neutral analogy pairs: Apart from gender specific analogy generation, we also generate analogies for gender neutral professions, and the results have been shown in Table IV.

For gender neutral analogy pairs, we can see that the percentages of analogies generated (out of 222 profession words) by 6 embeddings are about the same. The overall percentage of non-relational analogies out of analogies generated increase, and the overall percentage of non-stereotypical analogies within generated analogies decrease when the number of PCs involved increases, except for word embedding debiased with PC=3.

We also see that, overall, the score for analogies rationality (the ability of the word embedding generating rational analogies) first increase and then decrease with the increase of PCs, while the overall scores for analogies stereotype increase with the increase of PCs.

The results from gender specific analogy generation and gender neutral analogy generation showed that the stereotypes embedded in word embeddings are decreasing with the increase of the PCs involved in the debiasing process. There is a steep decrease in percentage of successful generated analogies from the original COHA word embedding to debiased word embeddings, but about half of the analogies the original COHA word embedding generated were also not rational, which makes generating high quality analogy pairs with semantic meanings not only an issue for debiased word embeddings but also an issue for word embeddings as a whole.

To measure the ability of the word embedding in preserving gender neutral word pairs, we use the word embedding to generate analogies for gender neutral words, and we see a steep decrease in the number of analogies generated. However, within the total number of analogies generated, we also see an increase in the rating for the analogy pairs for embeddings debiased within embeddings debiased with the first 3 PCs, and a decrease in the rating when the number of PCs involved exceeded 4.

We see the same pattern in the inter-coder reliability analysis where the agreement among the coders seems to be decreasing once the number of PCs involved exceeded 4, which could be another indication that the word embedding quality is deteriorating that the raters start to have different opinions on some of the analogies generated.

For profession analogy pairs, we also see a similar pattern where with the increase of the PCs involved, the analogies generated initially become more rational and non-stereotypical at first, then the analogies start to become less rational and more stereotypical again.

However, we notice from the analogy pairs that, when the number of PCs used to debias word embedding is large, it is more likely for the stereotypes to be flipped. For example, the analogy pair "actress: actor" in the original COHA embedding might become "actor:actress" on the female-male gender direction after debiasing. We also realize that there can be reverse gender bias in the word embedding if too many PCs are used for debiasing. For example, if the profession pair "clerk: sheriff" on the female-male gender direction is considered gender biased against female, then the profession pair "sheriff: clerk" on the female-male gender direction should be considered gender biased against male. Such bias may be easily overlooked.

D. Sentiment Classification

From the results that display in Table V and Table VI, it shows that accuracy, recall rate, true negative and true positive

TABLE III: **Gender specific analogy pair ratings** We report some highlighted statistics of the ratings for gender specific analogies generated by the original COHA embedding and the 5 debiased embeddings.

	Analogies generated	Non-rational analogies	Score for generated analogies	Score for all analogies	inter-coder reliability
Raw	86.00%	41.86%	**2.33**	**2.33**	**0.73**
Debiased(PC=1)	48.00%	29.17%	**3.18**	**1.78**	**0.61**
Debiased(PC=2)	46.00%	26.09%	**3.00**	**1.60**	**0.71**
Debiased(PC=3)	40.00%	15.00%	**2.17**	**1.01**	**0.59**
Debiased(PC=4)	36.00%	22.22%	**1.96**	**0.82**	**0.59**
Debiased(PC=5)	32.00%	25.00%	**2.15**	**0.80**	**0.60**

TABLE IV: **Gender neutral analogy pair ratings** We reporte the average ratings for gender neutral analogies generated by the original COHA embedding and the 5 debiased embeddings.

	Analogies generated	Non-rational analogies	Non-stereotypical analogies	Score for analogies rationality	Score for analogies stereotype
Raw	51.80%	20.00%	32.17%	**2.34**	**2.26**
Debiased(PC=1)	50.00%	21.62%	38.74%	**2.36**	**2.84**
Debiased(PC=2)	50.90%	20.35%	34.51%	**2.41**	**2.63**
Debiased(PC=3)	50.45%	25.89%	38.39%	**2.22**	**2.84**
Debiased(PC=4)	50.00%	19.82%	35.14%	**2.41**	**2.85**
Debiased(PC=5)	50.45%	25.89%	31.25%	**2.22**	**2.86**

TABLE V: **Linear Regression Classifier Result**

	Accuracy	Recall Rate	True Negative	True Positive
Raw	80.54%	81.26%	81.28%	79.80%
Debiased(PC=1)	82.20%	81.57%	81.92%	82.48%
Debiased(PC=2)	82.42%	81.81%	82.16%	82.68%
Debiased(PC=3)	82.44%	81.57%	81.92%	82.96%
Debiased(PC=4)	82.10%	81.28%	81.64%	82.56%
Debiased(PC=5)	82.32%	81.53%	81.88%	82.76%

TABLE VI: **SVM Classifier with RBF Kernel Result**

	Accuracy	Recall Rate	True Negative	True Positive
Raw	80.54%	80.54%	81.02%	80.07%
Debiased(PC=1)	82.20%	82.19%	82.36%	82.03%
Debiased(PC=2)	82.42%	82.41%	82.59%	82.25%
Debiased(PC=3)	82.44%	82.42%	82.44%	82.44%
Debiased(PC=4)	82.10%	82.09%	82.15%	82.05%
Debiased(PC=5)	82.32%	82.31%	82.37%	82.26%

rate has slightly increased in various degree by apply debiased word embeddings for both classifiers.

For Linear Regression, the original biased word embedding has an accuracy of 80.54% in which contains 81.28% true negative rate and 79.80% true positive rate, while all debiased words embedding has reached over 82% accuracy. In all 5 debiased word embeddings, debiased PC=3 performs the best with a 82.44% accuracy. Similar with SVM classifier, by applying debiased word embeddings, with the same data corpus, classifier's accuracy can increase by about 2%. Debiased words embeddings shows a stable trend that debiased PC=3 eliminated word discrimination to the greatest extent compares to other four debiased word embeddings in the evlauaton of sentiment classification. With the least accuracy performance which is PC=4, the number of accuracy for both classifiers increased by 1.56% compare to with the original biased word embedding. It can conclude by adjusting word embedding to eliminate biases that included in PC=1 to PC=5, classifier's performance has improved but in various degree.

VIII. Discussion

We find that incorporating more principal components into the gender subspace do not always improve the debiasing performance of the debiased word embeddings or their performance in downstream NLP tasks. Moreover, it could impair the word embedding's ability to preserve important linguistic properties like semantic meaning and rationality.

We have listed the limitations here. We take the average vector of the multiple principal components as their representative vector in gender space. Though it sacrifices some information from different principal components, the significant results from our experiments shows that it captures the gender information well.

In our work, We use euclidean distance to calculate the distance between the two words in analogy pair in order to preserve the semantic similarities in analogy pairs. While euclidean distance is standard approach to capture the distance between vectors in space, it is also known not to be the best metric for distance between vectors in high dimensional space.

In addition, We realize that the word embedding we use has limitations itself, for there are many names from fictions which are associated with the roles they bear, thus when generating analogy pairs, some of them appear as the analogies. There also seem to be many words that do not have analogies in the word embedding. Even though our word embedding can generate analogies for some of the profession words we input, the success rate is not as high as we expected.

To quantify the rationality and stereotypes of analogy pairs, we have 3 raters rate for the analogy pairs generated by our word embeddings. It is scientific for we are able to mitigate potential cognitive human errors that can be attributed to one coder. However, the fault tolerance is not high considering that we only have 3 raters.

In future work, researches should take different representations for multiple principal components into consideration and compare the results and impacts of those representations. Researchers could try different metrics to count for the distance

between vectors in high dimensional space. For the disadvantages we experienced that are associated with the word embedding we used, researchers could either try eliminating those fictional names or try experimenting on different word embeddings and compare the results. When conducting inter-coder reliability analysis, researchers could have more raters rate for analogy pairs generated to increase fault tolerance and accuracy of the ratings. Researchers should also compare the influence of principal components on different static word embeddings considering the differences among them. Moreover, researchers could test with principal components that are generated from different word pairs. Last but not the least, researchers could observe the influence of multiple principal components in more NLP tasks in the future.

IX. Conclusion

To clarify the influence of incorporating more principal components into the gender subspace, we evaluate the word embeddings debiased with different principal components in several intrinsic and extrinsic tasks. Based on our experiments, we found that incorporating more principal components into the gender subspace do not always improve the quality of the debiased word embeddings or their performance in downstream NLP tasks.

References

[1] A. Bakarov, "A survey of word embeddings evaluation methods," *arXiv preprint arXiv:1801.09536*, 2018.

[2] W. Bin, W. Angela, C. Fenxiao, W. Yunchen, and K. CC, Jay, "Evaluating word embedding models: Methods and experimental results," 2019.

[3] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[4] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science 356(6334):183–186.*, 2017.

[5] M. Davies, *The corpus of historical american english: COHA*. BYE, Brigham Young University, 2010.

[6] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," *arXiv preprint arXiv:1903.03862*, 2019.

[7] Z. Jieyu, Z. Yichao, L. Zeyu, W. Wei, and C. Kai-Wei, "Learning gender-neutral word embeddings," in *In Proceedings of EMNLP)*, 2018, p. 4847–4853.

[8] S. Karve, L. Ungar, and J. Sedoc, "Conceptor debiasing of word representations evaluated on weat," *arXiv preprint arXiv:1906.05993*, 2019.

[9] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," 1977.

[10] d. M. Laurens, van and H. Geoffrey, "Visualizing data using t-sne. journal of machine learning research," *machine learning research, 9(Nov):2579–2605*, 2008.

[11] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *ACL*, 2011.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 298–307.