So Can We Use Intrinsic Bias Measures or Not?

Sarah Schröder[a], Alexander Schulz[b], Philip Kenneweg[c] and Barbara Hammer[d]
CITEC, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany

Keywords: Intrinsic Bias Measures, Pretrained Language Models, Template-Based Evaluation.

Abstract: While text embeddings have become the state-of-the-art in many natural language processing applications, the presence of bias that such models often learn from training data can become a serious problem. As a reaction, a large variety of measures for detecting bias has been proposed. However, an extensive comparison between them does not exists so far. We aim to close this gap for the class of intrinsic bias measures in the context of pretrained language models and propose an experimental setup which allows a fair comparison by using a large set of templates for each bias measure. Our setup is based on the idea of simulating pretraining on a set of differently biased corpora, thereby obtaining a ground truth for the present bias. This allows us to evaluate in how far bias is detected by different measures and also enables to judge the robustness of bias scores.

1 INTRODUCTION

Biases in machine learning models, but especially in pretrained language models (PLM) are complex in nature and difficult to assess with respect to all possible bias origins and potential harms in applications (Shah et al., 2019). While the ability to detect bias in such PLMs is crucial for transparency and also constitutes a first step for mitigating bias related problems, there exists a variety of different methods for measuring bias in such models (Caliskan et al., 2017; May et al., 2019; Bolukbasi et al., 2016; Gonen and Goldberg, 2019; De-Arteaga et al., 2019). In particular, they can be grouped by intrinsic and extrinsic evaluation strategies, where the former ones directly work with the intrinsic text embeddings while the latter aim for evaluation with an additional down stream task.

In this work we focus on intrinsic bias measures and biases that are learned during pretraining, because such pretrained models are the basis in many applications and biases learned in the pretraining phase can persist in later applications. The class of intrinsic bias measures is, however, diverse, including different cosine based scores (Caliskan et al., 2017; May et al., 2019; Bolukbasi et al., 2016) and Neighbor, Clustering and Classification tests (Gonen and Goldberg, 2019) among others. A unified, systematic and fair comparison between them remains challenging.

In the present work, we conduct an experiment where we simulate pretraining on a variety of biased corpora and assess how these biases manifest in the model. Not only do we consider binary gender bias, as frequently done in related work, but also assess multi-attribute biases at the example of religious and ethnicity biases. We evaluate many intrinsic bias measures from the literature in terms of their ability to capture biases in the model, compatibility to each other and their robustness. Compatibility is often impacted by different testing scenarios (e.g. target words, templates) of the different bias measures (Seshadri et al., 2022). We remove this variable by using a unified test case for all measures, to specifically find out how the scores themselves relate to each other.

Our contributions are: (i) We propose a novel test framework for intrinsic bias measures w.r.t. pretraining of language models, which enables a fair comparison between different measures, regarding their performance to measure bias and their stability. (ii) Thereby, we create and publish a benchmark that is significantly larger and has more variety than other template-based approaches like BEC-Pro (Bartl et al., 2020) or the SEAT test cases (May et al., 2019). (iii) Our benchmark includes the possibility of using multi-group attributes, which we demonstrate for ethnicity and religion. Thereby we gain insights into how binary and multi-group biases manifest in the models and are captured by intrinsic measures. (iv) We perform an experimental evaluation, demonstrating differences in stability and overall performance.

[a] https://orcid.org/0000-0002-7954-3133
[b] https://orcid.org/0000-0002-0739-612X
[c] https://orcid.org/0000-0002-7097-173X
[d] https://orcid.org/0000-0002-0935-5591

Schröder, S., Schulz, A., Kenneweg, P. and Hammer, B.
So Can We Use Intrinsic Bias Measures or Not?.
DOI: 10.5220/0011693700003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 403-410
ISBN: 978-989-758-626-2; ISSN: 2184-4313

For reproducibility's sake we publish our code, including templates, target words and config files[1].

2 RELATED WORK

In literature, there exist many intrinsic bias measures, including WEAT (Caliskan et al., 2017), SEAT, (May et al., 2019), Direct Bias (Bolukbasi et al., 2016), RIPA, (Ethayarajh et al., 2019), Neighbor, Clustering and Classification tests (Gonen and Goldberg, 2019) as well as the log probability bias score (Kurita et al., 2019). We recap those in the next section as their evaluation is at the core of the present work. Further, there also do exist extrinsic bias measures which aim to evaluate bias in a downstream task such as occupation classification, where an example is Bias in Bios (De-Arteaga et al., 2019).

A family of work investigates in how far observations with intrinsic measures transfer to downstream tasks. For instance, it has been shown that WEAT correlates with extrinsic bias metrics only in very restricted settings (Goldfarb-Tarrant et al., 2020) or that intrinsic bias measures in general have a limited correlated with extrinsic metrics (Kaneko et al., 2022). Similarly debiasing before fine-tuning does not effectively reduce biases in the downstream task (Kaneko et al., 2022). As reasons they state that the language models re-learn bias in the fine-tuning step due to flawed training data. Another work (Steed et al., 2022) investigates the bias transfer hypothesis (that biases from PLM affect downstream tasks) using two downstream data sets. The authors show that most of the downstream bias can be explained by fine-tuning. On the other hand, certain debiasing measures (re-sampling and scrubbing of identity terms) w.r.t. the downstream tasks only effectively removes biases when the model was not pretrained (i.e. never contained biases from the start). This in turn indicates that mitigating biases in PLM is not sufficient but must also not be neglected.

It has been shown that templates used for bias measurements have a large impact on the results, i.e. modifying a template without changing the semantic meaning can alter the measured bias significantly (Seshadri et al., 2022). The authors point out that many works use a limited amount of templates, which makes scientific claims less reliable.

Similarly, the compatibility of bias measures from the literature, including the test scenario (used words, templates, attributes) has been found limited. (Delobelle et al., 2021). This authors argue that the "semantically bleached" templates suggested by other work do in fact influence the biases measured when inserting the targets/attributes of interest, hence are not as "bleached". Furthermore, the way sentence representations are created (e.g. CLS embedding vs. mean pooling) influences bias measurements.Basically, a major issue with (intrinsic) bias measures are the incompatible testing conditions.

To summarize bias defintions from other works, a unifying bias framework has been proposed (Shah et al., 2019) that includes four definitions of bias origins (semantic bias in embeddings, selection bias, over-amplification by the model and label bias) and two definitions of predictive biases/ bias consequences (outcome disparity and error disparity).

While (Schröder et al., 2021) aim to compare different bias measures, they focus more on theoretical aspects of bias scores and are, furthermore, restricted to cosine based scores.

3 BIAS EVALUATION METHODS

In the following, we depict the intrinsic bias scores from the literature which we compare in our experiments. Some of these offer both a bias score for single words (e.g. the gender bias of secretary), others only a bias score that determines bias over sets of words. This is often a mean word bias or some contrasting bias measure that checks whether stereotypical groups exist.

3.1 SEAT

SEAT (May et al., 2019) is an extension of the Word Embedding Association Test (Caliskan et al., 2017) for sentence representations. WEAT and SEAT are among the most commonly used intrinsic bias measures. In our experiments we apply SEAT since we use contextualized embeddings and hence need to compare bias measures in the context of sentences. We evaluate both the word dissimilarity score $s(\vec{w},A,B)$ and the effect size $d(X,Y,A,B)$. Since WEAT and SEAT are only defined for binary bias evaluations, we also use the generalized WEAT (GWEAT) (Swinger et al., 2019), which proposes a score for an arbitrary number of groups $g(X_1,A_1,...,X_n,A_n)$. Again we apply it to sentence representations in the way of SEAT.

3.2 DirectBias and RIPA

The DirectBias (DB) (Bolukbasi et al., 2016) determines a bias subspace in the embedding space, which

[1]https://github.com/HammerLabML/PLMBiasMeasureBenchmark

is defined by bias attributes, such as gender words. Then any correlation of neutral words with this subspace is considered a bias. RIPA (Ethayarajh et al., 2019) is closely related to the DirectBias as it uses the same strategy. Contrary to the DirectBias, which normalizes word vectors, RIPA considers the vector length. In our experiments we apply both measures to sentence representations instead of words, in the same way as SEAT. Both can report a word-wise bias score as well as a mean bias for a set of words.

3.3 Clustering, Classification and Neighbor Test

In order to evaluate the effect of debiasing on word embeddings, the clustering, classification and neighbor test (Gonen and Goldberg, 2019) were introduced. These bias tests detect whether words with similar gender stereotypes cluster together in the embeddings space. The neighbor test evaluates how many of the closest k neighbors of a word have the same stereotype. The clustering and classification test measure the accuracy of k-Means clustering and a SVM classifier for separating two groups of stereotypical words. All of these test require the user to define stereotypical groups beforehand. In the paper, the authors sort words by WEAT's dissimilarity score $s(\vec{w},A,B)$ into stereotypical male/female groups. We extend the classification and neighbor test to cases with multi-group biases. For the classification test, we simply use a multi-class classifier, for the neighbor test we count the amount of neighbors that are part of the same stereotypical group. While k-Means itself could also handle multiple clusters, the authors used accuracy to evaluate the clustering performance, which cannot be used with more than two groups.

3.4 Log Probability Scores

There exist a variety of log probability scores, which are closely related to the masked language objective of BERT. Initially (Kurita et al., 2019) proposed the log probability bias score (LPBS), which queries the probability of BERT to replace a mask token with a male over a female term (or vice versa) given a context like a stereotypical male/female occupation. Thereby, the authors could show that BERT does contain human-like biases. Similar scores in literature are the CrowS-Pairs Score (Nangia et al., 2020), StereoSet Score (Nadeem et al., 2021) or All Unmasked Likelihood score (Kaneko and Bollegala, 2021).

4 EXPERIMENTAL DESIGN

Our goal in the following is to be able to compare and judge different intrinsic bias measures regarding their performance to measure bias and their robustness. For this purpose, we aim to achieve a setting, where biases are fully observable: We simulate various demographic distributions from which we derive pretraining corpora and train LMs. To keep computation time minimal and enable training of many models, we use a pretrained Hugginface model (bert-base-uncased) for Masked Language Modeling (MLM) and train it on each of our corpora for a few epochs. Figure 1 illustrates our approach. As a baseline for biases manifesting in the PLM, we directly query our model for the probabilities to replace mask tokens with certain demographic attributes. The details are explained in Section 4.3. Our test cases include gender, ethnicity and religious biases regarding occupations. For ethnicity we consider 5 and for religion 4 different groups, which is crucial because related work often neglects settings with more than two groups.

4.1 Generating Training Data

Many works that evaluate bias measures either only use one or few PLMs or, based on one PLM, mitigate or amplify existing biases to generate a variety of biased models. To achieve a more diverse setting and simulate the whole cycle of data acquisition, pretraining and bias manifestation, we generate demographic distributions, specifically the probabilities of certain demographic groups being associated with occupations. These associations are randomized and do not have to align with actual biases in humans and society. This has the advantage that we can distinguish if the PLM actually learns the biases presented during our training instead of sticking to biases it learned beforehand.

Given a list of target words (here occupations), attributes (gender, religion, ethnicity) and groups per attribute (e.g. male/female for gender), we create a probability distribution per target and attribute $p(group|target,attribute)$, e.g. how likely does a cook belong to a certain ethnicity (if ethnicity is mentioned). Algorithm 1 describes our approach. We used the groups mentioned in Table 1.

Table 1: Demographic groups used in our experiments.

Attributes	Groups per Attribute
Gender	male, female
Ethnicity	black, white, asian, hispanic, arab
Religion	christian, muslim, jewish, hindu

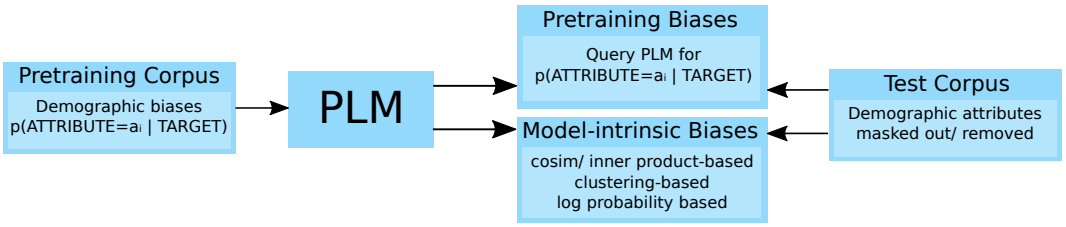

Figure 1: Our experimental setup to evaluate intrinsic bias measures with regard to the pretraining of language models.

Data: $groups, minP < maxP$
Result: $groups, probs$
$n \leftarrow size(groups)$;
$minP \leftarrow minP/n$;
$maxP \leftarrow maxP/n$;
$P \leftarrow 1.0$;
$groups \leftarrow shuffle(groups)$;
$probs = []$;
while $i < n-1$ **do**
  $maxBound \leftarrow min(maxP, P - minP * n)$;
  $p \leftarrow randomUniform(minP, maxBound)$;
  $P \leftarrow P - p$;
  $AddItem(probs, p)$;
end
$AddItem(probs, P)$;

Algorithm 1: This algorithm generates a random demographic distribution for a list of groups. **minP** < 1.0 is a strict lower bound for the group probabilities. The higher, the less likely are extreme differences between the demographic groups. **maxP** is the maximal upper bound. If **maxP** < 1.0 an equal distribution is impossible and there will be exactly one majority group with $p(group) > \frac{1}{n}$.

We create a unique demographic distribution for each combination of $minP \in \{0.05, 0.1, 0.15, 0.2\}$ and $maxP \in \{0.8, 0.85, 0.9, 0.95\}$. Due to the choice of both parameters, each group will be represented with $p(group) > 0$ and there will be exactly one majority group (the last one chosen by the algorithm) with $p(group) > \frac{1}{n}$ given n groups.

To create datasets based on these distributions, we use a selection of template sentences, containing a placeholder for target words and at least one placeholder for a protected attribute. There might be multiple attributes or multiple mentions of the same attribute in one template. Formally, a template sentence S contains a target placeholder and attribute placeholders A_i with i denoting the type of attribute (e.g. gender/ethnicity). Contrary to other works, we do not limit ourselves with "semantically bleached" templates, as it has been shown that even those templates influence the measured biases (Seshadri et al., 2022). Instead, we chose a larger number of templates including many with additional context, assuming that averaging over a variety of templates improves the stability of bias measurements. Some exemplary templates are shown in Table 2.

Table 2: Exemplary templates and the terms that can be inserted as attributes.

i saw the OCCUPATION at the RELIGION2 (church, mosque, synagogue, mandir)
the OCCUPATION enjoyed GENDER1 lunch (his, her)
the OCCUPATION often visits GENDER1 family in ETHNICITY8 (his, her)/(africa, europe, asia, mexico, qatar)

Given a template S and a list of target words T, we create sentences $S_t \forall t \in T$. For each attribute placeholder A_i in S_t we choose a fitting group j and insert the corresponding group attribute $a_{i,j}$ based on $p(a_{i,j}|t, A_i)$. For MLM training we mask out one attribute instead of random words to ensure that BERT specifically learns to infer these. If multiple attributes are present, we create a training sample per attribute, where this one is masked out. The resulting corpora contain 29800 training samples derived from 100 occupations as target words and 165 templates. Among our training templates, 60 include an ethnicity reference, 145 a gender and 45 a religion reference.

4.2 Training and Model Validation

On each of our generated training datasets, we retrain BERT for 5 epochs using the Masked Language Objective. We repeat this for 5 iterations, resulting in 5 BERT instances trained on the same demographic distribution, which will be relevant in our robustness experiments. To ensure that our models learned the biases well enough, we test the Pearson correlation of the unmasking probabilities $p_{unmask}(a_{i,j}|t, A_i)$ (see Section 4.3) with the group probabilities in the training data $p_{data}(a_{i,j}|t, A_i)$. If the correlation is below a threshold of 0.85, we repeat the training (starting from the pretrained BERT again). After a maximum of 5 training runs, we keep the best performing model, to keep computation time limited.

We utilize 113 test templates to generate our test data similarly to the training data, i.e. inserting each

target word into each template. However, we do not insert any demographic attributes, but instead mask out one attribute that was tested for and replace all other attributes by a neutral term. This results in 17600 test sentences. Given these, we query the unmasking probabilities of all groups and average over each target. The exact same test sentences are used for the evaluation of the different intrinsic bias scores (see Section 5.2).

4.3 Baseline: Unmasking Bias

To evaluate intrinsic bias scores in terms of a pretrained model, we choose a measure directly linked to the MLM objective and hence similar to the LPBS, called **unmasking bias**. Given a template such as "GENDER is a TARGET", we replace GENDER by a mask token and TARGET by a target word for which we want to determine the gender bias. Instead of gender one could also use a protected attribute with more than two groups such as religion or ethnicity. Then, we query BERT for the probabilities of certain protected groups to replace the mask token normalized by the probability to choose any of group associated with that attribute. In this example, one could compare the probabilities of "he" and "she".

More formally, given a template s, a target words t and an attribute A_i (e.g. ethnicity), we are interested in the probability that our trained language model associates this target with a group $j \in [1,n]$ (e.g. black), where n is the number of groups for A_i (e.g. 5 for ethnicity, in our case). For this purpose, we insert t and query the model for the probability of the corresponding group attribute $a_{i,j}$ to be inserted into s:

$$p_{unmask}(a_{i,j}|s,t,A_i) = \frac{P([MASK] = a_{i,j}|s,t,A_i)}{\sum_{k=1}^{n} P([MASK] = a_{i,k}|s,t,A_i)}. \quad (1)$$

By using a large set of templates S, we achieve a robust estimate of the probability that group j is associated with target t in the model:

$$p_{unmask}(a_{i,j}|t,A_i) = \frac{1}{|S|} \sum_{s \in S} p_{unmask}(a_{i,j}|s,t,A_i). \quad (2)$$

One advantage of this formulation over LPBS is that $p_{unmask}(a_{i,j}|t,A_i)$ does not depend on the number of protected groups.

Now we can use this probability in order to judge in how far our model has learned the biases present in the data: From our training data, we estimate $p_{data}(a_{i,j}|t,A_i)$ by the relative frequencies of the according group attribute and target word occurring together (we have generated the training using such templates). Then we could measure the dissimilarity between $p_{unmask}(a_{i,j}|t,A_i)$ and $p_{data}(a_{i,j}|t,A_i)$.

Most bias scores, however, report only an aggregated score for one attribute, aggregating over the groups. In order to be comparable to this setup, we also aggregate over the groups using the Jensen-Shannon divergence (Lin, 1991) between $p(a_{i,j}|t,A_i)$ and the uniform distribution which represents no bias. We don't use KL because we want symmetry.

Calculating this for both $p_{unmask}(a_{i,j}|t,A_i)$ and $p_{data}(a_{i,j}|t,A_i)$, we obtain scalar scores for each of the two and for each t. We finally utilize the Pearson correlation over all t for a measure of similarity between the bias in the data and the bias learned by the model. Accordingly, we also compare the similarity between the bias in the data and a bias score.

5 EXPERIMENTAL RESULTS

In our experiments, we first evaluate the reliability of our template-based approach w.r.t. the number of templates. Afterwards, we apply our setup to compare different bias measures, first when considering only biases of individual target words, then using the full bias scores. Finally we investigate the robustness of our scheme and some bias scores when training different models on the same biased data set.

5.1 Assessing the Reliability of Our Template-Based Evaluation

There is work suggesting that template based approaches for bias evaluation are unreliable (Seshadri et al., 2022), which is mostly due to the fact that a limited amount of templates is used. For instance, (Bartl et al., 2020) propose BEC-Pro, a dataset for bias evaluation which is created from only 5 different templates. To circumvent this problem, we use a much larger number of over 100 templates (see Section 4) for training and testing, respectively.

Furthermore, we evaluate the reliability of our evaluation with respect to changing template numbers in the following. We run a statistical test to estimate the significance of biases obtained with our templates: We compute the unmasking probabilities for

Table 3: Mean and standard deviation of p-values reported by our permutation test over all models compared to number of test templates with the respective attribute.

	Gender	Religion	Ethnicity
p-value	0.01 $\pm$0.01	0.08 $\pm$0.03	0.14 $\pm$0.03
templates	82	34	53

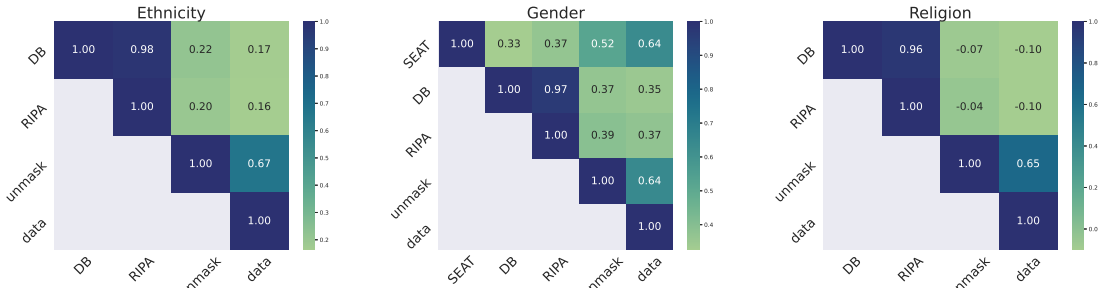

Figure 2: Pearson Correlation of word bias scores for ethnicity, gender and religious bias.

each combination of template sentence, target word and protected attribute and take the mean probabilities per target and attribute combination. Then we run a permutation test for 1000 iterations under the hypothesis that the ranking of group probabilities does not change if computed over a subset (90%) of templates. We report the p-values in Table 3.

These show that in the gender case, which involved the largest number of templates, we achieve statistical significant results. In the multi group cases we observe p-values larger than 0.05. Hence results are not as reliable. We assume that this is due to the smaller number of test templates and the increased complexity of bias for more than two groups.

5.2 Comparison of Bias Measures

5.2.1 Word Biases

For each protected attribute, we first evaluate the correlation of different word bias scores. This includes only the cosine scores and the data and unmasking probability, where we report the Jensen-Shannon divergence (see Section 4.3).

The results are illustrated in Figure 2. We report relatively high correlations with the data biases for our baseline ('unmask') with any attribute. The cosine scores correlate rather well with the baseline and data biases in the gender case, but not in the multi attribute cases (SEAT is only defined for the binary gender case). Furthermore, the DirectBias and RIPA correlate strongly with each other, which was expected due to their similar definitions.

This contradicts observations from previous work, which reports that cosine scores failed to capture biases in some scenarios (Goldfarb-Tarrant et al., 2020; Kurita et al., 2019) or produced inconsistent results (May et al., 2019). However, it has been observed that the compatibility of different bias measures is strongly limited by the different test cases, i.e. when comparing word biases to sentence biases or computing biases with different kinds of templates (Delobelle et al., 2021). Our findings underline that unifying test cases makes different bias scores more compatible.

The fact that we report much higher correlations in the gender case could mean that a binary concept such as gender bias is in fact represented in a linear way, which satisfies the assumptions of cosine based scores, while non-binary concepts could be represented in a more complex way which cosine based scores cannot account for entirely.

5.2.2 Overall Bias Scores

In practice, often a bias score over whole concepts (set(s) of words) instead of single words is reported, which is the case e.g. for WEAT/SEAT. Hence, we evaluate these bias scores compared to the mean unmasking bias over all target words (see Section 4.3). The DirectBias and RIPA apply an "objective" mean word bias. On the other hand, SEAT, GWEAT, the clustering, classification and neighbor test require the target words to be sorted into groups based on their stereotypes. When users choose such groups one can assume that they have certain knowledge about the stereotypes that can be expected. However, their expectations might not exactly align with the biases in the data/model, either due to subjective views, limited knowledge or sampling bias of the dataset compared to biases in the real world. To model this, we derive the groups from the actual demographic probabilities in our datasets after adding Gaussian noise ($\mu = 0, \sigma = 0.3$). For each occupation, we simply choose the group with the highest co-occurrence probability in the dataset.

Figure 3 shows the correlations of the bias measures and the data biases. Overall, correlations are slightly lower than for the word bias scores. Other than that the results are similar in the sense that we observe stable correlations with the data bias for the baseline ('unmask') for all three attributes and observe the best correlations in the gender case. We find that SEAT and the neighbor bias achieve the best Pearson correlation both with the data and unmasking bias and also correlate with each other.

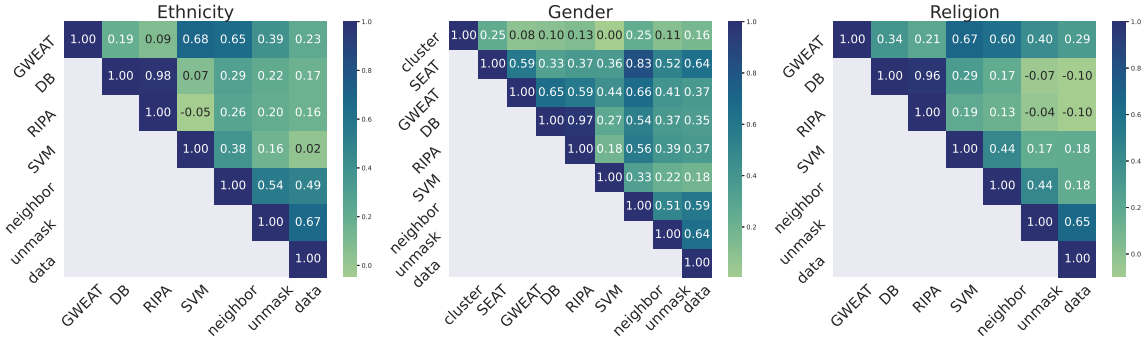

Figure 3: Pearson Correlation of bias scores for ethnicity, gender and religious bias.

Table 4: Consistency over experiment iterations. Mean percentage differences of scores reported in models with the same demographic distributions. Lower values indicate consistency.

	SEAT	DB	RIPA	unmask
Ethnicity	-	0.501	0.518	0.111
Gender	1.358	0.358	0.361	0.291
Religion	-	0.486	0.483	0.147

GWEAT shows a similar behavior to SEAT, but overall lower correlations. The clustering and classification ('SVM') achieve the lowest correlations with the data biases. Noteworthy the differences between gender and the multi group cases are comparably low for GWEAT and the neighbor test. Other than that, we find high correlations between GWEAT and the classification ('SVM') bias.

5.3 Robustness of Bias Measures

In addition to the general performance and compatibility of bias measures, we also want to investigate their robustness. For this purpose, we test how stable bias measures are, given the same initial bias distribution. We compare the word bias measures over the 5 iterations (5 different models) trained with the same bias distribution. Results are aggregated over all bias distributions and shown in Table 4.

The fact that the unmask baseline achieves very small percentage differences indicates that the different model instances have learned the according bias robustly. Overall the considered bias measures vary more than the baseline, where SEAT varies more than DirectBias and RIPA. For DirectBias and RIPA we see slightly more stable results in the gender case, which could indicate that an increasing number of groups also makes bias measurements more difficult and thus less robust. For SEAT we see large discrepancies between biases measures over similar models, which is due to the fact that SEAT reports rather low biases in general. However, we also observe that the signs of biases reported by SEAT frequently change.

6 CONCLUSION

In the present work we have proposed a novel test scenario where different bias measures can be compared using the same test sentences. We have showed that template-based approaches can produce reliable results if a sufficient number of templates is employed. In addition to binary attribute cases, we have also evaluated multi-attribute cases and showed large differences between them regarding the detection of bias. Furthermore, we showed that cosine and intrinsic scores actually can attest for biases quite well, given our specific pretraining scenario. We have shown that the more complex bias tests based on classification and clustering do not perform better in detecting bias than the simple SEAT score. We have further investigated the robustness of word biases and have shown that SEAT performed inferior as compared to DirectBias and RIPA.

Limitations of our work include the not perfect baseline performance of the pretrained model, i.e. our trained models do represent the bias perfectly (the unmask correlation does not go over 70%). This could make it somewhat more difficult for bias methods to detect the bias. Here, more training data could yield an improvement. Further, our evaluation is limited to pretraining as such and implications for downstream tasks cannot be directly transferred. However, pretrained models can be used directly without fine-tuning, so this setup remains valid.

This work opens up new interesting directions of investigations, including direct evaluations in settings where fine-tuning the embedding is not possible, such as similarity ranking or training only a classification head. Here biases from pretraining are likely to persist. The investigation of intersectional groups is a

further promising application of our setup, because our template-based approach naturally allows to include multiple protected groups also in the test set. Finally, a further extension of the benchmark, w.r.t. templates for the multi-attribute attributes or other targets than occupations could improve the evaluation further.

ACKNOWLEDGEMENTS

We gratefully acknowledge the funding by the German Federal Ministry of Economic Affairs and Energy (01MK20007E) and by the Ministry of Culture and Science of the state of North Rhine-Westphalia in the project "Bias aus KI-Modellen".

REFERENCES

Bartl, M., Nissim, M., and Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

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

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Delobelle, P., Tokpo, E. K., Calders, T., and Berendt, B. (2021). Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., and Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862.

Kaneko, M. and Bollegala, D. (2021). Unmasking the mask - evaluating social biases in masked language models. *CoRR*, abs/2104.07496.

Kaneko, M., Bollegala, D., and Okazaki, N. (2022). Debiasing isn't enough!–on the effectiveness of debiasing mlms and their social biases in downstream tasks. *arXiv preprint arXiv:2210.02938*.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. *CoRR*, abs/1903.10561.

Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Schröder, S., Schulz, A., Kenneweg, P., Feldhans, R., Hinder, F., and Hammer, B. (2021). Evaluating metrics for bias in word embeddings. *CoRR*, abs/2111.07864.

Seshadri, P., Pezeshkpour, P., and Singh, S. (2022). Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

Shah, D., Schwartz, H. A., and Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Steed, R., Panda, S., Kobren, A., and Wick, M. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., and Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.