What are the biases in my word embedding?

Nathaniel Swinger* Lexington High School
Maria De-Arteaga* Carnegie Mellon University
Neil Thomas Heffernan IV Shrewsbury High School
Mark DM Leiserson University of Maryland
Adam Tauman Kalai Microsoft Research

December 27, 2019

Warning: This paper includes stereotypes and terms which are offensive in nature. It is important, however, for researchers to be aware of the biases contained in word embeddings.

Abstract

This paper presents an algorithm for enumerating biases in word embeddings. The algorithm exposes a large number of offensive associations related to sensitive features such as race and gender on publicly available embeddings, including a supposedly "debiased" embedding. These embedded biases are concerning in light of the widespread use of word embeddings. The associations are identified by geometric patterns in word embeddings that run parallel between people's names and common lower-case words and phrases. The algorithm is highly unsupervised: it does not even require the sensitive groups (such as gender, race, or religion) to be pre-specified. This is desirable because: (a) it may not always be easy to identify all vulnerable groups a priori; and (b) it makes it easier to identify biases against *intersectional* groups, which depend on combinations of sensitive features. The inputs to our algorithm are a list of target tokens, e.g. names, and a word embedding. It outputs a number of Word Embedding Association Tests (WEATs) that capture various biases present in the data. We illustrate the utility of our approach on publicly available word embeddings and lists of names, and evaluate its output using crowdsourcing. We also show how removing names may not remove potential proxy bias.

1 Introduction

Bias in data representation is an important element of fairness in Artificially Intelligent systems (Barocas et al., 2017; Caliskan, Bryson, and Narayanan, 2017; Zemel et al., 2013; Dwork et al., 2012). We consider the problem of *Unsupervised Bias Enumeration* (UBE), discovering biases automatically from an unlabeled data representation. There are multiple reasons why one might want such an algorithm. First, social scientists can use it as a tool to study human bias, as data analysis is increasingly common in social studies of human biases (Garg et al., 2018; Kozlowski, Taddy, and Evans, 2018). Second, identifying bias is a natural step in "debiasing" representations (Bolukbasi et al., 2016). Finally, it can help in avoiding systems that perpetuate these biases: problematic biases can raise red flags for engineers, while little or no bias can be a useful green light indicating that a representation is usable. This awareness may be useful for identifying problems or even suggesting that a representation should not be used in a certain application. While deciding which biases are problematic is ultimately application specific, UBE may be useful in a "fair ML" pipeline.

We design a UBE algorithm for word embeddings, which are commonly used representations of tokens (e.g. words and phrases) that have been found to contain harmful bias (Bolukbasi et al., 2016). Researchers linking these biases to human biases proposed the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan, 2017). The WEAT draws its inspiration from the Implicit Association Test (IAT), a widely-used approach to measure human bias (Greenwald, McGhee, and Schwartz, 1998). An IAT $\mathcal{T} = (X_1, A_1, X_2, A_2)$ compares two sets of *target tokens* X_1 and X_2, such as female vs. male names, and a pair of opposing sets of *attribute tokens* A_1 and A_2, such as workplace vs. family-themed words. Average differences in a person's response times when asked to link tokens that have anti-stereotypical vs. stereotypical relationships have been shown to indicate the strength of association between concepts. Analogously, the WEAT uses vector similarity across pairs of tokens

* Indicates equal contribution.

Word2Vec trained on Google news			fastText trained on the Web			GloVe trained on the Web		
w2v F8	**w2v F11**	**w2v F6**	**fast F10**	**fast F7**	**fast F5**	**glove F8**	**glove F7**	**glove F5**
illegal immigrant	aggravated robbery	subcontinent	n*****	jihad	s******	turban	cartel	pornstar
drug trafficking	aggravated assault	tribesmen	f*****	militants	maid	saree	undocumented	hottie
deported	felonious assault	miscreants	dreads	caliphate	busty	hijab	culpable	nubile

Table 1: Terms associated with name groups (see Tables 3 and 6 for name groups **w2v F8**, etc.) generated from three popular pre-trained word embeddings that were rated by crowd workers as both most offensive and aligned with societal biases. These associations do *not* reflect the personal beliefs of the crowd workers or authors of this paper. See Appendix A for a discussion of the bleep-censored words.

in the sets to measure association strength. As in the case of the IAT, the inputs for a WEAT are sets of tokens $\mathcal{T}$ predefined by researchers.

Our UBE algorithm takes as input a word embedding and a list of target tokens, and *outputs* numerous tests $\mathcal{T}_1, \mathcal{T}_2, \ldots$, that are found to be statistically significant by a method we introduce for bounding false discovery rates. A crowdsourcing study of tests generated on three publicly-available word embeddings and a list of names from the Social Security Administration confirms that the biases enumerated are largely consistent with human stereotypes. The generated tests capture racial, gender, religious, and age biases, among others. Table 1 shows the name/word associations output by our algorithm that were rated most offensive by crowd workers.

Creating such tests automatically has several advantages. First, it is not feasible to manually author all possible tests of interest. Domain experts normally create such tests, and it is unreasonable to expect them to cover all possible groups, especially if they do not know which groups are represented in their data. For example, a domain expert based on the United States may not think of testing for caste discrimination, hence biases that an embedding may have against certain Indian last names may go unnoticed. Finally, if a word embedding reveals no biases, this is evidence for lack of bias. We test this by running our UBE algorithm on the supposedly debiased embedding of Bolukbasi et al. (2016).

Our approach for UBE leverages two geometric properties of word embeddings, which we call the "parallel" and "cluster" properties. The well-known parallel property is that differences between two similar token pairs, such as Mary−John and Queen−King, are often nearly parallel vectors. This suggests that among tokens in a similar topic or category, those parallel to name differences may represent biases, as was found by Bolukbasi et al. (2016) and Caliskan, Bryson, and Narayanan (2017). The cluster property, which were previously unaware of, is that the (normalized) vectors of names and words cluster into semantically meaningful groups. For names, the clusters capture social structures such as gender, religion, and others. For words, clusters of words include word categories on topics such as food, education, occupations, and sports. We use these properties to design a UBE algorithm that outputs WEATs.

Technical challenges arise around any procedure for enumerating biases. First, the combinatorial explosion of comparisons among multiple groups parallels issues in human IAT studies as aptly described by Bluemke and Friese (2008): "The evaluation of multiple target concepts such as social groups within a multi-ethnic nation (e.g. White vs. Asian Americans, White vs. African Americans, African vs. Asian Americans; Devos and Banaji, 2005) requires numerous pairwise comparisons for a complete picture". We alleviate this problem, paralleling that work on human IATs, by generalizing the WEAT to n groups for arbitrary n. The second problem, for any UBE algorithm, is determining statistical significance to account for multiple hypothesis testing. To do this, we introduce a novel rotational null hypothesis specific to word embeddings. Third, we provide a human evaluation of the biases, contending with the difficulty that many people are unfamiliar with some groups of names.

Beyond word embeddings and IATs, related work in other subjects is worth mention. First, a body of work studies fairness properties of classification and regression algorithms (e.g. Dwork et al., 2012; Kearns et al., 2017). While our work does not concern supervised learning, it is within this work that we find one of our main motivations–the importance of accounting for intersectionality when studying algorithmic biases. In particular, Buolamwini and Gebru (2018) demonstrate accuracy disparities in image classification highlighting the fact that the magnitude of biases against an intersectional group may go unnoticed when only evaluating for each protected feature independently. Finally, while a significant portion of the empirical research on algorithmic fairness has focused on the societal biases that are most pressing in the countries where the majority of researchers currently conducting the work are based, the literature also contains examples of biases that may be of particular importance in other parts of the world (Shankar et al., 2017; Hoque et al., 2017). UBE can aspire to be useful in multiple contexts, and enable the discovery of biases in a way that relies less on enumeration by domain experts.

2 Definitions

A d-dimensional word embedding consists of a set of tokens $\mathcal{W}$ with a nonzero vector $\boldsymbol{w} \in \mathbb{R}^d$ associated with each token $w \in \mathcal{W}$. Vectors are displayed in boldface. As is standard, we refer to the *similarity* between tokens v and w by the cosine of their vector angle, $\cos(\boldsymbol{v}, \boldsymbol{w})$. We write $\overline{\boldsymbol{v}} = \boldsymbol{v}/|\boldsymbol{v}|$ to be the vector normalized to unit-length associated with any vector $\boldsymbol{v} \in \mathbb{R}^d$ (or 0 if $\boldsymbol{v} = 0$). This enables us to conveniently write the similarity between tokens v and w as an inner product, $\cos(\boldsymbol{v}, \boldsymbol{w}) = \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{w}}$. For token set S, we write $\overline{\boldsymbol{S}} = \sum_{v \in S} \overline{\boldsymbol{v}}/|S|$ so that $\overline{\boldsymbol{S}} \cdot \overline{\boldsymbol{T}} = \mathrm{mean}_{v \in S} \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{w}}$ is the mean similarity between pairs of tokens in sets S, T. We denote the set difference between S and T by $S \setminus T$, and we denote the first n whole numbers by $[n] = \{1, 2, \ldots, n\}$.

2.1 Generalizing Word Embedding Association Tests

We assume that there is a given set of possible targets $\mathcal{X}$ and attributes $\mathcal{A}$. Henceforth, since in our evaluation all targets are names and all attributes are lower-case words (or phrases), we refer to targets as names and attributes as words. Nonetheless, in principle, the algorithm can be run on any sets of target and attribute tokens. Caliskan, Bryson, and Narayanan (2017) define a WEAT statistic for two equal-sized groups of names $X_1, X_2 \subseteq \mathcal{X}$ and words $A_1, A_2 \subseteq \mathcal{A}$ which can be conveniently written in our notation as,

$$s(X_1, A_1, X_2, A_2) \stackrel{\text{def}}{=} \left(\sum_{x \in X_1} \overline{\boldsymbol{x}} - \sum_{x \in X_2} \overline{\boldsymbol{x}} \right) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{A}}_2).$$

In studies of human biases, the combinatorial explosion in groups can be avoided by teasing apart *Single-Category* IATs which assess associations one group at a time (e.g. Karpinski and Steinman, 2006; Penke, Eichstaedt, and Asendorpf, 2006; Bluemke and Friese, 2008). In word embeddings, we define a simple generalization for $n \geq 1$, nonempty groups $X_1, \ldots, X_n$ of arbitrary sizes and words $A_1, \ldots, A_n$, as follows:

$$g(X_1, A_1, \ldots, X_n, A_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \stackrel{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\boldsymbol{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

Note that g is symmetric with respect to ordering and weights groups equally regardless of size. The definition differs for $n = 1$, otherwise $g \equiv 0$.

The following three properties motivate this as a "natural" generalization of WEAT to one or more groups.

Lemma 1. *For any $X_1, X_2 \subseteq \mathcal{X}$ of equal sizes $|X_1| = |X_2|$ and any nonempty $A_1, A_2 \subseteq \mathcal{A}$,*

$$s(X_1, A_1, X_2, A_2) = 2|X_1|\, g(X_1, A_1, X_2, A_2)$$

Lemma 2. *For any nonempty sets $X \subset \mathcal{X}$, $A \subset \mathcal{A}$, let their complements sets $X^c = \mathcal{X} \setminus X$ and $A^c = \mathcal{A} \setminus A$. Then,*

$$g(X, A) = 2g(X, A, \mathcal{X}, \mathcal{A}) = 2\frac{|X^c|}{|\mathcal{X}|}\frac{|A^c|}{|\mathcal{A}|} g(X, A, X^c, A^c)$$

Lemma 3. *For any $n > 1$ and nonempty $X_1, X_2, \ldots, X_n \subseteq \mathcal{X}$ and $A_1, A_2, \ldots, A_n \subseteq \overline{\boldsymbol{\mathcal{A}}}$,*

$$g(X_1, A_1, \ldots, X_n, A_n) = \sum_{i \in [n]} g(X_i, A_i) - \sum_{i,j \in [n]} \frac{g(X_i, A_j)}{n}$$

Lemma 1 explains why we call it a generalization: for $n = 2$ and equal-sized name sets, the values are proportional with a factor that only depends on the set size. More generally, g can accommodate unequal set sizes and $n \neq 2$.

Lemma 2 shows that for $n = 1$ group, the definition is proportional the WEAT with the two groups X vs. all names $\mathcal{X}$ and words A vs. $\mathcal{A}$. Equivalently, it is proportional to the WEAT between X and A and their compliments.

Finally, Lemma 3 gives a *decomposition* of a WEAT into n^2 single-group WEATs $g(X_i, A_j)$. In particular, the value of a single multi-group WEAT reflects a combination of the n association strengths between X_i and A_i and n^2 disassociation strengths between X_i and A_j. As discussed on the literature on IATs, a large effect could reflect a strong association between X_1 and A_1 or X_2 and A_2, a strong disassociation between X_1 and A_2 or X_2 and A_1, or some combination of these factors. Proofs are deferred to Appendix B.