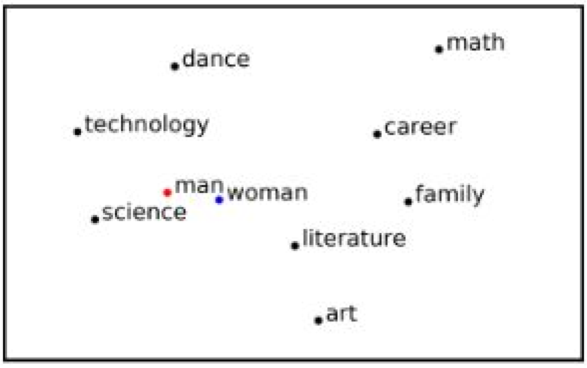
CS 626 Project Report
Debiasing BERT embeddings

Team 22
Drumil Trivedi 170020016
Sarvesh Mehtani 170050107
Anshul Nasery 170070015

Problem Statement

As natural language processing methods are increasingly deployed in real-world scenarios such as healthcare, legal systems, and social science, it becomes necessary to recognize the role they potentially play in shaping social biases and stereotypes. Previous work has revealed the presence of social biases in widely used word embeddings involving gender. These biases arise because of the data and training paradigm of large scale language models like BERT. Such biases can manifest in various forms, the most common among them being that the embeddings of certain attribute words are more similar to the embeddings for male terms than female terms. As an example shown below, the BERT embeddings for *man* are closer to words like *science* and *technology*, while *woman* is closer to words like *family* and *art.*



In this project, we try to tackle this problem in a post-hoc manner, i.e. the question we address is whether we can come up with a way to modify the word embeddings of BERT by finetuning the model through a new layer such that the new embeddings mitigate this correlation while being meaningful enough.

Prior Work

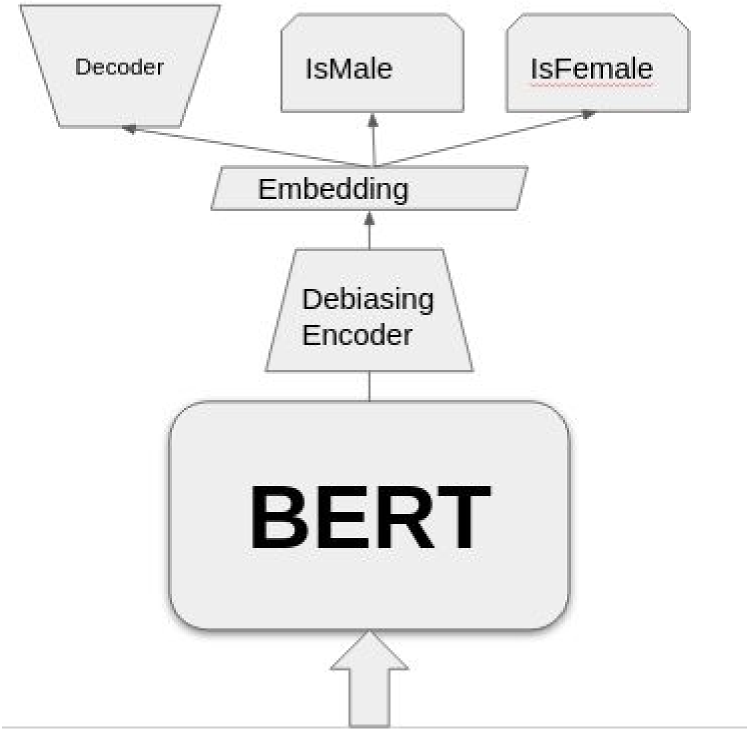
The task of debiasing word embeddings has seen some work in recent years. We quote two previous papers here which we believe are the most relevant to our project-
1. Liang et al. tried to compute a “gender subspace” of BERT embeddings by looking at the PCA of the embeddings for gendered terms. They then got the projection of all embeddings perpendicular to this subspace in order to get unbiased sentence embeddings.
2. Kaneko et al used an auto-encoder approach on top of GloVe embeddings to debias them, by making the encodings of non-gendered terms perpendicular to gendered terms. However they only look at single words without care of context.

Our work tries to address the shortcomings of both of these approaches by considering contextual embeddings in an auto-encoder based setup.

Our Approach

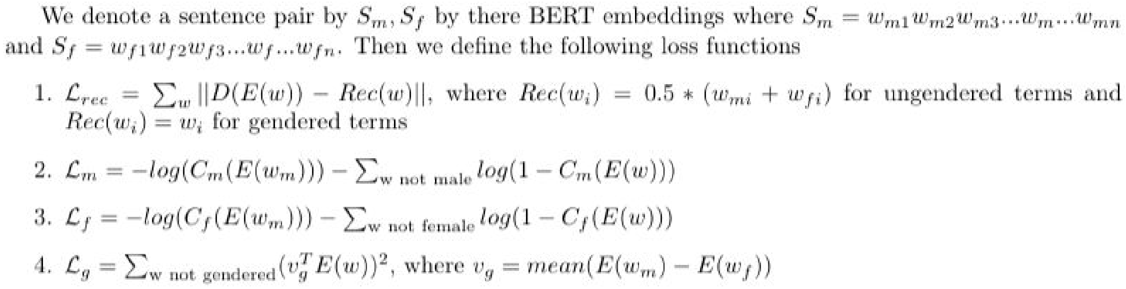
We present our approach below. At a high level, we augment the BERT embeddings by passing them through a “debiasing” layer. The original embeddings are then reconstructed in an autoencoder setup, and other losses are incorporated to make sure that the embeddings are informative enough and yet debiased.

Model



In the above image, encoder and decoder are single layer neural networks. IsMale and IsFemale are 2 layer classifiers which try to distinguish if the presented word embedding is a male word or a female word. The embedding returned for each word is the output of the Debiasing Encoder on each word embedding outputted by BERT.

Losses



We denote a sentence pair by S_m, S_f by there BERT embeddings where $S_m = w_{m1}w_{m2}w_{m3}...w_m...w_{mn}$ and $S_f = w_{f1}w_{f2}w_{f3}...w_f...w_{fn}$. Then we define the following loss functions

1. $\mathcal{L}_{rec} = \sum_w ||D(E(w)) - Rec(w)||$, where $Rec(w_i) = 0.5 * (w_{mi} + w_{fi})$ for ungendered terms and $Rec(w_i) = w_i$ for gendered terms
2. $\mathcal{L}_m = -log(C_m(E(w_m))) - \sum_{\text{w not male}} log(1 - C_m(E(w)))$
3. $\mathcal{L}_f = -log(C_f(E(w_m))) - \sum_{\text{w not female}} log(1 - C_f(E(w)))$
4. $\mathcal{L}_g = \sum_{\text{w not gendered}} (v_g^T E(w))^2$, where $v_g = mean(E(w_m) - E(w_f))$

The motivation behind these losses is as follows-

1. Tries to make sure that the learned embeddings are meaningful enough that we can reconstruct the original embedding (or an approximation thereof) from them. Further, due to the definition of Rec(w), we also ensure that the context leak of the gendered term into the

ungendered terms is mitigated.
2 and 3. Make sure that the embeddings of the male and female terms are informative enough to be distinguished as male and female words, by passing the embeddings through the IsMale and IsFemale classifiers and applying a binary cross entropy loss to make sure that the male and female word are correctly classified while the other words are not given the male/female tag.
4. Makes sure that the embeddings of ungendered words are perpendicular to the average embeddings of the male and female words. Here v_g is computed once every epoch.

Train Data

We follow Liang et. al. to generate training data from existing text datasets.
For the datasets yelp reviews, sst, wikitext, pom, meld, news_200, we select sentences which have one gendered term. Then we construct parallel sentences by replacing the gendered term with its equivalent alternate. For example, for the sentence *My 3 year old son needed a haircut this past summer and the lure of the 7 kids cut signs got me in the door,* we introduce a sentence pair with the other sentence being *My 3 year old daughter needed a haircut this past summer and the lure of the 7 kids cut signs got me in the door.* I.e. we replace *son* with *daughter.* Further, we discard sentences which have multiple gendered terms for simplicity. We also have a length filter of 30 words on the sentences. All this leads to a corpus of 6500 sentence pairs. Details of each sub-corpus are shown in the table below

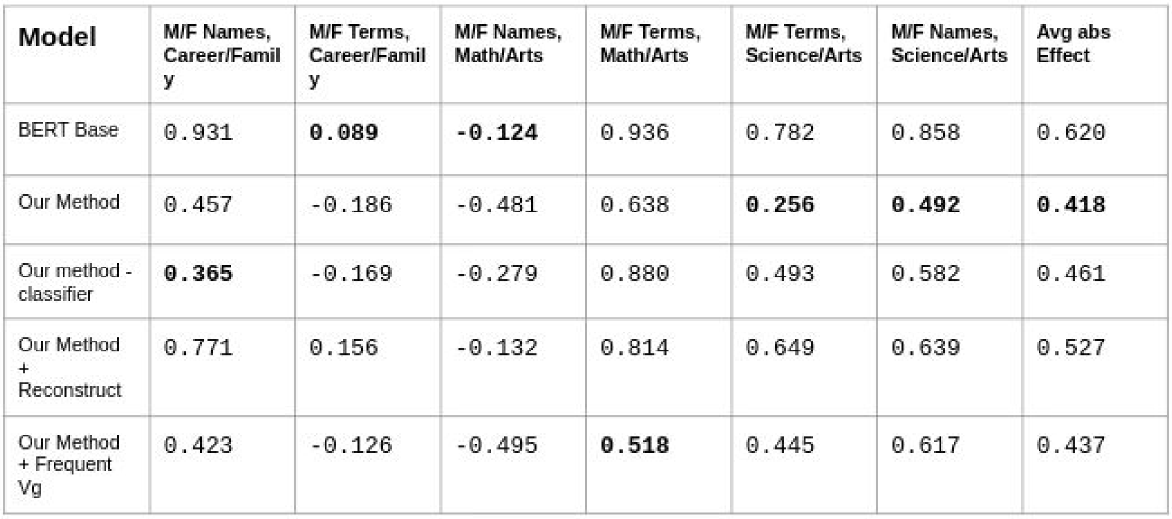
Dataset	Type	Topics	Formality	Sample
Yelp	Written	Shop/other review	informal	My 3 year old son/daughter needed a haircut this past summer and the lure of the 7 kids cut signs got me in the door
SST	written	movie reviews	informal	2“｛his/her｝fans walked out muttering words like horrible and terrible,but had so much fun dissing the film that they didn’t mind the tick
MELD	spoken 8.1	comedy TV-series	informal	“that’s the kind of strength that I want in the｛man/woman｝I love!”
POM	spoken	opinion videos	informal	and｛his/her｝family is, like, incredibly confused

Wiki Text	Written	Everything	Formal	The soldiers would be allowed to march away as {men/women} leaving under orders not as conquered soldiers.

Results

We follow the prior work and report the results of our model on the metric WEAT score. We compare this score for the original BERT and our debiased method. WEAT score is a metric which looks at the difference of the cosine similarity between male terms and certain attributes and female terms and the same attributes. A WEAT score close to 0 denotes debiased embeddings.

WEAT Score



Model	M/F Names, Career/Family	M/F Terms, Career/Family	M/F Names, Math/Arts	M/F Terms, Math/Arts	M/F Terms, Science/Arts	M/F Names, Science/Arts	Avg abs Effect
BERT Base	0.931	**0.089**	**-0.124**	0.936	0.782	0.858	0.620
Our Method	0.457	-0.186	-0.481	0.638	**0.256**	**0.492**	**0.418**
Our method - classifier	**0.365**	-0.169	-0.279	0.880	0.493	0.582	0.461
Our Method + Reconstruct	0.771	0.156	-0.132	0.814	0.649	0.639	0.527
Our Method + Frequent Vg	0.423	-0.126	-0.495	**0.518**	0.445	0.617	0.437

We see that our model does better on average than BERT across all variants. We note that we have failed to debias certain categories. We suspect this happened because of the type of data we used to debias sentences, which did not have terms from certain fields.

Further, we also present certain ablation studies to see how the various components of our model contribute to the debiasing.

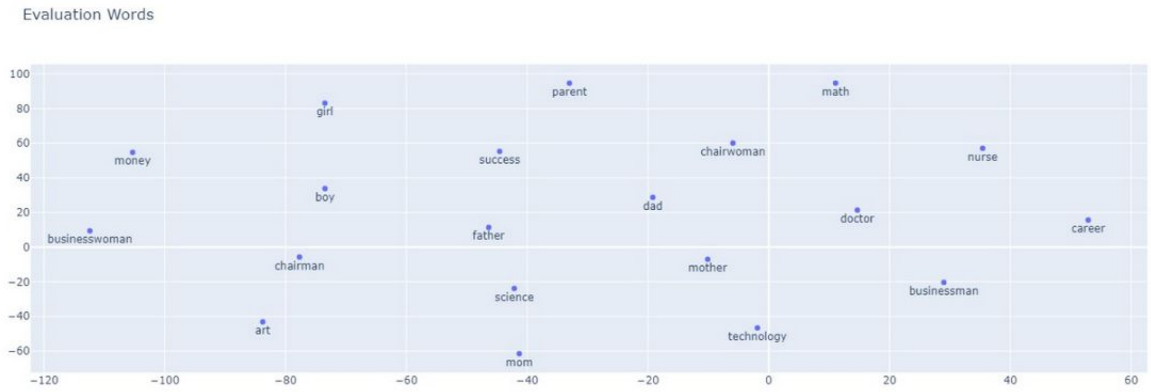
Row 3 refers to an ablation where we do not include the classifier losses, with the hope that the reconstruction loss would be enough to encode the gendered information in the embeddings. However we observe that this is not the case, leading to a large drop in the performance, leading us to believe that it is important that the embeddings of gendered words to contain enough information about their gender.

We also perform an ablation to see if our method of using the average embedding across pairs of sentences for reconstruction is good. We discard this average and instead only use the BERT embedding directly in the reconstruction loss. We find that this deteriorates the performance substantially across all metrics, leading us to conclude that our reconstruction

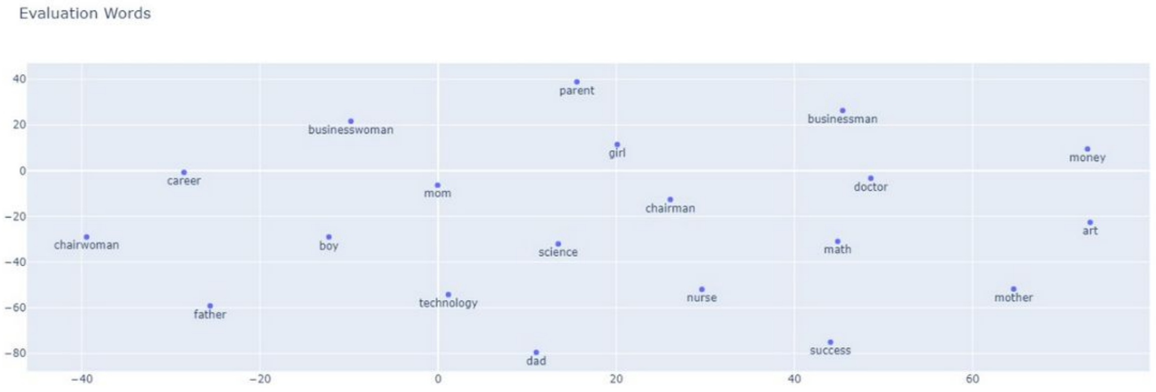
loss indeed contributes to better debiasing. This is also intuitive since we are forcing our model to ignore the gendered context for non-gendered words.
Finally, we conduct an ablation where we compute the average gendered embedding multiple times each epoch. We see that this leads to a slight drop in the performance, due to training instabilities.

TSNE Plots and Shortcomings

We present the TSNE plots of our embeddings and BERT base embeddings for some words.



Bert Embedings



Our model

We see that words like *Science* are equidistant from male and female terms for our model while they had a male preference in BERT. However, our model is not perfect, which can be seen in the word *art* which seems to have a female bias both in BERT and our model's embeddings. Similarly *career* seems to be more male biased. This is also seen in the table of results.

Conclusion and Future Work

We present a method for post-hoc debiasing of contextual embeddings. We also present ablations to investigate why our model works as well as display some cases where it does not work. One shortcoming of the current work is that it does not consider the context for evaluation of embeddings, we plan to work on coming up with a metric for this in the future.

References

[1] Towards Debiasing Sentence Representations, Liang, Paul Pu, Li, Irene Mengze, Zheng, Emily, Lim, Yao Chong Salakhutdinov, Ruslan, Morency, Louis-Philippe, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020

[2] Gender-preserving Debiasing for Pre-trained Word Embeddings, "Kaneko, Masahiro and Bollegala, Danushka", "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics","2019",