Debiasing Word Embeddings by Removing Gender Projection Effectively

Naveksha Sood
Indiana University Bloomington
soodn@iu.edu

Jhalak Acharya
Indiana University Bloomington
jhacharya@iu.edu

Abstract

Emergence of word embedding has revolutionized the world of computational linguistics. Word embeddings allow us to represent words in number space as vectors and perform algebraic operations on them, making it much easier for computers to understand the words and the relationships between them. But, these embeddings are often sourced from large amounts of unfiltered data present on the internet which is susceptible to inherent bias and stereotypes of real world such as gender bias or racial bias. Evidently the models based on these embeddings are also prone to bias. In an attempt to avoid gender bias seep into models, in this paper, we discuss a method to remove gender bias from word embeddings effectively by selecting a good gender axis. In our experiments, we try to predict the gender that occupies a certain profession using logistic regression model and support vector machine on original and debiased Fasttext embeddings. Results reveal that originally while some professions associated strongly to males such as doctor or females such as nurse; after selecting a good gender axis and removing that projection from our data, our models no longer reflect or at least reflect reduced bias in gender neutral professions. This study is an effective milestone in making word embeddings and eventually the models based on these embeddings more fair and hence, more reliable.

1 Motivation

Bias and stereotypes have been a prevalent part of our world. Hundreds of years have gone into identifying biases in real life and making our world an inclusive space is still work in progress. While we may still be years away from an equal world, we can still make efforts to not let that bias and stereotypes be as large a part of rapidly growing artificial intelligence (AI) space.

AI, or more specifically, machine learning (ML) relies heavily on data from real world. In image space, ImageNet data, marked the emergence of deep learning revolution in 2015, and it relied on copious amounts of crowd sourced image data. Unsurprisingly, the real world stereotypes seeped into the data as well as the models built on the data. Parallely, in natural language space, emergence of word embeddings such as wordtovec, glove, fasttext, trained on large amount of unfiltered data, on one hand, set the stage for major advancements in the field of natural language processing (NLP), and on the other, was susceptible to all kinds of gender, racial, religion-based bias as well as toxicity that we see in real world. Moreover, when pretrained embeddings are used for an ML application, the bias seeps into the model and may also get amplified. Consequently, this is reflected in the real-world implications of the outcomes of these models.

There have been several infamous instances of bias ML models that perpetuate bias in the society. In the recent past, credit-card approval and loan approval have been automated based on attributes of a person. Instances of these systems to prefer wealthier, fairer applicants purely due to the lack of enough training data for the people of color have surfaced (1). In another instance, tech-giant Amazon developed an AI tool to automate the recruitment process where resumes were screened to automate the process of screening candidates. However, due to the dominance of the males in the tech industry and Amazon's workforce which comprised of almost 60% males then, penalized women's candidates since it had learnt this gender bias and had to be soon scrapped (2). Another tech-giant, formerly known as Facebook, used the bias in these embedding for targeted advertising basis gender where the women were shown more stereotyped female advertisements such as those of

secretaries and nurses (3). These instances substantiate the need for debiasing the data on which ML models are trained to prevent any form of gender bias from leaking into the model.

Our primary goal in this paper is to investigate a hard debiasing technique with a focus on appropriate selection of a gender axis to debias the word embeddings. We have used Fasttext pretrained word embeddings to observe the effect of biasing and debiasing word embeddings. Further, we have analyzed accuracy of different models trained using Logistic Regression and Support Vector Machine to analyze the effect of debiasing word embeddings using the task of predicting gender given a profession. The results for these tasks have been detailed in the further sections for both labelled and unlabelled data.

2 Methodology

2.1 Bias in Word Embeddings

Word embeddings are arrays of numbers which are learnt to represent words. They can be considered as vectors in an n-dimensional space where n in the dimension of each word embedding. Since, these are vectors they can be projected in an n-dimensional space with each vector pointing in a certain direction. These vectors support mathematical operations, such as addition and subtraction. For instance, ideally,

$mother + father + child = family$

should be true.

Similarly, $he - she$ gives us information about $gender$. Likewise, $man - woman$ also gives us information about $gender$. Gendered pair of words are expected to lead to gender vector, however, if such a trend can be seen for gender-neutral words, it can be inferred that there exists inherent bias in these word embeddings.

In our experiments, we calculated the difference between appropriate and stereotyped gendered pairs of words using fasttext embeddings and plot the distance using the cosine metric as shown in figure 1. The vector for $man - woman$ appears closer to $he - she$, $boy - girl$, and so on which is understandable; however, such similarity or closeness can also be seen associated with $man - woman$ and stereotyped pairs of $surgeon - nurse$, $engineer - ballerina$ and $carpentry - sewing$. These substantiates the existence of such gender bias in fasttext word embeddings.

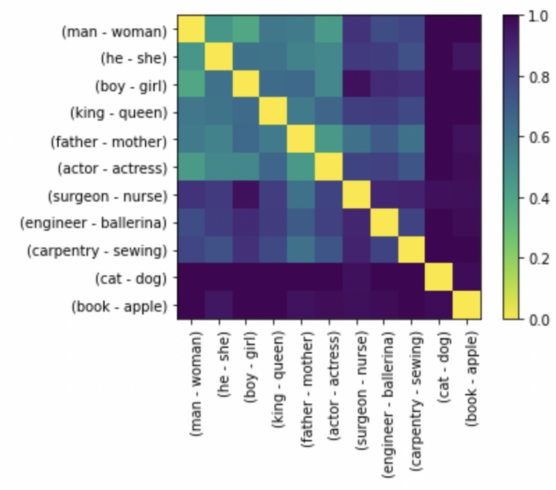

Figure 1: Cosine distance between fasttext embeddings of word pairs

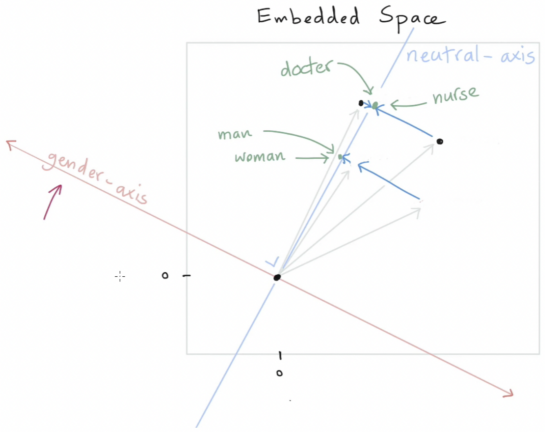

Figure 2: Finding gender axis in embedded space

2.2 Debiasing Mechanism

Now that we can see that there exists gender-related information in gender-neutral pretrained word embeddings, we try to project all the word embeddings in a space orthoganal to the gender axis, thereby removing the gender projection from the data as shown in figure 2. In order to get the gender axis, we calculate the average difference between appropriate gendered pairs of words. Theoretically, to get a better estimate of gender axis, we must take average across several pairs of gendered words which are similar in all other regards. We explore this hypothesis in our experiments.

2.3 Finding the Gender Axis

For our experiments detailed in the next sections and to determine the effect of the selection of a good gender axis on debiasing word embeddings, we have considered the following 2 cases:

- Case 1: Taking the average difference between only 2 gender appropriate pairs: *man–woman* and *he–she*.

- Case 2: Taking the average difference between 14 gendered pairs: $man - woman$, $he - she$, $him - her$, $masculine - feminine$, $boy - girl$, $actor - actress$, $king - queen$, $father - mother$, $son - daughter$, $husband - wife$, $brother - sister$, $bachelor - spinster$, $uncle - aunt$ and $manservant - maidservant$.

While Case 1 gives us a representative of a gender axis, the expectation is that in Case 2 the gender axis would eventually be closer to the true gender axis, analogous to the Law of Large Numbers (4).

2.4 Data

We considered a list of most popular 223 male stereotyped/appropriate professions, 222 female professions and 320 gender neutral professions, along with 1000 general words and 1440 gendered words. We use Fasttext embeddings for obtaining their vector representations.

3 Experiments Results

3.1 Cosine distance after debiasing word embeddings

We find the 10 most similar words to several words in the original embeddings and after projecting the embeddings to a gender neutral space by finding gender axis in two ways as mentioned in section 2.3. We observe that after debiasing the word king, queen moves closer to king from a distance of 0.293 to 0.251 for case 1 and 0.193 for case 2, and becomes the most similar word to king as shown in figure 3 and 4.

	idx	king	king_score	king_debias	king_debias_score
0	1	king	0.000	king	0.000
1	2	kings	0.245	queen	0.251
2	3	queen	0.293	kings	0.255
3	4	prince	0.350	prince	0.373
4	5	princes	0.428	princes	0.439
5	6	emperor	0.447	princess	0.451
6	7	lord	0.448	queens	0.453
7	8	duke	0.462	emperor	0.480
8	9	princess	0.485	lord	0.485
9	10	lords	0.492	duke	0.485

Figure 3: King - Removing gender projection found using Case 1

Similarly, nurse did not appear in the 10 most similar words to doctor but after debiasing it appears in the list of 10 most similar words at the 9th position for case 1 and 5th position for case 2 as shown in figure 5 and 6.

	idx	king	king_score	king_debias	king_debias_score
0	1	king	0.000	king	0.000
1	2	kings	0.245	queen	0.193
2	3	queen	0.293	kings	0.263
3	4	prince	0.350	prince	0.376
4	5	princes	0.428	princess	0.412
5	6	emperor	0.447	queens	0.415
6	7	lord	0.448	princes	0.442
7	8	duke	0.462	empress	0.457
8	9	princess	0.485	emperor	0.483
9	10	lords	0.492	duke	0.484

Figure 4: King - Removing gender projection found using Case 2

	idx	doctor	doctor_score	doctor_debias	doctor_debias_score
0	1	doctor	0.000	doctor	0.000
1	2	physician	0.224	physician	0.229
2	3	gynecologist	0.295	gynecologist	0.289
3	4	pediatrician	0.308	pediatrician	0.309
4	5	pharmacist	0.332	pharmacist	0.335
5	6	surgeon	0.338	surgeon	0.347
6	7	neurologist	0.339	dermatologist	0.347
7	8	cardiologist	0.347	neurologist	0.349
8	9	dermatologist	0.347	nurse	0.354
9	10	psychiatrist	0.354	cardiologist	0.357

Figure 5: Doctor - Removing gender projection found using Case 1

	idx	doctor	doctor_score	doctor_debias	doctor_debias_score
0	1	doctor	0.000	doctor	0.000
1	2	physician	0.224	physician	0.227
2	3	gynecologist	0.295	gynecologist	0.280
3	4	pediatrician	0.308	pediatrician	0.309
4	5	pharmacist	0.332	nurse	0.329
5	6	surgeon	0.338	pharmacist	0.332
6	7	neurologist	0.339	dermatologist	0.342
7	8	cardiologist	0.347	surgeon	0.344
8	9	dermatologist	0.347	neurologist	0.346
9	10	psychiatrist	0.354	cardiologist	0.354

Figure 6: Doctor - Removing gender projection found using Case 2

Hence, Case 1 shows the effectiveness of our debiasing approach and case 2 shows the effectiveness of selecting an appropriate gender axis. However, our understanding is ossified by the following experiments which we conducted.

3.2 Classification of professions using original and debiased word embeddings

This set of experiments include classification of a profession as male or female oriented profession. We trained a logistic regression (LR) and support vector machine (SVM) and evaluated our results for two case scenarios on orginal and debiased embeddings. We expect that if debiasing, in fact, has occurred, the model should be able to predict a profession as male or female with almost equal probability or at the least, the difference in prediction probabilities should reduce; i.e. our model should be fairly confused while assigning a gender to a profession.

For a complete and cohesive analysis of debiasing, we tested our models in 3 ways:

- Scenario 1: Training the model on original embeddings and testing them on original embeddings.
- Scenario 2: Training the model on original embeddings and testing them on debiased embeddings.
- Scenario 3: Training the model on debiased embeddings and testing them on debiased embeddings.

To further evaluate the impact of our debiasing technique, we used the test dataset of gender neutral professions and analyze the probability of predicting a profession as male and female.

4 Results

The results for our experiments on fasttext embeddings are summarized in the table 1 and 2.

	Scenario 1	Scenario 2	Scenario 3
LR	0.90	0.69	0.85
SVM	0.90	0.73	0.87

We can observe from the results on case 1 that when a model is trained with the original word embeddings, it is successful in predicting the stereotyped or labelled gender associated with it on the validation set with an accuracy of 0.90 for both LR and SVM. Further, when such a model is used to determine the gender of a profession with debiased embeddings, there is a drop in accuracy to 0.69 and 0.73 for LR and SVM respectively. However, in this case when a model is both trained and tested with debiased embeddings we can observe a result of 0.85 and 0.87 using LR and SVM respectively.

	Scenario 1	Scenario 2	Scenario 3
LR	0.88	0.59	0.68
SVM	0.89	0.60	0.59

The results obtained on case 2 are similar to those obtained in case 2 although there is a larger drop in accuracy to 0.59 and 0.60 respectively for LR and SVM when a model trained on original embeddings is validated on debiased embeddings. Using a better gender axis as in this case, we observe there is a much larger drop in accuracy to 0.68 and 0.59 for LR and SVM respectively when we train a model on debiased embeddings and validate it using debiased embeddings.

Furthermore, we used our debiased model and debiased embeddings on the test dataset to observe the probabilities with which a profession is assigned a gender.

	Profession	Female BiasedDB	Male BiasedDB	Prediction BiasedDB	Female DebiasedDB	Male DebiasedDB	Prediction DebiasedDB
0	accountant	0.177285	0.822715	male	0.318647	0.681353	male
1	acquaintance	0.262851	0.737149	male	0.354868	0.645132	male
2	actor	0.040965	0.959035	male	0.614853	0.385147	female
3	actress	0.999314	0.000686	female	0.959351	0.040649	female
4	adjunct_professor	0.858274	0.141726	female	0.686219	0.313781	female
...	...	...	...	...	...	...	...
315	warrior	0.107379	0.892621	male	0.570477	0.429523	female
316	welder	0.067933	0.932067	male	0.070504	0.929496	male
317	worker	0.329413	0.670587	male	0.229706	0.770294	male
318	wrestler	0.064184	0.935816	male	0.643966	0.356034	female
319	writer	0.425793	0.574207	male	0.590540	0.409460	female

Figure 7: Probability of a profession being a male or female using LR

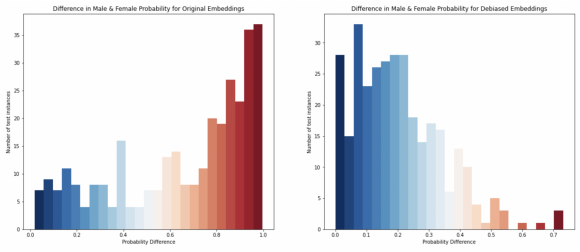

Figure 8: The change in probability distribution

Prior to debiasing, we observe the results are left-skewed showing that for most of the test examples there was a vast difference between probabilities of a profession being male or female oriented. However, after debiasing, we observe an exact opposite right-skewed plot which exhibits the small difference in probabilities of a profession being male or female for most of our test examples. This is evidence that the model associates almost equal probabilities to a profession being male or female and is not confident of its prediction.

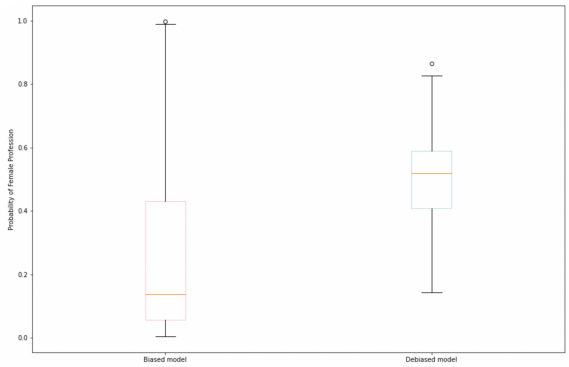

Figure 9: Increase in likelihood of a profession being associated with a female

We observe that with the biased model, the average probability of female and male prediction is 0.26 and 0.74 respectively and the average difference in probabilities is 0.65 which is high showing the confidence of the model in its predictions.

On the contrary, we observe almost equal probability of predictions for male and female as 0.50 and 0.498 respectively. Further, there is a significant drop in the average difference of probabilities to 0.20 which depicts the less confidence or the confusion of our model in its predictions thereby, substantiating our claims for the efficacy of our debiasing algorithm.

5 Conclusion

The results uphold our claims for the effectiveness of our debiasing technique. When we use debiased embeddings for validating our model, we see that the model struggles in predicting gender. Further, we have also observed that when a model is trained with debiased embeddings and validated with debiased embeddings there is a sharp drop in accuracy of predictions however, it is still higher than the previous case which may be indicative of some loss of information which does not necessarily have to be gender related. For either of the models, we obtain better results when we use a more comprehensive approach towards devising our gender-axis. We obtain a better gender-direction when we choose multiple sets of appropriated gendered-pairs. Consequently, this leads to a better neutral axis and better projections of our word embedding on that neutral axis. This signifies the effect of choosing an appropriate gender axis whilst magnifying the need for the selection of the same. When the models are tested using debiased embeddings on a debiased model, we obtain effective results with almost equal average probabilities of predictions for a profession being male and female.

For future work, we would like to explore other forms of non-binary bias such as races or religions and the effectiveness of this approach towards removing that kind of bias. We would also like to explore the results of debiasing on a more complicated downstream tasks such as sentence classification, resume parser and so on. It would also be worthwhile to compare this debiasing approach to other contemporary works. Overall, this seems to be a small step in the right direction and we are intrigued to expand its scope further.

6 References

Aron Klien, Reducing bias in AI based financial services, 2020.

Jeffery Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018.

Sam Biddle, Research says facebook's ad algorithm perpetuates gender bias, 2021.

Law of large numbers, Wikipedia. https://en.wikipedia.org/wiki/Law_of_large_numbers.

7 Code

You can replicate all our experiments and results using the code given at the link below. https://github.com/navekshasood/GenderDebiasing