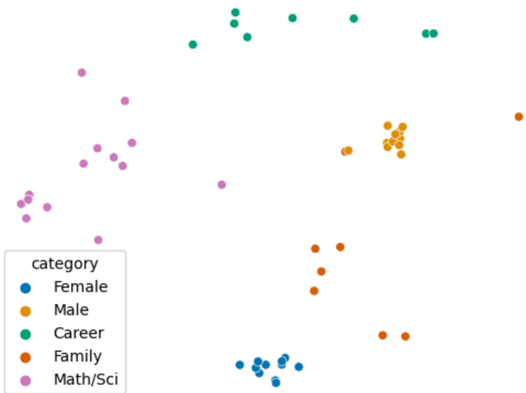
Intrinsic Bias Metrics Do Not Correlate with Application Bias

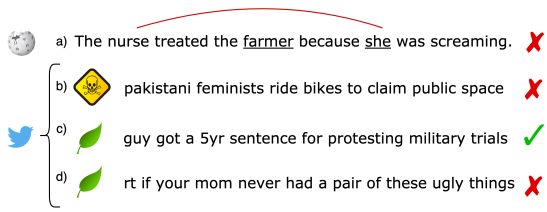
Seraphina Goldfarb-Tarrant[*†], **Rebecca Marchant**[*†]
Ricardo Muñoz Sanchez[*†], **Mugdha Pandya**[*†], **Adam Lopez**[†‡]
[†]University of Edinburgh
[‡]Rasa Technologies GmbH

Abstract

Natural Language Processing (NLP) systems learn harmful societal biases that cause them to extend and proliferate inequality widely, as they are deployed in more and more situations. To address and combat this, the NLP community has come to rely on a variety of metrics to identify and quantify bias in black-box models, which are used to monitor model behaviour and to guide efforts at debiasing. Some of these metrics are *intrinsic*, and are measured in word embedding spaces, and some are *extrinsic*, which measure the bias present downstream in the tasks that the word embeddings are plugged into. This research examines whether intrinsic metrics (which are easy to measure) correlate well to extrinsic metrics (which reflect real world bias). We measure both intrinsic and extrinsic bias across hundreds of trained models covering different tasks and experimental conditions and find that there is *no reliable correlation* between these metrics that holds in more than extremely specific settings. We advise that efforts to debias embedding spaces be always also paired with measurement of downstream model bias, and suggest that that community direct more effort into making downstream measurement simpler and easier.



(a) Intrinsic bias metrics measure the geometry of embedding spaces and reveal pervasive imbalances between types of identity words and their neighbouring words. This displays gender bias in our embeddings via the much larger distance of the female terms from the career-related terms.



(b) Sample development data from our downstream coreference and toxicity detection systems. Extrinsic bias metrics measure disparities in performance, such as rates of false negatives (a, d) and false positives (b) between different gender (or other vulnerable identity) groups. These examples display performance disparity between examples referencing females vs. males.

Figure 1: The relationship between *intrinsic* bias metrics from word embeddings and *extrinsic* bias metrics in applications is underexplored, and uncertain.

1 Introduction

Awareness of bias in Natural Language Processing (NLP) systems has rapidly increased as more and more systems are discovered to perpetuate societal unfairness at massive scales. There has been an accompanying surge in research into measuring and mitigating these biases, but this research suffers from lack of consistent metrics and methods to discover and measure bias. Instead, work on bias is 'rife with unstated assumptions' (Blodgett et al., 2020) and relies on metrics for bias that are easy to measure rather than ones that are most meaningful in revealing and tackling real world bias.

Over one third of all the bias papers surveyed in Blodgett et al. (2020), the most comprehensive recent survey of bias in NLP, deal with bias in embedding spaces. This makes the study of embedding bias a plurality, over twice as common as any other research area in the subfield of bias in NLP. And yet there is an untested assumption present in research into bias in embeddings. As is

[*] Equal contribution. Correspondence to s.tarrant@ed.ac.uk

visualised in Figure 1, bias in embedding spaces is measured with *intrinsic* bias metrics, the most common of which is the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), which measures bias based on the internal geometry of the embedding space. Those embeddings are then incorporated into downsteam applications, where bias is measured via *extrinsic* metrics that measure differences in performance for different types of test data. Research and engineering work on adjusting bias in embedding spaces relies on the unstated hypothesis that debiasing an embedding space will fix or reduce bias in an NLP application in which it is later used. There is, however, *no* research that confirms this hypothesis: whether it holds, whether it is reliable and predictable, and in which cases it may or may not be.

This leaves NLP bias research in a precarious position. Research into the utility and performance of word embeddings has extensively studied the relationship of intrinsic metrics, such as analogies and semantic similarity (Hill et al., 2015), to extrinsic metrics, such as accuracy and F1, across multiple languages and models (Glavas et al., 2019; Ruder et al., 2019). Bias research lacks the same type of systematic study, and thus as a field we are exposed to three large risks: 1) making misleading claims about the fairness of our systems, 2) concentrating our efforts at debiasing in suboptimal areas, and most importantly, 3) feeling a false sense of security as if we are making more progress on the issue than we in fact are. Our bias research can be rigorous and innovative, but without a solid understanding of the meaning of the metrics upon which we evaluate it, we cannot come to dependable or reliable conclusions.

This paper presents the first comparison of intrinsic and extrinsic metrics of bias, to analyse whether there is a correlation between them, and in which cases it holds. It addresses the following research question:

Is there a reliable relationship between the most commonly used intrinsic bias metric for embeddings (WEAT) and bias as measured in downstream applications?

To answer this, we analyse the relationship between intrinsic and extrinsic bias metrics for English (en) and Spanish (es), for two common embedding algorithms (word2vec and fastText) and for two downstream tasks (coreference resolution and hatespeech detection).

As additional contributions to these findings, we release new WEAT metrics for Spanish, and a new gender-annotated test set for hatespeech detection for English, both of which we created in the course of this research.

We find that while there is a moderately high correlation between these metrics in some select conditions (certain embeddings and bias modification methods and WEAT tests), in most conditions there is no correlation and in some it is negative. There is no unifying clear reason for the changes in these correlations across experiments. This suggests that the ethical scientist or engineer cannot rely on intrinsic measurements when investigating and mitigating bias, but must focus on the harms of specific applications and test bias effects more directly.

2 Background

As this research examines the relationship between intrinsic bias metrics of embeddings and extrinsic bias metrics of applications, we here give some background on both of these.

Intrinsic bias metrics measure similarity between words in embedding space. All variants on intrinsic bias metrics depend on hand engineered wordlists that represent concepts (such as male, female, old, young, pleasant, unpleasant) and on relative similarity between them. The most commonly used metric is WEAT (Caliskan et al., 2017), which we use here.

It measures the difference in the mean cosine similarity between two sets of target words X, Y and of attribute words A ,B where the difference measures the imbalance in associations between the wordlists and the concepts they represent. In the case of gender bias, these would be target sets such as *programmer, engineer, ...* vs. *dancer, sculptor, ...* and attribute sets of male terms (brother, father) vs. female terms (sister, mother). There is a *test statistic*, which is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (1)$$

where

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b}) \quad (2)$$

in practice, this is normalised by the standard deviation to get the *effect size* which is what we use in our experiments.[1]

WEAT was initially developed to show that the Implicit Association Test (IAT) from the field of psychology (Greenwald et al., 1998) can be replicated via word embeddings measurements. There are thus 10 original tests are chosen to replicate the tests presented to human subjects in IAT. These measure different kinds of biased associations, such as African-American names vs. White names and pleasant vs. unpleasant terms, and female terms vs. male terms with career vs. family words. WEAT was later repurposed as a *predictor* of bias in embedding spaces, via a somewhat muddy logical journey. WEAT has since been translated into 6 other languages (XWEAT) (Lauscher and Glavas, 2019), and extended to operate on full sentences (May et al., 2019) and on contextual language models (Kurita et al., 2019). When WEAT is used as a metric, papers report the effect size of the subset of tests relevant to the task at hand, each separately.

Extrinsic bias metrics measure bias in applications, via some variant of performance disparity between groups. For instance, a speech recognition system is unfair if it has higher error rates for African-American dialects (Tatman, 2017), meaning that systems perform less well for those speakers. A hiring classification system is unfair if it has more false negatives for women than for men[2], meaning that more qualified women are accidentally rejected than are qualified men. There are two commonly used metrics to quantify this possible performance disparity: Predictive Parity (Hutchinson and Mitchell, 2019), which measures the difference in precision for a privileged and non-privileged group, and Equality of Odds (Hardt et al., 2016), which measures the difference in recall between those groups (see Appendix A for formal definitions).

As the above examples show, the metric that best identifies bias in a system varies based on the task. For different applications, false negatives may be more harmful, for others false positives may be. For instance, for our first task of coreference for stereotypical vs. anti-stereotypical referents, a false negative is more problematic to fairness than a false positive. An error where a model fails to identify an anti-stereotypical coreference chain (women as CEOs) is more harmful to the underprivileged class than a false positive. For our second task, hate speech detection, it is unclear which is more harmful. An increase in false positives for one group can systematically censor certain content, as has been found in the greater false positives of hate speech detection for African-American Vernacular English (AAVE) (Sap et al., 2019; Davidson et al., 2019), whereas systematic false negatives can allow abuse of minority populations that are targets of hate speech but were not well represented in original training data collection, or are for other reasons harder to detect for a given model.

3 Methodology

We examine the relationship of WEAT metrics to two external bias metrics across two common embedding algorithms, two tasks, and two languages, using both pre-processing and post-processing methods to change amounts of bias in the embeddings. A full table of experiment conditions can be found in Table 1.

In each experiment, we train an embedding, measure the bias according to WEAT, and measure the bias in the downstream task that uses that embedding. We then modify the bias (both $\uparrow$ and $\downarrow$), and again measure WEAT on the modified embedding and also the downstream bias in the target task. When we have sufficient data points, we correlate the two distributions (via Pearson correlation and analysis with scatterplots).

If WEAT can be used as a reliable proxy for application bias, we expect there to be a moderately positive correlation between WEAT and performance disparity, and we expect this correlation to be consistent regardless of embedding algorithm or of method of modifying bias. If there is *not* a consistent relationship in all experimental conditions, or if the correlation sometimes trends negatively, then the field will have to conclude that there is a real but limited scope that is appropriate to use these instrinsic metrics and that we must apply them only in the cases where we have a good understanding of their predictive power and find it reliable. If there is no relationship across any individual experiments, we must conclude that the NLP community has developed a false sense of security from these metrics, and that our efforts at debiasing require the development of other intrinsic metrics that are more reliable, and/or that bias efforts need to incorporate measurement on downstream applications.

[1] See Appendix B for full details on WEAT normalisation in practice; we follow standard usage.

[2] Article at: `https://tinyurl.com/y6c6clzu`

To investigate this relationship we vary all of the below.

Embedding Algorithms. We use both fastText (Bojanowski et al., 2017) and Skip-gram word2vec (Mikolov et al., 2013) embeddings, as two of the most common embedding algorithms. Word representations in fastText are composed of both a word and the ngrams, or subwords, that it contains, which may cause bias to be acquired and encoded differently in fastText vs. in word2vec (as was found in Lauscher and Glavas (2019), discussed in more detail in Section 4).

We choose to focus on these, rather than on larger Transformers (Vaswani et al., 2017), because they are widely available in many toolkits and used in many real-world applications. It will be easiest to find clear relationships between metrics in simpler models, and introducing contextualisation adds significant complexity. Bias is not improved in large contextual language models (Zhao et al., 2019, 2020; Gehman et al., 2020; Sheng et al., 2019), so this work remains important. If clear trends do not emerge in this first controlled study, it is likely to be even more difficult to find them with more architectural layers and configurable hyperparameters.

Downstream tasks. We use three tasks that appear often in bias literature: Coreference resolution for English, hate speech detection for English, and hate speech detection for Spanish. To make the scenarios as reaistic as possible, we use a common, easy to implement and high performing architecture for each task: the E2E coreference system of Lee et al. (2017) and the the CNN of Kim (2014), which has been used in high-scoring systems in recent hate speech detection shared tasks (Basile et al., 2019). The pretrained embeddings are plugged into the neural network, frozen, and then the NN is trained as usual; the standard setup to leverage pretraining.

Languages. As above, we experiment on both English and Spanish. It is important to take a language with pervasive gender-marking (Spanish) into account, as previous work has shown that grammatical gender-marking has a strong effect on gender bias (McCurdy and Serbetci, 2017; Gonen et al., 2019; Zhou et al., 2019). We use Spanish for only hate speech however, because gender marking makes a challenge-set style coreference evaluation trivial to resolve and not a candidate for detection of gender bias.[3]

Datasets. To train embeddings, we use domain-matched data for each downstream task. For coreference we train on Wikipedia data, and for hate-speech detection we train on English tweets and Spanish tweets, respectively. Details of our datasets and preprocessing are in Appendix C. Our English Coreference system is trained on OntoNotes (Weischedel et al., 2017) and evaluated on Winobias (Zhao et al., 2018), a Winograd-schema style challenge set designed to measure gender bias in coreference resolution. English hate speech detection uses the abusive tweets dataset of Founta et al. (2018), and is evaluated on the test set of ten thousand tweets, which we have hand labelled as targeted *male*, *female*, and *neutral* (we release this labelled test set for future work). Spanish hate speech detection uses the data from the hate speech detection on twitter shared task of Basile et al. (2019), which contains labels for comments directed at women and directed at migrants. Further details of these datasets are in Appendix D.

Debiasing and Overbiasing Previous work on bias in embeddings studies methods to *reduce* embedding bias. However, since this work measures the relationship between intrinsic and extrinsic metrics as bias changes, we must generate many datapoints for each experiments, and take the novel approach of both decreasing *and* increasing bias. We measure the baseline bias level, via WEAT, for each embedding trained normally on the original corpus. We then adjust the bias up or down, remeasure WEAT, and measure the change in the downstream task.

We choose two methods from previous work that are capable of both debiasing and overbiasing: one preprocessing method that operates on the training data before training, and one post-processing method that operates on the embedding space once it has been trained. This is important as both kinds of methods may be used in practice; a sizeable company with their own proprietary data will train embeddings from scratch, whereas a smaller one will rely on publicly available pretrained ones. We also expect that they may interact differently with different embedding algorithms. With dataset balancing, when certain terms are subsampled or oversampled,

[3] This fact is the premise behind the work of Stanovsky et al. (2019) who use the increased explicit marking in translation to reveal bias.

Task	Data	Lang	Embeddings	Bias-Modification	Intrinsic Metric	Extrinsic Metric	Bias Type
Coreference	Ontonotes/WinoBias	en	fastText, word2vec	dataset balancing, Attract-Repel	WEAT 6,7,8	Precision gap (predictive parity), Recall gap (equality of odds)	Gender
Hate speech	Twitter	en					Gender
Hate speech	Twitter	es			XWEAT 7+8 (new), XWEAT Migrants (new)		Gender, Migrants

Table 1: A table of each of the variables included in the study for each task, from training data to kind of bias tested.

so are their surrounding words, since dataset balancing is done on full sentences. It is therefore comprehensive, but has some unknown elements. Targeted postprocessing, in contrast, is more predictable as it can directly influence certain terms (rather than sentence chunks) but may be a weaker method of modifying bias precisely because it is so targeted and may have less of a broad effect.

As a preprocessing method, we use dataset balancing (Dixon et al., 2018), which consists of sub-sampling the training data to be more equal with respect to some attributes.[4] For instance, if we are adjusting gender bias, we identify pro-stereotypical sentences[5] such as 'She was a talented housekeeper' vs. anti-stereotypical sentences, such as '*He* was a talented housekeeper' or 'She was a talented analyst'. We sub-sample and reduce the frequency of the pro-stereotypical collocations to debias, and sub-sample the anti-stereotypical conditions to overbias.

As a post-processing method applied to already trained embeddings, we use the Attract-Repel (Mrksic et al., 2017) algorithm. This algorithm was developed to use dictionary wordlists (synonyms, antonyms) to refine semantic spaces and get higher quality distinctions that might not be available from collocations alone. It aims to move similar words (synonyms) close to each other and dissimilar words (antonyms) farther from each other, while keeping a regularisation term to preserve original semantics as much as possible. We have used a variant of this algorithm for bias modification. Lauscher et al. (2020) have the most similar approach to this (also inspired by Attract-Repel), though with constraints implemented somewhat differently. For post-processing, we use the same pro- and anti-stereotypical wordlists as we use in dataset-balancing, and use the algorithm to, in the debiasing case, increase distance between pro-stereotypical combinations (she, housekeeper) and decrease distance between anti-stereotypical combinations(she, analyst or he, housekeeper). In the overbiasing case we do the reverse. Wordlists used for bias-modification, as well as hyperparameters for the bias modification algorithm, are in Appendix E.

WEAT & Bias modification wordlists. Both WEAT and bias modification methods depend on seed wordlists. WEAT uses wordlists to measure relationships between words in the space, and bias modification depends on identifying words to sub or supersample (for databalancing), or to adjust (for Attract-Repel). These are closely related to each other, and aligned by type of bias, such that we measure WEAT tests for gender bias with embeddings modified via gender bias wordlists (themselves derived from WEAT lists, as detailed below) and WEAT tests for migrant bias with embeddings modified for migrant bias.

WEAT wordlists are standardised, and for English we use WEAT terms as they are with one modification. WEAT 6 (career/family vs. male/female) uses proper names as gender terms, whereas the other two tests use more standard gender terms (she, her, he, him, mother, father). This is an artifact of replicating IAT, which introduces a confound in their comparability – if the WEAT tests have different patterns of correlation, we don't know whether this is because of the difference in the way gender bias patterns for career/family vs. for arts/science or whether it patterns differently because of proper names vs. gender terms. This is exacerbated in our case where proper names are treated even more differently than usual both in twitter (where @mentions stand in for proper names) and in the Winobias metric that we use (where professions are used instead of proper names precisely because names contain gender information and the challenge set intends to be ambiguous). We use the gender terms for WEAT 7/8 as the terms for 6, but otherwise leave terms as is.

For Spanish XWEAT, we make major modifica-

[4]We constrain the sub-sampling not materially change the dataset size, and limit it to removing $< 5\%$ of the original.

[5]Stereotypes as defined by Zhao et al. (2018) and by Caliskan et al. (2017), who use the U.S. Bureau of Labor Statistics and the Implicit Association Test, respectively.

tions and use entirely new terms. Original XWEAT was translated from English very literally, which causes two problems. First, many of the terms do not make sense in a Spanish speaking community – names included in the original, like *Amy*, are names in Spanish and thus were untranslated, but are uncommon and have upper class connotations not intended in the original test. Another example is how *firearms* is translated as *arma de fuego*, which while technically a correct literal translation, is not in use to describe weapons. The second issue is that nouns on the wordlists for both abstract math and science concepts as well as abstract art concepts are almost entirely grammatically female. For instance, *ciencia* (science), *geometría* (geometry), as well as *escultura* (sculpture) and *novela* (novel) are all grammatically female. It is well established that for languages with grammatical gender, words that share a grammatical gender have embeddings that are closer together than words that do not (Gonen et al., 2019; McCurdy and Serbetci, 2017). So when WEAT in English was translated into XWEAT in Spanish (Glavas et al., 2019), the terms were imbalanced with regard to grammatical gender, which makes the results misleading. We balance the lists, and often replace the abstract nouns with corresponding adjectives and professions, which can take male or female form, (*escultor* and *escultora* (sculptor, male and female), *científico* and *científica* (scientific, male and female), such that we can use both versions to account for the grammatical gender effect. Finally, we don't want to look at only gender bias, but we also want to examine bias against migrants. Metrics for intrinsic bias must be targeted to the type of harm expected in the downstream application, and there is not an out-of-the-box WEAT test for this. So we create a new WEAT test for bias against migrants in Spanish. Similar to the setup of tests for racial bias in original WEAT (which was based on biases expected to be in America in the English speaking populace) we have lists of names associated with migrants vs. non-migrants, and compare them with lists of pleasant and unpleasant terms. The names are based on work of Salamanca and Pereira (2013), who have a study ranking names as lower vs. upper class, and class status is closely correlated to whether a person is a migrant. We select a subset of names in which the majority in the stiudy agree on the class. Pleasant and unpleasant terms exist in WEAT and XWEAT, but we modify them and create our own version to balance grammatical gender. All WEAT sets are included in Appendix B.

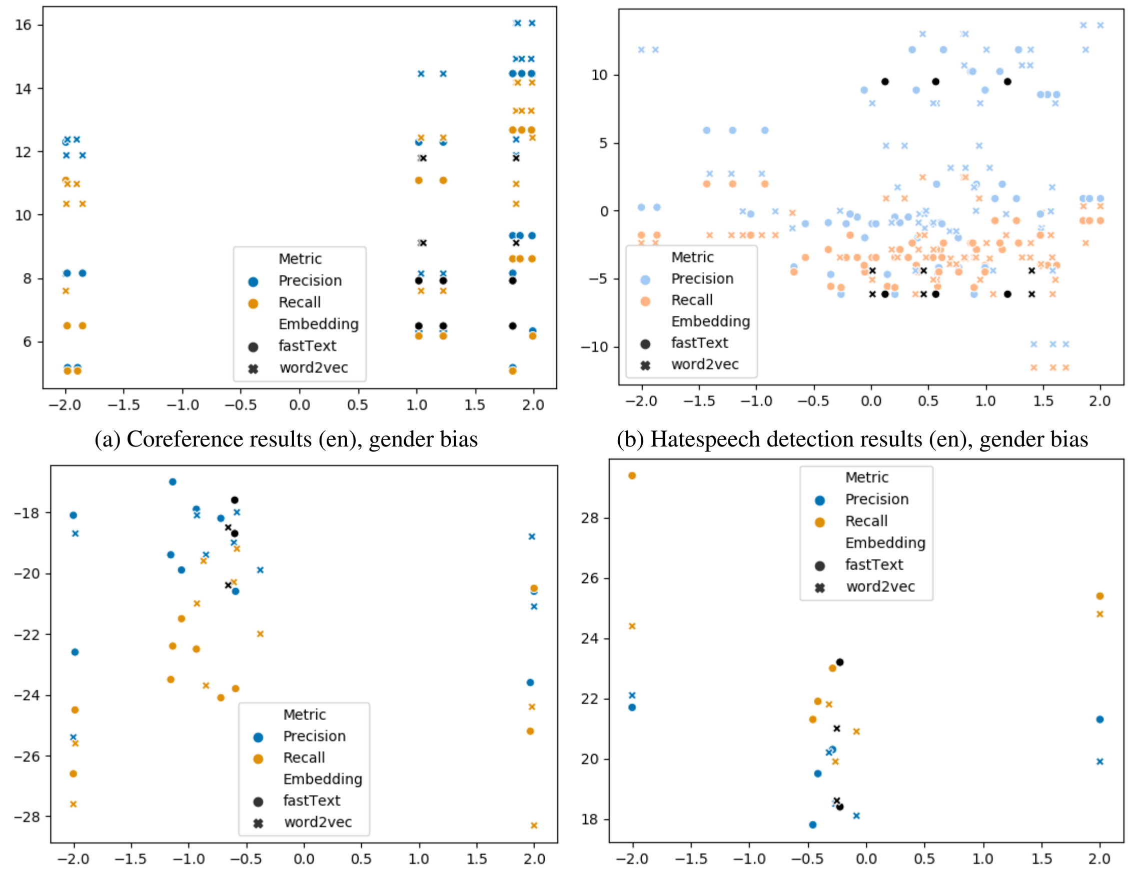
Bias modification wordlists follow the approach of Lauscher et al. (2020) and use a pretrained set of embeddings[6] to expand the set of WEAT words to their 100 unique nearest neighbours. For all experiments, we take the union of all WEAT terms, expand them, and use this expanded set for both dataset balancing and for Attract-Repel. For gender bias in coreference and hate speech, we use terms that are male vs. female and are career, math, science, vs. family, art. For gender bias and migrant bias in hate speech, we compare male/female or migrant/non-migrant with pleasant-unpleasant term expansions.

4 Results

Figure 2 displays one scatterplot for each task and bias-type combination for all experiment conditions. If as a field we want to be able to broadly use WEAT metrics for any given bias research, these graphs should each show a clear and a positive correlation. None of them do. There are no trends in correlation between the metrics that hold in all cases regardless of experimental detail, for any of the tasks. We additionally examined whether there are correlations restricted to individual conditions by breaking out graphs for each WEAT test, embedding algorithm, extrinsic metrics, and bias modification method. The select cases where positive correlations are present are discussed below. All breakout graphs are included in Appendix F.

Coreference (en): Gender The coreference task has the clearest relationship of all three tasks, with a significant moderate positive correlation for both Precision and Recall for word2vec. The overall trends are muddied by the data for fastText, which does not have a significant correlation under any conditions. Both that coreference would display the strongest trends, and that fastText would display the weakest, are as expected. The Winobias coreference task is as directly matched to the WEAT tests themselves as it is possible to be - since both use common career words to measure bias. So the relationship between the two metrics should be clearest here, where moving female terms closer to certain career terms should most directly help a system resolve anti-stereotypical coreference chains. And it

[6]We use the models available in `spacy.io`



(a) Coreference results (en), gender bias

(b) Hatespeech detection results (en), gender bias

(c) Hatespeech detection results (es), gender bias

(d) Hatespeech detection results (es), migrant bias

Figure 2: All data points (Performance Gap x, WEAT y) for each of the 4 tasks: gender bias in co-reference (en), gender bias in hatespeech detection (en), gender bias in hatespeech detection (es), and migrant bias in hatespeech detection (es). Original embeddings (before modification) shown in black. There is no correlation that holds independently of experimental conditions (embedding type, bias modification method, WEAT test).

is. We additionally expected fastText to behave less predictably because of its use of subwords. When subwords are used, word representations are more broadly interconnected. For example, the representation of the word *childish* is by design also made up of the representations for *child* and *ish*, but also all bigrams and trigrams it contains (*c*, *ch*, *chi*, etc). Is it difficult to predict how changing the composition of a training corpus will affect all words that contain *ch* in them. or this reason, fastText may be initially more resistant to encoding biases than word2vec, as was found in Lauscher and Glavas (2019), but may also be more complicated to debias. This has implications for extending this work to contextual models, which always use some form of subword unit.

Hatespeech (en): Gender The baseline level of bias (shown in black in Figure 2) in English hatespeech differs by embedding type, but only for precision. Initial models (with unmodified embeddings) using fastText have 10 additional points of precision for male-targeted hatespeech than for female-targeted. However initial models using word2vec have the opposite bias and have 4 less points of precision for male-targeted than female targeted hatespeech. For recall, the two embedding algorithms are equivalent, with 6 less points for male-targeted hatespeech. In fact, in the recall metric there is an early indication of unreliability of the the relationship between WEAT and extrinsic bias metrics, because there is a spread of different WEAT results that map to nearly the same difference in recall.

Figure 3 shows the trends broken out by WEAT measure and embedding for precision and recall, as well as by pre- vs. post-processing. Based on the spread of data points, it is easy to see that there is overall more effect on precision gap when embeddings are modified, whereas recall performance gap occupies a narrower band over a wide spread of WEAT metrics. There are no trends across any experiment conditions, save when isolating by debiasing method. Before breaking out by bias modification method, it seems that there is no relationship between the two metrics. This holds for precision. However, for recall, the aggregate appears this way is because there is a moderate *positive* correlation for postprocessing methods, but a moderate *negative* correlation for preprocessing. This holds for both embedding algorithms, though both positive and negative correlations are stronger for fastText. The absolute variance in recall is much smaller than for precision, but this is still significant for each embedding algorithm individually and for both grouped together.

Hatespeech (es): Gender and Migrant For hatespeech in Spanish, we examine two kinds of bias separately - gender bias and bias against migrants, in Figure 2c and Figure 2d. The only trends in correlation are when broken out by preprocessing vs. postprocessing, in Figures 4a and 4b, where it is also visible that it is much more difficult to modify bias for Spanish when preprocessing vs. when postprocessing.

There are many less data points for Spanish hatespeech than for English hatespeech, and they are less spread out. This is for a number of reasons that compound each other, and that underscore the difficulty of expanding supposedly language-agnostic techniques beyond English, even to high resourced languages like Spanish. First, we have only one WEAT test for each type of bias, since we made our own that carefully balanced grammatical gender, after rectifying the issues with the existing translated versions (see discussion above(todo: convert this to section reference)). Second, bias modification is more difficult - the richer agreement system in Spanish means that there are more surface forms of what would be one word in English, and it is more challenging to change all of them. Finally, the language model used for nearest neighbour expansion of wordlists (todo: convert this to section reference to wordlists) produces predominantly formal register words from news or scientific articles, presumably due to a less varied makeup of its training data than the English model. This makes them less well suited to debiasing twitter data, specifically, and there were no readily available models that had more casual register. For bias against migrants specifically, there is the additional challenge that wordlists are predominantly based on proper names, which are much rarer in twitter (with tends to use @ mentions instead) than in other media.

Gender bias is much easier to identify in Spanish than some other kinds of bias [perhaps better represented in the training data or via specific slurs, give some context] and has high performance scores with F1 in the high 80s. So there is always, in an *absolute* sense, no gender bias (all values are negative). There are no overall trends when this is modified to be more or less extreme. There is a moderate negative correlation between the two metrics for precision only when looking at only fastText embeddings, and this is true of both preprocessing and postprocessing as bias modification methods. There are no evident trends in any other experimental conditions.

Migrant bias is clearly challenging to identify, with much lower F1 in the low 60s. There are similarly no trends save in a select few experimental conditions. There is a positive correlation for recall between migrant bias and performance gap for preprocessing for fastText only (word2vec displays no relationship for recall). This fits the expectation that fastText is more sensitive to preprocessing than post-processing (todo: reference sec explaining this), though in the gender bias case it is equally sensitive to both, so it is hard to draw conclusions. To confuse the situation further, the only trends in precision are present in word2vec, and are negative correlations, and exist across both bias modification conditions.

5 Discussion

The broad results of this research show that we cannot blindly rely on an intrinsic WEAT metric as informative regarding extrinsic bias. So an NLP scientist or engineer has limited options when investigating and mitigating bias. They must a) find the specific set of wordlists, embedding algorithms, downstream tasks, and bias modification methods that are together predictive of bias for the given task, language, and model or b) implement full systems to test bias effects more directly, even if they work on language models and embeddings.

While the latter may seem onerous, it is not necessarily more so than exhaustively searching for a configuration where embedding bias metrics are maximally predictive.

This underscores the importance of making good downstream bias measures available, as either approach will require these. More datasets that are collected need to be annotated with subgroup demographic and identity information – there are very few available. More research needs to be done on the creation of good multipurpose challenge sets, both to use on their own merits as much more targeted and granular measures of bias, and as a substitute for test sets with identity labels when none are available. It is only when one or both of these things are readily available that we can see the true measure of the efficacy of our debiasing efforts.

6 Conclusion

We have examined the relationship of the intrinsic bias metric WEAT to the extrinsic bias metrics of Equality of Odds and Predictive Parity, for multiple tasks and languages, and determined that positive correlations between them exist only in very restricted settings. In many cases there is no correlation at all, or is even a negative one. While intrinsic metrics such as WEAT remain good descriptive metrics for computational social science, and for examining bias in human texts, we urge that the NLP community not rely on them for measuring model bias. We instead urge that they focus on careful consideration of downstream applications and the creation of datasets and challenge sets that enable measurement at this stage.

Acknowledgements

We thank Andreas Grivas, Kate McCurdy, Yevgen Matusevych, Elizabeth Nielsen, Ramon Sanabria, Ida Szubert, and Sabine Weber for comments on earlier drafts of this paper.

References

Valerio Basile, C. Bosco, E. Fersini, Debora Nozza, V. Patti, F. Pardo, P. Rosso, and M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval@NAACL-HLT*.

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *ArXiv*, abs/2005.14050.

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

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

T. Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *ArXiv*, abs/1905.12516.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AIES '18*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, I. Leontiadis, Jeremy Blackburn, G. Stringhini, Athena Vakali, M. Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *ArXiv*, abs/1802.00393.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP*.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *ACL*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? *ArXiv*, abs/1910.14161.

A. Greenwald, D. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*.

Felix Hill, Roi Reichart, and A. Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *FAT* '19*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Keita Kurita, N. Vyas, Ayush Pareek, A. Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Anne Lauscher and Goran Glavas. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In **SEM@NAACL-HLT*.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *AAAI*.

Kenton Lee, Luheng He, M. Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *NAACL-HLT*.

K. McCurdy and Oguz Serbetci. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *ArXiv*, abs/2005.08864.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nikola Mrksic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gasic, Anna Korhonen, and Steve J. Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Gastã Salamanca and Lidia Pereira. 2013. PRESTIGIO Y ESTIGMATIZACIÃ“N DE 60 NOMBRES PROPIOS EN 40 SUJETOS DE NIVEL EDUCACIONAL SUPERIOR. *Universum (Talca)*, 28:35 – 57.

Maarten Sap, D. Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *ArXiv*, abs/1909.01326.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *ACL*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

R. Weischedel, E. Hovy, M. Marcus, and Martha Palmer. 2017. Ontonotes : A large training corpus for enhanced processing.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *ACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *ArXiv*, abs/1904.03310.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *EMNLP/IJCNLP*.

A Bias Metric Definitions

Here will be math for the bias metrics

B WEAT Wordlists

Here will be included the full WEAT wordlists

C Preprocessing

Here will be data about how we preprocessed

D Training Data

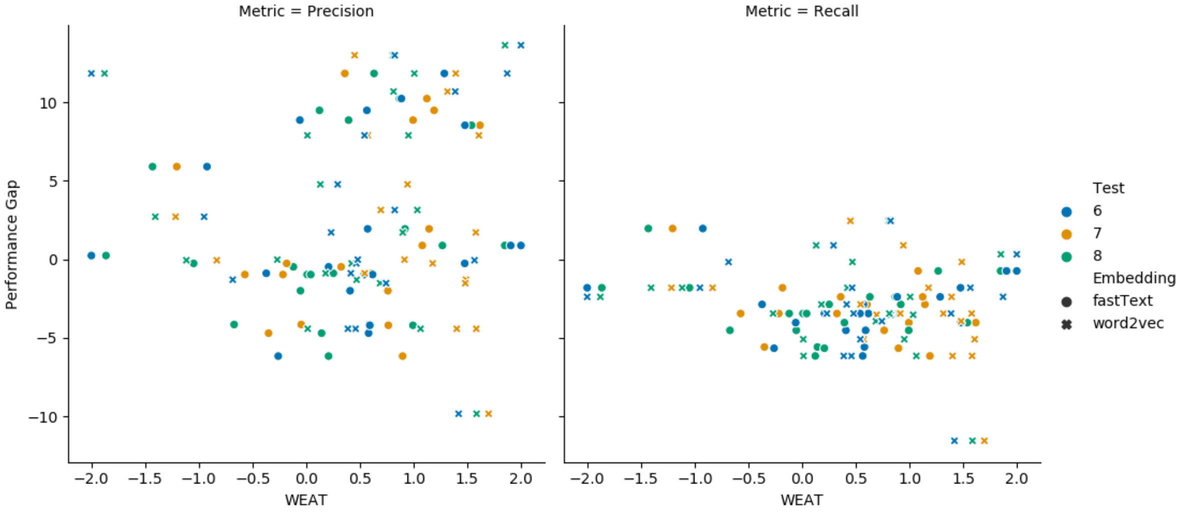
Here will be data about the training datasets

E Bias Modification

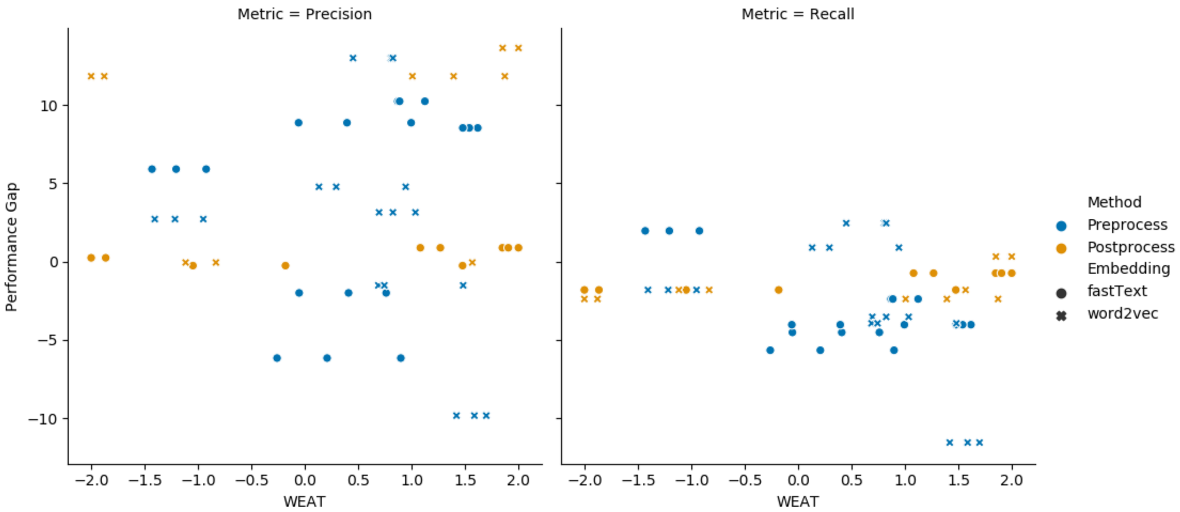
Here will be further details of exactly how we modified bias

F Further Results Graphs

Below are breakouts of graphs by different WEAT tests, embedding type, and bias modification method.

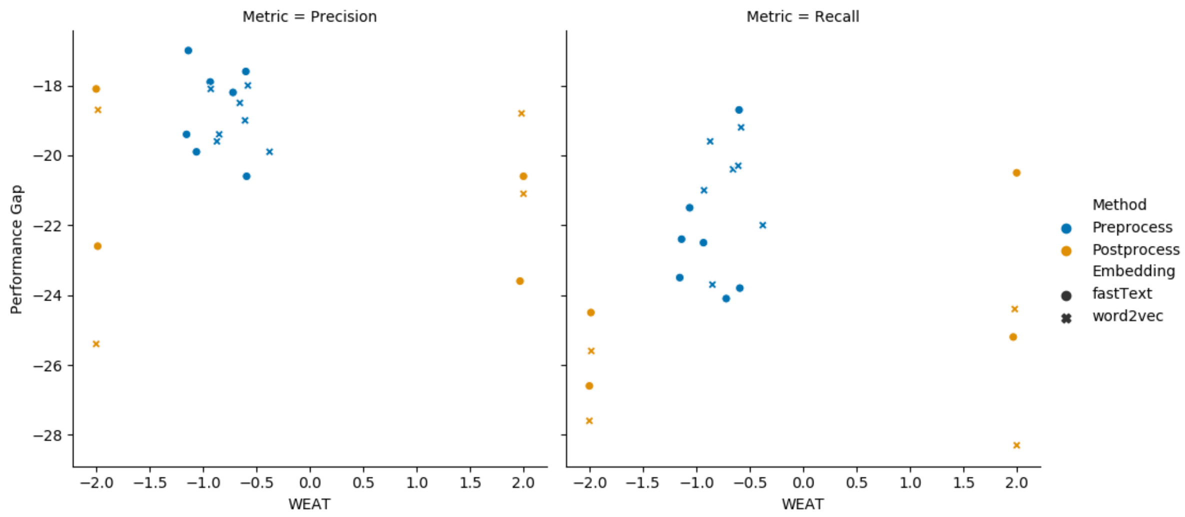


(a) Hatespeech (en) results broken out by different standard WEAT tests for gender bias.

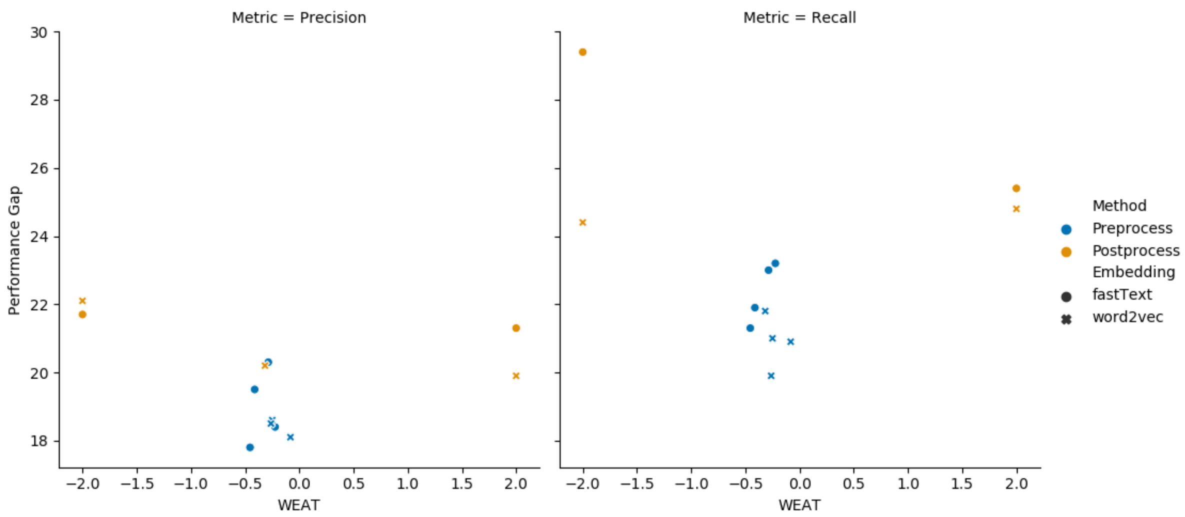


(b) Hatespeech (en) results broken out by bias modification method (pre- vs. post-processing).

Figure 3: Hatespeech Detection (en): Breakout of performance gap by WEAT test type and by embedding, as well as by pre vs. post-processing, for both precision and recall.



(a) Hatespeech (es) results for gender bias metrics broken out by bias modification method.



(b) Hatespeech (es) results for migrant bias metrics broken out by bias modification method.

Figure 4: Hatespeech Detection (es): Breakout of performance gap by pre- vs. post processing, for gender bias and for migrant bias.