WEFE: The Word Embeddings Fairness Evaluation Framework

Pablo Badilla[1,2] , **Felipe Bravo-Marquez**[1,2] and **Jorge Pérez**[1,2]
[1]Department of Computer Science, Universidad de Chile
[2]Millennium Institute for Foundational Research on Data, IMFD-Chile
{pbadilla, fbravo, jperez}@dcc.uchile.cl

Abstract

Word embeddings are known to exhibit stereotypical biases towards gender, race, religion, among other criteria. Several *fairness metrics* have been proposed in order to automatically quantify these biases. Although all metrics have a similar objective, the relationship between them is by no means clear. Two issues that prevent a clean comparison is that they operate with different inputs, and that their outputs are incompatible with each other. In this paper we propose WEFE, the word embeddings fairness evaluation framework, to encapsulate, evaluate and compare fairness metrics. Our framework needs a list of pre-trained embeddings and a set of fairness criteria, and it is based on checking correlations between fairness rankings induced by these criteria. We conduct a case study showing that rankings produced by existing fairness methods tend to correlate when measuring gender bias. This correlation is considerably less for other biases like race or religion. We also compare the fairness rankings with an embedding benchmark showing that there is no clear correlation between fairness and good performance in downstream tasks.

1 Introduction

Word embeddings are dense vector representations of words trained from document corpora [Mikolov *et al.*, 2013b; Pennington *et al.*, 2014]. They have become a core component of natural language processing (NLP) downstream systems because of their ability to efficiently capture semantic and syntactic relationships between words [Goldberg, 2017]. A widely reported shortcoming of word embeddings is that they are prone to inherit stereotypical social biases (regarding gender, ethnicity, religion, as well as other dimensions) exhibited in the corpora on which they are trained [Garg *et al.*, 2018; Caliskan *et al.*, 2017]. These biases usually show some attributes (e.g., professions, attitudes, traits) being more strongly associated with one particular social group than another. An illustrative example is the vector relationship between words "man" and "woman" being similar to the relationship between words "computer programmer" and "homemaker" [Bolukbasi *et al.*, 2016].

The problem of how to quantify the mentioned biases is currently an active area of research [Caliskan *et al.*, 2017; Garg *et al.*, 2018; Sweeney and Najafian, 2019; Manzini *et al.*, 2019; Gonen and Goldberg, 2019; Ethayarajh *et al.*, 2019; Zhou *et al.*, 2019; Swinger *et al.*, 2019], and several different *fairness metrics* have been proposed in the literature in the past few years. Although all metrics have a similar objective, the relationship between them is by no means clear. Two issues that prevent a clean comparison is that they operate with different inputs (pairs of words, sets of words, multiple sets of words, and so on), and that their outputs are incompatible with each other (reals, positive numbers, $[0, 1]$ range, etc.). Moreover, fairness metrics are usually proposed coupled with a specific de-bias method [Bolukbasi *et al.*, 2016; Manzini *et al.*, 2019]. This implies that one de-bias method exhibiting good results with respect to one fairness metric, not necessarily exhibits the same results with respect to a different metric.

In this paper we propose the Word Embeddings Fairness Evaluation (WEFE) framework for measuring fairness in word embeddings by comparing different metrics. Given a list of pre-trained word embedding models, WEFE calculates a fairness-based ranking of these models by encapsulating existing fairness metrics. In order to do so, we propose an abstract view of a fairness metric receiving a list of input *queries*, each query formed by *target* and *attribute* words. The target words describe the social groups in which fairness is intended to be measured (e.g., women, white people, Muslims), and the attribute words describe traits or attitudes by which a bias towards one of the social groups may be exhibited (e.g, pleasant vs. unpleasant terms). We show how this abstraction allows us to cleanly compare diverse fairness metrics. The detailed explanation can be found in Section 3.

We conduct a case study in which we rank various publicly available pre-trained word embeddings using WEAT [Caliskan *et al.*, 2017], RND [Garg *et al.*, 2018], and RNSB [Sweeney and Najafian, 2019] as fairness metrics. Our results show that for the case of gender bias, fairness rankings produced by different metrics tend to be correlated with each other. This correlation is substantially weaker when we consider other bias dimensions such as ethnicity and religion. This is somehow expected as gender bias is, arguably, the

one that has received the most attention by the community, and most of the metrics have been proposed specially for the case this bias. But this also provides evidence that more work is needed to propose fairness metrics able to consistently rank embeddings for dimensions beyond gender. We also compare fairness rankings against quality rankings obtained from the Word Embeddings Benchmark (WEB) [Jastrzkebski *et al.*, 2017], observing that fairness and quality rankings are not necessarily correlated with each other.

The rest of the paper is organized as follows. In Section 2 we review existing work in word embeddings bias. In Section 3 we describe our framework in detail. In Section 4 we conduct a case study in which our framework is applied to various pre-trained word embeddings models. The main findings and conclusions are discussed in Section 5.

2 Related Work

There are several architectures and training techniques that can be used for learning word embeddings from document corpora. The great majority of them are based on the distributional semantics hypothesis: words that appear in similar contexts tend to have similar meanings. Consequently, similar words tend to be mapped to closely located vectors.

The word2vec library implements two popular architectures: skip-gram and continuous bag of words, and two optimization techniques: negative sampling and hierarchical softmax [Mikolov *et al.*, 2013b; Mikolov *et al.*, 2013a]. Other popular embedding models are fastText [Bojanowski *et al.*, 2017], GloVe [Pennington *et al.*, 2014], Lexvec [Salle *et al.*, 2016; Salle *et al.*, 2016], and ConceptNet [Speer *et al.*, 2017]. ConceptNet goes beyond distributional semantics by incorporating knowledge graph relationships into the learning process. It is important to remark that any different parameterization of these models (e.g., the input corpus, number of vector dimensions, training method) will lead to different vectors.

As previously pointed out, word embeddings are prone to perpetuate biases and prejudices contained in the corpora on which they are trained. Below we review previous studies in this field.

The Word Embedding Association Test (WEAT), proposed by Caliskan et al. [2017], is based on the Implicit-association test (IAT) used in psychology. WEAT measures the degree of association between two sets of target words and two sets of attribute words. Target words represent social groups and attribute words represent attitudes (such as pleasant and unpleasant) as well as professions and occupations. This metric is calculated by performing arithmetic operations between the embeddings vectors of the words from each set. The results of this work reveal biases regarding ethnicity (in relation to pleasantness) and gender (in relation to occupations).

In a similar way, the Relative Norm Distance (RND) metric was proposed by Garg et al. [2018] to study temporal biases in diachronic corpora. Embeddings trained on different periods of time were evaluated using the proposed metric in relation to gender and ethnic biases. The results revealed that certain adjectives and occupations became more closely related to certain social groups over time. In contrast to WEAT, RND compares the embeddings of two sets of group words against a single set of neutral words. Group words represent social groups (e.g., gender, ethnicity) and neutral words correspond to words that are not intrinsically related to any social group (e.g, firefighter, doctor).

The Relative Negative Sentiment Bias (RNSB) metric proposed by Sweeney and Najafian [2019] relies on a sentiment lexicon of positive and negative words for measuring bias. The approach trains a logistic regression on the word embeddings matching the words of the lexicon, which is then applied to a set of national origin identity terms such as American, Mexican, and Canadian. The metric is calculated as the Kullback-Leibler (KL) divergence between the negative sentiment probability of the identity terms (after normalization) and a uniform distribution. The novelties of this metric are: 1) it can be directly applied to more than 2 social groups, and 2) it is used to rank different pre-trained embedding models according to fairness.

There have also been attempts to automatically reduce bias in pre-trained word embeddings. Bolukbasi et al. [2016] observed that there is one direction in the embedding space that largely captures gender. The proposed de-biasing approach sets gender neutral words (e.g., occupations) to zero in the subspace generated by the gender direction. However, Gonen and Goldberg [2019] have argued that this approach only hides the bias but does not eliminate it completely.

3 Framework

In this section we formally define the WEFE framework. WEFE works over pretrained word embeddings. We assume that a word embedding model $\mathbf{M}$ is simply a function mapping a word w to a vector $\mathbf{M}(w)$ in $\mathbb{R}^d$, where d is called the *dimension* of the embedding. For the rest of this article, and when the embedding model is clear from the context, words will not be explicitly distinguished from their corresponding embedding vectors.

3.1 WEFE Building Blocks

We follow the previous work on bias by modeling it as an indication of strong association between certain attributes (e.g., health occupations) and social groups (e.g., females). We next formally describe the main parts of WEFE.

Target set. A target word set (denoted by T) corresponds to a set of words intended to denote a particular social group, which is defined by a certain criterion. This criterion can be any character, trait or origin that distinguishes *groups of people* from each other e.g., gender, social class, age, and ethnicity. For example, if the criterion is gender we can use it to distinguish two groups, *women* and *men*. Then, a set of target words representing the *women* social group could contain words like "she", "woman", "girl", etc. Analogously, the target words for the *men* social group could include "he", "man", "boy", etc. It should be noticed that constructing target sets of words that represent groups of people is a subjective procedure.

Attribute set. An attribute word set (denoted by A) is a set of words representing some attitude, characteristic, trait, occupational field, etc. that can be associated with *individuals* from any social group. For example, the set of *science*

attribute words could contain words such as "technology", "physics", "chemistry", while the *art* attribute words could have words like "poetry", "dance", "literature". As for the case of target words, constructing attribute sets of words is a subjective procedure.

Query. A *query* is a pair $Q = (\mathcal{T}, \mathcal{A})$ in which $\mathcal{T}$ is a set of target word sets, and $\mathcal{A}$ is a set of attribute word sets. That is $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ where every T_i is a target word set, and $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ where every A_j is an attribute word set. For example, consider the target word sets

$$\begin{aligned} T_{\text{women}} &= \{\text{she}, \text{woman}, \text{girl}, \ldots\}, \\ T_{\text{men}} &= \{\text{he}, \text{man}, \text{boy}, \ldots\}, \end{aligned}$$

and the attribute word sets

$$\begin{aligned} A_{\text{science}} &= \{\text{math}, \text{physics}, \text{chemistry}, \ldots\}, \\ A_{\text{art}} &= \{\text{poetry}, \text{dance}, \text{literature}, \ldots\}. \end{aligned}$$

Then the following is a query in our framework

$$Q = (\{T_{\text{women}}, T_{\text{men}}\}, \{A_{\text{science}}, A_{\text{art}}\}). \tag{1}$$

Queries are the main building blocks used by fairness metrics to measure bias of word embedding models. But before we explain how fairness metrics work in our context, we need to introduce some further technicalities.

Templates and Subqueries. A *query template* is simply a pair $(t, a) \in \mathbb{N} \times \mathbb{N}$. We say that query $Q = (\mathcal{T}, \mathcal{A})$ satisfies a template (t, a) if $|\mathcal{T}| = t$ and $|\mathcal{A}| = a$. For example, the query in equation (1) above, satisfies the template $(2, 2)$. A template can also be used to produce all subqueries that satisfy the template. Formally, given query $Q = (\mathcal{T}, \mathcal{A})$ and template $s = (t, a)$, we denote by $Q(s)$ as the set of all queries $Q' = (\mathcal{T}', \mathcal{A}')$ such that $\mathcal{T}' \subseteq \mathcal{T}$, $\mathcal{A}' \subseteq \mathcal{A}$, and Q' satisfies template s, that is $|\mathcal{T}'| = t$ and $|\mathcal{A}'| = a$. For example, given the query Q in equation (1) above, the template $(2, 1)$ produces two subqueries

$$\begin{aligned} Q_1 &= (\{T_{\text{women}}, T_{\text{men}}\}, \{A_{\text{science}}\}) \\ Q_2 &= (\{T_{\text{women}}, T_{\text{men}}\}, \{A_{\text{art}}\}) \end{aligned}$$

and then $Q(s) = \{Q_1, Q_2\}$. As we later show, templates can be used to solve the *input mismatch* of fairness metrics.

Fairness Metrics. Intuitively, a fairness metric is a function that quantifies the degree of association between target and attribute words in a word embedding model. In our framework, every fairness metric is defined as a function that has a query and a model as input, and produces a real number as output. As we have mentioned in Section 2, several fairness metrics have been proposed in the literature. But, using our terminology, not all of them share a common input template for queries. Thus, we assume that every fairness metric comes with a template that essentially defines the shape of the input queries supported by the metric. For instance, as we later show in Section 4.1, a metric such as WEAT [Caliskan *et al.*, 2017] has a $(2, 2)$ template, while the RND metric [Garg *et al.*, 2018] has a $(2, 1)$ template.

Formally, let F be a fairness metric with template $s_F = (t_F, a_F)$. Given an embedding model $\mathbf{M}$ and a query Q that satisfies s_F, the metric produces the value $F(\mathbf{M}, Q) \in \mathbb{R}$ that quantifies the degree of bias of $\mathbf{M}$ with respect to query Q.

We still have the problem of how to interpret the value $F(\mathbf{M}, Q)$. Although it depends on every particular metric, we assume that the metric is equipped with a total order relation $\leq_F$ that establishes what is to be considered as *less biased*. That is, if we fix a query Q and we consider two different models $\mathbf{M_1}$ and $\mathbf{M_2}$, then $F(\mathbf{M_1}, Q) \leq_F F(\mathbf{M_2}, Q)$ states that model $\mathbf{M_1}$ *is less biased than model* $\mathbf{M_2}$ when measuring bias with respect to query Q. Notice that with this order relation, we can prescind of actually interpreting the value given by F, and just use it to compare embedding models, which is exactly what the ranking part of WEFE does.

3.2 WEFE Ranking Process

Next, we will show how to rank by fairness the embeddings models using multiple queries and multiple fairness metrics. Our starting point is composed of three sets:

- a set $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_r\}$ of predefined queries where each Q_i represents a particular bias test over a certain criterion,
- a set $\mathcal{M} = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n\}$ of pre-trained word embedding models, and
- a set $\mathcal{F} = \{F_1, \ldots, F_m\}$ of fairnes metrics, where every F_i comes with its particular template $s_i = (t_i, a_i)$ and order relation $\leq_{F_i}$.

Creating the scores matrix. Lets fix a fairness metric $F \in \mathcal{F}$ and assume that $s = (t, a)$ is its associated query template. The first step is to update $\mathcal{Q}$ by adding all subqueries that satisfies the template. Formally, we create the new set

$$\mathcal{Q}_F = Q_1(s) \cup Q_2(s) \cup \cdots \cup Q_r(s)$$

where $Q_i(s)$ is the set of all subqueries of Q_i that satisfies template $s = (t, a)$. We note that $\mathcal{Q}_F$ is usually bigger than $\mathcal{Q}$ (they coincide if the template of the metric is satisfied by all the original queries in $\mathcal{Q}$).

Now for a fixed embeddings model $\mathbf{M} \in \mathcal{M}$ we can compute the value $F(\mathbf{M}, Q)$ for every $Q \in \mathcal{Q}_F$. We can think of these values as a row vector of fairness scores, where every component of the vector corresponds to a different query. We repeat this process for every model $\mathbf{M}_i \in \mathcal{M}$ to construct a scores matrix associated with the fairness metric F. This matrix is of dimensions $|\mathcal{M}| \times |\mathcal{Q}_F|$.

Creating the rankings. The next step is to create the ranking. First, we aggregate each of the scores by embedding model (for each row). To do this, we need to choose an aggregation function that is consistent with the metric F. In particular, we need to ensure that the aggregation satisfies the following monotonicity property with respect to $\leq_F$. Let x, y, x' and y' be arbitrary values in $\mathbb{R}$, and assume that $x \leq_F y$ and $x' \leq_F y'$. Then it must hold that $\text{agg}(x, x') \leq_F \text{agg}(y, y')$. For most of the metrics that we use in our case study, an aggregation function such as the mean of the absolute values of the scores would satisfy this property. But for more complicated metrics deciding on a good aggregation function might not be a trivial matter.

After aggregating the scores we end up with a column vector of size $|\mathcal{M}|$ over which we can use $\leq_F$ to construct a ranking for all the embeddings in $\mathcal{M}$. For us, this ranking is represented by another column vector which values are a permutation of the values in $\{1, 2, \ldots, \mathcal{M}\}$ stating the index for each embedding model in the generated ranking. This ranking is generated in an ascending way, that is, smaller scores get the top positions.

Gathering rankings in a final matrix. Finally, we can repeat the previous process for each one of the fairness metrics in $\mathcal{F}$ to obtain a final matrix of size $|\mathcal{M}| \times |\mathcal{F}|$ containing the ranking indexes of every embedding model for every metric. In our case study, we use this matrix to study correlations among the fairness rankings produced by different metrics and for different sets of queries.

There are several aspects of the process that should be noticed. First, the dimensions of the final matrix ($|\mathcal{M}| \times |\mathcal{F}|$) is independent of the queries used to define the bias that we are considering. Moreover, every column in this matrix represents a fairness metric as a permutation of the same set of integers ($\{1, 2, \ldots, \mathcal{M}\}$). These two aspects allow us to effectively compare all different fairness metrics even though they can receive different forms of queries as inputs, and produce different scores as outputs. We also notice that we can compare all metrics without needing to actually change any of its particularities. Finally, any other meaningful ranking of embeddings can be added to this matrix and the correlations and comparisons can still be computed. In our case study, we add a performance ranking obtained from the Word Embedding Benchmark [Jastrzkebski *et al.*, 2017]

4 Case Study

In this section we instantiate our framework to conduct a case study in which seven publicly available word embedding models are compared according to four fairness metrics. These metrics are described in detail in the next section. We first briefly describe the embedding models and queries.

Embedding Models. The following are the pre-trained embedding models that we consider: 1) conceptnet, 2) fasttext-wikipedia, 3) glove-twitter, 4) glove-wikipedia, 5) lexvec-commoncrawl, 6) word2vec-googlenews, and 7) word2vec-gender-hard-debiased (also trained on Google News) [Bolukbasi *et al.*, 2016].

Queries and query sets. We consider a total of 25 queries satisfying the $(2, 2)$ template, all of them built upon previous work. From them we construct three query sets $\mathcal{Q}_{\text{gender}}$ with 7 queries, $\mathcal{Q}_{\text{ethnicity}}$ with 9 queries, and $\mathcal{Q}_{\text{religion}}$ with 9 queries. For the sake of space we cannot describe the content of each query, but we next list the previous work from which we form all of them. We take the attribute word sets *pleasant*, *unpleasant*, *math* and *arts* from [Caliskan *et al.*, 2017]; the target sets *ethnicity-surnames*, *male* and *female*, and attribute words related to *intelligence*, *appearance*, *sensitive* and *occupations* were taken from [Garg *et al.*, 2018]; the attribute word set *religion* was taken from [Manzini *et al.*, 2019]; *positive* and *negative* sentiment attribute words were taken from the Bing Liu lexicon [Hu and Liu, 2004].

4.1 Specific Fairness Metrics

Next, we describe the four fairness metrics we consider in this case study from the point of view of WEFE.

Word Embedding Association Test (WEAT)

Proposed by Caliskan et al. [2017] the WEAT metric receives two sets T_1 and T_2 of target words, and two sets A_1 and A_2 of attribute words. Thus, in our terminology, it always expects a query of the form $Q = (\{T_1, T_2\}, \{A_1, A_2\})$ and then its associated template is $s_{\text{WEAT}} = (2, 2)$. Its objective is to quantify the strength of association of both pair of sets through a permutation test. Given a word embedding w, WEAT defines first the measure $d(w, A_1, A_2)$ as

$$\big(\operatorname{mean}_{x \in A_1} \cos(w, x)\big) - \big(\operatorname{mean}_{x \in A_2} \cos(w, x)\big)$$

where $\cos(w, x)$ is the cosine similarity of the word embedding vectors. Then for a query $Q = (\{T_1, T_2\}, \{A_1, A_2\})$ the WEAT metric is defined as

$$F_{\text{WEAT}}(\mathbf{M}, Q) = \sum_{w \in T_1} d(w, A_1, A_2) - \sum_{w \in T_2} d(w, A_1, A_2)$$

The idea is that the more positive the value given by F_{WEAT}, the more target T_1 will be related to attribute A_1 and target T_2 to attribute A_2. On the other hand, the more negative the value, the more target T_1 will be related to attribute A_2 and target T_2 to attribute A_1. The ideal score is 0. Thus, the order induced by WEAT is such that $x \leq_{F_{\text{WEAT}}} y$ iff $|x| \leq |y|$.

WEAT Effect Size (WEAT-ES)

This metric represents a normalized measure that quantifies how far apart the two distributions of association between targets and attributes are. It also receives queries with template $s_{\text{WEAT-ES}} = (2, 2)$. Then $F_{\text{WEAT-ES}}(\mathbf{M}, Q)$ is computed as:

$$\frac{\operatorname{mean}_{w \in T_1} d(w, A_1, A_2) - \operatorname{mean}_{w \in T_2} d(w, A_1, A_2)}{\operatorname{std}_{w \in T_1 \cup T_2} d(w, A_1, A_2)}$$

Since the ideal is also 0, we define $\leq_{F_{\text{WEAT-ES}}}$ just as $\leq_{F_{\text{WEAT}}}$.

Relative Norm Distance (RND)

Proposed by Garg et al. [2018], it receives queries with template $s_{\text{RND}} = (2, 1)$. Given a query $Q = (\{T_1, T_2\}, \{A\})$ the metric $F_{\text{RND}}(Q)$ is computed as

$$\sum_{x \in A} \Big(\| \operatorname{avg}(T_1) - x \|_2 \; - \; \| \operatorname{avg}(T_2) - x \|_2 \Big)$$

where $\| \cdot \|_2$ represents the Euclidean norm, and $\operatorname{avg}(T)$ is the vector resulting from averaging all the vectors in T. That is, RND averages the embeddings of each target set, then for each of the attribute words, calculates the norm of the difference between the average target and the attribute word, and then subtracts the norms. The more positive (negative) the relative distance from the norm, the more associated are the sets of attributes towards group two (one).

The optimal value here is 0, and thus as for WEAT we let $x \leq_{F_{\text{RND}}} y$ iff $|x| \leq |y|$.

Queries set by criteria Model name	Gender WEAT	WEAT-ES	RND	RNSB	Ethnicity WEAT	WEAT-ES	RND	RNSB	
conceptnet-numberbatch 19.08-en dim=300	2 (0.37)	2 (0.20)	2 (0.01)	**1 (0.02)**	**1 (0.46)**	**1 (0.14)**	2 (0.03)	**1 (0.03)**	
fasttext-wiki-news dim=300	5 (0.71)	4 (0.47)	3 (0.02)	2 (0.02)	3 (0.49)	5 (0.20)	3 (0.06)	2 (0.04)	
glove-twitter dim=200	3 (0.50)	3 (0.41)	5 (0.13)	5 (0.23)	6 (0.75)	6 (0.42)	5 (0.16)	5 (0.07)	
glove-wiki-gigaword dim=300	4 (0.66)	7 (0.84)	6 (0.18)	6 (0.29)	7 (1.00)	7 (0.58)	6 (0.26)	4 (0.07)	
lexvec-commoncrawl W+C dim=300	6 (0.79)	5 (0.71)	7 (0.33)	7 (0.32)	2 (0.47)	2 (0.15)	7 (0.73)	7 (0.17)	
word2vec-gender-hard-debiased dim=300	**1 (0.16)**	**1 (0.08)**	**1 (0.00)**	3 (0.03)	4 (0.52)	3 (0.19)	**1 (0.03)**	3 (0.05)	
word2vec-google-news dim=300	7 (0.90)	6 (0.82)	4 (0.08)	4 (0.14)	5 (0.53)	4 (0.19)	4 (0.15)	6 (0.12)	
Queries set by criteria Model name	Religion WEAT	WEAT-ES	RND	RNSB	Overall WEAT	WEAT-ES	RND	RNSB	WEB
conceptnet-numberbatch 19.08-en dim=300	4 (0.96)	**1 (0.11)**	2 (0.05)	2 (0.07)	2 (0.61)	**1 (0.15)**	2 (0.03)	2 (0.04)	**1**
fasttext-wiki-news dim=300	**1 (0.84)**	3 (0.16)	3 (0.13)	**1 (0.04)**	3 (0.68)	3 (0.26)	3 (0.07)	**1 (0.03)**	2
glove-twitter dim=200	2 (0.84)	2 (0.15)	6 (0.44)	5 (0.18)	4 (0.71)	4 (0.32)	5 (0.25)	5 (0.15)	7
glove-wiki-gigaword dim=300	7 (1.18)	7 (0.27)	5 (0.33)	3 (0.10)	7 (0.97)	7 (0.54)	6 (0.26)	4 (0.14)	6
lexvec-commoncrawl W+C dim=300	3 (0.94)	6 (0.21)	7 (0.89)	6 (0.22)	5 (0.73)	5 (0.33)	7 (0.65)	7 (0.23)	4
word2vec-gender-hard-debiased dim=300	6 (1.05)	5 (0.19)	**1 (0.03)**	4 (0.17)	**1 (0.61)**	2 (0.16)	**1 (0.02)**	3 (0.09)	5
word2vec-google-news dim=300	5 (1.04)	4 (0.19)	4 (0.20)	7 (0.31)	6 (0.82)	6 (0.37)	4 (0.15)	6 (0.19)	3

Table 1: Final matrices obtained after applying our framework for several metrics, embedding models, and three different query sets. Rankings plus absolute values for each metric are included.

Relative Negative Sentiment Bias (RNSB)

We consider a straightforward generalization of this metric [Sweeney and Najafian, 2019][1]. RNSB receives as input queries with two attribute sets A_1 and A_2 and two or more target sets, and thus has a template of the form $s = (N, 2)$ with $N \geq 2$. Given a query $Q = (\{T_1, T_2, \ldots, T_n\}, \{A_1, A_2\})$ and an embedding model $\mathbf{M}$, in order to compute the metric $F_{\text{RNSB}}(\mathbf{M}, Q)$ one first constructs a binary classifier $C_{(A_1,A_2)}(\cdot)$ using set A_1 as training examples for the negative class, and A_2 as training examples for the positive class. After the training process, this classifier gives for every word w a probability $C_{(A_1,A_2)}(w)$ that can be interpreted as the degree of association of w with respect to A_2 (value $1 - C_{(A_1,A_2)}(w)$ is the degree of association with A_1). Now, we construct a probability distribution $P(\cdot)$ over all the words w in $T_1 \cup \cdots \cup T_n$, by computing $C_{(A_1,A_2)}(w)$ and normalizing it to ensure that $\sum_w P(w) = 1$. The main idea behind RNSB is that the more that $P(\cdot)$ resembles a uniform distribution, the less biased the word embedding model is. Thus, one can compute $F_{\text{RNSB}}(\mathbf{M}, Q)$ as the distance between $P(\cdot)$ and the uniform distribution $U(\cdot)$. RNSB uses the KL-divergence to compute that distance. As before, the optimal value is 0. Since it cannot deliver negative values, we let $x \leq_{F_{\text{RNSB}}} y$ iff $x \leq y$.

4.2 Results

Using the WEFE ranking process described in Section 3.2 together with the three query sets, the four fairness metrics and the seven embedding models described above, we obtained three scores matrices that are shown in Table 1, one for each query set $\mathcal{Q}_{\text{gender}}$, $\mathcal{Q}_{\text{ethnicity}}$ and $\mathcal{Q}_{\text{religion}}$. Additionally, we created a fourth matrix (Overall in Table 1) by applying our framework to query set $\mathcal{Q} = \mathcal{Q}_{\text{gender}} \cup \mathcal{Q}_{\text{ethnicity}} \cup \mathcal{Q}_{\text{religion}}$ containing all our queries using an aggregation function that performs a weighted average of the different query sets (the weights correspond to the cardinality of each query set).

We add an additional column to this last matrix obtained by running the Word Embedding Benchmark (WEB) on our embedding models. WEB rankings are obtained by adding up the rankings produced by all the metrics implemented by the benchmark. Notice that WEB metrics are ranked in descending order unlike the metrics evaluated in WEFE.

In addition, we generate correlation matrices between the rankings using Spearman's rank correlation coefficient (Figure 1). These matrices allow us to state whether or not the rankings are aligned with each other according to the criteria evaluated. Such agreement would enable us to establish whether the rankings obtained are reliable.

If we focus on the gender results in Table 1, we can observe a clear tendency for word2vec-gender-hard-debiased and conceptnet to be at the top of the ranking. We can also observe that the debiased version of word2vec outperforms the non-debiased version across all metrics. Another noteworthy result derived from Figure 1, are the high correlations observed between all metrics for gender. This implies a consistency between metrics with respect to gender bias.

For the case of ethnicity, although conceptnet consistently outperforms other models (with rankings, 1, 1, 2, and 1) the differences in terms of absolute scores with the closest competitor are very small. Moreover, for ethnicity the correlations between each ranking are significantly lower than for gender (Figure 1). Something similar happens for the case of religion in which not only the correlations between the rankings are lower but in which it is no clear at all what method is the best. Another observation worth reporting is that for the case of both ethnicity and religion the differences between word2vec and word2vec-gender-hard-debiased are considerably less noticeable in terms of absolute values than in the case of gender.

Other results that are somewhat consistent across the

[1] In the original RNSB proposal, attribute sets of words are always associated with positive and negative lexicons, and in the experiments target sets are only made by singletons.

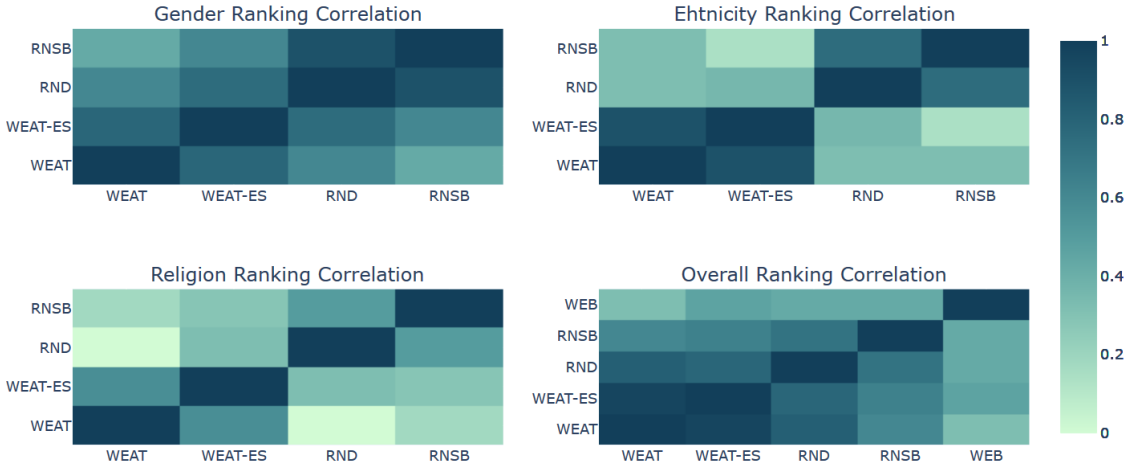

Figure 1: Spearman correlation matrix of rankings by different measures.

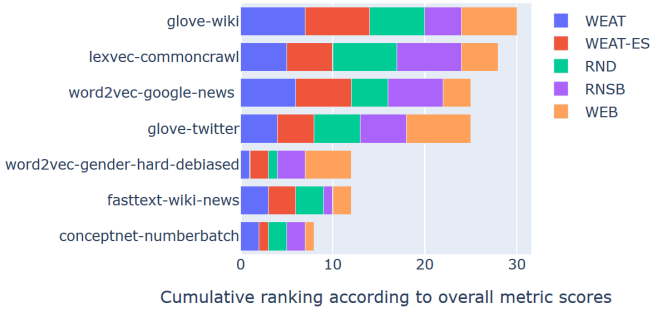

Figure 2: Accumulated rankings for the overall results plus WEB.

three tables and the four metrics, are that models, glove-twitter, glove-wikipedia, lexvec and word2vec-google-news, are rarely found in top ranking positions.

Unlike the results for ethnicity and religion, the overall matrix shows a more consistent behavior (Table 1). Conceptnet and fastext take the first two places in all metrics. In addition, the low scores obtained by the above-mentioned models are maintained. Similarly to the gender matrix, the four fairness metrics exhibit clear positive ranking correlations (Figure 1) in the overall matrix.

In relation to the rankings obtained from the Word Embedding Benchmark (WEB), although conceptnet and fasttext maintain their leading positions (Table 1), there is no clear correlation between WEB and WEFE rankings (Figure 1). For example, lexvec, which is poorly ranked among WEFE scores, is in the middle of the WEB ranking. In the case of word2vec and its gender debiased variation, their positions in WEB and WEFE rankings are swapped. This suggests that the gender debiasing method proposed in [Bolukbasi *et al.*, 2016] can affect the performance of the embedding model in word similarity and analogy tests.

These misalignenments can be further analyzed in Figure 2. The figure displays the rankings obtained by the overall WEFE rankings and WEB results using cumulative graph bars (i.e., the larger the size of a bar the lower its position in the corresponding ranking). The figure allows for easy detection of models with good WEFE rankings and poor WEB rankings, such as word2vec gender-hard-debiased version and glove-twitter, as well as the opposite effect: high bias and good WEB performance, such as word2vec and lexvec.

5 Conclusions and Future Work

Our framework allows to cleanly compare bias measurements by abstracting away several component such as target and attribute sets, queries, templates, fairness scores, and order relations among those scores. When applying our framework to specific pre-trained embeddings and fairness metrics we were able to spot some differences among these metrics. In particular, we show that the most widely used fairness metrics are not always correlated beyond the gender dimension. This gives evidence of the difficulty of measuring bias for aspects such as religion or ethnicity, and thus more research is needed in that direction. In addition, we were able to check that there is no direct correlation between the performance of the embeddings and the bias they contain.

One important subjective aspect in our framework is the design of the queries (target and attributes) used to test bias. We followed closely the previous work when selecting the words composing every query, but this election may definitely impact the rankings obtained. One important line for future research is to define a standard set of queries that can be shared with all the research community.

We have released WEFE as an open source toolkit[2] along with tutorials to reproduce this and other previous studies.

Acknowledgments

This work was funded by the Millennium Institute for Foundational Research on Data (IMFD). Pérez is also supported by Fondecyt grant 1200967.

[2]https://wefe.readthedocs.io/en/latest/

References

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

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

[Ethayarajh *et al.*, 2019] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics.

[Garg *et al.*, 2018] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[Goldberg, 2017] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[Gonen and Goldberg, 2019] Hila Gonen and Yoav Goldberg. 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)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[Jastrzkebski *et al.*, 2017] Stanislaw Jastrzkebski, Damian Lesniak, and Wojciech Marian Czarnecki. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *CoRR*, abs/1702.02170, 2017.

[Manzini *et al.*, 2019] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

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

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Salle *et al.*, 2016] Alexandre Salle, Aline Villavicencio, and Marco Idiart. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–424, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017.

[Sweeney and Najafian, 2019] Chris Sweeney and Maryam Najafian. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, 2019.

[Swinger *et al.*, 2019] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311, 2019.

[Zhou *et al.*, 2019] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5275–5283, Hong Kong, China, November 2019. Association for Computational Linguistics.