Unpacking Bias: An Empirical Study of Bias Measurement Metrics, Mitigation Algorithms, and their Interactions

María José Zambrano, Felipe Bravo-Marquez
Department of Computer Science, University of Chile
National Center for Artificial Intelligence (CENIA)
Millennium Institute for Foundational Research on Data (IMFD)
mzambran@dcc.uchile.cl, fbravo@dcc.uchile.cl

Abstract

Word embeddings (WE) have been shown to capture biases from the text they are trained on, which has led to the development of several bias measurement metrics and bias mitigation algorithms (i.e., methods that transform the embedding space to reduce bias). This study identifies three confounding factors that hinder the comparison of bias mitigation algorithms with bias measurement metrics: (1) reliance on different word sets when applying bias mitigation algorithms, (2) leakage between training words employed by mitigation methods and evaluation words used by metrics, and (3) inconsistencies in normalization transformations between mitigation algorithms. We propose a very simple comparison methodology that carefully controls for word sets and vector normalization to address these factors. We conduct a component isolation experiment to assess how each component of our methodology impacts bias measurement. After comparing the bias mitigation algorithms using our comparison methodology, we observe increased consistency between different debiasing algorithms when evaluated using our approach.

Keywords: Fairness, Embeddings, Bias

1. Introduction

Word embedding (WE) models, which are mappings from discrete words to dense continuous vectors, have been shown to reflect gender, racial, and religious stereotypes from the corpus on which they are trained (Bolukbasi et al., 2016), (Manzini et al., 2019). To address this issue, two types of solutions have emerged: 1) metrics for quantifying bias levels, such as WEAT (Bolukbasi et al., 2016) and ECT (Dev and Phillips, 2019)), and 2) mitigation algorithms aimed at reducing bias within the model, such as Hard Debias (Bolukbasi et al., 2016) and Half-Sibling Regression (Yang and Feng, 2020).

When it comes to benchmarking different mitigation methods, an intuitive approach is to measure the bias in pre-trained WE models before and after applying each mitigation method, using the metrics mentioned above. However, previous work has not systematically compared these methods, and there are significant discrepancies and interdependencies between methods and metrics that can affect the reliability of the results.

In this paper, we identify three primary issues that contribute to inaccurate comparisons between bias mitigation algorithms in their original implementation settings: 1) inconsistency in the selection of input words for debiasing, 2) overlap between words used for learning mitigation algorithms, and bias measurement metrics, and 3) vector normalization performed by certain algorithms. We propose a simple methodology to address these concerns and promote a more robust comparability.

Our approach enforces the use of identical word sets, introduces constraints to manage overlap between methods and metrics, and adopts consistent vector normalization transformations. Furthermore, we present a case study comparing the impact of different mitigation algorithms on various metrics using both the default approach and our proposed methodology.

We perform all experiments on gender[1] bias because most of mitigation algorithms are specifically designed to address this type of bias. We use the glove-wiki-gigaword-300 as our target word embedding model and evaluate its bias using various metrics. We assess bias before and after applying transformations from different bias mitigation algorithms, considering different settings.

We utilize the available bias mitigation methods and metrics provided in the WEFE framework (Badilla et al., 2020) for our experiments. Specifically, we focus on six bias measurement metrics: WEAT (Caliskan et al., 2017), WEAT ES (Caliskan et al., 2017), RND (Garg et al., 2018), RNSB (Sweeney and Najafian, 2019), ECT (Dev and Phillips, 2019), and RIPA (Ethayarajh et al., 2019), and four bias mitigation algorithms: Hard Debias (HD) (Bolukbasi et al., 2016), Double Hard Debias (DHD) (Wang et al., 2020), Repulsion At-

[1] We recognize that addressing gender as a binary variable is a delicate issue. Further exploration of this matter is detailed in our ethics statement.

traction Neutralization (RAN) (Kumar et al., 2020), and Half Sibling Regression (HSR) (Yang and Feng, 2020).

The rest of the paper is organized as follows. In Section 2 we offer an overview of previous work that is relevant to our research. Then, in Sections 3 and 4, we describe the relevant issues for comparing bias mitigation algorithms and propose how to address them. In Section 5, we comprehensively present our experiments, including their experimental setups, the comparative results of bias mitigation algorithms, both with and without the application of our proposed methodology. This section compares these two sets of results and discusses the variations observed when implementing our methodology. It also includes an analysis of isolated components that dissects our proposed methodology to evaluate the impact of each component. Finally, in Section 6, we present the conclusions drawn from our research.

2. Related Work

As mentioned above, word embedding models tend to capture and reflect biases present in the data they are trained on. Research in this field has focused primarily on the development of bias measurement metrics aimed at quantifying the biases contained in word embedding models, as well as bias mitigation algorithms designed to reduce these biases.

In this section, we begin by presenting the metrics used to measure bias in our study. We provide a concise description of the principles underlying these metrics and how they measure bias in word embeddings. We then elaborate on bias mitigation algorithms, explaining their methodologies for bias mitigation within these models.

2.1. Metrics

Previous research in the field has introduced various bias measurement metrics for word embedding models. These metrics share a common goal of quantifying the bias contained in these models but employ distinct methodologies to achieve this objective. In general, they measure the association between words that define a bias group and words typically associated with that group. Below, we provide brief descriptions of the six bias measurement metrics employed in our study.

The Word Embedding Association Test (WEAT), as proposed by Caliskan et al. (2017), is a metric designed to assess the level of association between two pairs of word sets. These pairs of word sets represent, on one hand, the social groups under examination, and on the other, various traits or professions that might be associated with these social groups. In addition to WEAT, the same authors proposed the WEAT Effect Size (WEAT ES) (Caliskan et al., 2017). This metric represents a normalized measure of how distinct the two distributions of associations are between a word set representing a social group and a word set containing traits that could be associated with it.

Relative Norm Distance (RND), as introduced by Garg et al. (2018), is a metric designed to capture the relative strength of association between a set of neutral words, which are not intrinsically related to any specific social group, and two sets of words that represent different social groups.

Relative Negative Sentiment Bias (RNSB), a metric proposed by Sweeney and Najafian (2019), is based on the idea that in the absence of bias, all words should exhibit similar levels of negativity. RNSB measures how negative words that define target social identity groups are.

The Relational Inner Product Association (RIPA) (Ethayarajh et al., 2019), calculates the dot product between a word that should be neutral to bias and a vector representing the difference between a pair of words that define the bias group.

Finally, Embedding Coherence Test (ECT) a metric proposed by Dev and Phillips (2019), focuses on measuring the degree of association between a set of professions and gendered word pairs.

These metrics are standardized and unified within a common interface, facilitating their interchangeability within the WEFE library (Badilla et al., 2020). In the library's metrics interface, each metric operates on a query, which consists of target and attribute word sets. These sets represent the social groups and words associated with them. By supplying a query that contains these sets along with a pre-trained word embedding model, one can measure bias within the model using these metrics.

2.2. Algorithms

To address the concern of bias in word embeddings, various bias mitigation algorithms have been developed. These algorithms aim to diminish the bias present in word embedding models through diverse approaches. In general, they focus on learning bias from words that define the social groups and adjust the embedding space to ensure that biased words are all at a similar distance from the bias space.

Bolukbasi et al. (2016) proposed Hard Debias (HD), an algorithm that identifies the direction responsible for capturing bias and then subtracts it from words that should be considered neutral.

Similarly, Wang et al. (2020) argued that word frequency in training corpora plays a significant role in bias, which limits the effectiveness of HD. To address this issue, they proposed Double Hard Debias (DHD) (Wang et al., 2020), which mitigates bias by subtracting the frequency direction that in-

fluences bias and then applying Hard Debias.

Another algorithm to consider is Half Sibling Regression (HSR) proposed by Yang and Feng (2020). HSR is founded on a confounding-noise-elimination approach. It employs causal inference techniques to identify and subtract spurious gender information from biased vectors.

Repulsion Attraction Neutralization (RAN) is a bias mitigation algorithm introduced by Kumar et al. (2020). With this algorithm, the authors claim not to only eliminate the bias present in word vectors, but also to alter the spatial distribution of its neighbours' vectors achieving a bias-free setting while maintaining minimal semantic offset.

The algorithms described above are implemented within a common interface in the WEFE library (Badilla et al., 2020), similar to the metrics, which allows them to be interchangeable. These algorithms operate on sets of words representing the bias groups, words to which the bias is to be mitigated, and words to ignore during the process.

3. Word Interaction

In this section, we examine the interaction between the word sets used in bias mitigation algorithms and bias measurement metrics, identify potential issues arising from this interaction, and propose mechanisms to mitigate them. There are five word sets involved in the process: 1) target and 2) attribute words from the metrics, and 3) gender-specific, 4) bias definition, and 5) objective words from the algorithms. We describe them below:

Target words are used to denote specific social identity groups defined by criteria such as gender, religion, or race. These criteria can include any characteristic, trait, or origin that distinguishes different groups of people from one another (Badilla et al., 2020).

Attributes include words that represent attitudes, traits, characteristics, occupations, among others. In a fair setting, these attributes should have equal associations with individuals from each social group (e.g., occupations, affective words) (Badilla et al., 2020).

Bias metrics typically operate by quantifying associations between a minimum of one attribute set and a minimum of two targets (e.g., male vs. female) and then contrasting these associations with a fair setting where attributes exhibit equal associations with each group.

Bias Definition refers to a set of word pairs derived from two contrasting identity groups utilized by mitigation algorithms to learn and address the intended bias direction. These words consistently represent male and female groups in bias definition methods (e.g., man-woman, he-she, girl-boy).

Gender Specific includes words that are associated with gender by definition but do not necessarily define the identity group (e.g., beard, womb, testosterone) (Bolukbasi et al., 2016). These words inherently contain gender-related connotations, so the bias mitigation process is not applied to them. Note that bias definition words are also included in this set.

Objective is the set of words to which the bias mitigation process is applied, which is usually the complement of the gender-specific set. These words are expected to be unrelated to the target identity groups.

Mitigation algorithms typically learn a transformation of the embedding space using bias definition words, which is then applied to the objective set, excluding gender-specific words.

The original implementations of the algorithms exhibit variability in the selection of words within sets. We argue that this variability introduces additional uncertainty into the observed bias changes that cannot be attributed solely to the algorithms themselves. To address this concern, we propose adopting a standardized set of words across algorithms, thereby controlling for this variable.

In our study, we construct the bias definition set by combining the male and female sets from (Garg et al., 2018) with the definition pairs from (Bolukbasi et al., 2016). The gender-specific set is taken from (Bolukbasi et al., 2016). For the objective word set, we consider the entire vocabulary of the model, excluding the words in the gender-specific set.

As was mentioned in Section 2.1, the metrics were already standardized by (Badilla et al., 2020), so we adopt their approach for the gender queries set, which incorporates word sets from (Caliskan et al., 2017), (Garg et al., 2018), (Hu and Liu, 2004), and (Manzini et al., 2019).

We can now proceed to analyze the second issue addressed in our study, which focuses on the overlap between the words utilized for learning mitigation algorithms and the bias measurement metrics. We present their intersections in Table 1.

Algorithms \ Metrics	Attributes (6,894)	Target (40)
Objective (398,559)	6,500	0
Bias Definition (44)	0	40
Gender Specific (1,449)	5	40

Table 1: Size of the intersections between word sets. The size of each set is given in brackets after its name.

We argue that the overlap between sets may hinder accurate bias measurement and comparison of bias mitigation algorithms. On the one hand, there are considerable similarities in the defini-

tion of certain sets used in both mitigation algorithms and metrics. It is crucial to avoid introducing inconsistencies by allowing words to be part of opposite sets. In addition, we hypothesize that words used for learning bias mitigation should be excluded from the evaluation to ensure generalization in the measurement, similar to the separation of training and test sets in standard supervised machine learning problems.

An instance illustrating the undesired intersections between the word sets is the intersection between Gender-Specific and Attributes. In this overlap, words such as "maid," "heroine," "mistress," "womanizer," and "hellion" are identified. These words are considered to carry gender bias by definition and are thus excluded from the debiasing process. However, they are still utilized as part of the bias measurement, despite not being included in the debiasing process itself.

To address this problem concerning the overlap of word sets, we propose implementing constraints on the intersection of the different word sets, as detailed in Table 3. The proposed word set constraints are as follows:

Objective/Attributes The attributes set should be entirely contained within the objective set to ensure that words expected to be unrelated to social identity groups (i.e., attribute words) are mitigated.

Objective/Target The target set should not overlap with the objective set. This is crucial because the target words inherently represent specific social identity groups, and applying mitigation techniques to them would directly impact their ability to represent those groups accurately.

Bias Definition/Attributes These sets are defined as opposites and hence, should not overlap. The bias definition set contains words that define social identity groups (e.g., male and female words), while attribute words are expected to be independent of these criteria.

Bias Definition/Target: Although both the bias definition and target sets contain words that define social identity groups, avoiding overlap between them is important. We expect that mitigation algorithms should generalize beyond the words used to learn the transformation. Assessing bias on the same words used for learning would lead to overly optimistic results. This restriction is analogous to the standard practice of separating training and test data in supervised machine learning.

Gender Specific/Target: The target set should be entirely contained within the gender-specific set to avoid bias mitigation on words that define social identity groups.

Gender Specific/Attributes: To maintain independence between gender and attributes, the attribute and gender-specific sets should not overlap. This ensures that attribute words, which are intended to be gender-neutral, can be accurately evaluated by the metric after mitigation. This constraint does not affect the generalization of the measurement as mitigation algorithms do not rely on the attribute set for learning the transformation.

To meet the above constraints, we propose to construct the word sets in the following way: First, we create a list of female/male word pairs for the target and bias definition sets without repeating words. Next, we expand the list to include additional word pairs that define the female and male groups. We search for synonyms in dictionaries to find suitable pairs, resulting in 24 word pairs presented in Table 2. The details regarding the sources of the words can be found in Appendix 8.1. Then, we manually rank these pairs based on their representation of the target groups. Then, to evenly distribute these pairs for our target and bias definition words, we assign odd ranking numbers to the bias definition words and even ranking numbers to the target words. Afterwards, for gender-specific words, we make sure that every word from the target word set is included in this set. We also delete from this set any word that is included in the attributes set. Finally, we consider the entire vocabulary of the embedding model as the objective word set after excluding any words present in the gender-specific set.

4. Vector normalization

Two of the algorithms included in our study, Hard Debias (HD) and Repulsion Attraction Neutralization (RAN), use vector normalization as a preprocessing step in their mitigation process. This involves normalizing all word vectors to the Euclidean norm before applying the respective algorithms. However, it has been found that vector length contains valuable information within word embeddings (Ethayarajh et al., 2019), as the matrix factorized by the embedding model cannot be reconstructed solely with normalized embeddings. To explore this further, we present the bias analysis of our target glove model before and after vector normalization in Table 4. The results show that vector normalization affects three metrics: RND, RNSB, and RIPA. This finding highlights that algorithms that use vector normalization and those that do not are not directly comparable.

To address the impact of vector normalization on bias measurement, we propose two approaches for a fairer comparison between algorithms: 1) normalize the model before bias mitigation for algorithms that do not perform it, or 2) reverse the normalization performed by any algorithm (e.g., HD, RAN) after mitigation by rescaling the resulting vectors to their original norm. We con-

Female	Male
she	he
woman	man
female	male
femenin	masculine
her	him
herself	himself
lady	gentleman
madam	sir
miss	mister
girl	boy
gal	guy
girlfriend	boyfriend
mother	father
mom	dad
wife	husband
grandmother	grandfather
daughter	son
sister	brother
aunt	uncle
niece	nephew
actress	actor
Mary	John
princess	prince
queen	king

Table 2: Twenty-four word pairs used as bias definition and target. The set is divided equally, with half used as bias definition and the other half as target, ensuring no overlap as proposed in our methodology.

Algorithms \ Metrics	Attributes	Target
Objective	|Attributes|	$\emptyset$
Bias Definition	$\emptyset$	$\emptyset$
Gender Specific	$\emptyset$	|Target|

Table 3: Proposed intersections between word sets of metric and mitigation algorithms.

Metrics \ Models	Glove	Glove Normalized
WEAT	0.8446	0.8446
WEAT ES	0.6556	0.6556
RND	0.1832	0.0252
RNSB	0.0859	0.0177
RIPA	0.2274	0.0344
ECT	0.8234	0.8190

Table 4: Comparison of the bias between the original model and its normalized version according to the metrics.

sider the second approach to be more appropriate as it preserves the information carried by vector length as recommended in (Ethayarajh et al., 2019). Nonetheless, we conduct experiments with both approaches for a comprehensive analysis.

5. Experiments

In this section, we provide a comprehensive overview of our experimental setup and present the results of our comparison, covering both our baseline methodology and the comparison performed using our proposed approach. We will begin by describing the experimental settings, followed by the presentation of the results, and a discussion of the findings.

5.1. Experimental Setting

For all our experiments, we utilize the glove-wiki-gigaword-300 model, which is accessible through Gensim[2], given that it is a model widely used in the field.

Our baseline setup consists of applying the four bias mitigation algorithms described in Section 2.2, namely, HD, DHD, RAN, and HSR. These algorithms are applied to the word embedding model chosen for our experiments, employing the default settings from their original implementations. Subsequently, we assess the model's bias levels both before and after mitigation, according to the metrics WEAT, WEAT ES, RND, RNSB, ECT, and RIPA.

For the Hard Debias method, we adopt the word sets proposed in the original implementation (Bolukbasi et al., 2016) for the definition of gender-specific words and bias definition. As our objective set, we consider the entire vocabulary of the model, excluding the words present in the gender-specific set.

Regarding the Repulsion Attraction Neutralization technique, the original proposal suggests a specific set of words to exclude from the debiasing process. Unfortunately, we do not have access to these sets, so we resort to using the same sets used for Hard Debias to define gender-specific words and bias definition. Similarly, we use the entire vocabulary of the model as our objective set, except for the words in the gender-specific set.

For the Double Hard Debias method, the original implementation uses the 1,000 most biased female words and the 1,000 most biased male words, identified by their similarity to the terms "she" and "he", respectively, as the objective set. Consistent with this setup, we replicate this configuration in our baseline experiment, ignoring gender-specific words and adopting the bias definitions proposed by (Bolukbasi et al., 2016).

In the original implementation of Half Sibling Regression (Yang and Feng, 2020), a set of 223 male and 223 female words is used as the bias definition and gender-specific set, available in their GitHub repository. The mitigation process is then

[2] https://github.com/RaRe-Technologies/gensim

performed on words not included in this set. We replicate this setup in our baseline experiment to ensure consistency.

We then repeat the process using our proposed methodology, which enforces the use of the same word sets and follows the constraints outlined in Section 3. Furthermore, to mitigate the concerns regarding normalization, we implement the two approaches delineated in Section 4: first, we reverse the normalization performed by the algorithms that employ it, and in a separate experiment utilizing our methodology, we normalize the word embedding model before the bias mitigation for algorithms that do not include this step in their operations.

5.2. Comparison of Algorithms

The results for all of our experiments are presented in Table 5. This includes our baseline comparison and the evaluations performed using our proposed methodology for both approaches: reversing the normalization and pre-normalization of the model prior to bias mitigation. In these tables, the arrow next to the metric name represents the desired direction of change in the metric. Each value of the change in the metric is accompanied by a ranking that indicates the performance of each algorithm according to that metric. In each subtable, the final column presents the standard deviation of metric variations for each method. Additionally, the last two rows include the average standard deviation across metrics denoted as $\overline{\sigma}$ and the corresponding p-value obtained from a two-sided, two-sample t-test comparing the average standard deviations between the different settings of the proposed methodology and the baseline. We utilize these p-values with a significance level of 0.005 to determine whether a given variation of our methodology effectively reduces variability in bias metrics across different debiasing methods with respect to the baseline.

In the original setting (first subtable of Table 5), the HD and RAN algorithms perform significantly better than DHD and HSR. However, when applying our proposed methodology (two last subtables of Table 5), we observe a reduction in the performance gap between the algorithms and a notable improvement in the performance of DHD when reducing bias in the model for both approaches for treating the normalization.

In addition, when examining the standard deviations of the results when reverting the normalization (second subtable of Table 5), we see a significant reduction (p-value of 0.04) in the variability of bias reduction among the algorithms. This suggests that by removing the variability introduced by word sets and vector normalization, we can more objectively evaluate the algorithms and more accurately understand their potential for bias mitigation.

Conversely, when we normalize the model before implementing the normalization (last subtable of Table 5), we observe a smaller and statistically insignificant reduction in the standard deviation (with a p-value of 0.08). When comparing both cases of our methodology, we note some changes in the results obtained by certain metrics that were unaffected by normalization when comparing the normalized model with the original (as seen in Table 4). This could imply that normalization might influence the operations performed by algorithms that do not typically employ it.

Our methodology is designed to compare algorithms in a controlled setting, ensuring that none is favored during the process. The results indicate that algorithms tend to reduce bias more similarly to each other when this approach is taken, as seen in Figure 1, both approaches of our methodology reduce variability in the results compared to the baseline.

5.3. Analysis of Isolated Components

In this section, we perform an isolated component analysis of our proposed methodology. Our aim is to methodically assess the influence of each step, which includes word set standardization, the management of word set overlap, and vector normalization, on the outcomes of our bias mitigation and measurement process. The results of this analysis are summarized in Table 6.

First, we assess the impact of solely standardizing the word sets utilized by the algorithms as detailed in Section 3. We apply the algorithms to the model while only considering this aspect of our methodology.

The experimental results reveal a $\overline{\sigma}$ difference of 0.07 $(0.162 - 0.092)$ between this setting and the baseline, suggesting that standardizing word sets effectively reduces variability between debiasing methods and the baseline. However, it is worth noting that this reduction in variability is not statistically significant at a significance level of 0.005 (p-value of 0.16).

One notable observation is the improvement in the bias mitigation performed by DHD observed in this setting, which is consistent with the results of the application of our methodology (as shown in Table 5). This confirms that a significant part of this improvement can be attributed to the standardization of the word sets. In the baseline, the algorithm works on a limited number of words, whereas other algorithms are applied to a larger set of words, making them not directly comparable.

The results obtained here highlight the importance of controlling word sets, as it enhances the comparability of algorithms in bias mitigation evaluations.

Baseline methodology					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.756 (1)	-0.058 (4)	-0.647 (3)	-0.677 (2)	0.320
WEAT ES (↓)	-0.519 (1)	-0.030 (4)	-0.145 (3)	-0.428 (2)	0.230
RND (↓)	-0.177 (1)	-0.010 (3)	-0.007 (4)	-0.176 (2)	0.097
RNSB (↓)	-0.094 (1)	-0.027 (3)	0.007 (4)	-0.092 (2)	0.043
RIPA (↓)	-0.221 (1)	-0.014 (4)	-0.197 (3)	-0.213 (2)	0.098
ECT (↑)	0.144 (1)	0.009 (3)	-0.55 (4)	0.132 (2)	0.185
					$\overline{\sigma}$: 0.162
Proposed methodology reversing normalization					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.376 (1)	-0.317 (3)	-0.236 (4)	-0.324 (2)	0.050
WEAT ES (↓)	-0.429 (1)	-0.283 (3)	-0.166 (4)	-0.328 (2)	0.094
RND (↓)	-0.031 (3)	-0.113 (1)	0.027 (4)	-0.038 (2)	0.049
RNSB (↓)	-0.008 (2)	-0.010 (1)	-0.0008 (3)	0.006 (4)	0.006
RIPA (↓)	-0.057 (3)	-0.002 (4)	-0.094 (1)	-0.064 (2)	0.033
ECT (↑)	0.061 (2)	0.027 (3)	-0.152 (4)	0.077 (1)	0.091
					$\overline{\sigma}$: 0.053
					p-value 0.04
Proposed methodology normalizing the model before debias					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.376 (2)	-0.386 (1)	-0.0125 (4)	-0.324 (3)	0.153
WEAT ES (↓)	-0.429 (1)	-0.426 (2)	-0.005 (4)	-0.328 (3)	0.173
RND (↓)	-0.008 (3)	-0.016 (1)	-0.0004 (4)	-0.010 (2)	0.005
RNSB (↓)	-0.003 (2)	-0.005 (1)	0.0004 (4)	-0.003 (3)	0.001
RIPA (↓)	-0.013 (2)	-0.013 (1)	-0.0009 (4)	-0.012 (3)	0.005
ECT (↑)	0.058 (2)	0.039 (3)	-0.009 (4)	0.078 (1)	0.032
					$\overline{\sigma}$: 0.061
					p-value 0.08

Table 5: Comparison of Metric Changes by Algorithms in the Original Setting and Proposed Methodology. The arrow next to the metric name indicates the desired direction of change in the metric. Each change in the metric value is accompanied by a ranking that reflects the algorithm's performance on that metric. The displayed p-value represents a two-sided, two-sample t-test comparing the average standard deviation across metrics between a setting of our methodology and the baseline.

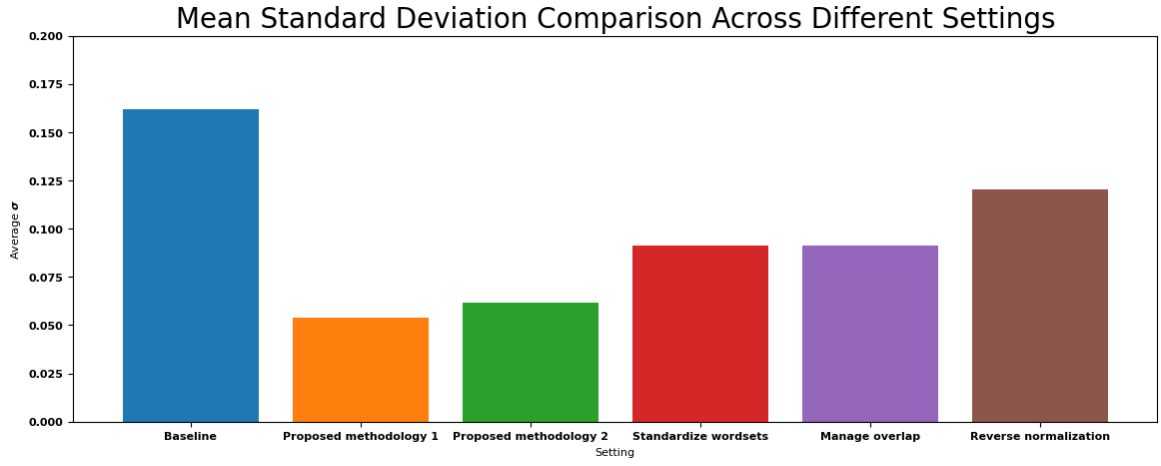

Figure 1: Comparison of the Mean Standard Deviation Across All Experiments. 'Proposed Methodology 1' represents the results when reversing normalization, and 'Proposed Methodology 2' when normalizing the model before debias.

Next, we investigate the isolated impact of managing the overlap between the word sets employed

Standardize word sets					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.6731 (2)	-0.6484 (3)	-0.385 (4)	-0.6774 (1)	0.122
WEAT ES (↓)	-0.4086 (2)	-0.3202 (3)	-0.1321 (4)	-0.4289 (1)	0.117
RND (↓)	-0.177 (1)	-0.0609 (3)	-0.0365 (4)	-0.1761 (2)	0.064
RNSB (↓)	-0.0705 (2)	-0.0371 (3)	0.0464 (4)	-0.0755 (1)	0.048
RIPA (↓)	-0.2154 (1)	-0.1457 (3)	-0.1046 (4)	-0.2133 (2)	0.046
ECT (↑)	0.1431 (1)	0.1171 (3)	-0.2201 (4)	0.1326 (2)	0.152
					$\overline{\sigma}$: 0.092
					p-value 0.16
Manage overlap between sets					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.4580 (1)	-0.0290 (4)	-0.2362 (3)	-0.3245 (2)	0.155
WEAT ES (↓)	-0.5555 (1)	-0.0229 (4)	-0.1660 (3)	-0.3287 (2)	0.198
RND (↓)	-0.3246 (1)	-0.0058 (3)	0.0278(4)	-0.3114 (2)	0.164
RNSB (↓)	-0.0537 (1)	0.0126 (4)	0.0107 (3)	-0.0528 (2)	0.032
RIPA (↓)	-0.1986 (1)	-0.0130 (4)	-0.0946 (3)	-0.1902 (2)	0.076
ECT (↑)	0.0850 (1)	0.0036 (3)	-0.1526 (4)	0.0801 (2)	0.096
					$\overline{\sigma}$: 0.120
					p-value 0.23
Reverse vector normalization					
Models / Δ Metrics	HD	DHD	HSR	RAN	σ
WEAT (↓)	-0.7561 (1)	-0.0587 (4)	-0.6479 (3)	-0.6774 (2)	0.277
WEAT ES (↓)	-0.5193 (1)	-0.0301 (4)	-0.1456 (3)	-0.4289 (2)	0.344
RND (↓)	-0.1527 (1)	-0.0100 (3)	-0.0076 (4)	-0.0956 (2)	0.061
RNSB (↓)	-0.0142 (2)	-0.0006 (3)	0.0240 (4)	-0.0156 (1)	0.015
RIPA (↓)	-0.1902 (2)	-0.0148 (4)	-0.1971 (1)	-0.1265 (3)	0.073
ECT (↑)	0.1458 (1)	0.0090 (3)	-0.2552(4)	0.1340 (2)	0.161
					$\overline{\sigma}$: 0.155
					p-value 0.92

Table 6: Comparison of Metric Changes by Algorithms when performing the isolation of components study. The arrow next to the metric name indicates the desired direction of change in the metric. Each change in the metric value is accompanied by a ranking that reflects the algorithm's performance on that metric. The displayed p-value represents a two-sided, two-sample t-test comparing the average standard deviation across metrics between an isolated component of our methodology and the baseline.

by metrics and algorithms, as described in Section 3. In this setting, our sole focus is on managing this overlap without altering the word sets in any other manner.

The results of our second analysis show that this specific component, when analyzed in isolation, manages to reduce the value of $\overline{\sigma}$ by 0.042 average standard deviation points ($0.162 - 0.12$). This reduction is less pronounced than in the previous analysis and is also not statistically significant (p value of 0.23) at the 0.005 level of significance. However, it is important to emphasize that the primary purpose of controlling word set overlap between algorithms and metrics is not only to make debiasing methods comparable, but also to ensure accurate measurement of bias reduction by removing dependencies between methods and metrics.

Finally, we investigate the influence of consistent vector normalization transformations on the results, by applying the algorithms and reversing the normalization performed by HD and RAN, while keeping all other settings consistent with our baseline.

This setting, only focusing on vector normalization, does not seem to have a significant impact on the results. While there is a reduction in variability, it is not statistically significant (p-value of 0.92). This lack of significant impact can be attributed to normalization only affecting certain metrics considered in the study. However, it is crucial to emphasize that vector normalization remains highly important when aiming for a fair comparison of the algorithms, as it ensures that differences in bias are not altered by the normalization of the vectors and are purely an effect of applying the algorithms.

This study has highlighted that while each part, when isolated, may not significantly impact the variability of the results, their collective application leads to more comprehensive and objective evaluations of these algorithms. As seen in Figure 1, all of the components reduce variability, but not as

much as the entire methodology. This not only emphasizes that all of the components together lead to a more comprehensive and objective evaluation of these algorithms, but each of them is essential when pursuing a fairer comparison among algorithms.

6. Conclusion

In this paper, we address concerns related to the comparison of bias mitigation algorithms, specifically focusing on the word sets used in the process and the pre-processing steps, including vector normalization. To address these concerns we introduced a methodology for comparing word embeddings bias mitigation algorithms by standardizing word sets, enforcing constraints between word sets, and controlling vector normalization. Our results show that when these variables are controlled, the performance of the algorithms becomes more consistent.

This phenomenon is common in NLP, where as a given problem gains popularity, a multitude of research papers emerge, each claiming the superiority of its proposed method based on its experimental results. However, it is often the case that when these methods are evaluated in a standardized, comparable, and impartial manner, the reported differences between them tend to dissipate, as emphasized by Melis et al. (2018). Our findings are consistent with this body of research. We also looked at each component of our methodology individually and found that while some components contribute to reduced variability, it is their combined effect that leads to a statistically significant reduction.

We hope that this straightforward approach will encourage the research community to evaluate bias mitigation methods in word embeddings systematically.

For future research, we plan to extend our methodology to contextualized embeddings and large language models. We aspire to expand our work to diverse languages and various forms of bias.

Ethics Statement

No human subjects were involved in this study. We are aware that the examples used in this paper contain stereotyped concepts and can potentially perpetuate harm or reinforce biases. We acknowledge the ethical implications of such content and emphasize the need to handle biased language in research responsibly. Our intention in using these examples is solely for the purpose of analyzing bias measurement metrics and mitigation algorithms in word embeddings. In our study, we used gender as a binary variable (male or female). We understand that this is a sensitive issue, but our main objective was to replicate previous research that uses gender as a binary variable. Nevertheless, we recognize the need to review and update this approach.

Acknowledgements

This work was supported by ANID Millennium Science Initiative Program Code ICN17_002 and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. María Joseé was founded by ANID Subdirección de Capital Humano/Magíster Nacional/2023 - 22230745.

7. Bibliographical References

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 430–436. International Joint Conferences on Artificial Intelligence Organization.

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*, volume 29. Curran Associates, Inc.

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

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

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):E3635–E3644.

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

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, pages 615–621.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Chris Sweeney and Maryam Najafian. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online. Association for Computational Linguistics.

Zekun Yang and Juan Feng. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9434–9441.

8. Appendix

8.1. Word Pairs

We construct the word pairs set by combining the word pairs proposed as definition pairs by (Bolukbasi et al., 2016), the female and male words proposed by (Garg et al., 2018), and additional words obtained from dictionaries. Here, we provide details on the origin of each word pair.
Definition pairs:

1. woman/man
2. girl/boy
3. she/he
4. mother/father
5. daughter/son
6. gal/guy
7. female/male
8. her/his
9. herself/himself
10. Mary/John

Female and male words:

1. she/he
2. daughter/son
3. hers / his
4. her/him
5. mother/father
6. woman/man
7. girl/boy
8. herself/himself
9. female/male
10. sister/brother
11. daughters/sons
12. mothers/fathers
13. women/men
14. girls/boys
15. females/males
16. sisters/brothers
17. aunt/uncle
18. aunts/uncles
19. niece/nephew
20. nieces/nephews

Words from dictionaries:

1. feminine/masculine
2. lady/gentleman
3. madam/sir
4. miss/mister
5. girlfriend/boyfriend
6. mom/dad
7. wife/husband

8. grandmother/grandfather
9. actress/actor
10. princess/prince
11. queen/king