Conceptual and methodological problems with bias detection and avoidance in natural language processing

Alicja Dobrzeniecka

2021-06-19

Contents

Chapter 1

Introduction

Placeholder

Chapter 2

Cosine similarity and bias detection

Placeholder

2.1 Word embeddings

2.2 Cosine similarity and distance

2.3 Cosine distance in a one-class bias detection

2.4 Cosine distance in a multi-class bias detection

2.5 Limitations of the approach

Chapter 3

Walkthrough with the religion dataset

Placeholder

3.1 Loading and understanding the dataset

3.2 First look at the empirical distributions

3.3 Looking at the islam-related words

3.4 Bayesian model structure and assumptions

3.5 Choosing predictors

3.6 Dataset-level coefficients

3.7 Model structure and assumptions

3.8 Protected classes in Reddit and Google embeddings

3.9 Dataset-level coefficients after debiasing

3.10 Protected classes after debiasing

Chapter 4

Discussion and summary

We propose the use of Bayesian methods to measure uncertainty in bias detection. There are a few advantages of this method. Including uncertainty enables one to directly observe the influence of sample sizes. Analyzing individual words and connection coefficients, one may notice how `neutral` words have smaller uncertainty intervals and `different` or `associated` quite the opposite. One of the reasons for such outcome is that we used approximately 230 neutral words and only between 11-25 (the number varies from class to class) stereotypical attributes from Manzini, Lim, Tsvetkov, & Black (2019) article. What is more, we also pay attention to distribution and details regarding anomalous values. With the use of simple visualizations that we introduced before, we were able to indicate suspicious cosine distance values. Additionally, we compare in details how the cosine distance values and uncertainty change after the debiasing. One can verify then how the individual vectors changed and if it what was expected. Our analysis with the use of Bayesian method gave us new ideas and hypothesis concerning not only the bias detection method but also the efficiency of debiasing itself.

We created a summary table for each of the datasets: Reddit, Reddit Debiased, and Google word embeddings. Let us first analyse the general observations from estimated coefficients mean introduced in 3.6. DATASET-LEVEL COEFFICIENTS. For Google embeddings the HPDI for all classes coefficients (associated, different, human, and none) has an interval that includes zero. This can lead one to a conclusion that the impact for associated, different, human and neutral attribute is, when averaged, quite similar. This indicates how including the uncertainty may change the use and interpretation of Manzini, Lim, Tsvetkov, & Black (2019) MAC metric. It seems that if one focuses only on differences between means of means, it is too simplistic. In case of Reddit word embeddings the situation is similar although HPDI interval is below 0 for Gender class when looking at `associated` and `different` mean coefficient. This can suggest that there is indeed slightly stronger impact of these attributes in cosine distance being smaller. One should also notice how `associated` and `different` coefficients have quite similar HPDI interval, the highest observed absolute difference is equal to only 0.1. This suggests that again the impact of associated attributes and difference ones is not clear at first sight when looking at averaged coefficients. Finally, let's compare the HPDI intervals for Reddit and Reddit debiased datasets. For Religion and Race dataset there is a minor shift (in absolute values the highest change is equal to approximately 0.1) of the mean coefficients towards zero. However for Gender dataset there is no significant change. This is of course the general look at the data, let's now analyze individual words.

In Reddit table one can observe that for Religion class the cosine distance results for the associated attributes are for approximately 60% of the words close to neutral attributes as well. This can suggest that in some cases words concerning humans can have higher similarity with some protected words independently if they are stereotypical or neutral-human words. One should also pay attention to the fact that for all of the `associated` and `different` attributes in Religion class the uncertainty interval overlaps at some point. What is even more surprising is that for protected words, such as `torah` `associated` attribute has the the cosine distance slightly over 1, which means no positive similarity! If the protected words that we chose do not have high similarity with harmful associated attributes, then one should consider at least three scenarios. The first one is that the choice of protected words and attributes may be corrupted. The second one is that the metric is not able to catch the hidden bias properly. The third one is that there is actually no bias between the words. Regardless of which scenario one considers, it is essential to take a look at the individual values before averaging them or aggregating in other ways. It seems that using Bayesian method can enhance the process of verifying the hypotheses concerning the choice of protected words and attributes.

Surprisingly in Gender class one can observe high cosine similarity values between some female stereotypical professions and male protected words. If a word stereotypically associated with females has low cosine distance to male protected words, then one should try to figure out the reasons for that. Cosine distance seems to capture the information on the co-occurrence of words and not on the semantic similarity strictly speaking.

In Google table ..

In Reddit debiased ..

Let's summarize the results of the bias detection methods analysis.

Additionally, one cannot be sure if the bias is still preserved after the debiasing. The fact that all of the cosine distances for protected words and harmful attributes moved to the right, does not mean that the bias is removed. It is shown in articles such as Gonen & Goldberg (2019), that the bias can hide in the vector geometry and preserve even after applying popular debiasing methods.

One should remember that there is no clear interpretation for the values obtained with MAC metric from Manzini, Lim, Tsvetkov, & Black (2019). One may assume that if the cosine distance is close to 1 then it is a desired outcome as it means, according to cosine distance assumptions, that there is almost no similarity between the words. However what does it mean to be close to 0? If the averaged cosine distance is equal to 0.8, then should we still debiase it? It is unclear what the criteria are. On one hand, it seems to be beneficial when the outcome is simplified as it is easier to compare results with one value per set. On the other hand, it is prone to misunderstanding of how to interpret the results and what threshold to assume.

The bottom line is that if we want to take bias seriously, so should we approach the uncertainty involved in our estimations. There is no replacement for proper statistical evaluation that does not discard information about the uncertaintly involved, larger word lists are needed, and visualisation of the results for particular protected classes provides much better guidance than chasing a single metric based on a means of means.

Bibliography:

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, *abs/1903.03862*. Retrieved from http://arxiv.org/abs/1903.03862

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. Retrieved from http://arxiv.org/abs/1904.04047