Visualizing the Debiasing of a Word Embedding

Debiasing Word Embeddings

Darryl York III, Angela Ellis, Aishwarya Varma, Aldo Polanco, Dr. Anna Rafferty*

Carleton College Computer Science

Introduction & Methods

We spent our term replicating the research conducted in the 2016 paper produced by Bolukbasi et al. titled, "Man is to Computer Programmer as Woman is to Homemaker"[1] . The focus of this paper was:

- To expose the implicit sexism found in text corpora and,
- To determine whether soft or hard debiasing algorithms are effective in eliminating the gender bias within word embeddings

Word Embeddings

An object for text analysis and text generation through mapping text into individual word vectors. Machine learning allows relationships between words and their surrounding text to be highlighted. Applications for word embeddings include consumer feedback parsing, spam detection, and information retrieval(ex. search engines). For our purposes, we use vector mathematics to expose biased relationships between words unrecognized or unproven through other forms of text analysis. We will specifically focus on the gender bias.

Our Dataset & PCA

PCA is a linear dimension reduction method used to reduce the dimensionality of our 300-dimensional word embedding. It does so by combining vector features, dropping the least important features and retaining their significant parts. Data is then mapped linearly while also maximizing the variance of the data.

Similarly to [1], we used Google's publicly accessible word2vec tool that takes a text corpus as input and outputs vectors of each word present. The w2v model used in our projects was trained on hundreds of google news articles.

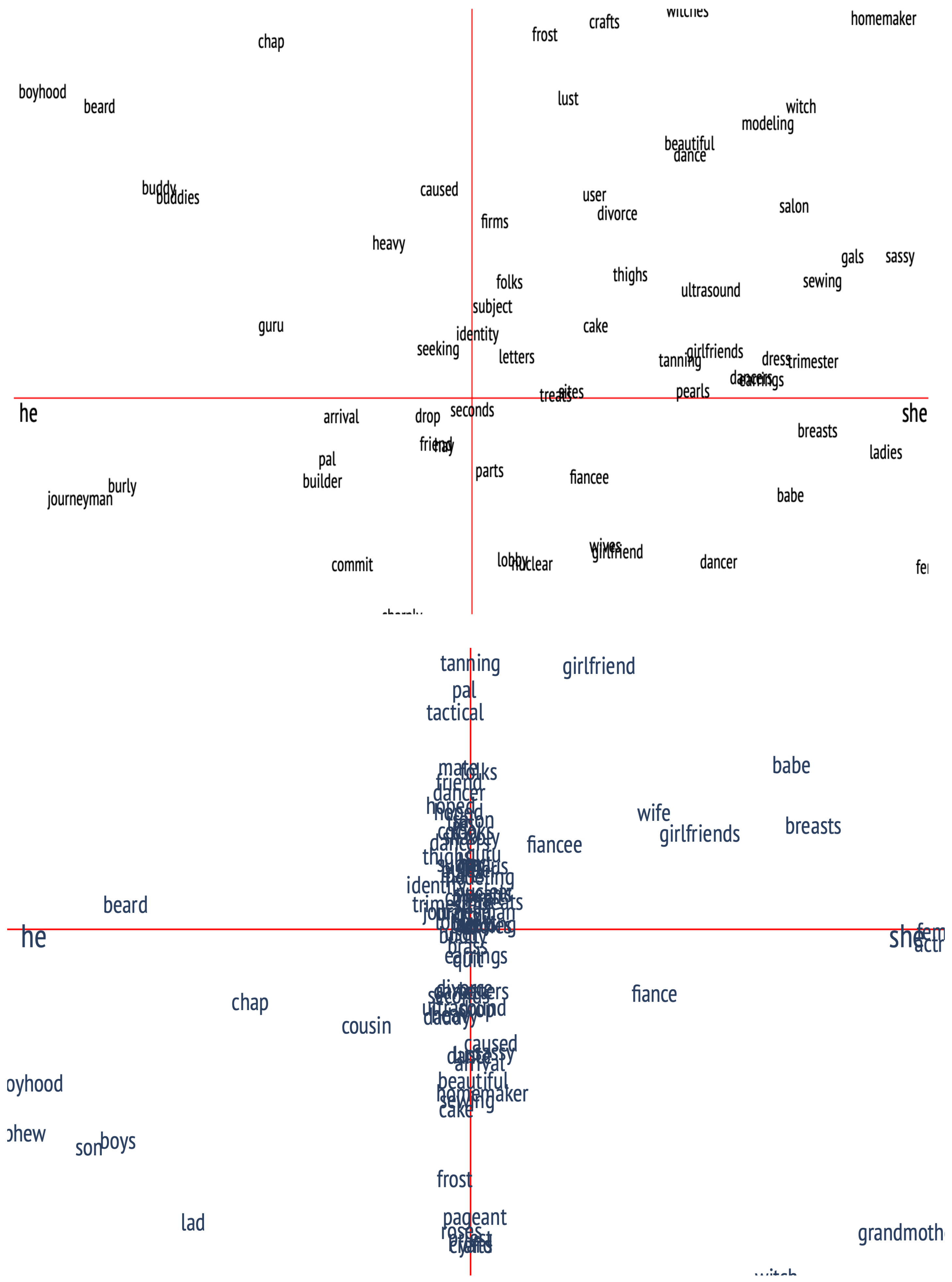

Figure 1: Biased vs Debiased

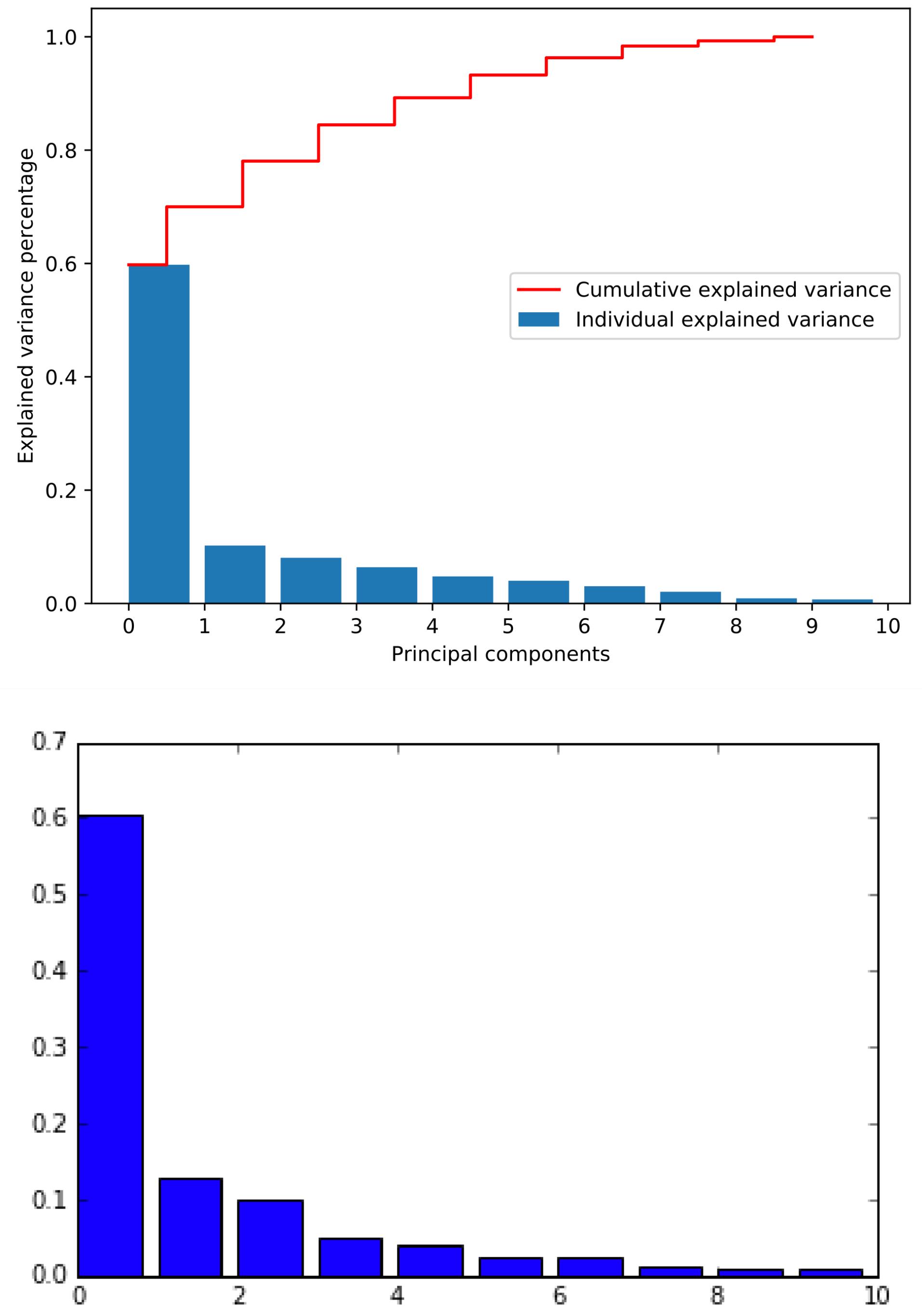

Figure 2: PCA Explained Variance

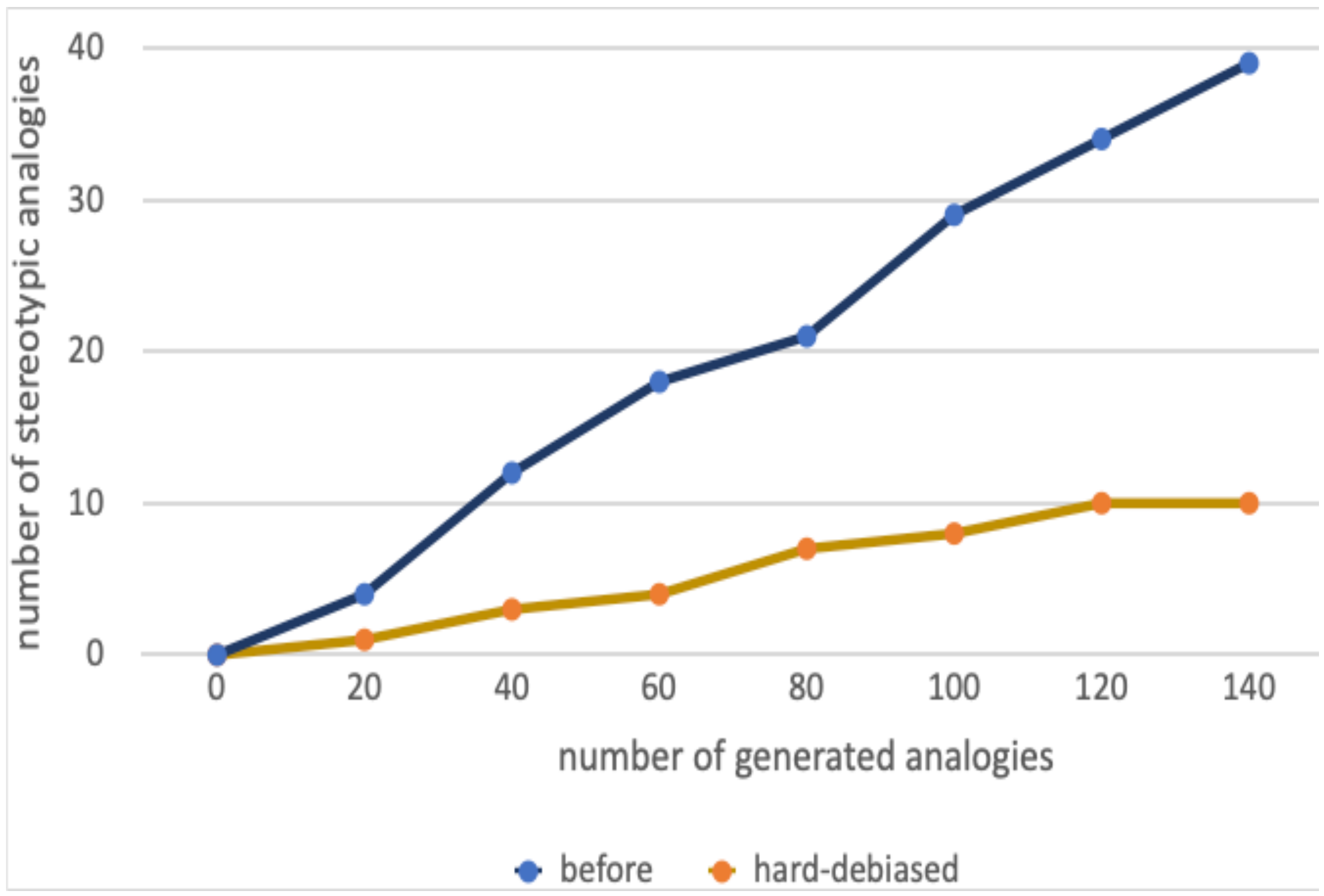

Figure 3: Stereotypic Analogies Generated

Results & Discussion

Replication of this paper was split into 4 sections including: **Vector Manipulation** to find a Gender Direction, **Gender Specific and Gender Neutral word distinction** through a Linear SVM, **Analogy Generation**, and **Data Visualization**.

Replicating the PCA explained variance in [1], we took a set of 10 gender pairs to determine the components of PCA needed to showcase the majority of variance in the dataset that accounted for gender. We see a consistent %70 variance across the top two components.

In the scatterplot replications, which are 2D representations of a subset of words found in the word embedding, the words are projected onto the gender direction found through PCA along the x-axis. Words closer related to he will be found on the left, while those more related to she are found on the right. The Debiased plot shows that our debiasing algorithm effectively removed the pull towards either gender direction!

Limitations

- ❖ Dataset complexity was too large for Google Colab GPU computations
- ❖ Coding intricacies not found in the paper
- ❖ Outdated Code Libraries

Conclusions

From our visualizations, there are clear differences in the analogies generated before and after debiasing, and debiasing of the word embedding is shown through scatterplot visualization.

As we can find the gender bias within a word embeddings, there are other biases available for analysis as well. Racial bias is of interest to me and in [1], it was recognized that embedding manipulation is possible for exposing racial stereotyping.

References

1.Tolga Bolukbasi et al., "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings" (arXiv, July 21, 2016), https://doi.org/10.48550/arXiv.1607.06520.