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

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. These 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 tokens. The algorithm is highly unsupervised: it does not even require the sensitive features to be pre-specified. This is desirable because: (a) many forms of discrimination–such as racial discrimination–are linked to social constructs that may vary depending on the context, rather than to categories with fixed definitions; 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.

We encourage readers to consult the more complete manuscript on the arXiv.

ACM Reference Format:
Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the Biases in My Word Embedding?. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3306618.3314270

1 INTRODUCTION

Bias in data representation is an important element of fairness in Artificially Intelligent systems [2, 7, 10, 24]. We consider the problem of *Unsupervised Bias Enumeration* (UBE): discovering biases automatically from an unlabeled data representation. There are multiple reasons why such an algorithm is useful. 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 [11, 16]. Second, finding bias is a natural step in "debiasing" representations [5]. Finally, it can help in avoiding systems that perpetuate these biases: problematic biases can raise red flags for engineers, who can choose to not use a representation or watch out for certain biases in downstream applications, while little or no bias can be a useful green light indicating that a representation is usable. 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 [5]. Researchers linking these biases to human biases proposed the Word Embedding Association Test (WEAT) [7]. The WEAT draws its inspiration from the Implicit Association Test (IAT), a widely-used approach to measure human bias [12]. 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 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, including racial, gender, religious, and age biases, among others.

Creating such tests automatically has several advantages. First, it is unfeasible for domain experts to manually author all possible tests of interest and cover all possible groups, especially if they do not know which groups are represented in their data. For example,

[*] Indicates equal contribution.

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314270

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. Automatically considering all possibilities also means that if no biases are revealed, this is evidence for lack of bias. We test this by running our UBE algorithm on the supposedly debiased embedding of [5].

Our approach for UBE leverages two geometric properties of word embeddings, which we call the *parallel* and *cluster* properties. The well-known parallel property indicates 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 [5] and [7]. The cluster property, which we were previously unaware of, indicates 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 accompany any procedure for bias enumeration. First, the combinatorial explosion of comparisons among many groups parallels issues in human IAT studies, as described by [4]: "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; 9] requires numerous pairwise comparisons for a complete picture". We alleviate this issue by paralleling work on human IATs, 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. We introduce a novel rotational null hypothesis specific to word embeddings. Third, we provide human evaluation of the biases, contending with the fact that many people are unfamiliar with some sets of names.

Beyond word embeddings and IATs, other related work is worth mention. A body of work studies fairness properties of classification and regression algorithms [e.g. 10, 15]. While our work does not concern supervised learning, we find one of our main motivations within this work–the importance of accounting for intersectionality when studying algorithmic biases. In particular, Buolamwini and Gebru [6] 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 [13, 22]. UBE can aspire to be useful in multiple contexts, and enable the discovery of biases without relying 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 $\overline{\boldsymbol{S}} \cdot \overline{\boldsymbol{T}} = \text{mean}_{v \in S,\, w \in T} \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{w}}$ 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. [7] 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,

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

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. 4, 14, 20]. 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:

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

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

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 to 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.

3 UNSUPERVISED BIAS ENUMERATION ALGORITHM

The inputs to our UBE algorithm are shown in Table 1. The output is m WEATs, each with n groups with associated sets of words and statistical confidences (p-values) in $[0, 1]$. Each WEAT has words from a single category, but several of the m WEATs may yield no significant associations.

At a high level, the algorithm follows a simple structure. It selects n disjoint groups of names $X_1, \ldots, X_n \subset \mathcal{X}$, and m disjoint categories of lower-case words $\mathcal{A}_1, \ldots, \mathcal{A}_m$. All WEATs share the same n name groups, and each WEAT has words from a single category $\mathcal{A}_j$, with t words associated to each X_i. Thus the WEATs can be conveniently visualized in a tabular structure.

For convenience, we normalize all word embedding vectors to be unit length. Note that we only compute cosines between them, and the cosine is simply the inner product for unit vectors. We now detail the algorithm's steps.

3.1 Step 1: Cleaning names and defining groups

We begin with a set of names[1] $\mathcal{X}$, e.g., frequent first names from a database. Since word embeddings do not differentiate between words that have the same spelling but different meanings, we first "clean" the given names to remove names such as "May" and "Virginia", whose embeddings are more reflective of other uses, such as a month or verb and a US state. Our cleaning procedure is similar but slightly more sophisticated to that of [7]. We train a linear Support Vector Machine [scikit-learn's LinearSVC, 19, with default parameters] to distinguish the input names from an equal number of non-names chosen randomly from the most frequent 50,000 words in the embedding. We then remove the 20% of names with smallest margin in the direction identified by the linear classifier.

We then use K-means++ clustering [from scikit-learn, 19, with default parameters] to cluster the normalized word vectors of the names, yielding groups $X_1 \cup \ldots \cup X_n = \mathcal{X}$. Finally, we define $\mu = \sum_i \overline{\boldsymbol{X}}_i / n$.

3.2 Step 2: Defining word categories

To define categories, we cluster the most frequent M lower-case tokens in the word embedding into m clusters using K-means++, yielding clusters of categories $\mathcal{A}_1, \ldots, \mathcal{A}_m$. The constant M is chosen to cover as many recognizable words as possible without introducing too many unrecognizable tokens. As we shall see, categories capture concepts such as occupations, food-related words, and so forth.

name	meaning	default
WE	word embedding	w2v
$\mathcal{X}$	set of names	SSA
n	number of target groups	12
m	number of categories	64
M	number of frequent lower-case words	30,000
t	number of words per WEAT	3
α	false discovery rate	0.05

Table 1: Inputs to the UBE algorithm.

3.3 Step 3: Selecting words $A_{ij} \subset \mathcal{A}_j$

A test $\mathcal{T}_j = (X_1, A_{1j}, \ldots, X_n, A_{nj})$ is chosen with disjoint $A_{ij} \subset \mathcal{A}_j$, each of size $t = |A_{ij}|$. To ensure disjointness,[2] $\mathcal{A}_j$ is first partitioned into n "Voronoi" sets $V_{ij} \subseteq \mathcal{A}_j$ consisting of the words whose embedding is closest to each corresponding center $\overline{\boldsymbol{X}}_i$, i.e.,

$$V_{ij} = \left\{ w \in \mathcal{A}_j \mid i = \arg\max_{i' \in [n]} \overline{\boldsymbol{w}} \cdot \overline{\boldsymbol{X}}_{i'} \right\}$$

It then outputs A_{ij} defined as the t words maximizing the following:

$$\max_{w \in V_{ij}} (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{w}} - \overline{\mathcal{A}}_j)$$

The more computationally-demanding step is to compute, using Monte Carlo sampling, the n p-values for $\mathcal{T}_j$, as described next.

3.4 Step 4: Computing p-values and ordering

To test whether the associations we find are larger than one would find if there was no relationship between the names X_i and words $\mathcal{A}$, we consider the following "**rotational null hypothesis**": the words in the embedding are generated through some process in which the alignment between names and words is random. This is formalized by imagining that a random rotation was applied (multiplying by a uniformly Haar random orthogonal matrix U) to the word embeddings but not to the name embeddings.

Specifically, to compute p-value p_{ij} for each (X_i, A_{ij}), we first compute a score $\sigma_{ij} = (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{A}}_{ij} - \overline{\mathcal{A}})$. We then compute $R = 10,000$ uniformly random orthogonal rotations $U_1, \ldots, U_R \in \mathbb{R}^{d \times d}$, drawn according to the Haar measure. For each rotation, we simulate running our algorithm as if the name embeddings were transformed by U (while the word embeddings remain as is). For each rotation U_r, the sets A_{ijr} chosen to maximize $(\overline{\boldsymbol{X}}_i U_r - \mu U_r) \cdot (\overline{\boldsymbol{w}} - \overline{\mathcal{A}}_j)$, and the corresponding V_{ijr} and the resulting σ_{ijr} are computed. Finally, p_{ij} is the fraction of rotations for which the score $\sigma_{ijr} \geq \sigma_{ij}$ (plus an add-1 penalty standard for Monte Carlo p-values).

Furthermore, since the algorithm outputs many (hundreds) of name/word biases, the Benjamini-Hochberg [1995] procedure is

[1] While the set of names is an input to our system, they could also be extracted from the embedding itself.

[2] If multiplicities are desired, the Voronoi sets V_{ij} could be omitted, optimizing $A_{ij} \subset \mathcal{A}_j$ directly.

w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F10	w2v F11	w2v F12
Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael	Randall	Dashaun	Keith
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses	Dashiell	Jamell	Gabe
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal	Randell	Marlon	Alfred
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai	Jordan	Davonta	Shane
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis	Chace	Demetrius	Stan
+581	+840	+692	+558	+890	+312	+391	+577	+120	+432	+393	+494
98% F	98% F	89% F	85% F	78% F	65% F	59% F	56% F	40% F	27% F	5% F	4% F
1983	1968	1978	1982	1993	1991	1985	1986	1989	1981	1984	1976
4% B	8% B	48% B	10% B	2% B	7% B	4% B	2% B	5% B	10% B	32% B	6% B
4% H	4% H	3% H	9% H	1% H	4% H	9% H	70% H	10% H	3% H	5% H	3% H
3% A	3% A	1% A	11% A	1% A	32% A	4% A	8% A	5% A	4% A	3% A	5% A
89% W	84% W	47% W	69% W	95% W	56% W	83% W	21% W	79% W	83% W	59% W	86% W

Table 2: Illustrative first names (greedily chosen) for $n = 12$ groups on the w2v embedding. Demographic statistics (computed a posteriori) are also shown though were not used in generation, including percentage female (at birth), mean year of birth, and percentage Black, Hispanic, Asian/Pacific Islander, and White.

used to determine a critical p-value that guarantees an α bound on the rate of false discoveries. Finally, to choose an output ordering on significant tests, the m tests are then sorted by the total scores σ_{ij} over the pairs determined significant.

4 EVALUATION

To illustrate the performance of the proposed system in discovering associations, we use a database of first names provided by the Social Security Administration (SSA), which contains number of births per year by sex (F/M) [1]. We use only years 1938-2017 and select only the names that appeared at least 1,000 times, which cover more than 99% of the data by population.

We use three publicly available word embeddings, each with $d = 300$ dimensions and millions of words: w2v, released in 2013 and trained on approximately 100 billion words from Google News [18], fast, trained on 600 billion words from the Web [17], and glove, also trained on the Web using the GloVe algorithm [21].

While it is possible to display the three words in each A_{ij}, the hundreds or thousands of names in each X_i cannot be displayed in the output of the algorithm. Instead, we use a simple greedy heuristic to give five "illustrative" names for each group, which are displayed in the tables in this paper and in our crowdsourcing experiments. The $k + 1^{\text{st}}$ name shown is chosen, given the first k names, so as to maximize the average similarity of the first $k + 1$ names to that of the entire set X_i. Hence, the first name is the one whose normalized vector is most central (closest to the cluster mean), the second name is the one which when averaged with the first is as central as possible, and so forth.

The WEATs can be evaluated in terms of the quality of the name groups and also their associations with words. A priori, it was not clear whether clustering name embeddings would yield any name groups or word categories of interest. For all three embeddings we find that the clustering captures latent groups defined in terms of race, age, and gender, as illustrated in Table 2 for $n = 12$ clusters. While even a few clusters suffice to capture some demographic differences, more clusters yield much more fine-grained distinctions. For example, with $n = 12$ one cluster is of evidently Israeli names (see column I of table 2), which one might not consider predefining a priori since they are a small minority in the U.S. Note that, although we do not have religious statistics for the names, several of the words in the generated associations are religious in nature, suggesting religious biases as well.

Emb.	# significant	% accurate	% offensive
w2v	235	72%	35%
fast	160	80%	38%
glove	442	48%	24%

Table 3: Summary statistics for the WEATs generated using the three embeddings ($n = 12$, $m = 64$). The total number of significant name/word associations, the fraction with which the crowd's choice of name group agreed with that of the generated WEAT (accuracy) among the top-12 WEATs, and the fraction rated as offensive.

4.1 Crowdsourcing Evaluation

We solicited ratings on the biases generated by the algorithm from US-based crowd workers on Amazon's Mechanical Turk[3] platform. The aim is to identify whether the biases found by our UBE algorithm are consistent with (problematic) biases held by society at large. To this end, we asked about society's stereotypes, *not* personal beliefs.

We evaluated the top 12 WEATs generated by our UBE algorithm for the three embeddings, considering $n = 12$ first name groups. Our approach was simple: after familiarizing participants with the 12 groups, we showed the (statistically significant) words and name groups of a WEAT and asked them to identify which words would, stereotypically, be most associated with which names group. A bonus was given for ratings that agreed with most other worker's ratings, incentivizing workers to provide answers that they felt corresponded to widely held stereotypes.

[3] http://mturk.com

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 4: Terms in associations 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.

This design was chosen over a simpler one in which WEATs are shown to individuals who are asked whether or not these are stereotypical. The latter design might support confirmation bias as people may interpret words in such a way that confirms whatever stereotypes they are being asked about. For instance, someone may be able to justify associating the color red with almost any group, a posteriori.

Note that the task presented to the workers involved fine-grained distinctions: for each of the top-12 WEATs, at least 18 workers would each be asked to match the significant $c \leq 12$ word triples to the c name groups (each identified by five names each). For example, workers faced the triple of "registered nurse, homemaker, chairwoman" with $c = 8$ groups of names, half of which were majority female, and the most commonly chosen group matched the one generated: "Janice, Jeanette, Lenna, Mattie, Marylynn." Across the top-12 WEATs over the three embeddings, the mean number of choices c was 8.1, yet the most commonly chosen group (plurality) agreed with the generated group 65% of the time (see Table 3). This is significantly more than one would expect from chance. The top-12 WEATs generated for w2v are shown in Table 5.

One challenge faced in this process was that, in pilot experiments, a significant fraction of the workers were not familiar with many of the names. To address this challenge, we first administered a qualification exam (common in crowdsourcing) in which each worker was shown 36 random names, 3 from each group, and was offered a bonus for each name they could correctly identify the group from which it was chosen. Only workers whose accuracy was greater than 1/2 (which happened 37% of the time) evaluated the WEATs. Accuracy greater than 50% on a 12-way classification indicates that the groups of names were meaningful and interpretable to many workers.

Finally, we asked 13-15 workers to rate associations on a scale of 1-7 of *political incorrectness*, with 7 being "politically incorrect, possibly very offensive" and 1 being "politically correct, inoffensive, or just random." Only those biases for which the most commonly chosen group matched the association identified by the UBE algorithm were included in this experiment. The mean ratings are shown in Table 3 and the terms present in associations deemed most offensive are presented in Table 4.

4.2 Potential Indirect Biases

Naively, one may think that removing names from a dataset will remove all problematic associations. However, as suggested by [5], indirect biases are likely to remain. For example, consider the w2v word embedding, in which *hostess* is closer to *volleyball* than to *cornerback*, while *cab driver* is closer to *cornerback* than to *volleyball.* These associations, taken from columns **F1** and **F11** of Table 5, might serve as a proxy for gender and/or race. For instance, if someone is applying for a job and their profile includes college sports words, such associations encoded in the embedding may lead to racial or gender biases in cases in which there is no professional basis for these associations. In contrast, *volunteer* being closer to *volunteers* than *recruits* may represent a definitional similarity more than a proxy, if we consider proxies to be associations that mainly have predictive power due to their correlation with a protected attribute. While defining proxies is beyond the scope of this work, we do say that $A_{ij}, A_{i'j}, A_{ij'}, A_{i'j'}$ is a *potential indirect bias* if,

$$(\overline{A}_{ij} - \overline{A}_{i'j}) \cdot (\overline{A}_{ij'} - \overline{A}_{i'j'}) > 0. \qquad (1)$$

One way to interpret this definition is that if the embedding were to match the pair of word sets $\{A_{ij}, A_{i'j}\}$ to the pair of word sets $\{A_{ij'}, A_{i'j'}\}$, it would align with the way in which they were generated. For example, does the embedding predict that *hostess-cab driver* better fits *volleyball-cornerback* or *cornerback-volleyball* (but this question is asked with sets of $t = 3$ words)?

We consider all possible fourtuples of significant associations, such that $1 \leq i < i' \leq n$ and $1 \leq j < j' \leq m$. In the case of w2v, 99% of 2,713 significant fourtuples lead to potential indirect biases according to eq. (1). This statistic is of 98% of 1,125 fourtuples and 97% of 1,796 fourtuples for the fast and glove embeddings, respectively. Hence, while names allow us to capture biases in the embedding, removing names is unlikely to be sufficient to debias the embedding.

5 LIMITATIONS

Absent clusters show the limitations of our approach and data. For example, even for large n, no clusters represent demographically significant Asian-American groups. However, if instead of names we use surnames [U.S. Census, 8], a cluster "Yu, Tamashiro, Heng, Feng, Nakamura, +393" emerges, which is largely Asian according to Census data. This distinction may reflect naming practices among Asian Americans [23].

6 CONCLUSIONS

We introduce the problem of Unsupervised Bias Enumeration. We propose and evaluate a UBE algorithm that outputs Word Embedding Association Tests. Unlike humans, where implicit tests are necessary to elicit socially unacceptable biases in a straightforward fashion, word embeddings can be directly probed to output hundreds of biases of varying natures, including numerous offensive and socially unacceptable biases. The racist and sexist associations

w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F11	w2v F12
	cookbook, baking, baked goods	sweet potatoes, macaroni, green beans			saffron, halal, sweets	mozzarella, foie gras, caviar	tortillas, salsa, tequila	kosher, hummus, bagel	fried chicken, crawfish, grams	beef, beer, hams
herself, hers, moms	husband, homebound, grandkids	aunt, niece, grandmother	hubby, socialite, cuddle	twin sister, girls, classmate	elder brother, dowry, refugee camp			bereaved, immigrated, emigrated	younger brother, twin brother, mentally r********	buddy, boyhood, fatherhood
hostess, cheerleader, dietitian	registered nurse, homemaker, chairwoman		supermodel, beauty queen, stripper	helper, getter, snowboarder	shopkeeper, villager, cricketer		translator, interpreter, smuggler		cab driver, jailer, schoolboy	pitchman, retired, pundit
	log cabin, library, fairgrounds	front porch, carport, duplex	racecourse, plush, tenements	picnic tables, bleachers, concession stand	locality, mosque, slum	prefecture, chalet, sauna		synagogues, constructions, hilltop	apartment complex, barbershop, nightclub	
	parish, church, pastoral	pastor, baptized, mourners	goddess, celestial, mystical		fatwa, mosques, martyrs	monastery, papal, convent	rosary, parish priest, patron saint	rabbis, synagogue, biblical		
volleyball, gymnast, setter	athletic director, winningest coach, officiating	leading rebounder, played sparingly, incoming freshman	hooker, footy, stud	sophomore, junior, freshman	leftarm spinner, dayers, leg spinner				cornerback, tailback, wide receiver	
sorority, gymnastics, majoring	volunteer, volunteering, secretarial	guidance counselor, prekindergarten, graduate		seventh grader, eighth grade, seniors	lecturers, institutes, syllabus		bilingual, permanent residency, occupations		incoming freshmen, schoolyard, recruiting	fulltime, professional, apprenticeship
		civil rights, poverty stricken, nonviolent			subcontinent, tribesmen, miscreants	xenophobia, anarchist, oligarchs	leftist, drug traffickers, undocumented	disengagement, intifada, settlers	blacks, segregation, lynching	
tiara, blonde, sparkly	knitting, sewing, beaded	brown eyes, cream colo..., wore	girly, feminine, flirty	brown hair, pair, skates	sari, turban, hijab				dreadlocks, shoulderpads, waistband	mullet, gear, helmet
					dirhams, lakhs, rupees	rubles, kronor, roulette	pesos, remittances, gross receipts	shekels, settlements, corpus		
		grandjury indicted, degree murder, violating probation		child endangerment, vehicular homicide, unlawful possession	chargesheet, absconding, interrogation	absentia, tax evasion, falsification	illegal immigrant, drug trafficking, deported		aggravated robbery, aggravated assault, felonious assault	
	volunteers, crafters, baby boomers	caseworkers, evacuees, attendants	beauties, celebs, paparazzi	setters, helpers, captains	mediapersons, office bearers, newsmen				recruits, reps, sheriffs	

Table 5: The top-12 WEATs output by our UBE algorithm on the w2v embedding. Columns represent name groups X_i from Table 2, rows represent categories A_j (e.g., a cluster of food-related words). Orange indicate associations where the crowd's most commonly chosen name group agrees with that of the generated WEAT. No significant biases generated for w2v F10.

exposed in publicly available word embeddings raise questions about their widespread use.

Acknowledgments. We are grateful to Tarleton Gillespie, other colleagues and the anonymous reviewers for useful feedback.

REFERENCES

[1] Social Security Administration. 2018. Baby Names from Social Security Card Applications - National Level Data. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data Accessed 5 July 2018.

[2] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. *Special Interest Group for Computing, Information and Society (SIGCIS)* (2017).

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[4] Matthias Bluemke and Malte Friese. 2008. Reliability and validity of the Single-Target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. *European journal of social psychology* 38, 6 (2008), 977–997.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[8] Joshua Comenetz. 2016. Frequently occurring surnames in the 2010 Census. *United States Census Bureau* (2016).

[9] Thierry Devos and Mahzarin R Banaji. 2005. American= white? *Journal of personality and social psychology* 88, 3 (2005), 447.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226.

[11] 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 (2018), E3635–E3644.

[12] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[13] Md Hoque, Rawshan E Fatima, Manash Kumar Mandal, Nazmus Saquib, et al. 2017. Evaluating gender portrayal in Bangladeshi TV. *arXiv preprint arXiv:1711.09728* (2017).

[14] Andrew Karpinski and Ross B Steinman. 2006. The single category implicit association test as a measure of implicit social cognition. *Journal of personality and social psychology* 91, 1 (2006), 16.

[15] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv preprint arXiv:1711.05144* (2017).

[16] Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *arXiv preprint arXiv:1803.09288* (2018).

[17] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] Lars Penke, Jan Eichstaedt, and Jens B Asendorpf. 2006. Single-attribute implicit association tests (SA-IAT) for the assessment of unipolar constructs. *Experimental Psychology* 53, 4 (2006), 283–291.

[21] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[22] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv preprint arXiv:1711.08536* (2017).

[23] Ellen Dionne Wu. 1999. âĂIJThey Call Me Bruce, But They Won't Call Me Bruce Jones:âĂİ Asian American Naming Preferences and Patterns. *Names* 47, 1 (1999), 21–50. https://doi.org/10.1179/nam.1999.47.1.21 arXiv:https://doi.org/10.1179/nam.1999.47.1.21

[24] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.