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

Alicja Dobrzeniecka

2021-06-16

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

It is worth diving deeper into some observations that arise while analyzing the results. As we pointed out, if there is no control of what cosine distance values individual word pairs have, then there is less understanding of the exact relations between words. If the protected words that we chose do not have high similarity with harmful stereotypes, then one should consider at least two 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. In both cases it is essential to take a look at the individual values before averaging them or aggregating in other ways.

One has to also remember about the specific situation present in gender dataset. Surprisingly one could observe there high cosine distances values between some female stereotypical professions and male protected words. This leads to new ideas of how to understand the bias origins and its proper detection. To start with, cosine distance seems to catch the information regarding words co-occurrence and not the semantic similarity. If a word stereotypically associated with females has high cosine distance value for male protected words, then the metric is not proper to establish the bias presence as it can be misleading. What we mean is that in same cases it can unnecessarily

label words as biased while omitting other more harmful associations. What is more, this highlights again the need for control group. One could check each stereotypical attribute with each protected word but not in order to average the values. One should rather plot the distributions and investigate how the metric measures vectors geometry and hidden semantic information.

Debiasing with the method provided by Manzini, Lim, Tsvetkov, & Black (2019) shows how it is unpredictable in terms of the results. In some cases the final cosine distance is indeed close to neutral words, but in other cases the change is minor. What is more, the inclusion of the uncertainty shows how in some cases it is not clear why a word is according to this method classified as anomalous. The fact that uncertainty for `assosiated` and `different` class is so high leads to the situation where in the boundaries of an uncertainty of some attributes, one can find the neutral words as well. It makes the usage and interpretation of this dataset and metric extremely complicated and unclear.

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. One could argue that maybe it is not clear how to measure bias presence and removal but the method for debiasing works fine. Notwithstanding, it is showed in researches such as Gonen & Goldberg (2019), that the bias can hide in the vector geometry and preserve even after applying popular debiasing methods. Therefore it seems to not be justified to claim that the method works properly without precise methods to verify this.

One should remember that there is no baseline for the interpretation of the averaged cosine distances in (**Manzini2019blackToCriminalToCriminal?**). 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 80, 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.

We propose the use of Bayesian method to introduce the uncertainty measure in bias detection. There are a few advantages when including this method in the bias analysis. As noted before in our example with religion dataset, one may obtain new insight into the bias issue after analyzing the mean estimates of the coefficients. There seems to be a minor difference for the mean concerning the `assosiated` class. This can lead to the conclusions that we do not have strong reasons to perceive lower cosine distance as significantly related to the fact that the attribute is associated.

One may dive deeper into this idea by investigating the connection coefficients for individual words with the use of the visualization. There one observe that indeed most of the words are clustered together independently of the class they belong to. It is of course partly due to the fact that we include uncertainty now. This example shows how understanding of the bias measurement may differ when switching to Bayesian method. One could argue that after this analysis one obtains more information on the dataset and can pursue with the further research more cautiously.

Chapter 5

Summary

The aim of this paper was to show the analysis of bias detection methods used in modern natural language processing projects.

In the further research it is worth to dive deeper into some issues or to examine new ideas that occurred during our work.

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