Exploring and Mitigating Gender Bias in GloVe Word Embeddings

Ma Francesca Luisa C Vera
Department of Computer Science
Stanford University
fvera@stanford.edu

Abstract

When societal biases are discovered within items, it is natural to consider ways in which it is possible to remove those biases. In the past, language has been shown to carry certain biases including those that perpetuate gender and racial stereotypes. Consequently, much research has been done on how to better counteract these biases within language. In conjunction with this research is natural language processing, and the growing use of technology to solve linguistic tasks has led to many considerations about the biases algorithms, models, and tools may carry. One example of such a tool is word embeddings, which give words corresponding numerical values. Within these embeddings are gender biases against occupations that ought to be gender-neutral, but are often stereotyped towards male or female genders. We demonstrate that within the GloVe word embeddings, these occupations are stereotyped since similarities of these occupations to embeddings for “he” and “she” produce clearly different results. The correlation coefficient for “masculine” occupations is 0.91 against the 0.83 coefficient for “feminine” occupations. To debias these words, we propose finding a gender subspace in which gender-neutral words are placed at the 0 position – while still maintaining the embeddings of gender-specific words such as “man” and “woman”. Because the new gender-neutral words are given embeddings such that they are equidistant between corresponding pairs of gender-specific words, it is expected that when looking at the stereotyped occupations, the embeddings of these occupations will be similar to males and females when tested. When we test for correlation coefficients against “masculine” and “feminine” occupations, we get 0.99 and 0.97 respectively, demonstrating how gender bias has been mitigated.

1 Introduction

In recent years, there has been increasing concern regarding the degrees of bias that might exist in technological tools used to perform tasks for societal problems. It is important to understand these biases because these tools may then perpetuate even further bias in the results of a task. One area that has drawn the attention of academics in recent years is gender bias in word embeddings. Word embeddings include a set of vectors that correspond to a specific word, and these vectors relate to each other in the same ways the words do. Thus, word embeddings are powerful because they can be used to solve natural language processing tasks that may require this transformation between words and numerical values. However, recent work has called into question whether or not these word embeddings promote societal biases including those related to race and gender. If this is the case, one critical responsibility of the users of these word embeddings is to ensure that these biases are mitigated before the embeddings are put to use.

There are many different sets of pretrained word embeddings available for public use, including the GloVe pretrained word embeddings. The GloVe word embeddings include sets that were trained on billions of tokens, some up to 840 billion tokens. It is available for download online, making it a popular source for word embeddings in the NLP space. When looking at such an influential tool, any semblance of bias can influence the results of a task gravely.

This paper aims to highlight some of the gender biases that exist in the GloVe word embeddings before adapting a past debiasing method to mitigate the biases in these embeddings. It also experiments with the concept of gender-specific words using GloVe word embeddings, exploring whether or not it is possible to classify a large set of embeddings as gender-specific or not.

2 Related Work

In 2016, Bolukbasi et al. released a paper, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, which pioneered this space of debiasing word embeddings.[1] Their work involved looking at the lists of occupations that most clearly exhibited gender stereotypes and looking at gender stereotyped he : she analogies. In this context, bias was defined as showing what should be gender neutral occupations or analogies are favored in the direction of one gender over another. Using w2vNEWS embeddings, they first identify a gender subspace before "neutralizing" and "equalizing". (Further explained in Section 4) For the purposes of this paper, this is the method I will be adapting and applying to the GloVe word embeddings.

Another important piece of work by Chakraborty et al., *Reducing gender bias in word embeddings*, uses GloVe vectors to .[2] Similarly to Bolukbasi et al., Chakraborty et al. look at modifying the actual embeddings so that they do not exhibit gender biases. To do so, they alter the settings in which the embeddings are trained, taking the co-occurrence matrix of an occupation and scaling these so that the co-occurrence probabilities of these occupations becomes 1. After using cosine similarity between embeddings and a gender direction, they also aim to reduce bias by adding a regularization term to the objective function that "penalizes" similarity to the gender direction, doing so for only the biased occupational words.

Recently, there has been growing interest in the applications of adversarial learning. In 2018, Zhang et al. released a paper, *Mitigating Unwanted Biases with Adversarial Learning*, which discusses fairness and attempts to create "more fair" algorithms in different societal settings.[3] One such setting was natural language processing, particularly gender bias in word embeddings. Using embeddings trained from Wikipedia to generate input data, Zhang et al. set up the "he : she" analogy test as a supervised learning task where the model would pick the word corresponding to "she" after being given a word analogous to "he." To debias the system, they added an "adversarial discriminator network" that would have trouble guessing the gender direction of an output y. Thus, the embeddings remain unchanged but the space they exist in differs such that they do not strongly perpetuate gender stereotypes from the analogy task.

3 Demonstrating Bias in GloVe Word Embeddings

The first step to mitigating bias in anything is to show that there currently exists some form of bias in the thing one is trying to debias. In this case, we aim to show that the GloVe word embeddings demonstrate an inherent gender bias.

Borrowing a method from Bolukbasi et al (mentioned in the related work section), we will show that when looking at occupations that are supposed to be gender-neutral, there are inherent gender biases built into the word embeddings of these occupations. To show these biases, we will map out the similarities these occupations have to the embeddings of "he" and "she" – if an embedding appears more similar to "he" than "she", it would appear that that occupation tends towards males – and vice versa. The occupations have been split into "he" occupations and "she" occupations, with the understanding that these are the "most biased" occupations in either a he or she direction. To calculate similarity, we use both the inner product of the embeddings and the cosine similarity. Finally, to show that gender biases exist across different types of embeddings, we use both the GloVe word embeddings pretrained on Wiki and pretrained on Common Crawl.

Plotting all the occupations and their respective similarities to "he" and "she" gives us a qualitative overview on how biased these words are. To quantify the results, we also look at the correlation coefficients for the "he" occupations against the "she" occupations, with coefficients that are nearly equal indicating a lack of bias.

3.1 Using occupations from Bolukbasi et al.

It makes sense that the results rely heavily on which predetermined "he" and "she" occupations are used, so we decided to experiment using different groups. The first set of occupations was taken from the Bolukbasi et al paper.[1] To determine these occupations, they found the most extreme occupations as projected onto a gender direction, and labeled these as "Occupational Stereotypes." The lists of occupations are as follows:

"He" Occupations: *["maestro", "skipper", "protege", "philosopher", "captain", "architect", "financier", "warrior", "broadcaster", "magician", "pilot", "boss"]*

"She" Occupations: *["homemaker", "nurse", "receptionist", "librarian", "socialite", "hairdresser", "nanny", "bookkeeper", "stylist", "housekeeper", "designer", "counselor"]*

These occupations were determined by Bolukbasi et al. as the ones that carried the greatest degree of gender bias within them. To verify, we turn to the GloVe word embeddings and our method of determining gender bias. When using both cosine and inner product to calculate similarity, the gender bias is very clear. Below (Table 1) are the results for the correlation coefficients after running our experiments on the Wiki-trained GloVe word embeddings. Because the coefficients for "he" occupations and "she" occupations are not nearly equal for either type of similarity, there is clearly some form of gender bias within the embeddings.

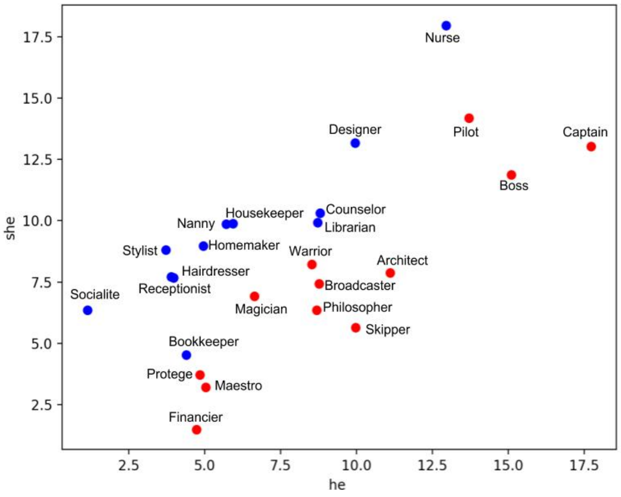
Table 1: Correlation coefficients using Wiki-trained Glove word embeddings

Correlation Coefficients	"He" Occupations	"She" Occupations
Cosine Similarity	0.9147	0.8322
Inner Product Similarity	0.9116	0.8820

An easier way to visualize gender bias is by plotting the occupation's similarities to "he" and "she" against each other. This can be seen in figures 1-4, which show the results for both the Wiki-trained GloVe word embeddings and the Common Crawl-trained GloVe word embeddings. Each red point represents a "he" occupation and each blue point represents a "she" occupation – with the charts showing which occupation corresponds to which point. For all charts, there is a very obvious split between occupations (blue and red points), which corresponds to the calculated coefficients.

Figure 1(left): Inner product similarities using Wiki-trained word embeddings

Figure 2 (right): Inner product similarities using Common Crawl-trained word embeddings



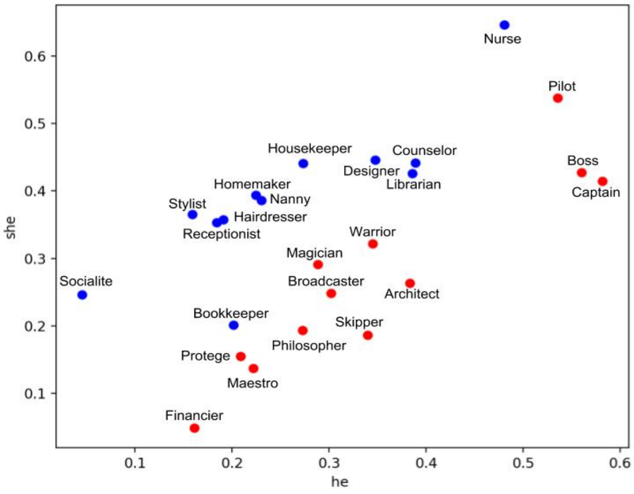
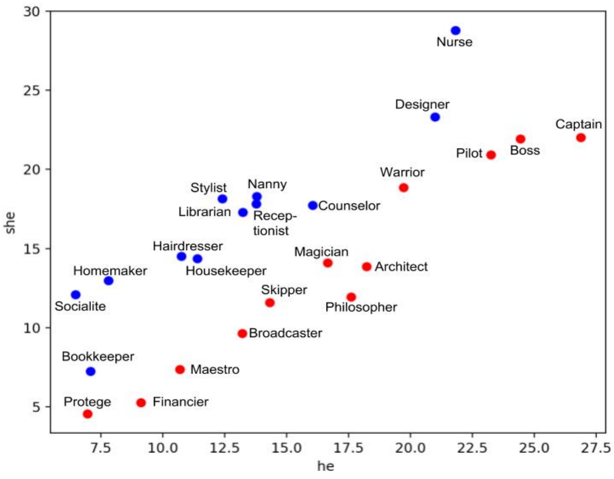
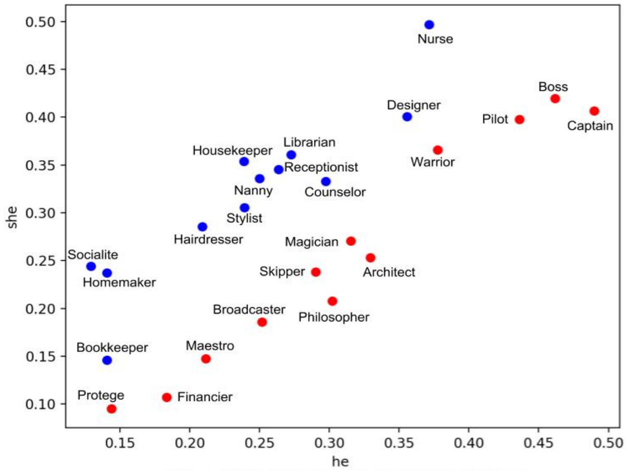


Figure 3 (left): Cosine similarities using Wiki-trained word embeddings

Figure 4 (right): Cosine similarities using Common Crawl-trained word embeddings





3.2 Using "most similar" occupations

We also looked at the occupations that had the closest similarities to "he" and "she" word vectors, and then plotted those occupations to view whether or not they carried as clear of a bias as the occupations from Bolukbasi et al. did. To find these occupations, we first downloaded a list of 1155 general occupations. For each occupation on the list, we calculated its similarity to "he" word embedding and "she" word embedding. After sorting the list, we found the twelve closest occupations to each embedding – deeming those "he" and "she" occupations. The results are as follows. In bold are words that also appeared in Bolukbasi et al.'s list. Underlined are words that appear in both "he" and "she" lists.

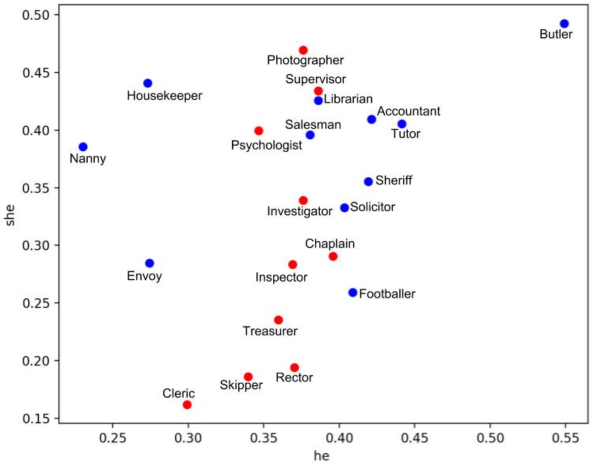
Table 2: List of occupations "most similar" to "he" and "she"

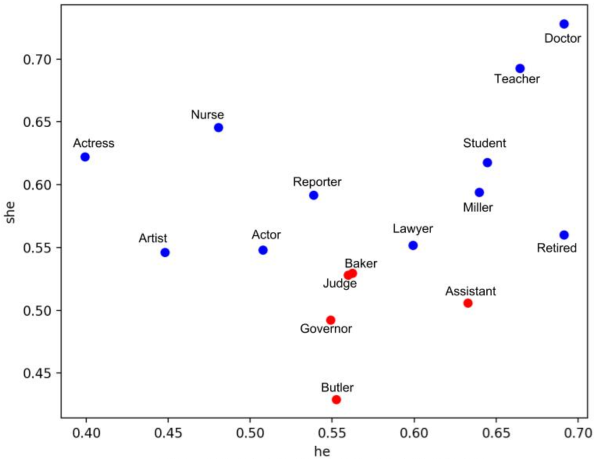
	"He" Occupations	"She" Occupations
Cosine Similarity	*["retired", "doctor", "teacher", "student", "miller", "assistant", "lawyer", "baker", "judge", "governor", "butler"]*	*["doctor", "**teacher**", "**nurse**", "actress", "student", "miller", "reporter", "retired", "lawyer", "actor", "artist"]*
Inner Product Similarity	*["cleric", "photographer", "**skipper**", "chaplain", "accountant", "inspector", "rector", "investigator", "psychologist", "treasurer", "supervisor"]*	*["**librarian**", "**housekeeper**", "**nanny**", "accountant", "sheriff", "envoy", "tutor", "salesman", "butler", "footballer", "solicitor"]*

The results are very interesting because there is a lot of overlap between the two lists and only few similarities with the Bolukbasi et al. list. The method also seemed to better at identifying biased "she" occupations than "he" occupations – although some anomalies like "actor" appear. There are several reasons for this outcome. Firstly, the results were influenced by the initial list of 1155 occupations; a different list of initial occupations for comparison would have yielded different results. Furthermore, Bolukbasi et al. added an extra element of a "gender direction" in their calculations whereas we only use cosine or inner product similarity.[1] That being said, it is encouraging that when plotted using the same method to measure bias as above, there is still some observed gender bias among occupations – although not as clear as before.

Figure 5 (left): Cosine similarities of occupations found through inner product

Figure 6 (right): Cosine similarities of occupations found through cosine





Consequently, we decided to continue using the Bolukbasi et al. occupations for our evaluation of debiasing.

4 Debiasing GloVe Word Embeddings

After demonstrating that the GloVe word embeddings carry some gender bias within them, especially for "gendered" occupations, we then tried to debias these word embeddings so that when looking at these occupations again, no clear bias would be reflected. As mentioned in the related work section, there have already been a few academic studies regarding the debiasing of word embeddings. For continuity, we adopt the method in Bolukbasi et al.'s paper since our evaluation was based on findings from that paper.

4.1 Approach for mitigating bias

The method from Bolukbasi et al. was based on two important steps:

1. Identifying the "gender subspace"
2. Neutralizing or Equalizing words appropriately to form a new set of embeddings

The first step of identifying the gender subspace involves identifying a direction in which the embedding carries some gender bias. To calculate this gender subspace, we first identify a gender direction by looking at "definitional" pairs that help it orient towards the genders. For those words that are not gender specific, the new embedding is the difference between the former embedding and gender direction, multiplied by the embedding dot the gender direction, divided by the gender direction dot itself. For words that are gender specific, their embeddings are maintained.

Next, we decide whether or not a word should be equalized or neutralized. If a word is gender neutral, we neutralize it such that they are at position 0 in the gender subspace. Equalize then looks at pairs of corresponding gender-specific words so that we can enforce any gender-neutral word to be equidistant from these corresponding pairs. From there, we collect all the altered embeddings to create a new set of word embeddings with limited bias.

4.2 Results of debiasing GloVe word embeddings

To evaluate our method of debiasing, we will use the same methods as we did for showing that the initial GloVe word embeddings carried some form of gender bias. This means that when we use the new debiased embeddings in our methods of evaluation, we should not see a clear divide between “he” and “she” occupations. This applies to plotting the corresponding similarities and having the same correlation coefficients.

Table 3: Correlation coefficients using debiased word embeddings

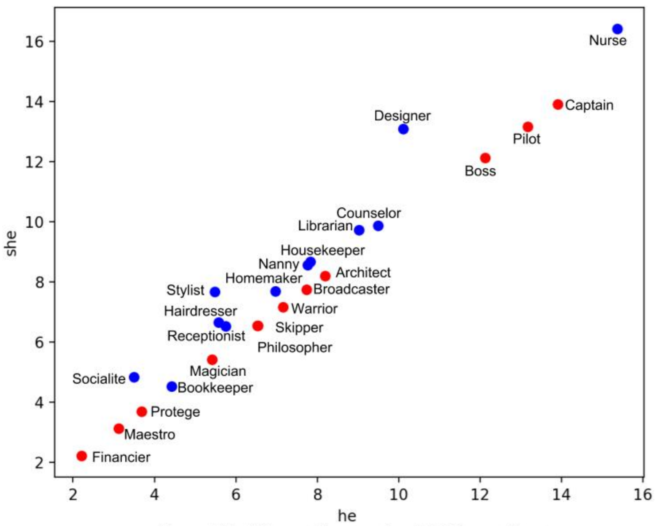
Correlation Coefficients	“He” Occupations	“She” Occupations
Cosine Similarity	0.9890	0.9688
Inner Product Similarity	0.9921	0.9723

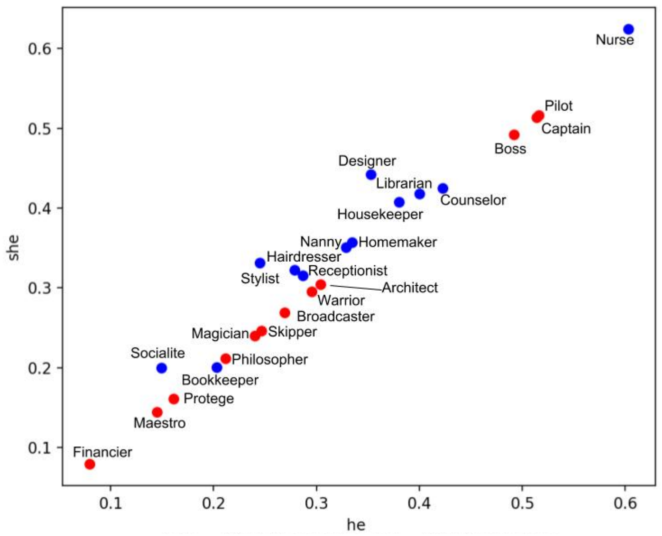
We observe in table 3 that the correlation coefficients for “he” occupations and “she” occupations when using the new debiased embeddings are nearly the same. This also holds true for when we use either cosine similarity or inner product similarity.

Now that we have verified numerically that the initial gender bias has been reduced, we turn to the qualitative methods of evaluation to see whether or not the differences between “he” and “she” occupations have changed visually. Figures 7 and 8 show the points for all gendered occupations when the similarities have been calculated with the new word embeddings. As we can see, unlike the previous plots where “she” occupations tended towards the “she” vector and “he” occupations tended towards the “he” vector, the ratio between similarities to “he” and “she” are nearly 1:1 for all occupations – putting what were once gendered occupations in the middle of the “he” and “she” genders.

Figure 7: Inner product similarities using newly debiased embeddings

Figure 8: Cosine product similarities using newly debiased embeddings





It is encouraging that both quantitative and qualitative methods of evaluation reflect some change in the word embeddings. This means that we were successful in our task of mitigating the gender bias in the GloVe wiki-trained word embeddings and can use the new debiased embeddings in other NLP tasks.

5 Extra Experiments (on Gender-Specific Words)

In addition to the task of debiasing a set of word embeddings, we were inspired to explore the concept of gender-specific words in a similar way to what was done in the Bolukbasi et al. paper. Some words such as “he”, “she”, “mom”, “dad” etc. are specific to gender and so do not carry gender biases. There are two interesting tasks that can be done regarding gender-specific words:

1. By using a pre-labeled set of gender-specific/non-gender-specific words and their embeddings, can we classify a broader list of words as either gender-specific/non-gender-specific?
2. Can we also use the existing set of gender-specific words to find additional gender-specific words from a larger list of general words?

The list of gender-specific words were taken from Bolukbasi et al. [1], which took a subset of 218 words from w2vNEWS and looked at their Wordnet definitions to determine if there was some element of gender inherent to the definition. Some of the words included: *['he', 'his', 'her', 'she', 'him', 'man', 'women', 'men', 'woman', 'spokesman', 'wife', 'himself', 'son', 'mother', 'father', 'chairman', 'daughter', ..., 'fatherhood', 'councilwoman', 'princes', 'matriarch', 'colts', 'ma', 'fraternities', 'pa', 'fellas', 'councilmen', 'dowry', 'barbershop', 'fraternal', 'ballerina']*. As Bolukbasi states, the words are highly "subjective" and encourages customization to the application.[1] One can imagine that in the application of occupations, words such as "councilwoman" will be more important to include than "colts". For the purposes of this paper, we will use this set of words to perform the two tasks at hand.

5.1 Classifying gender-specific words

The first task we experimented on was the classification of words as either gender-specific or non-gender-specific. To do this, we trained a linear SVC using C = 1.0 on a subset of GloVe word embeddings where any gender-specific word in the subset was labeled "1", and the rest were considered gender-neutral so labeled "0". According to Bolukbasi et al.'s results on the same task, the binary accuracy is expected to be "well over 99% due to the imbalanced nature of the classes." In the results below, we find that we are able to achieve the same score on different numbers of iterations. Another metric used by Bolukbasi et al. was the F-score from 10-fold cross-validation. They achieved 0.627. Our results vary much more in this metric. Although we are able to beat the score when there are only 5000 words used in the training subset, it is expected that this is nowhere close to the 50 000 used by Bolukbasi et al. Thus, we significantly underperform in this regard when the number of words used to train the model increases.

Table 4: Results from classification of gender-specific words task

Number of Words in Subset	Accuracy	10-fold Cross-Validation Score
5000	0.9996	0.7998
7500	0.9994	0.6998
10 000	0.9994	0.6597
20 000	0.9994	0.4998
100 000	0.9994	0.4998

5.2 Identifying gender-specific words

The second task involves finding even more gender-specific words from an initial set of gender-specific "seed" words. To do so, we first train a linear SVC on a subset of GloVe word embeddings and labels where any gender-specific word that appears is labeled "1". In the previous subsection, we demonstrate that setting up a classifier of this sort achieves high accuracy and reasonable cross-validation scores at certain numbers of iterations. After training on this model, we found the model's coefficient and intercept. Subsequently, for each word embedding in the larger list of general words, the dot product of that embedding was taken with the model's coefficient and then compared to the intercept, resulting in a list of words taken from the list supposedly with some gender-specificity.

This method yielded surprising results. There is some evidence of success as our model was able to extract the following gender-specific words: *[“macho”, “dude”, “gentleman”, “dads”, “guys”, “gunman”, “man”, “mommy”, “guy”, “woman”, “spokesman”]* among a few others from the model. However, some gender-neutral words such as “kid” and “somebody” were added to the list. An interesting observation is that the model also pulled some occupations as gender-specific, including “politician” and “cop” – hinting at the occupational gender biases that were discussed earlier in the paper.

6 Conclusion

It is important to be conscious of the inherent biases that might exist in technology because using these tools will perpetuate the inherent biases they hold in whatever tasks the tools are trying to perform. Word embeddings are no exception to this concept, and this work demonstrates how the GloVe word embeddings, a popular set of word embeddings, also carry biases with regards to gender stereotypes.

The ability to understand bias, determine where it exists, and then mitigate it is important in better understanding the biases that crop up in our world. In the context of NLP, our methods successfully create a new set of word embeddings that have limited gender bias. There is potential for further research using the foundation we built by testing the techniques in this paper against some standard NLP tasks such as the analogy task. The analogy task is useful in determining whether the embeddings encourage gender stereotypes within analogies.

Note that our work in debiasing changes the word embeddings directly. However, there has been work in debiasing that alters the gender bias space rather than altering the embeddings themselves. Consequently, it would be worthwhile to explore if this can be done using GloVe word embeddings. Although this method does not leave users with a new set of debiased embeddings to use, it demonstrates that gender biases have been recognized on many fronts and there exist successful attempts to mitigate these biases in the technological world. Overall, there is a lot of promising work out there that aims to mitigate societal biases within technology today.

Acknowledgments

Thank you to the entire CS224n staff for support and guidance on this project.

References

[1] Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349– 4357. 2016.

[2] Chakraborty, T.; Badie, G.; Rudder, B.. Reducing gender bias in word embeddings. 2016.

[3] Zhang, B. H.; Lemoine, B.; Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In *arXiv preprint arXiv:1801.07593*. 2018.

[4] Pennington, J.; Socher R.; Manning C. D.. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532-1543. 2014.