The Mexican Janitor and the American Custodian

Anthony Capunay
University of California, Berkeley
acapunay@berkeley.edu

We investigate bias in Glove word embeddings against Mexicans. We show that lower-ranked professions are more related to the word embedding $\overrightarrow{mexican}$ than to the word embedding $\overrightarrow{american}$, and higher-ranked professions are more related to the word embedding $\overrightarrow{american}$ than to the word embedding $\overrightarrow{mexican}$. We also show that a positive-negative antonym pair is likely to have the positive antonym closer to $\overrightarrow{american}$ and the negative antonym closer to $\overrightarrow{mexican}$. We then debias the word embeddings using Bolukbasi et al. [2]'s Hard Debias algorithm and show that the word embeddings maintain their usefulness with a simple classification model.

1 Introduction

The creation of word embeddings such as word2vec and Glove have helped revolutionize Natural Language Processing (NLP). It has been shown that these embeddings contain useful properties such as having related words close together, for example, city names will be close together in the word embedding space. Another useful property of these embeddings is that pairs of words can be used to form analogies, for example, if we take $\overrightarrow{king}$ (the embedding for the word king) and subtract $\overrightarrow{man}$ and add $\overrightarrow{woman}$, we approximately get $\overrightarrow{queen}$.

Unfortunately, it has also been shown that these word embeddings contain biases present in society. For example, Bolukbasi et al [2] showed that $\overrightarrow{computer\ programmer}$ - $\overrightarrow{man}$ + $\overrightarrow{woman} \approx \overrightarrow{homemaker}$, i.e., the word embeddings encode a similarity between the pair man, computer programmer to the pair woman, homemaker.

These biases can be harmful to under-represented groups. Suppose a Google Maps-like website sorts the order in which it shows restaurants by the sentiment score of the reviews. If the word embeddings encode some negative meaning into words related to under-represented groups, then reviews having phrases like "arab food", "mexican food", "chinese restaurant" may get lower sentiment scores than "american food" or "french food". For restaurant owners from under-represented groups, getting sorted lower on Google Maps may affect their business revenue and therefore their personal finances and well-being.

In this paper, we focus on bias in word embeddings against Mexicans. In recent years, many political figures have used this under-represented group as a scapegoat for other problems affecting the United States. We want to make sure to expose bias in word embeddings against this under-represented group, if it exists, so as to lessen the harm against this under-represented group.

2 Data

We will mainly use Glove as the word embeddings to analyze. We will look at word2vec only to compare the bias against Mexicans in Glove to the bias against Mexicans in word2vec.

Glove. For the Glove word embeddings, we took the 100-dimensional embeddings with the 400K vocabulary. These word embeddings were built on Wikipedia documents and Gigaword 5.

word2vec. For the word2vec word embeddings, we took the top 600K word embeddings. These word embeddings are 300-dimensional and are built on the Google News corpus.

Note that both of these word embeddings are built on top of news articles (with word2vec being fully built on news articles and Glove partially built on news articles), and as pointed out by Bolukbasi et al [2], one would expect biases to be only slightly present in the embeddings given that journalists wrote the training data. But as we will show, the bias against Mexicans present in these embeddings is worrisome, more so in the Glove embeddings than in the word2vec embeddings though.

3 Analysis of Bias

In this section, we look at where word embeddings for professions fall on the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis. We also look at positive-negative antonym pairs, which we define as a pair in which one antonym has a positive connotation (e.g. obedient) and the other has a negative connotation (e.g. disobedient), and compare the position of each member of the pair on the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis. Finally, we look at whether the embeddings create steoreotypic analogies of the form $\overrightarrow{american}$ is to [profession] as $\overrightarrow{mexican}$ is to ___.

3.1 Professions on $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis

In this section, we wanted to see whether there was a propensity for high-earning and/or highly prestigious professions to be closer to $\overrightarrow{american}$ than to $\overrightarrow{mexican}$. Conversely, we wanted to see whether the low-earning and/or less prestigious professions were closer to $\overrightarrow{mexican}$ than to $\overrightarrow{american}$. This would indicate to us that the word embeddings are including a bias for the word mexican that makes it less prestigious/valuable than the word american.

The list of professions we take is: businessman, manager, legislator, maid, waiter, waitress, janitor, doorman, custodian, gardener, landscaper, stonemason, governor, doctor, nurse, attorney, lawyer, dentist, astronaut, plumber, barber, hairdresser, cashier, dishwasher, nanny, manicurist, bartender, carpenter, programmer, CEO, VP, executive, and accountant.

In order to quantify whether a profession vector is closer to $\overrightarrow{mexican}$ or $\overrightarrow{american}$, we determine its position on the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis by projecting the profession vector onto the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis. The results are shown in Fig.1. The red dotted line across the blue axis is the midpoint in the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis. Ideally, any profession would fall at the point in the blue axis that intersects the red dotted line. But note that prestigious jobs like CEO, Executive, and astronaut lie on the $\overrightarrow{american}$ side and less prestigious jobs like cashier, dishwasher, and bartender lie closer to the $\overrightarrow{mexican}$ side.

Another interesting thing to note is that custodian lies closer to the $\overrightarrow{american}$ side and janitor lies closer to the $\overrightarrow{mexican}$ side. Often, people think custodian and janitor are similar professions, and many times, one of the words is used in place of the other, although custodian seems to have a more positive connotation than janitor. So, out of a pair of words that have very similar definitions, the one with the positive connotation lies on the $\overrightarrow{american}$ side and the one with negative connotation lies on the $\overrightarrow{mexican}$ side.

We also do the same calculations for word2vec (Fig.2) and find that in word2vec, the professions closer to the $\overrightarrow{mexican}$ side also fall in similar positions than in their Glove counterparts. But, the professions closer to the $\overrightarrow{american}$ side actually tend to fall at the midpoint. In other words, the less prestigious professions still fall closer to the $\overrightarrow{mexican}$ side in word2vec, but the more prestigious professions don't lean to either side of the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis. So word2vec has less bias in this sense than Glove.

3.2 Antonym Pairs

In this section, we explore pairs of words that have opposite meanings, i.e., we look at antonyms. More specifically, we look at positive-negative antonyms, which we define as those antonyms where one word has a positive meaning and the other word has a negative meanings, for example, sober and drunk. We take the difference of the positions of the projections of the word embeddings on the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis, more specifically, we take the projection onto $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis of the positive word embedding and subtract the projection onto $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis of the

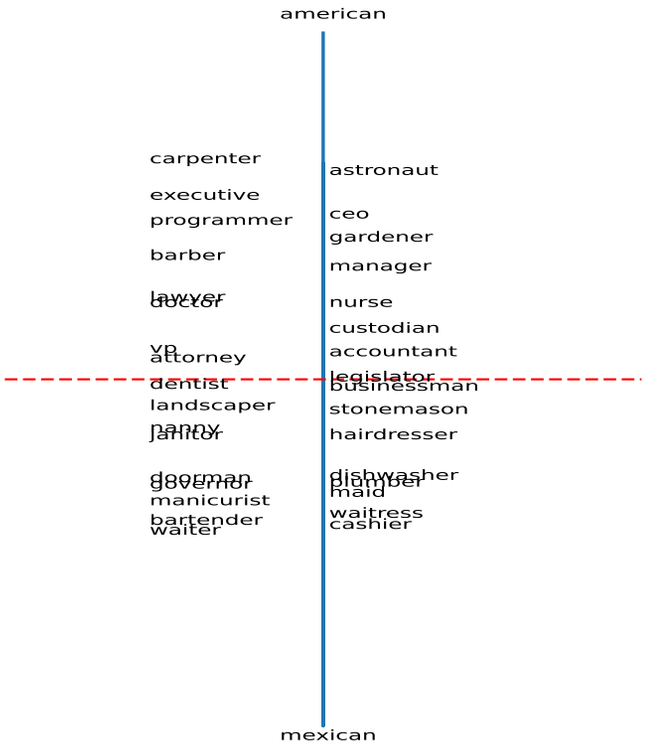

Fig. 1. Positions of different professions on the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis.

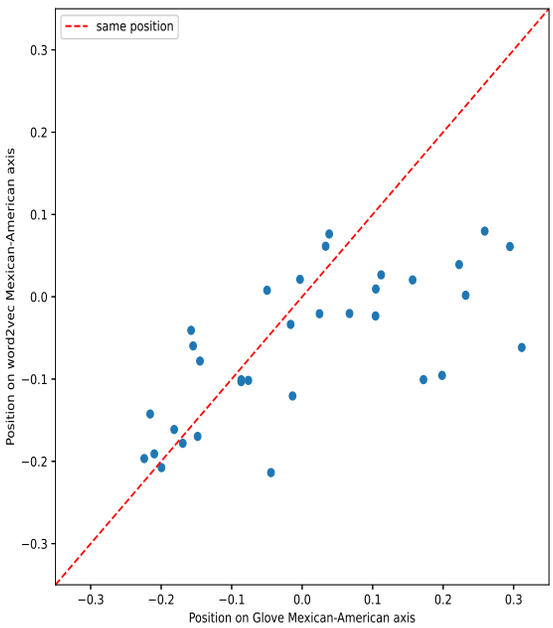

Fig. 2. Comparison of positions of professions on $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis in Glove and word2vec.

negative word embedding. The lower or more negative this difference, which we call displacement, the more the difference between how the positive word relates to the word american and how the negative word relates to the word mexican.

We take 25 random pairs of antonyms and plot the displacement in Fig.3. We see that when we move from the projection onto the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis of the positive antonym to the projection onto the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis of the negative antonym, we always have to move towards the $\overrightarrow{mexican}$ side of the axis, i.e., the positive words, in the case of these antonyms, are always more related to american than the negative antonyms, or in more disappointing terms, the negative antonyms are always more related to mexican than the positive antonyms. The largest displacements are seen on the most offensive antonym pairs, such as the displacements on sober to drunk and intelligent to stupid.

When looking at the displacements for the same positive-negative antonym pairs on word2vec (Fig.4), we see that the displacements have a different behavior than Glove. The word2vec displacements are as likely to be from the $\overrightarrow{american}$ to $\overrightarrow{mexican}$ side than from the $\overrightarrow{mexican}$ to $\overrightarrow{american}$ side. In this sense, word2vec doesn't seem to have the same bias as Glove.

3.3 Analogies

Similar to Bolukbasi et al. [2], we wanted to see whether the bias in the embeddings would produce stereotypical analogies for the $\overrightarrow{american}$-$\overrightarrow{mexican}$ bias. We tried different sets of analogies, but did not find any steoreotypical analogies produced due to the $\overrightarrow{american}$-$\overrightarrow{mexican}$ bias. For example, the analogy *american* is to *businessman* as *mexican* is to ___, returned *businessman*. The same was true of the words manager, legislator, doctor, dentist, astronaut, and basically any other profession that we tried.

4 Bias and Debiasing Algorithm

Bolukbasi et al. [2] showed a way to measure bias in a word embeddings depending on a bias direction. Their metric, $DirectBias_c$ is defined as:

$$DirectBias_c = \frac{1}{|N|} \sum_{w \in N} |cos(\vec{w}, g)|^c \qquad (1)$$

where c is a parameter to determine how strict to be when measuring bias, $\vec{w}$ is a word embedding, and g is the bias direction. Note that since we're working with normalized vectors, then $cos(\vec{w}, g)$ is a measurement of how large the projection of the word embedding w is on the bias direction g, i.e., how much bias w contains. Therefore, the metric calculates the average bias of a set of words N.

DirectBias measurement. In our case, when we take our set N to be the 33 professions we listed in 3.1 and $c = 1$ (same as Bolukbasi et al. [2], so we can compare our numbers) then we get $DirectBias_1 = 0.12$. Note that Bolukbasi et al found w2vNEWS to have a DirectBias measurement

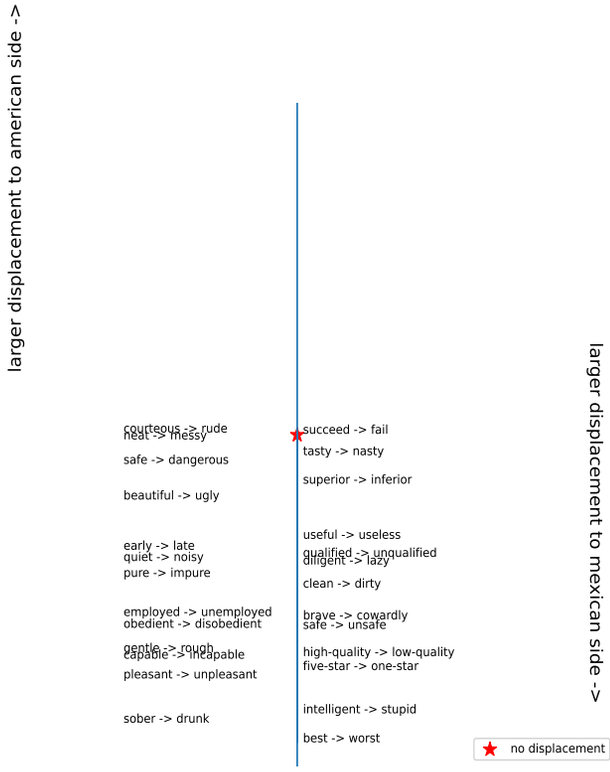

Fig. 3. Displacement of positive-negative antonym pairs

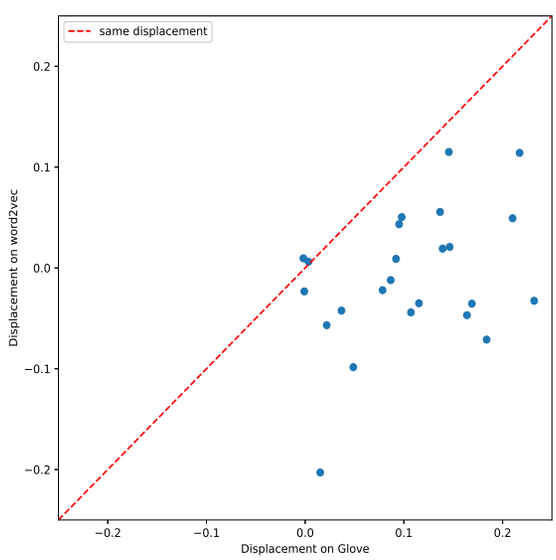

Fig. 4. Comparison of displacements in Glove and word2vec

of 0.08 when taking their list of 327 professions. Our numbers are not directly comparable, since Bolukbasi et al take a much larger set of professions so that their number may be closer to the gender-*DirectBias* of professions than our number to the full mexican-*DirectBias* of professions.

Application of Debiasing algorithm. We applied the Hard Debias debiasing algorithm from Bolukbasi et al [2]

to the $\overrightarrow{american}$-$\overrightarrow{mexican}$ bias. There are two main components of the Hard Debias algorithm: neutralize and equalize. Neutralize basically removes the projection of a word embedding onto the bias direction by subtracting out the projection from the word embedding. Note that this step makes the $DirectBias = 0$. The equalize step takes pairs of words and makes sure that every neutral word is equidistant to each word in the pair.

The debiasing algorithm neutralizes all words, except words that should not be debiased, for example, in the case of gender, there are gender-specific words like grandfather and grandmother that should not be debiased. In our case, we did not include any words not to debias since it does not seem that there are words that should be more mexican and others that should be more american.

The debiasing algorithm also takes pairs of words to define the bias subspace. For example, in the case of the gender bias, he-she and him-her are two sets of words that can be used to determine the gender bias subspace. In our case, we only included the pair mexican-american, for the same reason as above, we do not think a word can be defined as "mexican" and another as "american", except those words themselves.

5 Effects of debiasing on downstream models

Many practitioners may be afraid that because the debiasing algorithm "moves" the word embeddings around, then the word embeddings may lose their usefulness when building models. We show, using a simple sentiment classifier, that word embeddings retain their usefulness even after we debiased using the Hard Debias algorithm.

5.1 Simple sentiment classifier

We implement the simple sentiment classifier described in Speer [4]. The main steps in building this sentiment classifier are the following:

1. Get a list of words with negative sentiments and a list of words with positive sentiments. We use the same lists as Speer.
2. Get the word embeddings for each of the previous words. The embeddings are used as input to a logistic regression model that predicts whether a word has positive or negative sentiment.
3. A word's sentiment score is defined as the difference of the log probability of having positive sentiment minus the log probability of having negative sentiment.
4. A sentence's sentiment score is defined as the average sentiments of all the words in the sentence.

This model may seem too simple and one may think that this is not a model that is realistically used in the real world, but the simpler a model is, the less the barrier to a practitioner understanding the model and therefore the more likely the practitioner is to implement the model. Also, the simpler the model, the more likely it is to already be applied in some industry. This is because legacy systems are built on older methods, and there were no deep learning methods for sentiment scores a few years ago.

5.2 Sentiment Scores before debiasing

Before debiasing the word embeddings, our model is able to achieve an accuracy of 0.90, so it is very accurate at being able to predict whether a word has a positive or negative sentiment.

But we also see that our model scores $\overrightarrow{mexican}$ with a sentiment of -0.33 and $\overrightarrow{american}$ with a sentiment of 1.37, i.e., the model thinks the word $\overrightarrow{mexican}$ is more likely to be a negative-sentiment word than a positive-sentiment word and $\overrightarrow{american}$ is more likely to be a positive-sentiment word than a negative-sentiment word, and the model built this belief on just the positions of the words in the word embedding space. This is another proof that the Glove embeddings have a built-in bias against Mexicans.

Also, since in this model a sentence's sentiment score is the average sentiment of the words in the sentence, then this means that exchanging the word american in a sentence for mexican will lower the sentiment score of the sentence, e.g., the sentence "This restaurant has great mexican food" will have a lower sentiment score than the sentence "This restaurant has great american food", even though these two sentences should have the same sentiment score. As pointed out in Section 1, this type of behavior can be detrimental to business owners of under-represented groups.

5.3 Sentiment Scores after debiasing

After debiasing the Glove word embeddings using the Hard Debias algorithm, we find that our simple model still has an accuracy of 0.90, i.e., our model's accuracy has not suffered due to the algorithm moving the word embeddings around.

Also, the word embeddings $\overrightarrow{mexican}$ and $\overrightarrow{american}$ both have the same sentiment score of 1.08. Therefore, in this case, whether a sentence contains the word mexican or american would make no difference on the sentiment score of a sentence, as it should be. This type of behavior makes sure that a model would not harm under-represented groups.

6 Conclusion

We have shown the Glove embeddings to have a substantial amount of bias against Mexicans, more specifically, less prestigious professions are more related to the word embedding $\overrightarrow{mexican}$ than to the word embedding $\overrightarrow{american}$. Also, we showed that when looking at positive-negative antonym pairs, the positive words are more associated with the american side of the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis than negative words, or in other words, the negative words are more associated with the mexican side of the $\overrightarrow{american}$-$\overrightarrow{mexican}$ axis than positive words.

We then applied the Hard Debias algorithm to the Glove word embedding to remove the $\overrightarrow{american}$-$\overrightarrow{mexican}$ bias, and we showed that it made a simple sentiment classifier fairer

and the simple sentiment classifier did not lose any accuracy after the word embeddings were debiased.

References

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III au2, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp, 2020.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016.

[3] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019.

[4] Robyn Speer. How to make a racist ai without really trying, Jul 2017.