Taking uncertainty in word embedding bias estimation seriously: a bayesian approach

Alicja Dobrzeniecka and Rafal Urbaniak

1 Cosine-based measures of bias

1.1 Word embeddings and their bias

Modern Natural Language Processing (NLP) models are used to complete various tasks such as providing email filters, smart assistants, search results, language translations, text analytics and so on. All of them need as an input words represented by means of numbers which is accomplished with word embeddings, in which particular lexical units are represented as vectors of real numbers. They are necessary to perform machine learning tasks that require the use of natural language. One of the ways to learn these vectors representations is to use simple neural network that based on the millions of real sentences examples assigns the numbers based on their co-occurrence and context. Using co-occurence results in a vector space that contains similar words closer to each other and dissimilar further away. It has been suggested [1–6] that in the learning process such models can learn implicit biases that reflect harmful stereotypical thinking (eg. that the word *she* is much closer in the vector space to the word *cooking* than the word *he*). This phenomenon may be undesired when using these vectors for further downstream tasks (eg. web search, recommendations, etc.). Before we move to more technical details regarding this topic, it is worth looking first at the theoretical background which may help to better grasp the concept of bias within machine learning algorithms. find downstream tasks and bias examples from papers

In the author focuses on the general issue, namely whether algorithms can in principle be value-free. cite Are algorithms Value-Free The author draws attention to the observation that in both scientific research and machine learning domain, induction plays a crucial role in the decision-making process. Induction is essential for both of these domains as "if ever considered literally just the facts, we would never be able to draw inductive conclusions" . Induction comes with certain risks which are 1) the problem with justification, and 2) the cite Gabrielle risk of getting things wrong. During the process of designing machine learning process some decisions have to always be made. As the author states, there is no such thing as the justification in the abstract. We may justify certain design decisions, data choice, etc. only by referring to the world's being certain cite Gabrielle way. Therefore if the induction is justified by contingent features that the world exhibits, which are intrinsically dependent on our current set of values, then our values indeed are the source of the justified inductive arguments. The other issue with induction is that as the machine learning algorithm always has access to a limited amount of information, it is prone to make mistakes from time to time, even if it's general accuracy is very high. What does it all mean in the context of bias within word embeddings? It seems that the presence of bias, value-laden algorithms is unavoidable due to the very nature of induction.

Assuming that the algorithms indeed due to the nature of their decision-making cannot fulfill the ideal of being value-free, it is even more justified to focus on ways how to properly detect harmful bias to limit its negative impact by the awareness of its presence. The aim of this paper is to contribute to this idea that using proper statistical tools may improve our understanding of what is happening inside machine learning models and especially within word embeddings.

1.2 General challenges

1.3 WEAT and MAC

One of the first measures in the discussion has been developed by [1]. First, the gender direction gd in is obtained by taking the differences of the vectors corresponding to ten different gendered pairs (such as $\overrightarrow{she} - \overrightarrow{he}$ or $\overrightarrow{girl} - \overrightarrow{boy}$) and then identifying their principal component (which is the vector obtained by projecting the data points on their linear combination in a way that maximizes the variance of the projections).[1] The gender bias of a word w is understood as its projection on the gender direction: $\vec{w} \cdot gd$. Given the supposedly gender neutral words N[2] and the gender direction gd the direct gender bias is defined as the average cosine similarity of the words in N from gd (c is a parameter determining how strict we want to be): get back to this!

$$\mathsf{directBias_c(N, gd)} = \frac{\sum_{w \in N} |\cos(\vec{w}, gd)|^c}{|N|} \tag{1}$$

The use of projections has been criticized for instance by [4], who point out that while the distance to the gender direction might be an indicator of bias, it is only one possible manifestation of it, and reducing the cosine distance to such a projection might be insufficient. For instance, "math" and "delicate" might be in equal distance to a pair of opposed explicitly gendered words, while being closer to quite different stereotypical attribute words. Further, it is observed in [4] that most word pairs preserve similarity under debiasing meant to minimize projection-based bias.[3]

A measure of bias in word embeddings which does not employ gender directions, the Word Embedding Association Test (WEAT), has been proposed in [2]. The idea is that the bias between two sets of target words, X and Y (we call them protected words), should be quantified in terms of the cosine similarity between the protected words and attribute words coming from two sets of stereotype attribute words, A and B (we'll call them attributes). For instance, X might be a set of male names, Y a set of female names, A might contain stereotypically male-related, and B stereotypically female-related career words. The association difference for a term t is:

$$\mathsf{s}(t, A, B) = \frac{\sum_{a \in A} \cos(t, a)}{|A|} - \frac{\sum_{b \in B} \cos(t, b)}{|B|} \tag{2}$$

then, the association difference between A a B is:

$$\mathsf{s}(X, Y, A, B) = \sum_{x \in X} \mathsf{s}(x, A, B) - \sum_{y \in Y} \mathsf{s}(y, A, B) \tag{3}$$

The effect size is computed by normalizing the difference in means as follows:[4], ^[A similar methodology is employed in [3]. The authors employ word embeddings trained on corpora from different decades to study the shifts in various biases. For instance, to compute the occupational embeddings bias for women the authors first compute the average vector of vector embeddings of words that represent women (e.g. *she*, and *female*), then calculate the Euclidean distance between this mean vector and words for occupations. Then they take the mean of these distances and subtract from it the analogously obtained mean for the average vector of vector embeddings of words that represent men. Formally they take the relative norm distance between X and Y to be:

$$\text{relative norm distance} = \sum_{v_m \in M} ||v_m - v_X||_2 - |v_m - v_Y||_2 \tag{4}$$

where the norm used is Euclidean, and v_X and x_Y are average vectors for sets X and Y respectively.]

[1] In the notebook associated with the paper, the authors simply use $\overrightarrow{she} - \overrightarrow{he}$ as a gender direction, though.

[2] We follow the methodology in assuming that there is a class of words that ideally should be identified as neutral, such as *ballpark, solution, lecture, science, book* can be identified. We will have a bit to say about this assumption when we describe our dataset construction.

[3] In [1] another method which involves analogies and their evaluations by human users on Mechanical Turk is also used. We do not discuss this method in this paper, see its criticism in [7].

[4] WEAT is a modification of the Implicit Association Test (IAT) [8] used in psychology, and it uses almost the same word sets, allowing for a *prima facie* sensible comparison with bias in humans. [2] argue that significant biases—thus measured—similar to the ones discovered by IAT can be discovered in word embeddings. [5] extended the methodology to a multilingual and cross-lingual setting, arguing that using Euclidean distance instead of similarity does not make much difference, while the bias effects vary greatly across embedding models (interestingly, with social media-text trained embeddings being less biased than those based on Wikipedia).

$$\mathsf{weat}(A,B) = \frac{\mu(\{\mathsf{s}(x,A,B)\}_{x\in X}) - \mu(\{\mathsf{s}(y,A,B)\}_{y\in Y})}{\sigma(\{\mathsf{s}(w,A,B)\}_{w\in X\cup Y})} \tag{5}$$

WEAT, however, has been developed to investigate biases corresponding to a pair of supposedly opposing stereotypes, and so the question arises as to how generalize the measure to contexts in which biases with respect to more than two stereotypical groups are to be measured. Such a generalization can be found in [6]. The authors introduce Mean Average Cosine similarity as a measure of bias (strictly speaking, in the paper they report cosine distances[5] rather than similarities). Let $T = \{t_1,\ldots,t_k\}$ be a class of protected word embeddings, and let each $A_j \in A$ be a set of attributes stereotypically associated with a protected word). Then:

$$\mathsf{s}(t_i,A_j) = \frac{1}{|A_j|}\sum_{a\in A_j}\cos(t,a) \tag{6}$$

$$\mathsf{mac}(T,A) = \frac{1}{|T|\,|A|}\sum_{t_i\in T}\sum_{A_j\in A}\mathsf{s}(t_i,A_j) \tag{7}$$

That is, for each protected word T and each attribute class, they first take the mean for this protected word and all attributes in a given attribute class, and then take the mean of thus obtained means for all the protected words.

Having introduced the measures, first, we will introduce a selection of general problems with this approach, and then we will move on to more specific but important problems related to the statistical significance of such measurement. We will focus on WEAT and MAC, as we want to put issues with the use of projections aside. WEAT will be useful in our criticism of the statistical methods involved, as it is simpler, so the explanation and visualizations will be more transparent (and *mutatis mutandis* this criticism will apply to MAC as well), and MAC will be useful in the development of our alternative method as its range of applicability is the widest.

1.4 Methodological problems with cosine-based measures of bias

One issue to consider is the selection of attributes for bias measurement. The word lists used in the literature are often fairly small (5-50). While the papers do employ statistical tests to measure the uncertainty involved, we will later on argue that these method are not proper for the goal at hand and show that a more appropriate use of statistical methods leads to estimates of uncertainty that are rather epistemologically pessimistic.

Check Manzani

sensitivity to choice of words

Let's think about MAC, using the case of religion-related stereotypes. In the original paper, words from all three religions were compared against all of the stereotypes. One reason this is problematic is that no distinction between cases in which the stereotype is associated with a given religion, as opposed to the situation in which it is associated with another one, is made. This is problematic, as not all of the stereotypical words have to be considered as harmful for all of the religions. One should investigate the religions separately as some of them may have stronger harmful associations that others.

crossref to a list of words an explanation

The interpretation of the results is also a challenge. In [6] we can find summaries of average cosine distances per group (such as gender, race, or religion). For instance, for religion, here is the relevant fragment of table:

add p-values in the table

kable the table

(MAC stand for mean average cosine similarity, although in reality the the table contains mean cosine distances). What may attract attention is the fact that the value of cosine distance in "Biased" category is already quite high (i.e. close to 1) even before debiasing. High cosine distance indicates low cosine similarity between values. One could think that average cosine similarity equal to approximately 0.141 is not significant enough to consider it as bias. However, the authors still aim to mitigate such "bias" to make the distance even larger. Methodologically the question is, on what basis is this small similarity still considered as a proof of the presence of bias, and whether these small changes are meaningful.

The underlying problem here is that in the paper there is no control group. One should also include control groups to have a way of comparing the results for a supposedly stereotyped group with the results for sets of neutral or human-related neutral words. In our approach later on, we distinguish between stereotypes associated with a given group, stereotypes associated with different groups, and introduce control groups: neutral words and stereotype-free human predicates.

bliskosc znaczeniowa a ko-okurencja

comparison should be made to actual frequencies, to separate bias caused by cosine similarty to upstream bias

unclear czy to sie przeklada na downstream effect

[5] By the cosine distance in the literature the authors mean 1-cosine similarity; note however that this terminology is slightly misleading, as mathematically it is not a distance measure, because it does not satisfy the triangle inequality, as generally $dist(A,C) \not\leq dist(A,B) + dist(B,C)$; we'll keep using this mainstream terminology though.

1.5 Metrics that pre-average are a bad guide

In contrast, statistical intervals will help us decide whether a given cosine similarity is high enough to consider the words to be more similar than if we chose them at random. We will use highest posterior density intervals, in line with Bayesian methodology.

Crucially, these approaches use means of mean average cosine similarities to measure similarity between protected word and harmful stereotypes. If one takes a closer look at the individual values that are taken for the calculations, it turns out that there are quite a few outliers and surprisingly dissimilar words. This issue will become transparent when we inspect the visualizations of individual cosine distances, following the idea that one of the first step to understand data is to look at it.

With such a method the uncertainty involved is not really considered which makes it even more difficult to give reasonable interpretations of the results. We propose the use of Bayesian method to obtain some understanding of the influence the uncertainty has on the interpretation of final results.

odleglosc nie musi uchwytywac relacji semantycznej zeby oceniac bias

zalozenie, bardziej ze jezeli odleglosci przekladaja sie na downstream, to warto patrzec na odleglosci, nawet bez glebszej filozoficznej interpretacji ich

$s(X,Y,A,B)$ is the statistic used in the significance test, and the p-value is obtained by bootstrapping: it is the frequency of $s(X_i,Y_i,A,B) > s(X,Y,A,B)$ for all equally sized partitions X_i,Y_i of $X \cup Y$. The effect size is computed by normalizing the difference in means as follows:

$$bias(A,B) = \frac{\mu(\{s(x,A,B)\}_{x\in X}) - \mu(\{s(y,A,B)\}_{y\in Y})}{\sigma(\{s(w,A,B)\}_{w\in X\cup Y})} \tag{8}$$

The t-tests they employ are run on average cosines used to calculate MAC.

2 Pre-averaging and manufactured certainty

2.1 General problem with pre-averaging

2.2 Simulations for WEAT

2.3 Simulations for MAC

3 Bayesian method

3.1 Bernstein approach

3.2 Chasing metrics

3.3 Bayesian method introduction

3.3.1 What is Bayesian statistics?

Bayesian statistics is a theory which enables one to assign epistemological uncertainty with the use of probability. It is a system that allows to express a degree of belief in a certain event. Bayesian methods use existing prior beliefs together with real data to receive posterior beliefs. Bayesian inference together with probability tools may be used to quantify the uncertainty in the inferences.

3.3.2 How does it differ from the classical statistics?

One could ask what is the difference between Bayesian statistics and the classical, frequentist approach. Broadly speaking the Frequentist statistics looks, as the name suggests, at the frequency of an event in a long run and based on that assigns its probability. In Bayesian statistics the approach is more epistemological as one may include his/her prior beliefs on the likelihood of an event and update it accordingly to the available data which results in the posterior probability of an event. What is more in a classical approach, the parameters are fixed whereas in Bayesian statistics the parameters are assumed to be random variables to which one assigns a probability distribution. In contrast with CIs, the posterior distributions (and their highest posterior density intervals HPDIs, the narrowest intervals containing a certain ratio of the area under the curve) are easily interpretable and have direct relevance for the question at hand.

3.3.3 How does the Bayesian model work?

Referring to the *Statistical rethinking* one may distinguish three steps that are necessary to design Bayesian model: 1. Data story: Think about the narrative on how the data can emerge. 2. Update: Let the model learn from the data. 3. Evaluate: Verify whether it is a good representation of the phenomenon.
A Bayesian model consists of the variables and the distributions that define these variables. The priors that one assigns provide the likelihood of each parameter before any data has been seen. Then using real data one may apply numerical techniques like Markov chain Monte Carlo to calculate the posterior distribution.

3.3.4 Interpreting model's results

The standard output from the model's learning may contain the following pieces of information.
WAIC -

3.3.5 What are hierarchical models?

Hierarchical models are also known as multilevel models, random effects, varying effects, or mixed effects models. In *Statistical rethinking* one may read that the central device of the multilevel models is to learn the strength of the prior from the data itself. Multilevel models may be regarded as adaptive regularization, meaning that the model tunes itself to get the most optimal skepticism. It is worth mentioning that with these types of models adding more parameters can sometimes lead to the effect where effective number of parameters (the penalty term in WAIC) is actually reduced. The overfitting may be more dependent on how the parameters are related to one another than with how many parameters one inputs into the model.

3.3.6 Coding example from our model

To build a model one may use original *Stan* which is a state-of-the-art platform for statistical modeling and high-performance statistical computation (https://mc-stan.org/). The simpler option is to use *Rethinking* package from the *Statistical rethinking* course. With one of the functions *ulam* one may build RStan models from the formulas in an easy manner.
Explain the model?

```
buildModel <- function(dataset){
  options(buildtools.check =
  function(action) TRUE )

  modelResult <- ulam(
    alist(
      distance ~ dnorm(mu,sigma),
      mu <- d[pwi] * different
      + a[pwi] * associated
      + h[pwi] * human
      + n[pwi] * none,

      d[pwi] ~ dnorm(dbar, dsigmabar),
      a[pwi] ~ dnorm(abar, asigmabar),
      h[pwi] ~ dnorm(hbar, hsigmabar),
      n[pwi] ~ dnorm(nbar, nsigmabar),

      dbar ~ dnorm(1,.3),
      abar ~ dnorm(1,.3),
      hbar ~ dnorm(1,.3),
      nbar ~ dnorm(1,.3),
      dsigmabar ~ dexp(2),
```

```
      asigmabar ~ dexp(2),
      hsigmabar ~ dexp(2),
      nsigmabar ~ dexp(2),
      sigma <- s[connection],
      s[connection] ~  dexp(2)
    ),

    data = dataset, chains=2 ,
    iter=8000 , warmup=1000,
    log_lik = TRUE, cores = 4
  )
}
```

3.3.7 Limitations

Each method has its own limitations and Bayesian statistics is no exception. One may think about a few limitations with this approach. One of them is that there is no final and correct way to choose a prior. At the same time prior beliefs can be explicitly stated, debated, or often found in previous research which may give good justification for a given set of priors. Finally, Bayesian method may be computationally expensive which may be a limiting factor when considering to use it in a more real-time environment.

3.4 Existing applications to NLP and perhaps to bias

3.5 Model

4 Results

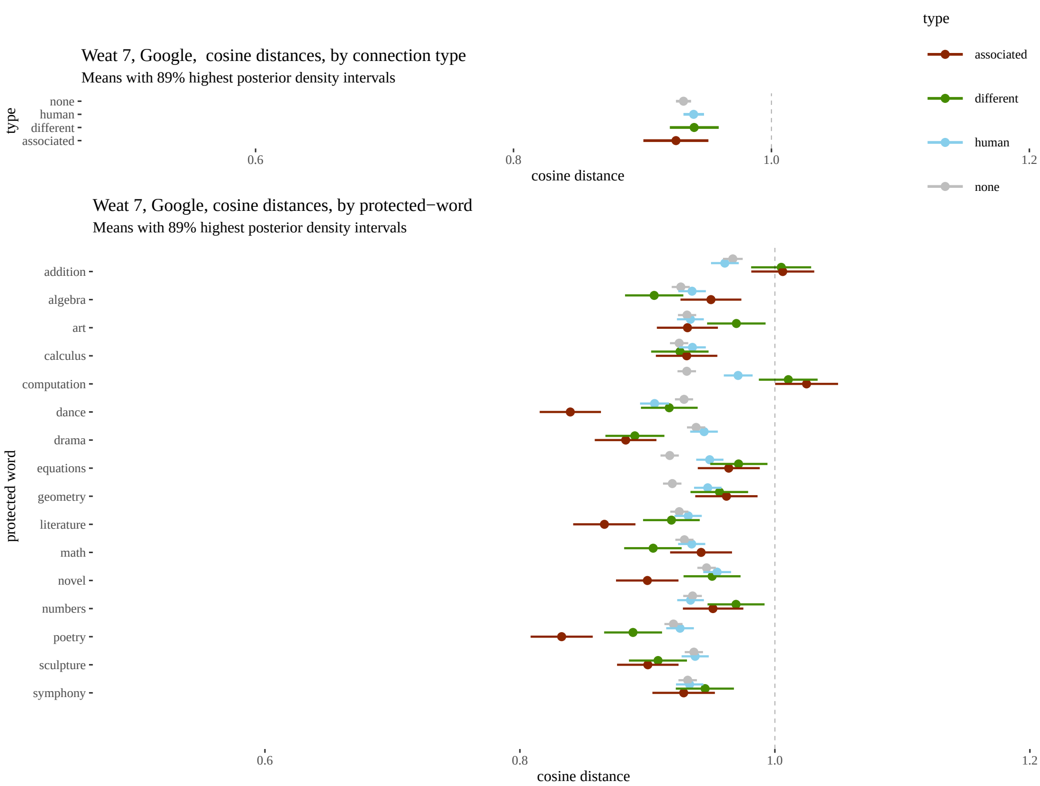

Figure 1: dsds

5 Discussion and summary

6 APPENDIX

6.1 Examples of WEAT and MAC calculations

Word lists

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR* abs/1607.06520, (2016). Retrieved from http://arxiv.org/abs/1607.06520

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. DOI:https://doi.org/10.1126/science.aal4230

[3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (April 2018), E3635–E3644. DOI:https://doi.org/10.1073/pnas.1720347115

[4] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. DOI:https://doi.org/10.18653/v1/N19-1061

[5] Anne Lauscher and Goran Glavas. 2019. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR* abs/1904.11783, (2019). Retrieved from http://arxiv.org/abs/1904.11783

[6] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 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

[7] Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics* 46, 2 (June 2020), 487–497. DOI:https://doi.org/10.1162/coli_a_00379

[8] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 101–115. DOI:https://doi.org/10.1037/1089-2699.6.1.101