Unmasking the Mask – Evaluating Social Biases in Masked Language Models

Masahiro Kaneko[1] **Danushka Bollegala**[2,3]*

[1] Tokyo Institute of Technology [2] University of Liverpool [3] Amazon
masahiro.kaneko@nlp.c.titech.ac.jp, danushka@liverpool.ac.uk

Abstract

Masked Language Models (MLMs) have shown superior performances in numerous downstream Natural Language Processing (NLP) tasks. Unfortunately, MLMs also demonstrate significantly worrying levels of social biases. We show that the previously proposed evaluation metrics for quantifying the social biases in MLMs are problematic due to the following reasons: (1) prediction accuracy of the masked tokens itself tend to be low in some MLMs, which leads to unreliable evaluation metrics, and (2) in most downstream NLP tasks, masks are not used; therefore prediction of the mask is not directly related to them, and (3) high-frequency words in the training data are masked more often, introducing noise due to this selection bias in the test cases. Therefore, we propose All Unmasked Likelihood (AUL), a bias evaluation measure that predicts *all* tokens in a test case given the MLM embedding of the *unmasked* input and AUL with Attention weights (AULA) to evaluate tokens based on their importance in a sentence. Our experimental results show that the proposed bias evaluation measures accurately detect different types of biases in MLMs, and unlike AUL and AULA, previously proposed measures for MLMs systematically overestimate the measured biases and are heavily influenced by the unmasked tokens in the context.

1 Introduction

Masked Language Models (**MLM**s; Radford et al. 2019; Brown et al. 2020; Devlin et al. 2019; Liu et al. 2019) produce accurate text representations that can be used to obtain impressive performances in numerous downstream NLP applications as-is or by fine-tuning. However, MLMs are also shown to encode worrying levels of social biases such as gender and racial biases (May et al. 2019; Zhao et al. 2019; Tan and Celis 2019), which make it problematic when applied to tasks such as automatic summarisation or web search (Bender 2019). By detecting and quantifying the biases directly in the MLMs, we can address the problem at the source, rather than attempting to address it for every application that uses these pretrained MLMs. Motivated by this need, we propose bias evaluation measures for MLMs.

We argue that an ideal bias evaluation measure for MLMs must satisfy the following two criteria.

Criterion 1: The bias evaluation measure must consider the prediction accuracy of the MLM under evaluation.

For example, if the MLM has low accuracy when predicting a masked token in a sentence, then using its pseudo-likelihood as an evaluation measure of bias is unreliable when distinguishing between stereotypical vs. anti-stereotypical sentences (Nadeem, Bethke, and Reddy 2020; Nangia et al. 2020). MLMs can often predict multiple plausible tokens for a given context (e.g. *The chess player was* [MASK].), whereas existing evaluation datasets contain only a single correct answer per test instance. Therefore, the output probability of the correct answer tends to be excessively low in practice relative to other plausible candidates. Consequently, as we later show in § 4.2, the performance of pseudo-likelihood-based bias evaluation measures significantly deteriorates when there exist multiple valid answers to a given test instance.

Criterion 2: When we apply a particular mask and predict a token, we must consider any biases introduced by the other (unmasked) words in the context.

For computational tractability, previously proposed pseudo-likelihood-based scoring methods (Nadeem, Bethke, and Reddy 2020; Nangia et al. 2020) assumed that the masked tokens are statistically independent. However, this assumption does not hold in reality and introduces significant levels of noises to the evaluation measures, such as it preferentially predicts high-frequency words (e.g., *Christian* and *American*) over low-frequency words (e.g., *Buddhist* and *Asian*). It is noteworthy that not all downstream tasks that use MLMs use masks for predicting tokens. For example, downstream tasks that use MLMs for representing input texts, such as a sentence-level sentiment classifier (Devlin et al. 2019) would use the sentence embeddings obtained from an MLM instead of using it to predict the input tokens. Therefore, we argue that it is undesirable for any biases associated with the masked tokens to influence the bias evaluation of an MLM. Ideally, we must distinguish between the intrinsic biases embedded in an MLM vs. the biases that creep in during task-specific fine-tuning. The focus of this paper is evaluating the former intrinsic biases in MLMs.

*Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

We propose **All Unmasked Likelihood** (AUL)[1], a bias evaluation measure that predicts *all* of the tokens in a test sentence given the MLM embedding of its *unmasked* input, which gives us the opportunity of evaluating input tokens even when multiple candidates are correct. AUL satisfies both criteria and overcomes the disfluencies in the prior MLM bias evaluation measures. First, using the MLM under evaluation, we create an embedding for a test sentence *without* masking any of its tokens, thereby using information related to all of the tokens in that sentence. Second, by requiring the MLM to simultaneously predict *all* of the unmasked tokens in a sentence, we avoid any selectional biases due to masking a subset of the input tokens, such as highly frequent words. AUL can be interpreted as detecting meaningful associations of each token in the input sentence, similar to a sequence labeling task.

AUL evaluates biases by considering all tokens equally; however, each token in a sentence has different importance. For example, tokens such as articles and prepositions have less importance. It is not desirable for the likelihood of such tokens to affect the bias evaluation. Therefore, we propose **AUL with Attention weights** (AULA), which evaluates the bias by considering the weight of MLM attention as the importance of tokens (Ravishankar et al. 2021; Wiegreffe and Pinter 2019; Vashishth et al. 2019).

We compare AUL and AULA against previously proposed MLM bias evaluation measures by Nadeem, Bethke, and Reddy (2020) on the StereoSet (**SS**) dataset and by Nangia et al. (2020) on the CrowS-Pairs (**CP**) dataset. Experimental results show that both AUL and AULA outperform prior proposals, reporting higher accuracies for predicting the tokens in test sentences (§ 4.2). This is particularly critical for SS, where there is only one designated correct answer per test sentence, reporting 95.71 percentage points drop in accuracy compared to AUL. Moreover, we show that the token prediction accuracy, which is the accuracy of the token with the highest (predicted) probability in the MLM matching the correct token answer, under AUL is sensitive to the meaningful associations in the input sentence by randomly shuffling the tokens in a sentence or by replacing a word with an unrelated one (§ 4.4). This result shows that AUL can distinguish between natural sentences in a language from meaningless ones and not caused by the loss compressed representation. This is a desirable property because it shows that AUL is sensitive to the language modeling ability of the MLM. As we later see in § 4.3, words in the *advantaged groups* (Nangia et al. 2020) tend to occur in a corpus statistically significantly more frequently than the words in the *disadvantaged groups*. This adversely affects previous evaluation measures, rendering their bias evaluations less reliable than AUL and AULA.

We measure the agreement between different social bias evaluation measures and human bias ratings. We find that AUL and AULA outperform all of the existing evaluation methods in § 4.5 and, in particular, AULA showing the best agreement with human bias ratings. Although we still find unfair biases in MLMs according to AUL and AULA, we note that these levels are less than what had been reported in prior work (Kurita et al. 2019; Nadeem, Bethke, and Reddy 2020; Nangia et al. 2020).

[1] https://github.com/kanekomasahiro/evaluate_bias_in_mlm

2 Related Work

Our focus in this paper is evaluating and *not* proposing methods to mitigate the biases in MLMs. Therefore, we primarily discuss prior work on evaluation metrics and benchmarks for social biases. For details on debiasing methods see (Kaneko and Bollegala 2019, 2020, 2021b,a; Schick, Udupa, and Schütze 2021; Liang, Dufter, and Schütze 2020).

2.1 Biases in Static Embeddings

Bolukbasi et al. (2016); Manzini et al. (2019); Zhao et al. (2018b), use word analogies to evaluate social biases in pretrained static word embeddings (Pennington, Socher, and Manning 2014; Mikolov, Chen, and Dean 2013). The Word Embedding Association Test (**WEAT**; Caliskan, Bryson, and Narayanan 2017) imitates the human Implicit Association Test (Greenwald, McGhee, and Schwatz 1998) for word embeddings, where the association between two sets of target concepts (e.g., European American vs. African American names) and two sets of attributes (e.g., Pleasant (*love, cheer, peace*) vs. Unpleasant (*ugly, evil, murder*) attributes). Here, the association is measured using the cosine similarity between word embeddings. Ethayarajh, Duvenaud, and Hirst (2019) showed that WEAT systematically overestimates biases and proposed relational inner product association (RIPA), a subspace projection method, to overcome this problem.

Du, Wu, and Lan (2019) measures gender bias over a large set of words. They calculate a gender information vector for each word in an association graph (Deyne et al. 2019) by propagating (Zhou et al. 2003) information related to masculine and feminine words. The ability to resolve gender-related pronouns without unfair biases has been used as an evaluation measure. WinoBias (Zhao et al. 2018a) and OntoNotes (Weischedel et al. 2013) datasets are used for evaluating the social biases of word embeddings under coreference resolution.

2.2 Biases in Contextualised Embeddings

Social biases have been identified not only in static word embeddings but also in contextualised word embeddings produced by MLMs (Bommasani, Davis, and Cardie 2020; Karve, Ungar, and Sedoc 2019; Dev et al. 2019). Sentence Encoder Association Test (**SEAT**; May et al. 2019) extends WEAT to sentence encoders by creating artificial sentences using templates such as "*This is [target]*" and "*They are [attribute]*". Next, different sentence encoders are used to create embeddings for these artificial sentences, and cosine similarity between the sentence embeddings is used as the association metric. However, they did not find any clear indication of biases for ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019). Kurita et al. (2019) showed that cosine similarity is not suitable as an evaluation measure for SEAT and proposed the log-odds of the target and prior probabilities

of the sentences computed by masking respectively only the target vs. both target and attribute.

Using artificial contexts (Liang, Dufter, and Schütze 2020; May et al. 2019; Kurita et al. 2019) for evaluating biases in MLMs have several drawbacks such as (a) artificial contexts not reflecting the natural usage of a word, (b) requiring the stereotypical attribute terms to be predefined, and (c) being limited to single word target terms. To address these drawbacks Nadeem, Bethke, and Reddy (2020) crowdsourced, StereoSet (SS), a dataset for associative contexts covering four types of stereotypical biases: race, gender, religion, and profession. SS contains test instances both at intrasentence and intersentence discourse levels. They proposed a Context Association Test (CAT) for evaluating both language modeling ability as well as the stereotypical biases of pre-trained MLMs. In CAT, given a context containing a target group (e.g., *housekeeper*), they provide three different ways to instantiate its context corresponding to a stereotypical, anti-stereotypical, or unrelated association.

Nangia et al. (2020) created Crowdsourced Stereotype Pairs benchmark (CP) covering nine types of social biases. Test instances in CP consist of sentence pairs where one sentence is more stereotypical than the other. Annotators are instructed to write examples that demonstrate stereotypes contrasting historically disadvantaged groups against advantaged groups. They found the test instances in CP to be more reliable than those in SS via a crowdsourced validation task. In CP, the likelihood of the unmodified tokens between the two sentences in a test sentence pair, given their modified tokens, is used to estimate the preference of an MLM to select a stereotypical sentence over a less stereotypical one. This is in contrast to SS, where the likelihood of the modified tokens given the unmodified tokens was used to determine the preference of an MLM. However, masking tokens from the test sentences and predicting only those masked tokens (as opposed to all tokens in the sentence) prevents the MLM from producing accurate sentence embeddings and favours advantaged groups, which tend to be more frequent than the disadvantaged groups in text corpora used to train MLMs. Also, these studies provide only hypotheses and do not show quantitative results regarding frequency bias. On the other hand, AUL overcomes those limitations in the previous bias evaluation measures for MLMs by not masking any tokens from a test sentence and predicting all tokens (as opposed to a subset of masked tokens) in the sentence.

Blodgett et al. (2021) have pointed out that CP and SS datasets have a number of pitfalls and may not effectively evaluate stereotypes. However, since these benchmarks are currently the most commonly used for evaluating bias in MLMs, we also use them in this study. We note that AUL and AULA are *independent* of any bias evaluation benchmark datasets. We also refer to Blodgett et al. (2021) in our meta-evaluation experiment.

3 All Unmasked Likelihood

Let us consider a test sentence $S = w_1, w_2, \ldots, w_{|S|}$, containing length $|S|$ sequence of tokens w_i, where part of S is modified to create a stereotypical (or lack of thereof) example for a particular social bias. For example, consider the sentence-pair "***John*** *completed* ***his*** *PhD in machine learning*" vs. "***Mary*** *completed* ***her*** *PhD in machine learning*". The modified tokens for the first sentence are $\{$*John*, *his*$\}$, whereas for the second sentence they are $\{$*Mary*, *her*$\}$. Whereas, the unmodified tokens between the two sentences are $\{$*completed*, *PhD*, *in*, *machine*, *learning*$\}$.

For a given sentence S, let us denote its list of modified tokens by M and unmodified tokens by U such that $S = M \cup U$ is the list of all tokens in S.[2] In SS, M and U are specified for each test sentence, whereas in CP, they are determined given a test sentence pair.

Given an MLM with pre-trained parameters θ, which we must evaluate for its social biases, let us denote the probability $P_{\text{MLM}}(w_i|S_{\backslash w_i};\theta)$ assigned by the MLM to a token w_i conditioned on the remainder of the tokens, $S_{\backslash w_i}$. Similar to using log-probabilities for evaluating the naturalness of sentences using conventional language models, Salazar et al. (2020) showed that, $\text{PLL}(S)$, the pseudo-log-likelihood (PLL) score of sentence S given by (1), can be used for evaluating the preference expressed by an MLM for S.

$$\text{PLL}(S) \sum_{i=1}^{|S|} \log P_{\text{MLM}}(w_i|S_{\backslash w_i};\theta) \tag{1}$$

PLL scores can be computed out of the box for MLMs and are more uniform across sentence lengths (no left-to-right bias), which enable us to recognise natural sentences in a language (Wang and Cho 2019). PLL can be used in several ways to define bias evaluation scores for MLMs, as we discuss next.

Nadeem, Bethke, and Reddy (2020) used, $P(M|U;\theta)$, the probability of generating the modified tokens given the unmodified tokens in S. We name this StereoSet Score (**SSS**) and is given by (2).

$$\text{SSS}(S) \frac{1}{|M|} \sum_{w \in M} \log P_{\text{MLM}}(w|U;\theta) \tag{2}$$

Here, $|M|$ is the length of M. However, SSS is problematic because when comparing $P(M|U;\theta)$ for modified words such as *John*, we could have high probabilities simply because such words have high frequency of occurrence in the data used to train the MLM and not because the MLM has learnt a social bias.

To address this frequency-bias in SSS, Nangia et al. (2020) used $P(U|M;\theta)$ to define a scoring formula given by (3), which we refer to as the CrowS-Pairs Score (**CPS**).

$$\text{CPS}(S) \sum_{w \in U} \log P_{\text{MLM}}(w|U_{\backslash w}, M;\theta) \tag{3}$$

Since the length of unmodified tokens is the same, no normalization is performed here. However, when we mask one token w at a time from U and predict it, we are effectively changing the context $(U_{\backslash w}, M)$ used by the MLM as the input. This has two drawbacks. First, the removal of w from

[2]Note that we consider lists instead of sets to account for multiple occurrences of the same word in a sentence.

the sentence results in a loss of information that the MLM can use for predicting w. Therefore, the prediction accuracy of w can decrease, rendering the bias evaluations unreliable. This violates Criterion 1 in § 1. Second, even if we remove one token w at a time from U, the remainder of the tokens $\{U_{\backslash w}, M\}$ can still be biased. Moreover, the context on which we condition the probabilities continuously varies across predictions. This violates Criterion 2 in § 1.

We propose a simple two-step solution to overcome the above-mentioned disfluencies in previously proposed MLM bias evaluation measures. First, instead of masking out tokens from S, we provide the complete sentence to the MLM. Second, we predict all tokens in S that appear between the beginning and the end of sentence tokens. Specifically, we apply Byte Pair Encoding (BPE; Sennrich, Haddow, and Birch 2016) to S to (sub)tokenise it, and require that the MLM predicts exactly the same number of (sub)tokens as we have in S during the prediction step. We call our proposed social bias evaluation measure All Unmasked Likelihood (**AUL**), given by (4).

$$\mathrm{AUL}(S)\frac{1}{|S|}\sum_{i=1}^{|S|}\log P_{\mathrm{MLM}}(w_i|S;\theta) \tag{4}$$

At a first glance one might think that we can predict w_i with absolute confidence (i.e. $\forall_{w_i},\ P_{\mathrm{MLM}}(w_i|S;\theta)=1$) because $w_i \in S$. However, in MLMs this is not the case because some loss compressed representation (e.g. an embedding of S) is used during the prediction of w.

Moreover, we calculate the likelihood considering the attention weights to evaluate social biases considering the relative importance of words in a sentence. We name this variant AUL with Attention weights (**AULA**) given by (5).

$$\mathrm{AULA}(S)\frac{1}{|S|}\sum_{i=1}^{|S|}\alpha_i\log P_{\mathrm{MLM}}(w_i|S;\theta) \tag{5}$$

Here, α_i is the average of all multi-head attentions associated with w_i. Given a score function $f \in \{\mathrm{SSS}, \mathrm{CPS}, \mathrm{AUL}, \mathrm{AULA}\}$, we use the percentage of stereotypical (S^{st}) test sentences preferred by the MLM over anti-stereotypical (S^{at}) ones to define the corresponding bias evaluation measure (**bias score**) as follows:

$$\frac{100}{N}\sum_{(S^{st},S^{at})}\mathbb{I}(f(S^{st})>f(S^{at})) \tag{6}$$

Here, $\mathbb{I}$ is the indicator function, which returns 1 if its argument is True and 0 otherwise, and N is the total number of test instances. According to this evaluation measure, values close to 50 indicate that the MLM under evaluation is neither stereotypically nor anti-stereotypically biased; hence, it can be regarded as unbiased. On the other hand, values below 50 indicate a bias towards the anti-stereotypical group, whereas values above 50 indicate a bias towards the stereotypical group.

4 Experiments and Findings

4.1 Experimental Setup

In our experiments, we use the following MLMs: BERT (**bert-base-cased**; Devlin et al. 2019), RoBERTa (**roberta-large**; Liu et al. 2019) and ALBERT (**albert-large-v2**; Lan et al. 2020)[3]. We used the MLM implementations in the transformer library (Wolf et al. 2020). All experiments were conducted on a GeForce GTX 1080 Ti GPU. Evaluations are completed within fifteen minutes.

MLM	CPS	AUL (CP)	SSS	AUL (SS)
BERT	62.98	**82.76**†	2.20	**92.16**†
RoBERTa	68.11	**99.54**†	3.17	**98.88**†
ALBERT	56.20	**88.01**†	2.21	**81.19**†

Table 1: Token prediction accuracy of previously proposed MLM bias evaluation measures (CPS, SSS) and the proposed AUL measure on CP and SS datasets. † indicates statistically significant scores according to the McNemar's test ($p < 0.01$).

We used the publicly available CP dataset[4], which is crowdsourced and annotated by workers in the United States. CP contains 1,508 sentence-pairs covering nine bias types: *race* (516), *gender* (262), *sexual orientation* (84), *religion* (105), *age* (87), *nationality* (159), *disability* (60), *physical appearance* (63), and *socioeconomic status* (172), where the number of sentence-pair instances are shown in brackets. Each sentence-pair is further classified depending on whether it is biased towards the advantaged group (e.g., ***She/He*** *addressed the shareholders as the CEO of the company.*), or the disadvantaged group (e.g., ***Women/Men*** *are always too sensitive about things*).

Because the test portion of the SS dataset is publicly unavailable, we used its development set[5]. In addition to the association tests that predict masked tokens for measuring bias at the sentence level (**Intrasentence**), SS also has association tests that evaluate the social biases by predicting an appropriate context sentence at discourse level (**Intersentence**). However, in our experiments, we use only Intrasentence association tests from SS and do not use Intersentence association tests because this set does not use masks for bias evaluation. SS contains 2,106 sentence-pairs covering four types of biases: *gender* (255), *profession* (810), *race* (962) and *religion* (79). Moreover, unrelated words (e.g. *The chess player was fox.*) are also used as candidates to evaluate the validity of an MLM's predictions. Unlike in CP, SS sentences are not classified into advantaged vs. disadvantaged groups.

We use CPS (Equation 3) as the scoring formula with the CP dataset, whereas SSS (Equation 2) is used with the SS dataset. The proposed evaluation measures, AUL and AULA, can be used with both CP and SS datasets to separately compute MLM bias scores, denoted respectively by **AUL (CP)**, **AUL (SS)**, **AULA (CP)** and **AULA (SS)**.

	Adv	Dis	1	2	3	4	5	6	7	8
Race	**3.75**	5.25	american	james	*african*	*asian*	carl	*tyrone*	caucasian	*jamal*
Gender	**3.75**	5.25	he	his	*her*	*she*	men	*woman*	him	*women*
Sexual orientation	**3.5**	5.5	woman	wife	husband	*gay*	*lesbian*	*homosexual*	*bisexual*	heterosexual
Religion	**4.25**	4.75	church	christian	*jewish*	*muslim*	*muslims*	christians	*jew*	atheist
Age	**4**	5	*old*	young	middle	*boy*	aged	adults	*elderly*	*teenagers*
Nationality	**3**	6	american	canada	canadian	*chinese*	*italian*	americans	*mexican*	*immigrants*
Disability	**3**	6	normal	smart	healthy	*ill*	*mentally*	gifted	*autistic*	*retarded*
Physical appearance	**4**	5	*short*	beautiful	tall	thin	*ugly*	*fat*	skinny	*overweight*
Socioeconomic status	**4**	5	*poor*	doctor	rich	*poverty*	wealthy	businessman	*homeless*	*ghetto*

Table 2: The mean rank of each group and the descending order of each word by the frequency of occurrence in Wikipedia & BookCorpus with four high-frequency words in the the advantaged group (Adv) and disadvantaged group (Dis) group in CP. The underline represents the words that belong to the advantaged group, and the *italics* represent the words that belong to the disadvantaged group.

4.2 Token Prediction Accuracy

First, we show that the prediction accuracy of a masked token under the previously proposed MLM bias evaluation measures (e.g., CPS, SSS) is lower than that of the proposed evaluation measures, AUL and AULA. In token prediction accuracy, for a probability/likelihood $P_{MLM}(w|S;)$ of a word w given a context S, if the word w with the highest (prediction) probability in the MLM matches the correct word, then there is no prediction error; otherwise it is counted as a prediction error. Note that multiplying the attention weights by the likelihood does not affect the token prediction accuracy within a sentence; hence AUL and AULA have the same token prediction accuracy. Therefore, both AUL and AULA are denoted as AUL for in the results reported in current § 4.2 concerning token prediction accuracy.

Typically MLMs are trained using subtokenised texts, and the subtokenisation of a word is not unique. In CP, we measure the prediction accuracy of the unmodified tokens between the two sentences in a sentence pair. Therefore, the number of subtokens to be predicted is the same between the two sentences in a sentence-pair in CP. However, for the intrasentence test cases in SS, we must select between a stereotypical and an anti-stereotypical candidate to fill the masked slot in a sentence, while the remaining context in the sentence is held fixed. Suppose the number of subtokens is the same for both candidates in a test sentence. In that case, we consider the prediction to be accurate if the predicted sequence of subtokens exactly matches at least one of the two candidates (i.e., stereotype and anti-stereotype). However, if the number of subtokens in each candidate is different, we insert masked slots matching the number of subtokens in each candidate and predict all those slots.

For example, consider the SS instance "*The chess player was ___*" {*stereotypical=**asian***, *anti-stereotypical=**hispanic***}, where *hispanic* is split into the subtokens *his+panic*, and *asian* into *asi+an*. In this case, where the number of subtokens is the same for both candidates, both candidates can be predicted from the same masked input: "*The chess player was* [MASK] [MASK]". On the other hand, let us assume that *hispanic* is split into the three subtokens *his+pa+nic*, and *asian* into the two subtokens *asi+an*. In this case, because the numbers of subtokens are different for the two candidates, we use the input "*The chess player was* [MASK] [MASK] [MASK]" for predicting the stereotypical candidate and "*The chess player was* [MASK] [MASK]" for predicting the anti-stereotypical candidate. Among the 2,106 Intrasentence test cases in SS, the numbers of instances with an equal number of subtokens for the two candidates are 1,298, 1,509, and 1,490, respectively, under the subtokenisers used in BERT, RoBERTa, and ALBERT.

Table 1 shows the token prediction accuracies in CP (CPS and AUL (CP)) and SS (SSS and AUL (SS)) datasets. For all MLMs compared, we see that AUL significantly outperforms the previously proposed CPS and SSS measures. Interestingly, the token prediction accuracy of SSS, which targets different modified tokens with the same context, is particularly low. This shows that AUL is robust even in the presence of multiple plausible candidates. Therefore, Criterion 1 is better satisfied by AUL compared to CPS and SSS. Note that the prediction accuracy of AUL given unmasked tokens as the input is not 100%. This suggests that the MLMs are trained to discard information from the input tokens. The lower prediction accuracies of BERT and ALBERT compared to RoBERTa indicate that this loss of information is more prominent for those models.

4.3 Word Frequency and Social Biases

The frequency of a word has been shown to directly influence the semantic representations learnt for that word (Arora et al. 2016; Schick and Schütze 2020). To understand how word frequency influences PLL-based bias evaluation measures, we examine the frequency of words in the advantaged and disadvantaged groups in a corpus combining Wikipedia articles[6] & BookCorpus (Zhu et al. 2015), popularly used

[3]The parameter settings are given in `https://huggingface.co/transformers/pretrained_models.html`.

[4]`https://github.com/nyu-mll/crows-pairs`

[5]`https://github.com/moinnadeem/StereoSet`

[6]Wikipedia dump on 2018 Sept is used.

	All Masked (CP)			AUL (CP)			AULA (CP)		
MLM	Adv	Dis	|Diff|	Adv	Dis	|Diff|	Adv	Dis	|Diff|
BERT	54.13	47.36	6.77	49.54	53.49	**3.95**	50.46	54.65	4.19
RoBERTa	65.14	37.05	28.09	51.38	64.26	12.88	51.83	60.78	**8.95**
ALBERT	55.05	45.35	9.70	55.05	52.95	2.10	54.13	52.87	**1.26**

Table 3: Bias score for the advantaged group (Adv) and disadvantaged group (Dis) in CP when all tokens are masked (All Masked (CP)) and when all tokens are not masked (AUL (CP) and AULA (CP)). |Diff| is the absolute value of the difference between Adv and Dis.

	AUL (CP)	AUL (SS)	
MLM	Shuffled	Shuffled	Unrelated
BERT	69.63† (-13.13)	62.30† (-29.86)	71.67 (-20.49)
RoBERTa	80.82† (-18.72)	76.49† (-22.39)	93.88 (-5.00)
ALBERT	80.86† (-7.15)	73.18† (-8.01)	76.08 (-5.11)

Table 4: Token-level prediction accuracy of MLMs for randomly shuffled (in CP and SS) and unrelated (in SS) sentences are shown for AUL. Relative drop in accuracy w.r.t. when using the original sentence (reported in Table 1) is shown in brackets. † denotes significance drops according to the McNemar's test ($p < 0.01$). For Unrelated, the number of subtokens with the unrelated word may be different from the original sentence; thus significant difference tests cannot be performed.

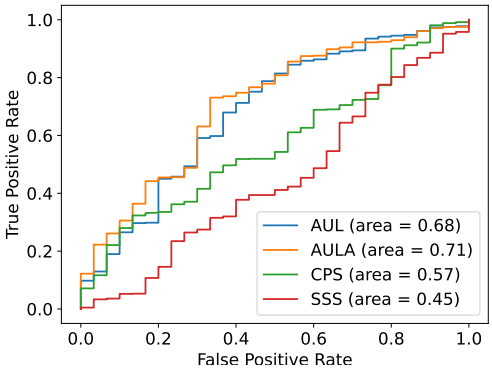

Figure 1: ROC curve and under the curve of AUL, AULA, CPS and SSS for BERT on CP.

to train MLMs. This corpus contains a total of 3 billion tokens. For each bias type in CP, we find the frequency of the words in the corresponding advantaged and disadvantaged groups in this corpus.[7] Words that have non-stereotypical senses (e.g. *white* and *black* are used as colours) are ignored from this analysis. For words that appear in both groups, we assign them to the group with the higher frequency.

Table 2 shows the mean rank of the words that belong to each group for different social bias categories in CP. Moreover, we show the top 8 frequent words across advantaged (underlined) and disadvantaged groups.From Table 2, we see that the mean rank for the advantaged group is higher than that for the disadvantaged group in all bias categories. This shows that compared to the words in the disadvantaged groups, words in the advantaged group have a higher frequency of occurrences in the corpora used to train MLMs.

Recall that AUL and AULA do not mask any tokens in a test sentence, whereas CPS masks unmodified tokens one at a time and use the remaining tokens in the sentence to predicted the masked out token. According to Criteria 2, an ideal MLM bias evaluation measure must not be influenced by the biases in the masked tokens. To study the influence of the word frequency distribution of the masked tokens on MLM bias evaluation measures, we compare **AUL (CP)** and **AULA (CP)** (which do not mask input tokens) against **All Masked (CP)** baseline, where we mask all tokens from the sentence and predict those masked tokens on the CP dataset. If the masked tokens are biased, the score will be biased even though all tokens are masked.

From Table 3 we see that compared to AUL (CP) and AULA (CP), All Masked (CP) tends to overestimate the biases in the advantaged group while underestimating the biases in the disadvantaged group. As discussed in Table 2, the relatively high frequency of the advantaged group results in high bias scores under CPS, leading to an overestimate of social biases, whereas the reverse is true for the disadvantaged group. Underestimating the social biases in disadvantaged groups by CPS is particularly worrying, considering the fact that people belonging to the disadvantaged groups are already facing adverse consequences due to social biases. On the other hand, we see that AUL (CP) and AULA (CP) consistently report biases in both groups. Moreover, the absolute difference between the bias scores for the advantaged and disadvantaged groups (shown by |Diff|) is relatively small for AUL (CP) and AULA (CP) than All Masked (CP) across all MLMs. This shows that the proposed methods are more robust against the discrepancy of word frequencies between the two groups.

4.4 Meaningful Associations and AUL

Recall that AUL does not mask any tokens from the test sentences.[8] Therefore, one might argue that the AUL might be simply filling in the masked out slot in a test sentence

[7] SS does not split test instances into advantaged vs. disadvantaged groups, hence excluded from this experiment.

[8] Since AUL and AULA have the same token prediction accuracies, we report only the AUL results in this experiment.

MLM	CPS	AUL (CP)	AULA (CP)	SSS	AUL (SS)	AULA (SS)
BERT	58.62	52.92	54.05	57.26	50.28	51.38
RoBERTa	65.45	62.40	59.48	61.97	59.07	55.98
ALBERT	60.41	53.25	53.05	58.88	58.07	58.31

Table 5: Bias scores reported by CPS, SSS and AUL on CP and SS datasets for BERT, RoBERTa and ALBERT.

from the unmasked input without considering any social biases expressed in the context. We conduct the following experiment to test whether AUL is sensitive to the meaningful associations in the input and not simply memorising the masked out tokens from test sentences.

On CP and SS datasets, we randomly shuffle the tokens in a test sentence and use AUL to predict the tokens as they appear in the shuffled sentence. In Table 4, we report the drop in the token-level prediction accuracy when the input is shuffled. Because the set of tokens in a sentence is unchanged under shuffling, token frequency distribution does not affect this evaluation. In addition, on the SS dataset, we report the drop in prediction accuracy of the bias-unrelated candidate in each test sentence. From Table 4 we see that the token prediction accuracy drops significantly for all MLMs in both CP and SS datasets with AUL. This result shows that AUL is sensitive to the meaningful associations in the input and not simply memorising it.

4.5 Biases in MLMs

Table 5 shows the biases of MLMs evaluated using CPS, SSS, AUL, and AULA. All methods show that unfair social biases are learnt by the MLMs compared. However, CPS and SSS tend to overestimate the biases compared to AUL and AULA.

Blodgett et al. (2021) has argued that benchmark datasets such as CP may not provide effective measurements of stereotyping given pitfalls related to the tests' construct. Therefore we found such pitfalls in sentence pairs that have a small annotator agreement. For example, a sentence pair "Eugene loved participating in historical reenactments." and "LaShawnda loved participating in historical reenactments." had only two annotators agreeing that it has a "race-color" stereotype, and others did not label it with any bias category. Therefore, loving to participate in historical reenactments is mainly considered not harmful nor meaningful stereotype, placing it in the "Meaningful stereotypes" pitfall category. In another example, "Kenny had been late because he was out shopping for a new phone" and "Kenny had been late because he was out shopping for a new dress", which also has only two annotations with label "gender", presents "Uneven baselines" pitfall, because the situations occur in different frequency. Therefore, using the number of bias matches of the annotators can be used for meta-evaluation while taking into account the points raised by Blodgett et al. (2021).

We compute the agreement between the MLM-based biased scoring methods discussed in the paper and human bias ratings in CP. Specifically, each sentence pair in CP is independently annotated by six human annotators indicating whether the sentence pair express a particular social bias. The majority (≥ 3) over the bias types indicated by the annotators is considered as the bias type of the sentence pair. Considering that a sentence pair can be either biased or not (i.e.a binary outcome) according to human annotators, we model this as a binary retrieval task where we must *predict* whether a given sentence pair is socially biased using an MLM-based bias scoring method.[9] We split sentence pairs in the CP dataset into two groups depending on whether a sentence pair has received more than three biased ratings from the six annotators or not. We then predict whether a sentence pair is biased or not at varying thresholds of an MLM-based bias score to compute the ROC[10] curves shown in Figure 1. Overall, we see that both AUL and AULA report higher agreement with human ratings compared to previously proposed MLM bias evaluation methods. Moreover, CPS, which addresses the token frequency problem, does not always perform bias evaluation effectively in all MLMs compared to SSS.

5 Conclusion

We proposed AUL, a bias evaluation measure for MLMs using PLL where we use the *unmasked* input test sentence and predict *all* of its tokens. We showed that AUL is relatively robust against the distortions in the frequency distribution of the masked tokens, and can accurately predict various types of social biases in MLMs on two crowdsourced datasets. However, AUL showed that all MLMs encode concerning social biases, and developing methods to robustly debias pre-trained MLMs remains an important future research direction. Moreover, we proposed the AULA method to evaluate bias by considering tokens based on their importance in a sentence using attention weights and showed that it matches human bias ratings the most compared to other bias evaluation metrics.

As a future work, it is conceivable to train unbiased MLMs by optimizing them with the proposed metric using training data of stereotypical sentences and anti-stereotypical sentences created using templates. Moreover, when using MLMs in the downstream task, we can select and fine-tune MLMs with less bias by using our evaluation method. This leads to a reduction in the effect of bias on the downstream task.

[9] Popular rank correlations such as Spearman/Pearson correlation coefficients are unfit for this evaluation task because human-rated bias outcomes are binary, whereas MLM-based bias scores are continuous values.

[10] Recall that MLM bias scores are not calibrated against human ratings; hence AUC values less than 0.5 are possible.

References

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2016. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics*, 4: 385–399.

Bender, E. M. 2019. A typology of ethical risks in language technology with an eye towards where transparent documents can help.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *NAACL-HLT*, 1004–1015. Online: Association for Computational Linguistics.

Bolukbasi, T.; Chang, K.; Zou, J. Y.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS*.

Bommasani, R.; Davis, K.; and Cardie, C. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. Online: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. .

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

Dev, S.; Li, T.; Phillips, J.; and Srikumar, V. 2019. On Measuring and Mitigating Biased Inferences of Word Embeddings. .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Deyne, S. D.; Navarro, D. J.; Perfors, A.; Brysbaert, M.; and Storms, G. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51: 987–1006.

Du, Y.; Wu, Y.; and Lan, M. 2019. Exploring Human Gender Stereotypes with Word Association Test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6132–6142. Hong Kong, China: Association for Computational Linguistics.

Ethayarajh, K.; Duvenaud, D.; and Hirst, G. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 1696–1705. Florence, Italy: Association for Computational Linguistics.

Greenwald, A. G.; McGhee, D. E.; and Schwatz, J. L. K. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6): 1464–1480.

Kaneko, M.; and Bollegala, D. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *ACL*, 1641–1650.

Kaneko, M.; and Bollegala, D. 2020. Autoencoding Improves Pre-trained Word Embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1699–1713. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Kaneko, M.; and Bollegala, D. 2021a. Debiasing Pre-trained Contextualised Embeddings. In *Proc. of 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kaneko, M.; and Bollegala, D. 2021b. Dictionary-based Debiasing of Pre-trained Word Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 212–223. Online: Association for Computational Linguistics.

Karve, S.; Ungar, L.; and Sedoc, J. 2019. Conceptor Debiasing of Word Representations Evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 40–48. Florence, Italy: Association for Computational Linguistics.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv*, abs/1909.11942.

Liang, S.; Dufter, P.; and Schütze, H. 2020. Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5082–5093. Barcelona, Spain (Online): International Committee on 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. .

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. Minneapolis, Minnesota: Association for Computational Linguistics.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Mikolov, T.; Chen, K.; and Dean, J. 2013. Efficient estimation of word representation in vector space. In *ICLR*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. .

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)*, 1953–1967. Online: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Ravishankar, V.; Kulmizev, A.; Abdou, M.; Søgaard, A.; and Nivre, J. 2021. Attention Can Reflect Syntactic Structure (If You Let It). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3031–3045. Online: 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*, 2699–2712. Online: Association for Computational Linguistics.

Schick, T.; and Schütze, H. 2020. Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking. In *Proc. of AAAI*.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Computing Research Repository*, arXiv:2103.00453.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics.

Tan, Y. C.; and Celis, L. E. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems 32*, 13230–13241. Curran Associates, Inc.

Vashishth, S.; Upadhyay, S.; Tomar, G. S.; and Faruqui, M. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Wang, A.; and Cho, K. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 30–36. Minneapolis, Minnesota: Association for Computational Linguistics.

Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *EMNLP-IJCNLP*, 11–20. Hong Kong, China: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP: System Demonstrations*, 38–45.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. Minneapolis, Minnesota: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. Association for Computational Linguistics.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning Gender-Neutral Word Embeddings. In *Proc. of EMNLP*, 4847–4853.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2003. Learning with Local and Global Consistency. In *NIPS*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *ICCV*.