16th May 2021

Identifying and Reducing Biases in Word Embeddings

Project members:

CHEUNG Lok Yee Joey

IP Wing Yan

Table of Content

1. Introduction

In natural language processing (NLP), word embeddings are used quantitatively, in the form of continuous vectors to represent words for text analysis. According to Mikolov et al. (2013) [1], these vector representations not only show the multiply degrees of similarity between words, but they also reflect the varieties of relationships between words by vector differences. For instance, a simple arithmetic can be done as *vector("King") + vector("Man") - vector("Women")* which would result in the vector of the word "Queen". They are indeed useful tools for real-world applications ranging from question retrieval to document ranking.

However, it is revealed that social biases are exhibited in word embedding algorithms. Gender bias, racial bias and age bias are types of stereotypical bias commonly noticed. While word embedding has been widely used in machine learning, there is an increasing risk of causing serious consequences such as prejudices in resume screening or underlying recidivism as shown in some real-life cases [2]. Since important decisions are made using these algorithms like criminal prediction and company's hiring process, it is vital that biases in word embedding are eliminated instead of being amplified, and future use could be fully accurate.

Our objectives of this project are to conduct exploratory and quantitative analysis on two types of biases (gender bias and racial bias) in word embedding in order to visualize and measure the seriousness of the problem. It would be intriguing to see how woman and man associate with different professions, as well as how different races associate with positive or negative impression. While there are different types of racial identities related to stereotyping such as white against black, American against Asian and so on, we take one of them as an example and focus on investigating European American against African American here. We would use GloVe as our word vectors source and obtained professions words from authorized websites. After so, we aim to start our experimentation on bias reduction and evaluate the de-biased results. Major processing approaches like data visualization, Word Embedding Association Test (WEAT), linear projection are applied.

2. Data Source

2.1 Data Description

This project used GloVe for our primary word embedding data source. GloVe is an unsupervised learning algorithm that captures global corpus statistics to obtain vector representation for words [3]. We made use of the largest number of word embeddings with word vectors pre-trained on "Common Crawl" corpus for a more representable result. The word embedding text file contains 2.2 million vocabularies and their vectors in 300 dimensions.

Professions and occupations words were collected from online dictionaries and educational website i.e. Enchanted Learning [4]. We used professions words as they are easily

understood and clarified by humans and capture common social biases. We investigated into the extent of the biases of these words towards gender direction as well as racial direction, and carried out de-bias process correspondingly.

During the investigation on gender bias and conduct Word Embedding Association Test (WEAT), male names and female names lists are the sets of target words representing two genders while career words and family words lists are sets of attribute words. Career and family are used as traditional mindset reflects that men tend to focus on their work and women are more likely to take care of the family.

Target words:

Male names: John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill

Female names: Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna

Attrib. words:

Career words : executive, management, professional, corporation, salary, office, business, career

Family words : home, parents, children, family, cousins, marriage, wedding, relatives

When it comes to WEAT for racial bias, we chose European American names and African American names to indicate sets of target words representing two different races. Pleasant words and unpleasant words are two sets of attribute words as racial stereotype are often based on a person's positive or negative traits and impressions. The two types of racial names are used as data for identifying racial direction in the visualization part as well.

Unpleasant: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit

Two sets of target words are of equal sizes in each bias case for the measurement of the association between sets of target words with sets of attribute words. All the data in WEAT are based on the literature study by Caliskan et al. (2017) [6].

2.2 Data Pre-processing

We applied data processing and cleaning before analysis. The professions words were stored in an excel file as "occupation_word_list.xlsx". For better illustrations, duplicates were removed. Then, we loaded the GloVe word embedding in Word2Vec format and got rid of some professions words that did not present in GloVe. Finally, we obtained a cleaned and processed occupation word list with 287 professions words. As shown below:

```
Out[6]: 287
```

In WEAT, male names and female names, career words and family words stored in excel files for gender bias part were pre-processed by checking their presence in GloVe. This resulted in 8 words each. Pre-processing is shown below:

```
Out[460]: 8
```

European American names and African American names, pleasant and unpleasant words were stored in excel files and were used in WEAT for the racial bias part. They were also pre-processed by checking their presence in GloVe. 18 words each in racial names and 25 words each in pleasant and unpleasant words were resulted. Pre-processing is shown below:

```
Out[32]: 18

In [39]: pleasant = pd.read_excel("pleasant.xlsx", names=['ple'], header=None)
         pleasant_words = [w for w in pleasant.ple if w in glove]
         len(pleasant_words)

Out[39]: 25

In [40]: unpleasant = pd.read_excel("unpleasant.xlsx", names=['unple'], header=None)
         unpleasant_words = [w for w in unpleasant.unple if w in glove]
         len(unpleasant_words)

Out[40]: 25
```

We mainly used Jupyter Notebook and Google Colab with Python 3 as our programming language. Some common NLP libraries were also employed namely Genism, Matplotlib, scikit–learn etc.

3. Approaches to Analytics

3.1 Data Visualization

Data visualization by using graphs enables us to identify patterns and provide a clear and efficient representation of the information. To understand the gender bias and racial bias in word embeddings (e.g. a word is closer to *he* than *she*) in terms of occupation, we adopt data visualization using scatter plots to observe the situation.

As suggested by Bolukbasi et al (2016) [5], we first define the gender direction using *he-she* since they are commonly used to classify gender and have fewer alternative meanings or interpretations. Racial direction would be derived by using names (reason explained below).

Cosine similarity of profession words and normalized vector in bias direction i.e. normalized *"she"* and normalized *"he"* (or European American names, African American names) were calculated and plotted on *he-she* (or EurAmer-AfrAmer) axis for visualization. High score means higher similarity. Since we realized some jobs appear to be similar in nature like policeman and cop, we then discarded either one of them to reduce redundancy.

Cosine similarity is calculated as follow [5]:

$$\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}.$$

where u: vector of bias direction, v: vector of profession word

Very often it is difficult to find a good word pair to indicate the bias direction, therefore we make use of names as an alternative to find the race direction, as suggested by Dev and

Phillips (2019) [7]. In fact, different racial groups tend to use certain names more frequently in real life and Greenwald et al. (1998) [8] has also found extreme impacts of race as indicated simply by name, thus it is reasonable to detect the race bias using names.

Mathematically, the bias direction is defined as follow:

$$v_{B,\text{names}} = \frac{s - m}{\|s - m\|},$$

where s = avg(si)/||avg(si)|| and m = avg(mi)/||avg(mi)||

For race, s is the vector direction derived from a set of common European American names and m is the vector direction derived from a set of African American names.

For gender, s can simply be '*he*' direction and m can simply be '*she*' direction.

3.2 Quantitative Measurement Using WEAT

WEAT was adopted for our quantitative measurement of gender bias [6]. It analogizes the terminology of the Implicit Association Test (IAT), a social psychology test accessing the implicit stereotype of humans. WEAT considers two sets of target words say {male names, female names} and two sets of attribute words say {career-oriented words, family-oriented words} which we used in our actual implementation in gender bias. For the case in racial bias, we took {European American names, African American names} as the target words while {pleasant words, unpleasant words} were used as attribute words.

In the calculation process, s(w,A,B) returns the association of a target word w with attribute words A,B. s(X,Y,A,B) returns the differential association of target words X,Y with the attribute words A,B :

In formal terms, let X and Y be two sets of target words of equal size, and A, B the two sets of attribute words. Let $\cos(\vec{a}, \vec{b})$ denote the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$.

- The test statistic is

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

where

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

Then, WEAT score is obtained by measuring the normalization of the distance between the two distributions of the above association. Mean and std-dev refer to the arithmetic mean and the sample standard deviation respectively. The score ranges between 2.0 and -2.0. The closer the score to zero, the less bias.

WEAT score calculation [6]:

$$\frac{\text{mean}_{x \in X} s(x,A,B) - \text{mean}_{y \in Y} s(y,A,B)}{\text{std-dev}_{w \in X \cup Y} s(w,A,B)}$$

3.3 Bias Reduction

In our de-bias process, we adopted Linear Projection method [7] which projects all words vectors orthogonally to bias direction. Proportions along bias direction (gender/racial direction) are subtracted from the vector of the word embeddings in this method. Resulting debiased word embeddings would lie very close to each other, if not almost identical. Noted that gender specific words such as actress were not debiased.

> **Linear Projection.** A better baseline is to project *all* words $w \in W$ orthogonally to the bias vector v_B.
>
> $$w' = w - \pi_B(w) = w - \langle w, v_B \rangle v_B.$$

We used this instead of the commonly seen Hard Debias method because it is much easier to adopt and is fully automated.

To evaluate the results, we carried out step 3.1 and 3.2 again for the debiased word embedding vectors and compare the results. In data visualization, our desired output would be having word embeddings lying approximately on 45-degree line which means none of the jobs is considered as female or male dominance or race prejudice. For quantitative measurement, our resulted WEAT scores were expected to be closer to zero than the original scores.

4. Analysis and Results

4.1 Gender Bias

4.1.1 Data Visualization

With the use of the imported genism model, we first calculated the cosine similarities of the professions words with normalized *"he" (v_he)* and normalized *"she" (v_she)* respectively. Higher score indicates a higher similarity between the vector of a profession word with *v_he* or *v_she.* The 10 most male-stereotyped and female-stereotyped words are selected and listed below:

In [13]:
```
# find out 10 most he-positive-extreme occupation words
htemp = []

for x in processed_job_list:
    htemp.append(cos_sim(v_he, glove[x]))

he = pd.Series(htemp, index=processed_job_list, name='he')
he.nlargest(n=10)
```

In [14]:
```
# find out 10 most she-positive-extreme occupation words
stemp = []

for x in processed_job_list:
    stemp.append(cos_sim(v_she, glove[x]))

she = pd.Series(stemp, index=processed_job_list, name='she')
she.nlargest(n=10)
```

Extreme *he* occupation		Extreme *she* occupation	
Occupations	Cosine similarity	Occupations	Cosine similarity
Judge	0.4816	Nurse	0.5138
Doctor	0.4675	Teacher	0.5122
President	0.4584	Doctor	0.5092
Soldier	0.4467	Actress	0.4971
Teacher	0.4437	Waitress	0.4606
Politician	0.4431	Judge	0.4486
Captain	0.4378	Singer	0.4399
Cop	0.4317	Dancer	0.4389
Policeman	0.4229	Student	0.4256
Professor	0.4152	Maid	0.4185

Then, we compared the above 17 distinct jobs with *he* and *she*. If it is closer to *he* in terms of cosine similarity, bias = blue, else bias = red:

Out[56]:

	he	she	bias
maid	0.235358	0.418461	red
student	0.373731	0.425599	red
cop	0.431660	0.388554	blue
waitress	0.284964	0.460631	red
judge	0.481602	0.448580	blue
soldier	0.446679	0.352279	blue
actress	0.265117	0.497068	red
captain	0.437806	0.306524	blue
president	0.458380	0.358608	blue
politician	0.443103	0.336180	blue
doctor	0.467547	0.509170	red
policeman	0.422934	0.351142	blue
professor	0.415242	0.393214	blue
dancer	0.307044	0.438895	red
nurse	0.319476	0.513825	red
singer	0.363594	0.439919	red
teacher	0.443727	0.512202	red

To visualize in a fair (same number of dots for both sides) and clear (no overlapping on the graph) manner, we only selected five of each to plot. In Figure 1, red dots are more likely to lie on the left side of the diagonal which represent these occupation words are having higher similarity with *she* than *he*. This indicates red dots occupations e.g. nurse, dancer which are more artistic and may require more patience, bias towards female. On the contrary, blue dots that mostly lie on the right, representing occupation words having higher similarity with *he* than *she.* Occupations like president, captain which are some leading roles, bias towards male.

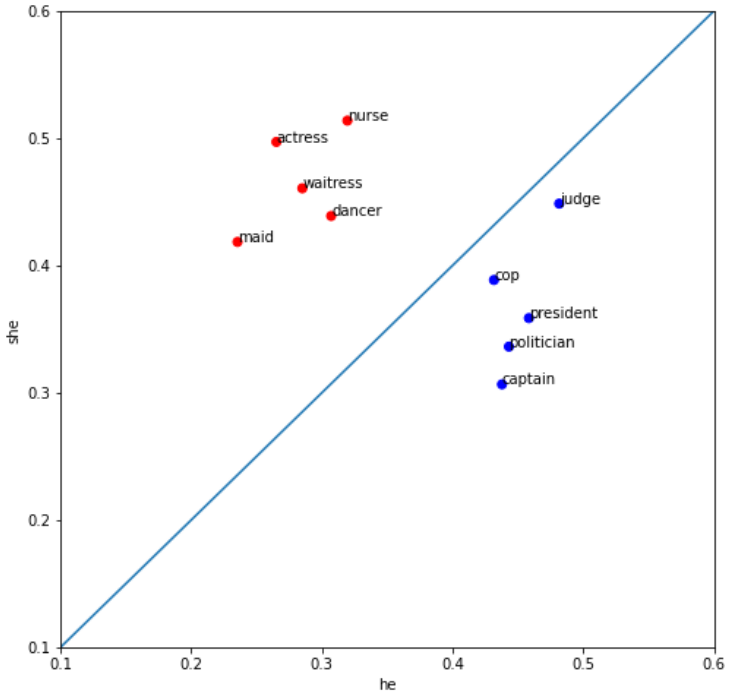


Figure 1: Occupation words projected onto he-she direction.

4.1.2 Quantitative Measurement

Following the formula shown in 3.2, we constructed below two functions to calculate the WEAT score.

In [9]:
```
def association(w, A, B):
    """
    Returns association of a target word w with attribute words, i.e. s(w, A, B)
    w: one target word vector
    A: attribute word vectors
    B: attribute word vectors
    """
    return np.mean(list(map(lambda a: cos_sim(w,a), A))) - np.mean(list(map(lambda b: cos_sim(w,b), B)))
```

In [10]:
```
def weat_nscore(X, Y, A, B):
    """
    Returns normalized WEAT score
    X: target word vectors
    Y: target word vectors
    A: attribute word vectors
    B: attribute word vectors

    """

    X_association = np.array(list(map(lambda x : association(x, A, B), X)))
    Y_association = np.array(list(map(lambda y : association(y, A, B), Y)))

    mean_diff = np.mean(X_association) - np.mean(Y_association)
    std = np.std(np.concatenate((X_association, Y_association)))

    return mean_diff / std
```

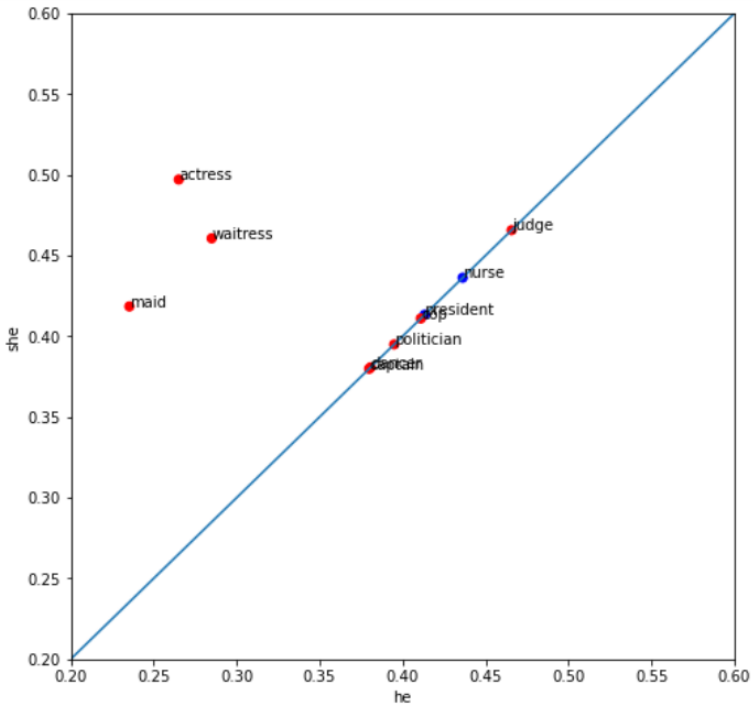
Target words	Attribute words	Original Finding (by Caliskan)	Our Finding
Male vs Female names	Career vs Family	1.81	1.87

```
In [22]: weat_nscore(glove[male_name_words], glove[female_name_words], glove[career_words], glove[family_words])
Out[22]: 1.8734031
```

We obtained the normalized WEAT score = 1.87 which is already a highly positive score. In other words, male is more inclined to work outside, and female is more inclined to look after the house. “Man are breadwinners while women are homemakers” is definitely wrong and biased nowadays.

4.1.3 Bias Reduction

For demo purpose, we only debiased a small set of gendered words including 7 out of the 10 plotted occupation words and attribute words used in WEAT. Noted that we did not debias all the 10 occupation words since the remaining three of them are gender specific terms i.e. actress, waitress and maid. After debiasing using linear projection (please refer to Appendix for details), all the seven gender neutral professions are now lying on 45-degree line. In other words, none of those jobs is significantly male or female dominant.



We conducted the WEAT again with the debiased attribute words. The score is successfully reduced from 1.87 to 1.37 as show below. The decrease may not seem very significant as the gender direction was simply defined by *he-she* but not by names.

```
In [33]: weat_nscore(glove[male_name_words], glove[female_name_words], debiased_careers, debiased_family)
Out[33]: 1.372129
```

4.2 Racial Bias

4.2.1 Data Visualization

We first compute the word directions of the two ethnicities - European American *v_euro* and African American *v_afr* by using the below function.

```
In [7]: def bias_direction(bias1, bias2):
            """
            Returns the two normalized vector directions and a normalized bias direction
            """

            twomeans = [sum(glove[w] for w in word)/len(word) for word in (bias1, bias2)]
            vec_directions = [v / np.linalg.norm(v) for v in twomeans]
            vB = vec_directions[0] - vec_directions[1]
            fin_vB = vB / np.linalg.norm(vB)

            return vec_directions[0], vec_directions[1], fin_vB
```

```
In [33]: # Determine the bias direction, which encodes the racial difference
         v_euro, v_afr, vB_racial = bias_direction(EuroAmerNameWords, AfrAmerNameWords)
```

Likewise, we computed the cosine similarities of the professions words with *v_euro* and *v_afr* respectively. The higher the similarity between the vectors of a professions word with *v_euro* or *v_afr,* the larger the score. The 10 most *EuroAmer*-positive-extreme and *AfrAmer*-positive-extreme occupation words are listed below.

```
In [34]: # find out 10 most EuroAmer-positive-extreme occupation words
         etemp = []

         for x in processed_job_list:
             etemp.append(cos_sim(v_euro, glove[x]))

         EuroAmer = pd.Series(etemp, index=processed_job_list, name='EuroAmer')
         EuroAmer.nlargest(n=10)
```

```
In [35]: # find out 10 most AfrAmer-positive-extreme occupation words
         atemp = []

         for x in processed_job_list:
             atemp.append(cos_sim(v_afr, glove[x]))

         AfrAmer = pd.Series(atemp, index=processed_job_list, name='AfrAmer')
         AfrAmer.nlargest(n=10)
```

***EuroAmer*-positive-extreme occupation**		***AfrAmer*-positive-extreme occupation**	
Occupations	Cosine similarity	Occupations	Cosine similarity
author	0.465899	saxophonist	0.285134
director	0.454183	singer	0.268677
writer	0.453935	percussionist	0.267285
professor	0.412403	socialite	0.266445
actor	0.404513	harpist	0.243948
actress	0.401434	quarterback	0.240530
singer	0.400027	flutist	0.240256
journalist	0.390697	dancer	0.236925
reporter	0.385549	policewoman	0.234511
artist	0.377582	actress	0.217354

Then, we compared the above 18 distinct jobs with *EuroAmer* and *AfrAmer*. If it is closer to *EuroAmer* in terms of cosine similarity, bias = blue, else bias = red:

Out[60]:

	EuroAmer	AfrAmer	bias
writer	0.453935	0.114762	blue
author	0.465899	0.093143	blue
socialite	0.234357	0.266445	red
policewoman	0.033310	0.234511	red
harpist	0.197487	0.243948	red
quarterback	0.306665	0.240530	blue
actor	0.404513	0.175432	blue
journalist	0.390697	0.201874	blue
actress	0.401434	0.217354	blue
saxophonist	0.239611	0.285134	red
flutist	0.187591	0.240256	red
director	0.454183	0.117487	blue
percussionist	0.212203	0.267285	red
professor	0.412403	0.129072	blue
artist	0.377582	0.120193	blue
dancer	0.270649	0.236925	blue
reporter	0.385549	0.162507	blue
singer	0.400027	0.268677	blue

To visualize in a fair (same number of dots for both sides) and clear (no overlapping on the graph) manner, we only selected four of each to plot. In Figure 2, red dots on the left of the diagonal represent profession words having higher cosine similarity with African Americans names. This indirectly infers that red dots occupations e.g. flutist, saxophonist which are music-related, are more likely to be African Americans. On the other hand, blue dots projected on the right of 45-degree line are jobs “closer” to Euro-American names. Examples like professor, author which may require higher education, are Euro-American stereotypical professions.

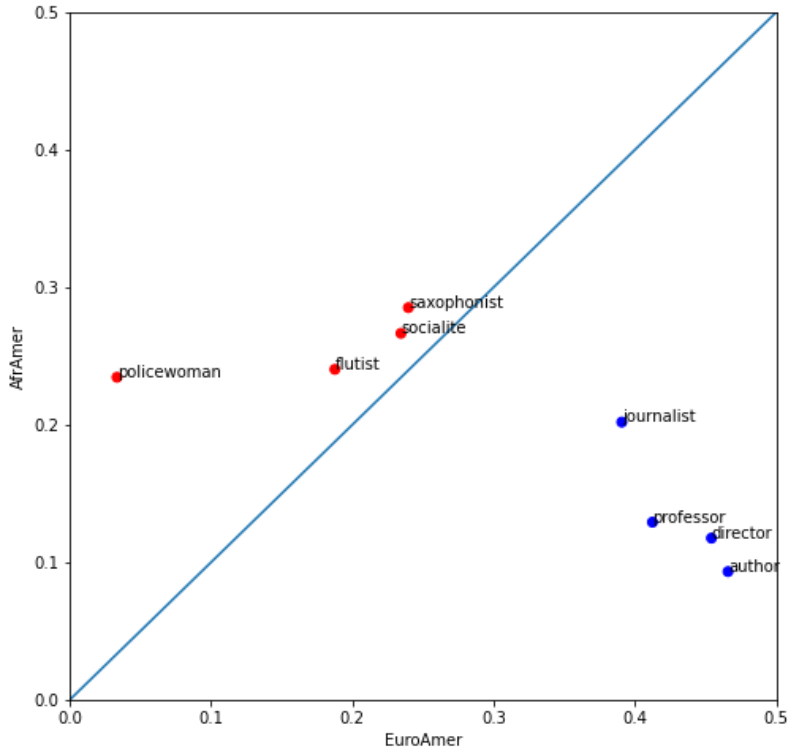


Figure 2: Occupation words projected onto *EuroAmer-AfrAmer* direction.

4.2.2 Quantitative Measurement

Using the functions defined above, we calculate the WEAT score to measure how European American and African American are associated with pleasant words such as freedom and unpleasant words such as poverty.

Target words	Attribute words	Original Finding (by Caliskan)	Our Finding
European-American vs African-American names	Pleasant vs Unpleasant	1.50	1.58

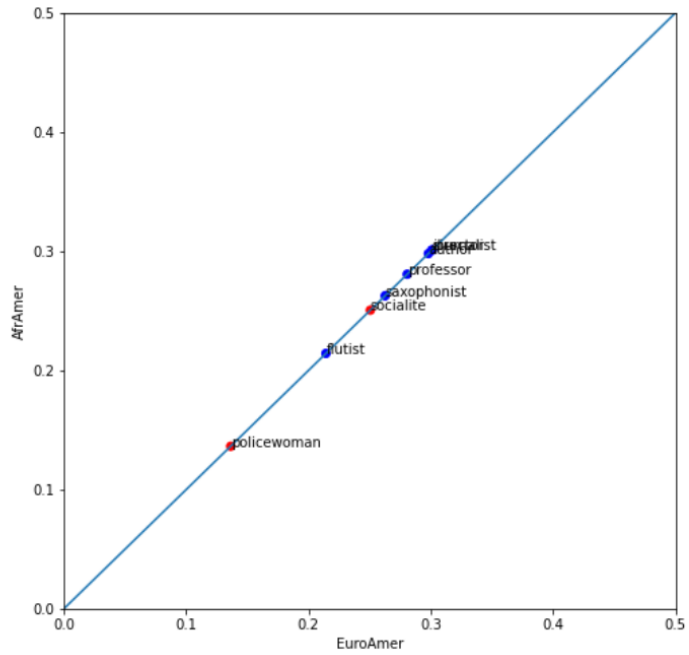
```
In [93]: weat_nscore(glove[EuroAmerNameWords], glove[AfrAmerNameWords], glove[pleasant_words], glove[unpleasant_words])
Out[93]: 1.5777843
```

A positive value indicates that European American names were more associated with pleasant than unpleasant words, compared to African American names.

4.2.3 Bias Reduction

For demo purpose, we only debias a small set of gendered words including the 8 plotted occupation words and attribute words used in WEAT. After debiasing using linear projection (please refer to Appendix for details), all the previous 8 profession words are now lying on

45-degree line. In other words, neither European American nor African American is more related to certain jobs.



We conducted the WEAT test again with the debiased attribute words. The score significantly reduces from 1.58 to 0.08! Since the bias direction this time is derived from the target words (European American names and African American names) directly.

```
In [101]: weat_nscore(glove[EuroAmerNameWords], glove[AfrAmerNameWords], debiased_pleasant, debiased_unpleasant)
Out[101]: 0.08318729
```

5. Conclusion

Word embedding is an important element for us to understand the bias in language. In this paper, we emphasized the visualization and quantitative measurement of bias in word embeddings. We managed to present the bias using scatter plot in a readable way. When occupation words were projected onto *he-she* and *EuroAmer-AfrAmer* direction, we could observe the tendency of bias toward either side. It was an interesting finding that more educated jobs are prone to prejudice against European Americans. Moreover, we also quantified the bias using WEAT. For instance, the tendency of male focusing on career and women on family were proved by a positive WEAT score. We have undoubtedly fulfilled our objectives.

These undesirable stereotypes in word embeddings could be problematic when they are being used in machine learning and NLP tasks. We have also managed to reduce part of the bias using Linear Projection method which is a simple and effective way to remove bias directions (gender/racial direction) from word vectors. The resulting vector was

unexpectedly good in racial debias part. Resulting vectors turned onto the diagonal, which represented their neutrality. Furthermore, WEAT score attained was close to zero, indicating a low association between {European American names, African American names} and {pleasant words, unpleasant words}.

For further study, we hope to broaden the scope of our analysis to examine other types of racial bias and other kinds of human bias such as age, or even other languages if possible. It is also important to define and collect a more comprehensive set of bias specific and bias neutral words in order to debias the whole word embedding for practical usage.

Bias in word embeddings merely reflect the bias situation in society. It is important that we do not only rely on reducing bias in computer systems to prevent the exacerbation of stereotypes in existing society.

6. Reference

[1] T. Mikolov, K. Chen, G. Corrado, & J. Dean (2013). “Efficient Estimation of Word Representations in Vector Space” https://arxiv.org/pdf/1301.3781.pdf

[2] A. Khademi, and V. Honavar. (2020). "Algorithmic bias in recidivism prediction: A causal perspective (student abstract)." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 10. https://doi.org/10.1609/aaai.v34i10.7192

[3] J. Pennington, R. Socher & C.D. Manning. (2014). “GloVe: Global Vectors for Word Representation” https://nlp.stanford.edu/pubs/glove.pdf

[4] *Job and Occupation Vocabulary Word List*. Enchanted Learning. (n.d.). https://www.enchantedlearning.com/wordlist/jobs.shtml

[5] T. Bolukbasi, K.W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai. (2016). “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings” *arXiv:1607.06520*, https://arxiv.org/pdf/1607.06520.pdf

[6] A.Caliskan, J. J. Bryson, and N. Arvind. (2017).“Semantics derived automatically from language corpora contain human-like biases.” Science 356.6334: 183-186. http://opus.bath.ac.uk/55288/

[7] S. Dev, J. Phillips. (2019). "Attenuating bias in word vectors." *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR https://arxiv.org/pdf/1901.07656.pdf

[8] A. G. Greenwald, D. E. McGhee, J. L. Schwartz. (1998).“Measuring individual differences in implicit cognition: the implicit association test.”, Journal of personality and social psychology https://www.uni-muenster.de/imperia/md/content/psyifp/aeechterhoff/wintersemester2011-12/attitudesandsocialjudgment/greenwaldmcgheeschwatz_iat_jpsp1998.pdf