Investigating the Relationship Between Model Size and Gender Bias in Contextualized Language Models

Lotte Thys

January 2025

Abstract

Contextualised word embeddings play a crucial role in modern natural language processing (NLP) systems and are becoming increasingly popular. Recent research has show that they also inherit and amplify biases present in the real-world data on which they are trained. While early research has explored bias in non-contextualized word embeddings, measuring bias in contextualized embeddings remains a challenge due to their dynamic word representations and the incompatibility of existing measures. This research project investigates the relationship between model size and gender bias in contextualized language models by evaluating three architectures—BERT, RoBERTa, and XLNet—at two different sizes. Five bias detection methods (SEAT, LPBS, CAT, CrowS-Pairs, and AUL) are employed to assess whether larger models exhibit increased stereotypical bias and whether bias correlates with language modeling ability. Results indicate that while some bias metrics show an increase in bias with model size, others display neutral or even inverse trends. This highlights incompatibility across bias measures. Similarly, the correlation between language modeling ability and bias varies across metrics. These findings emphasize the need for a nuanced approach when assessing fairness in language models and call for further refinement of bias measurement techniques.

1 Introduction

In recent years, word embeddings have become a fundamental component of natural language processing (NLP) systems. In order to be able to model language, these word embeddings are trained on large amounts of real-world text. This training data will inevitably contain biases and stereotypes that are prevalent in society. During training, these biases are picked up by word embeddings. Early research into bias in non-contextualized word embeddings showed that these word embeddings reproduce and amplify the biases that are present in their training data (Bolukbasi et al. (2016), Caliskan et al. (2017)). Caliskan et al. (2017) developed the Word Embedding Association Test (WEAT) as a measure of bias in non-contextualized word embeddings.

In non-contextualized word embeddings such as Word2vec (Mikolov et al. (2013)) and GLoVe (Pennington et al. (2014)), a word is represented by a single

vector, regardless of context. This means that each time a word occurs in a text, it is projected onto the same vector. In contextualized word embeddings (e.g.: BERT, Devlin et al. (2019); XLNet, Yang et al. (2019); RoBERTa, Liu et al. (2019)), each occurrence of a word is represented by a different vector, depending on its surrounding context. This allows models to better handle polysemy (words with multiple meanings). This innovation has resulted in improved performance in most NLP tasks and benchmarks. Consequently, contextualized word embeddings have become more and more popular.

The fact that words do not have a singular representation in non-contextualized word embeddings means that WEAT cannot be used to measure bias in these models. Previous research has shown that these contextualized embeddings also exhibit bias (May et al. (2019), Kurita et al. (2019), ...).

Several different measures have been proposed to quantify the bias in these models, a selection of which I review in section 2. There is a wide variety of types of bias that these different measures investigate. Some focus only on gender bias or racial bias. Others consider multiple types of bias. For this project, I focus specifically on gender bias to limit the scope of the project and to make comparison between bias measures more consistent.

One of the proposed measures is the Context Association Test (CAT), introduced in Nadeem et al. (2021). CAT combines a bias metric (which they call the *stereotype score*) with a score that evaluates the ability to model language. In an experiment, the authors of the paper evaluated several different models in at least two sizes each. One of their results was that as the model size increases for a given architecture, the language modeling score and the stereotype score both increase. They also found a strong correlation between a model's language modeling ability and its stereotype score. This is a concerning trend that piqued my interest.

Delobelle et al. (2022) surveyed the literature on fairness metrics and investigated the compatibility of different bias metrics for contextualized language models. They concluded that many of these existing metrics do not correlate with each other. This raises the question whether the trends concerning model size and bias found in Nadeem et al. (2021) also hold for other bias measures.

In this research project, I further investigate the relationship between model size and bias by comparing the base and large versions of three different contextualized language models. I compare five different bias scoring methods to see how they correlate with model size. In addition, I look at the correlation between the language modeling score of Nadeem et al. (2021) and the different bias scores, to investigate the correlation found in Nadeem et al. (2021) between language modeling ability and stereotypical bias. By using multiple bias scores, this experiment provides a broader perspective on how model size influences bias across different evaluation frameworks.

The research question this project aims to answer is twofold:

- In general, do other bias scores increase when model size increases, like the stereotype score from Nadeem et al. (2021) does?
- Does the statement "as a model becomes stronger, so does its stereotypical bias" from Nadeem et al. (2021) hold true with other bias scores?

2 Literature Review

2.1 WEAT (Word Embedding Association Test)

The Word Embedding Association Test (WEAT) (Caliskan et al. (2017)) is a bias detection metric for non-contextualized word embedding based on the psychological measure of the Implicit Association Test Greenwald et al. (1998). WEAT measures the association between two sets of target words on the one hand (e.g., typically male and typically female gendered professions) and two sets of attribute words on the other (e.g., male and female gender words).

Let X and Y be equal-size sets of target word embeddings and A and B be two sets of attribute words. The WEAT test statistic is

$$s(X, Y, A, B) = \Big[\sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)\Big]$$

where

$$s(w, A, B) = [\text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)]$$

In other words, $s(w, A, B)$ calculates how much more word w is associated with attributes in A than with attributes in B (using cosine similarity), and $s(X, Y, A, B)$ measures the difference in magnitude in these associations for words in X as opposed to words in Y.

A permutation test on $s(X, Y, A, B)$ is used to compute the significance of the association between (X, Y) and (A, B).

$$p = \Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

where the probability is calculated over the space of partitions (X_i, Y_i) of $X \cup Y$ where X_i and Y_i are of equal size

The effect size is

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)}$$

A larger effect size corresponds to more severe pro-stereotypical representations, controlling for significance. WEAT was developed for non-contextualized word embeddings and relies on a singular vector representation for each word.

2.2 SEAT (Sentence Encoder Association Test)

May et al. (2019) introduces the Sentence Encoder Association Test (SEAT), a bias detection metric that extends WEAT for use with contextualized language models. SEAT compares sets of sentences instead of sets of words by applying WEAT on the vector representation of a sentence, retrieved from a sentence encoder. To perform a word-level test using a sentence encoder, template sentences are used. Some example template sentences are "This is <word>.", "<word> is here.", "This will <word>.", and "<word> are things.". These sentences were selected to be semantically bleached so that the sentence would convey little meaning apart from the meaning of the inserted word. Using SEAT, May et al. (2019) experimentally tested seven different sentence encoders, including BERT. They find that the contextualized sentence encoders exhibit less bias than previous models do. However, they suggest that the templates might not

be as semantically bleached as expected or that cosine similarity might be an inadequate measure of similarity for sentence encoders.

Delobelle et al. (2022) show that the choice of templates can indeed impact the fairness evaluations of metrics that use templates, and that these templates are not completely "semantically bleached". They urge caution when using templates.

2.3 LPBS (Log Probability Bias Score)

Kurita et al. (2019) introduce the Log Probability Bias Score (LPBS). This bias detection measure uses templates containing slots for both an attribute word and a target word (e.g., "[TARGET] is a [ATTRIBUTE]"). LPBS corrects for the prior probability of the target word. This prior probability is the probability with which the model selects the target word for a given template sentence when the attribute word is masked, i.e. how likely the word is in the model, given a template sentence without the attribute word. LPBS is defined as the difference between the increased log probability score for two targets (e.g. he/she), where the increased log probability score for a target-attribute pair is calculated as

$$\log \frac{p_{tgt}}{p_{prior}}$$

with p_{tgt} being the probability that the model assigns the target word to the sentence where the [TARGET] is masked, but the attribute is left in the sentence, and p_{prior} being the probability that the model assigns the target word to the sentence where both [TARGET] and [ATTRIBUTE] are masked.

Kurita et al. (2019) computed LPBS for the set of attributes and target words from Caliskan et al. (2017). They calculated the mean LPBS for each attribute and measured statistical significance with a permutation test.

2.4 CAT (Context Association Test)

Nadeem et al. (2021) introduced the Context Association Test (CAT) to measure stereotypical bias in four domains (one of which is gender). This test is based on a crowd-sourced dataset they created called StereoSet. StereoSet consists of target terms that are each given three contexts; one stereotypical, one anti-stereotypical, and one unrelated. In addition to measuring the level of bias of a model, CAT also measures its language modeling ability. The stereotypical and anti-stereotypical contexts are used to measure the level of stereotypical bias. The unrelated context is used to measure the language modeling ability. There are two types of CAT. Intrasentence CAT test the model on the sentence level, whereas intersentence CAT does so on the discourse level. The idealized CAT score (icat) is defined as

$$icat = lms * \frac{min(ss, 100 - ss)}{50}$$

where lms is the language modeling score and ss is the stereotype score. The lms of a target term is the percentage of instances in which the model prefers a meaningful association (stereotype or anti-stereotype) over a meaningless (unrelated) one. The overall lms is the average lms for all target terms. Therefore, the ideal lms score is 100. The ss of a target term is the percentage of instances

in which the model prefers a stereotypical association over an anti-stereotypical one. The model should not lean more towards stereotypical associations, nor anti-stereotypical ones. Thus, the ideal *ss* score is 50. The ideal icat score is 100.

Nadeem et al. (2021) critique May et al. (2019) and Kurita et al. (2019) for the use of artificial context (i.e., template sentences) as these do not reflect the natural usage of a word. They also critique the need to predefine stereotypical attribute terms and the fact that May et al. (2019) do not include multi-word terms like "construction worker".

2.5 CrowS-Pairs (Crowdsourced Stereotype Pairs)

Nangia et al. (2020) introduce the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs). It covers nine types of bias, one of which is gender bias. The CrowS-Pairs dataset is a crowd-sourced dataset that consists of pairs of sentences. In each pair, one sentence is either stereotypical or anti-stereotypical about a disadvantaged group, and the other one is about a contrasting group. The sentences in a pair are minimally distant, meaning that the only words that differ between the sentences are those that identify the group being spoken about. An example of such a pair of sentences is "**John** ran into **his** old football friend" and "**Shaniqua** ran into **her** old football friend". Where the bold words, the only words that differ between the sentences, identify a man in the first sentence and a woman in the second one. Nangia et al. (2020) define the unmodified part U of a sentence S as the tokens that occur in both sentences of the pair and the modified part M of a sentence as the tokens that do not overlap with the other sentence in the pair. To compute a score of a model for this benchmark, the percentage of instances where the model prefers the more stereotyping or less anti-stereotyping sentence is calculated. There is a possibility that the modified tokens of one sentence are more represented in the model than the modified tokens of the other sentence. Taking the example above, "John" is likely to occur more often in the data on which the model is trained than "Shaniqua". Nangia et al. (2020) want to control for this imbalance in the scoring of models. To control for this imbalance, they estimate the probability of the unmodified tokens conditioned on the modified tokens, calculating the preference of the modified (group specific) part of one sentence over that of the other based on the context of the unmodified part of the sentence. To approximate this probability, they adapt the pseudo-log-likelihood MLM scoring (Wang and Cho (2019); Salazar et al. (2020)). For each sentence, one unmodified token u_i at a time is masked until all unmodified tokens have been masked

$$\text{score}(S) = \sum_{i=0}^{|C|} \log P(u_i \in U | U_{\setminus u_i}, M, \theta)$$

The metric for a model is the percentage of instances where here the model prefers the more stereotyping or less anti-stereotyping sentence. The ideal score is 50

This metric is similar to CAT in that both metrics use crowd-sourced datasets and use the percentage of instances where the model chooses the more stereotyping option to assign a score. The main differences are that (1) they use different datasets, (2) CAT's icat score also takes into account the language modeling

abilities of the model, and (3) the CrowS-Pairs metric uses an adapted pseudo-log-likelihood score which takes to correct for the frequency of occurrence of words in the training data.

2.6 AUL (All Unmasked Likelihood)

Kaneko and Bollegala (2022) introduced All Unmasked Likelihood (AUL). They argued that previous pseudo-log-likelihood based measures fail to take into account that MLMs can predict multiple plausible tokens when given a context sentence with a masked token. In combination with the fact that the datasets for previous measures only have one correct answer for every test instance, this results in low probabilities for the correct answers. Kaneko and Bollegala (2022) also argued that the previous pseudo-log-likelihood methods assumed that the masked tokens are statistically independent, which they are not in practice. They introduced AUL as a pseudo-log-likelihood based bias evaluation metric that does not have these problems. AUL predicts all of the tokens given the whole unmasked sentence, where CAT predicts the modified tokens given the unmodified ones and CrowS-Pairs predicts the unmodified tokens one by one given the modified tokens and the other unmodified tokens.

$$\text{AUL}(S) := \frac{1}{|S|} \sum_{i=1}^{|S|} \log P_{MLM}(w_i|S;\theta)$$

The bias score is defined as

$$\frac{1}{N} \sum_{(S^{st}),(S^{at})} \mathbb{I}(AUL(S^{st}) > AUL(S^{at}))$$

This is the percentage of stereotypical (S^{st}) test sentences preferred over anti-stereotypical (S^{at}) ones. $\mathbb{I}$ is the indicator function, which returns 1 if its argument is True and 0 otherwise. The ideal score for a model is 50%.

3 Experimental setup

3.1 Models

In my experiment, I used the base and large versions of three different pretrained language models: BERT (Devlin et al. (2019)), RoBERTa (Liu et al. (2019)), and XLNet (Yang et al. (2019)) I use the cased version of each model. I analyze three models to ensure that any trends I might find in terms of model size are not unique to a single architecture. The sizes of the models are shown in Table 1.

	BERT	RoBERTa	XLNet
Base	110M	125M	110M
Large	340M	355M	340M

Table 1: Size of the used language models, in total number of parameters

3.2 Bias Measures

3.2.1 SEAT

May et al. (2019) use SEAT with a number of different tests (sets of target and attribute words) to investigate different biases. The tests I used are those that focus on gender bias, which are listed in Table 2. The sets of words in these tests were also used in Caliskan et al. (2017) and come from previous psychological and sociological studies. The scores I report here are the means of the effect sizes of these six tests. May et al. (2019) also introduce tests for the intersectional *Angry Black Woman* stereotype and for *double binds* (contradictory or unsatisfiable expectations of femininity and masculinity). These are also biases that have a gendered element. I have opted not to include them for this project because these are more intersecting and complex biases and the other measures I will use to not consider these specific intersectionalities.

Test Name	Target Concepts	Attributes
Test 6	Male Names/Female Names	Career/Family
Test 6b	Male Terms/Female Terms	Career/Family
Test 7	Math/Arts	Male Terms/Female Terms
Test 7b	Math/Arts	Male Names/Female Names
Test 8	Science/Arts	Male Terms/Female Terms
Test 8b	Science/Arts	Male Names/Female Names

Table 2: The Target concepts and attributes for the SEAT tests focused on gender bias

3.2.2 LPBS

Like SEAT, LPBS tests use two sets of target words and two sets of attribute words. They also use the tests from Caliskan et al. (2017). However, if inserted into the first slot of the templates, many of the words from the Caliskan tests would render the sentences grammatically incorrect. Therefore, Kurita et al. (2019) fixed the first slots to pronouns or common indicators for either gender and slotted the Career/Family, Math/Arts, and Science/Arts words from the Caliskan test into the second slot. I implemented these tests in the same manner, resulting in the template sentences found in 3. The scores I report here are means of the effect sizes of the three tests (tests 6, 7 and 8).

Target words	Template
he / she	T likes A
boys, men / girls, women	T like A
he / she	T is interested in A

Table 3: The Target words and templates for the LPBS tests focused on gender bias. (T: Target, A: Attribute word from Career/Family, Math/Arts, and Science/Arts sets)

3.2.3 CAT

The score associated with CAT is the *icat* score. This is not solely a bias measure, as this score combines a language modeling score with the stereotype score. Therefore, I opted not to use this score. Instead, I use the *stereotype score* (*ss*) on its own as the bias score for CAT, as this is a better basis for comparison to the other bias measures.

The StereoSet dataset that is used for CAT contains tests for four different types of bias. Since this project is focused on gender bias, I only considered the tests that concern gender bias. I also limited my implementation to the Intrasentence tests, since the other bias measures are also limited to sentence-level tests.

For the two scores discussed previously, the neutral score is 0. That is to say, if a model would not stereotype nor anti-stereotype, it would get a score of 0. For the CAT *ss* the neutral score is 50. To account for this disparity and to make comparison easier, I have defined the *CAT bias score* as being $ss - 50$. This is the score I report here.

In addition to the *CAT bias score*, I also make use of the CAT *language modeling score* on its own. This is the score I use to examine the second part of my research question.

3.2.4 CrowS-Pairs

The score associated with CrowS-Pairs is similar to the CAT stereotype score in that its neutral value is 50. So, like with the CAT bias score, the score I report for CrowS-Pairs is $score - 50$.

The CrowS-Pairs dataset accounts for nine types of bias. Again, I only selected the tests that look for gender bias.

3.2.5 AUL

As with the previous two scores, the neutral score for AUL is 50. So the scores I report here are $AUL - 50$.

Kaneko and Bollegala (2022) use the datasets from Nadeem et al. (2021) and Nangia et al. (2020) to calculate their score. I have done the same, only considering the tests for gender bias. This results in two different scores for each model. AUL *cp* is the AUL score when using the CrowS-Pairs dataset and AUL *ss* is the AUL score when using the StereoSet dataset.

4 Results

	SEAT	LPBS	CAT bias score	CAT lms	CrowS-Pairs	AUL cp	AUL ss
$BERT_{base}$	0.477	0.694	11.482	82.503	7.630	3.050	-0.200
$BERT_{large}$	0.138	0.673	14.036	83.102	9.160	1.530	3.330
$RoBERTa_{base}$	0.850	0.381	3.657	71.750	4.960	6.870	14.710
$RoBERTa_{large}$	0.449	0.370	6.872	75.816	8.400	8.020	12.750
$XLNet_{base}$	0.199	0.278	-3.459	69.385	0.380	-0.380	4.900
$XLNet_{large}$	-0.042	-0.051	3.994	74.159	-0.380	-0.380	5.690

Table 4: Experiment results. Recall that the results shown for CAT bias score, CrowS-Pairs and AUL are not the usual scores for these measures, but have been lowered by 50 points to share a neutrality point with the other measures. This makes it easier to notice when a measure deems the model to be anti-stereotyping

Table 4 shows the results of the experiment I performed. Figure 1 shows trend lines for increasing the model sizes for each measure. Figure 2 shows how increasing the size of the model increases the size of the language modeling score. Figure 3 shows a correlation matrix of the different measures with the addition of correlation with model sizes and correlation with the CAT language modeling score from Nadeem et al. (2021)

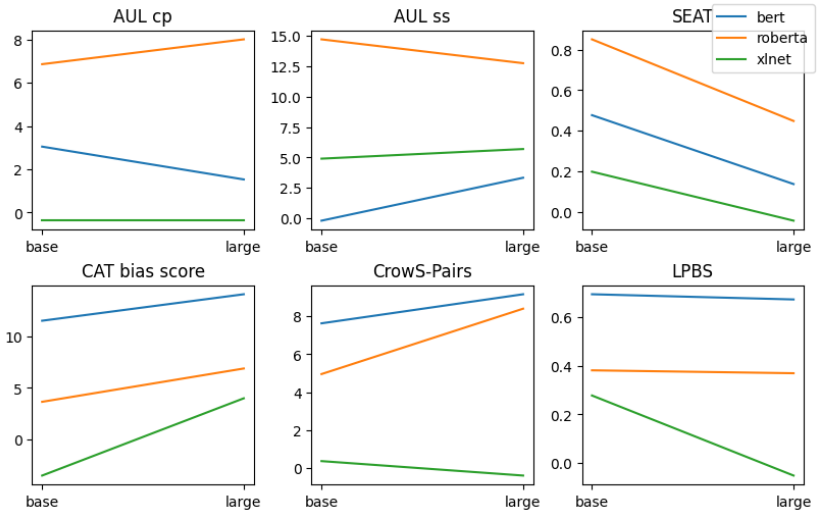

Figure 1: Shows the impact of increasing the size of the models for each bias measure. The left hand side of each sub-graph shows the base version of the models, the right hand side shows the large version.

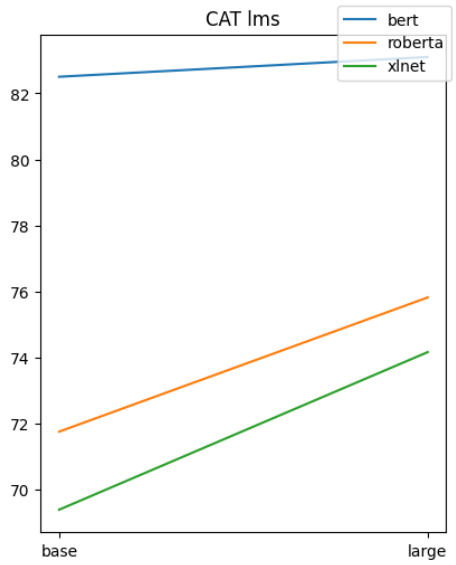

Figure 2: Shows how increasing the size of the model impacts the language modeling ability according to the language modeling score of Nadeem et al. (2021)

.

5 Discussion of Results

5.1 Increase in Bias with Increase in Model Size?

It is clear from Figure 1 the an increase in model size does not necessarily lead to an increase in bias. It was to be expected that their would not be a universal pattern from the findings of Delobelle et al. (2022). The CAT bias score does clearly increase as model size increases, regardless of the model architecture, as stated in Nadeem et al. (2021). None of the other measures show a universal increase, however. In fact, for the SEAT measure the opposite is true. There, the measured bias decreases as model size increases. For the other four measures there is no pattern of universal increase or decrease.

There is also no clear pattern for any single model. There are no models that increase in bias for the majority of the measures, or any that decrease.

We cannot make any definitive statement about an increase or decrease of gendered bias in these language models from these results. We can however say that the increase in model size does not lead to an increase in bias across measures, to answer part one of my research question.

5.2 Increase in Language Modeling Ability with Increase in Model Size?

Figure 2 shows that there is a tendency among all tested models for the language modeling score of Nadeem et al. (2021) to increase as the model size increases. I would caution against claiming that the language modeling ability increases as the model size increases from this data alone, though. That claim can not be

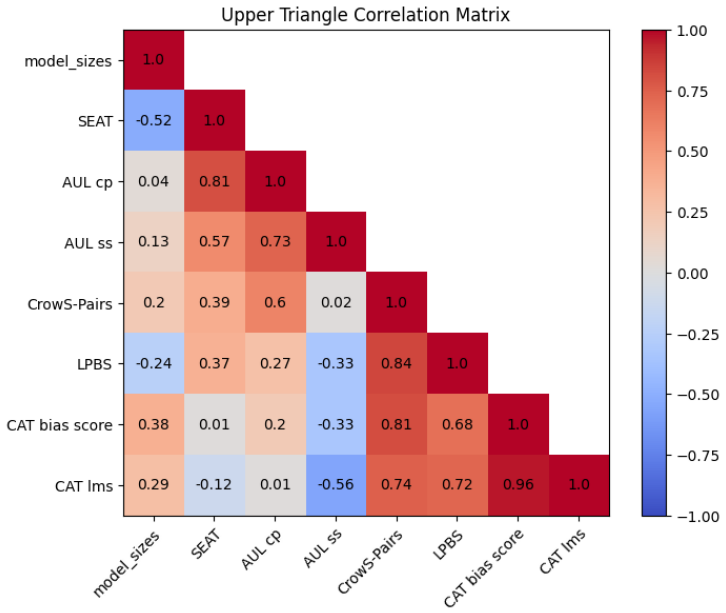

Figure 3: Correlations between different measures. The first column shows correlations with model size. The bottom row shows correlations with the language modeling score.

proven with six data point from a single measure. However, previous has shown the same trend.

Interesting to note in these results is that if we look at the correlation matrix in Figure 3, the model size and language modeling score (CAT lms) seem not to be strongly correlated, even though we see a clear increase in lms as model size increases in Figure 2. This is influenced by the fact that BERT has a smaller number of features than RoBERTa in both the base size and large size, but has a significantly larger lms score for both sizes.

5.3 Increase in Bias with Increase in Language Modeling Ability?

The second part of my research question concerned the idea that bias and language modeling ability are correlated, as they are in Nadeem et al. (2021). Looking at bottom row in Figure 3, we can see that the CAT bias score and the language modeling score (CAT lms) are indeed strongly correlated, but this is not true for all bias measures. The bias measure from CrowS-Pairs and LPBS are also correlated to the CAT lms. The other measures are not correlated to the CAT lms at all. SEAT and AUL ss are in fact negatively correlated with CAT lms.

Again, this non-correlation is to be expected given that Delobelle et al. (2022) found that many metrics are not compatible and do not correlate.

To answer the research question, we can say with certainty that an increase in language modeling ability (as measured using the CAT lms) does not necessarily

mean an increase in bias for all bias measures.

6 Code

The code for this project is available at
`https://github.com/L-Thys/research-project-1`

7 Conclusion

This study explored the relationship between model size and gender bias in contextualized language models, using multiple architectures (BERT, RoBERTa, and XLNet) in different sizes and five different bias detection methods. The results indicate that increasing model size does not consistently lead to an increase in bias across all measures. While the CAT bias score aligns with previous findings that larger models exhibit greater stereotypical bias, other metrics, such as SEAT, show the opposite trend, while some reveal no clear pattern. This discrepancy underscores the lack of alignment between different bias measurement techniques, as highlighted by previous research.

Furthermore, while language modeling ability generally improves with increased model size, its correlation with bias varies depending on the bias metric used. Some measures show a positive correlation, supporting prior findings from Nadeem et al. (2021), whereas others do not, suggesting that the relationship between bias and model performance is more complex than initially assumed.

These findings further emphasize the need for a more standardized and comprehensive approach to bias measurement in contextualized language models.

References

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. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Delobelle, P., Tokpo, E., Calders, T., and Berendt, B. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Kaneko, M. and Bollegala, D. (2022). Unmasking the mask – evaluating social biases in masked language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11954–11962.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In Costa-jussà, M. R., Hardmeier, C., Radford, W., and Webster, K., editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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 Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wang, A. and Cho, K. (2019). BERT has a mouth, and it must speak: BERT as a Markov random field language model. In Bosselut, A., Celikyilmaz, A., Ghazvininejad, M., Iyer, S., Khandelwal, U., Rashkin, H., and Wolf, T., editors, *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.