Man is to Computer Programmer as Woman is to Homemaker?

Debiasing Gender Stereotypes in Word Embeddings

Aishwarya Varma, Aldo Polanco, Darryl York III, Angela Ellis

Carleton College Computer Science

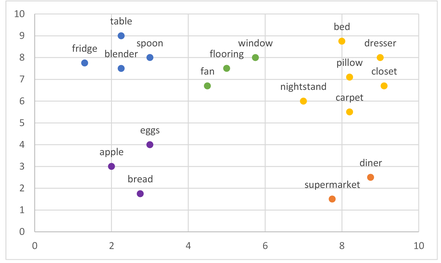
Background

We replicated the findings of Bolukbasi et al.'s "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," where they attempt to remove gender bias from word embeddings. Our goals were to replicate their findings, compare our results, and test for robustness.

Word Embeddings

Word embeddings are tools that are used to represent words as vectors. Embeddings can be visualized as a virtual space containing a bunch of words; the closer the words are to each other in the embedding, the more similar they are in meaning. The paper uses Google News Word2Vec, a tool which takes Google News articles as input and outputs each word in the text as vectors. The full embedding contains 3 million words and has 300 dimensions. These tools have been used in many machine learning and natural language processing tasks, including parsing through consumer feedback, spam detection, and information retrieval (e.g. search engines).

Many word embeddings contain gender bias (amongst other types of bias) because they are derived from human sources. For example, if we search for a word-pair relationship in our embedding that is equivalent to "man" and "computer programmer," we would receive "woman" and "homemaker" in return. This bias is harmful, as neither of these professions are inherently gender specific. Consequently, because of their widespread use, it's imperative that word embeddings are stripped of their bias, as it could cause gender stereotypes to be amplified.



Visual representation of a word embedding

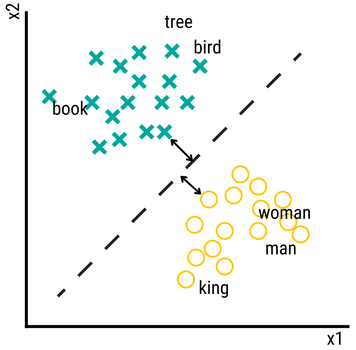
Methods

Bolukbasi et al. structured their work into three phases:

1. identify where the gender bias lies in the embedding
2. determine words that are appropriately gendered in the embedding
3. neutralize all other words **not** found in the above list in the embedding using the debiasing algorithm and equalize gender specific words

In order for the debiasing step to occur, we had to determine which words were gender specific (GS) and gender neutral (GN). For example, "king" and "queen" are appropriately associated with "man" and "woman," so they are GS. "Computer programmer" and "homemaker," however, *should* be GN.

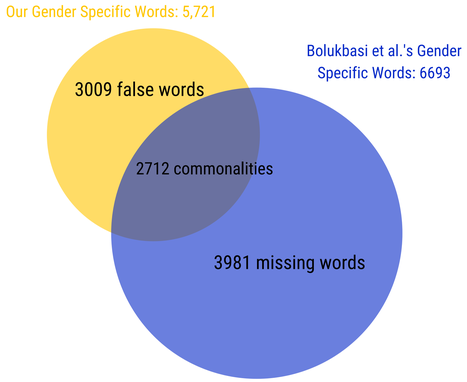
It would be impossible to manually determine which words are GS and GN, simply because the embedding consists of 3 million words. To address this, we used a linear classifier, using the Linear SVC class from sklearn, a free Python machine learning library, to categorize all of the words into two classes: GS and GN. We trained the Linear SVC on a 27,000 word subset of the embedding, of which we knew which words were GS and GN. To predict whether a new word is GS or GN, the Linear SVC fits a line between the two classes such that the margin of difference is maximized. The larger the margin, the more confident we can be that a word is either GS or GN.



Linear Classifier: Maximizing the difference between GS and GN words

Results

Our program outputted a list of 5,721 GS words, of which ~47% are in common with the list generated by Bolukbasi et al. However, ~53% of the words in our list are not in *their* list, and 3,981 words (~59% of their list) are not present in *our* set.



A diagram representing the overlap and comparison between my results and Bolukbasi et al.'s results

Ours	Bolukbasi et al.
lad	jester
sperm	daddies
stud	aux
sir	gray_beard
prostate_cancer	Sweetie
fathers	Princesses
bachelor	breasted
ex_girlfriend	bros
lesbians	homeboy
councilman	womenfolk
actresses	MVP_Peyton_Manning
gentlemen	sugar_daddy

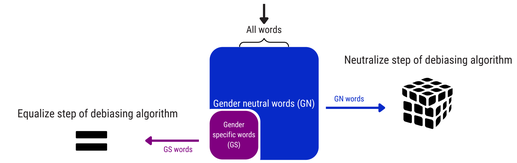
A small sample of GS words generated from both our linear classifiers

Discussion

While these results might seem skewed, it's important to note that there were nonsensical GS words found in both lists(i.e. "Senate_Permanent_Subcommittee" in theirs and "guh" in ours). For the most part, both of our programs produced words that were GS, with a few exceptions.

Next Steps

This list of GN words (i.e. everything that was not in the list of GS words) would then be inputted into the debiasing algorithm to be neutralized. This means that any words in the GN list that previously had a gender bias would have no specific affiliation with either gender after the algorithm runs its course.



Limitations

Throughout the duration of the project, there were several limitations that affected the success of our project:

- limited computational power (even with Colab Premium)
- 10-week term
- only looked at gender bias (and not other types of bias)
- didn't explore word embeddings outside of the Google News source
- relied on the paper's subjective analysis of what is deemed "gendered"
- limited experience with this topic and machine learning

Sources

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Jul 21, 2016. Man is to Computer Programmer as Woman is to Homemaker: Debiasing Word Embeddings. Microsoft Research New England, Cambridge, MA.

T. Mikolov. (2013). Word2vec. https://github.com/tmikolov/word2vec (accessed Sept. 20, 2022).