Lipstick on a Pig:

Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

TABLE OF CONTENTS

01

THESIS

CENTRAL CLAIM:

The word embeddings used in NLP algorithms have consistently demonstrated gender bias. While there are new methods for debiasing these embeddings, they aren’t effective enough since they hide bias, rather than actually removing it.

02

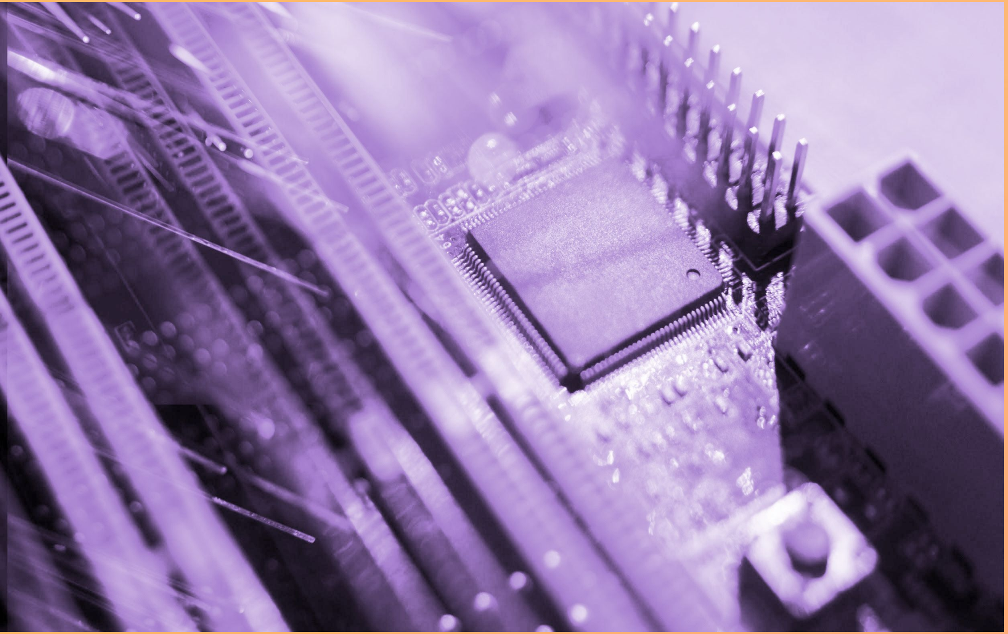
CURRENT DEBIASING METHODS

METHOD 1

HARD DEBIASING

- Post processing debiasing method - manipulates vectors after training.
- Makes neutral words equidistant from gendered ones.
- Removes the gender direction from neutral words.

METHOD 2



GN-GloVe

- Aims to debias word embeddings **during** training.
- Changes the **loss** of the model.
- Uses 2 groups of f/m words and makes them **differ in the last coordinate** -> that's the key idea.
- This allows to **exclude** the last coordinate.
- Representation of neutral words is **orthogonal** to the gender direction -> their dot product should be 0.

THE MAIN PROBLEM:

Saying that a word is “debiased” when it is only equidistant from two gendered words is inadequate. This is because even when this is true, words associated with certain gender stereotypes will cluster together.

Both of these methods use this definition

03

EXPERIMENTS

THE NUMBERS

50,000 Most frequent words

47,698 Words for GN-GloVe

26,189 Words for Hard-Debiased

Setup

Bias of a word is computed by taking its **projection** on the **gender direction**.
Association between sets of words is quantified using WEAT (estimating the probability that a random **permutation** of the target words, e.g. professions, would be **close** to the attributes sets).

Experiments

500 most biased words from each group were clustered using k-means, then accuracy of **alignment** with gender was computed for both of the embeddings.
It was suggested to measure bias by approximating the **percentage** of f/m words among k nearest **neighbors** of the target word, it was implemented for a list of professions.
Correlation between this and the original measure was computed.
Predicted the **gender** and evaluated its **regularization** on the remainders using an RBF-kernel SVM.

04

RESULTS

RESULTS (CNTD)

CLUSTERING

male and female biased words cluster together. (High percentage of alignment)

CONCLUSION

When clustering is considered in the definition of bias, these embeddings are still fairly biased after debiasing.

BIAS PRESENT

HIgh accuracy of predicting gender of debiased words based on most biased words.

ASSOCIATIONS

Association found between female/male concepts and stereotyped words

05

CONCLUSION

DISCOVERIES

Words with strong initial gender bias are easy to cluster together even after “debiasing”.

1.

Words that receive implicit gender from social stereotypes tend to group with other such words of the same gender.

2.

The implicit gender of words with prevalent previous bias is easy to predict based on their vectors alone.

3.

Gender-direction provides a way to measure gender-association of a word, but does not determine it.

06

OUR TAKEAWAYS