What is Your Favorite Gender, MLM?
Gender Bias Evaluation in Multilingual Masked Language Models

Anonymous ACL submission

Abstract

Bias is a disproportionate prejudice in favor of one side against another. Due to the success of transformer-based Masked Language Models (MLMs) and their impact on many NLP tasks, a systematic evaluation of bias in these models is needed more than ever. While many studies have evaluated gender bias in English MLMs, only a few works have been conducted for the task in other languages. This paper proposes a multilingual approach to estimate gender bias in MLMs from 5 languages: Chinese, English, German, Portuguese, and Spanish. Unlike previous work, our approach does not depend on parallel corpora coupled with English to detect gender bias in other languages using multilingual lexicons. Moreover, a novel model-based method is presented to generate sentence pairs for more robust analysis of gender bias, compared to the traditional lexicon-based method. For each language, both the lexicon-based and model-based methods are applied to create two datasets respectively, which are used to evaluate gender bias in an MLM specifically trained for that language using one existing and 3 new scoring metrics. Our results show that the previous approach is data-sensitive and not stable as it does not remove contextual dependencies irrelevant to gender. In fact, the results often flip when different scoring metrics are used on the same dataset, suggesting that gender bias should be studied on a large dataset using multiple evaluation metrics for best practice.

1 Introduction

With the advent of attention (Vaswani et al., 2017) leading to BERT (Devlin et al., 2019), large language models are deployed to perform important tasks in society (Bender et al., 2021). In turn, there have been many studies in the field of Natural Language Processing (NLP) trying to improve upon the existing Masked Language Models (MLMs) (Tan and Bansal, 2019; Clark et al., 2020). MLMs are used not only for predicting the masked token but also successfully utilized in the field of Natural Language Understanding. While more complex models and bigger pre-trained data increase the accuracy of downstream NLP tasks, they also create room for bias.

There has been growing interest in detecting and reducing bias in the field of NLP due to the prevalence of language models in applied society (Bender et al., 2021). Detecting gender disparity has especially gained popularity in multiple domains: studies ranging from finding human-like bias from famous sentence encoders such as BERT (Kurita et al., 2019) to providing new benchmarks for evaluating gender bias in coreference resolution (Zhao et al., 2018a). All the previous work exposed that language technologies can produce undesirable bias (Blodgett et al., 2020).

Bolukbasi et al. (2016) uses word analogies to evaluate gender bias in pre-trained static word embeddings. Studies have tried to evaluate gender bias in contextualized word embeddings by providing word pairs that differentiate whether the words contain gender information or bias (Zhao et al., 2018b), or by having an occupation template and inspecting if the MLMs are more likely to predict the [Mask] associated with the occupation as he or she (Liang et al., 2020). Recent studies have come up with innovative methods for debiasing MLMs. Webster et al. (2020) debiases MLMs by re-balancing the evaluation corpus and switching bias attribute words within the data set. Bommasani et al. (2020) uses mathematical derivation to normalize sentence vectors from MLMs. Even though the gender bias of English-based MLMs has been looked into, gender bias in multi-lingual MLMs is underexplored.

In this paper, we first draw out the limitations of a previous work that evaluates the gender bias of MLMs in multiple languages. In §3, we present our enhanced methods that extract and make the pairs of sentences for gender bias evaluation. Then, we create a list of gender words in 5 different languages to extract sentences with gender words in §4. We probe why our proposed method is more consistent and retains more data from the original corpus compared to the previous works in §5 and §6, especially when the corpus is initially skewed.

We make two primary contributions in this paper. (1) We construct Multilingual Gender Lexicon (MGL), which does not depend on a parallel corpus, but can detect sentences with gendered words in English, German, Spanish, Portuguese, and Chinese. (2) We present two metrics and methods that only make meaningful comparisons in quantifying gender bias by using MGL to extract sentences with gendered words. Our rigorous methodology of evaluating gender bias in multilingual MLMs will raise the standards for bias evaluation in MLMs.

2 Related Work

Several research detected bias in language models. Nangia et al. (2020) presents a data set, CrowS-Pairs, which quantifies bias in Masked Language Models (MLMs). CrowS-Pairs is a collection of single sentences with masked attribute words that potentially generate social bias (e.g., *She is a [Mask]* where the mask is an occupation that may incur stereotypical bias). With this data set, Nangia et al. (2020) calculates the likelihood from their predictions to evaluate racial, gender, and religious bias. Nadeem et al. (2021) measures bias in a similar way by masking the modified token (e.g., *[Mask] is a nurse.*) and computing the likelihood from their predictions across multiple language models like BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021). Despite the great work, these benchmark data can only be tested in English. Moreover, these benchmarks lack a clear definition of what is being measured, leaving room for ambiguity and assumptions about 'bias' (Blodgett et al., 2021).

Expanding from the previous works, Ahn and Oh (2021) utilizes pairs of sentences, with the first sentence masking only the attribute word and the second sentence masking both the modified token and the attribute word. This paper defines their own metric named the Categorical Bias (CB) score, which shows the variance of log-likelihood. This paper proposes that CB score can be interpreted as an effect size of the attribute word. In addition to this novelty, Ahn and Oh (2021) analyzes ethnic bias across 6 languages, making an effort to generalize bias evaluation. All the previously mentioned studies are limited in that they require human-written sentences with annotations about the bias expressed within the sentence. This manual, simplified method of creating a bias evaluation data fails to capture the natural usage of the language and can even be exploited when the model finds a simple loophole around the set of rules (Durmus et al., 2022). Our work is distinguished in that it requires little annotation but is not evaluated on a fabricated, unnatural data set.

Recently, studies have attempted to evaluate bias in multilingual MLMs. Closest to our work, Kaneko et al. (2022) analyzes eight languages with a data set that requires little annotation in English. Kaneko et al. (2022)'s bias evaluation method requires a parallel corpus between English and the target language as the data set. Using a set of male and female-related words in English, the study evaluates bias in the model by calculating a likelihood with All Unmasked Likelihood with Attention weights (AULA; Kaneko and Bollegala (2022)). While Kaneko et al. (2022) quantifies gender bias of MLMs using a parallel corpus, this approach makes overarching assumptions to naively extract sentences with gender information. Our proposed methodology does not require a parallel corpus nor make this assumption but inspects for bias in each language with our own set of male and female-related words.

3 Bias Evaluation Methodology

3.1 Baseline

Kaneko et al. (2022) proposed a Multilingual Bias Evaluation (MBE) score to evaluate gender bias generated by Masked Language Models (MLMs) in different languages. We adopt the MBE score as our baseline score. With a parallel corpus, the MBE score captures gender bias in three steps.

First, MBE inspects for sentences containing a female or male noun in the English corpus. Here, the words are collated from the list of gender nouns from Bolukbasi et al. (2016) and common first names from Nangia et al. (2020). MBE assumes that a gender word is present in the translated sentence and extracts the corresponding sentence in the target language. After extraction, the sentences are categorized into T_f and T_m, representing female and male sentences.

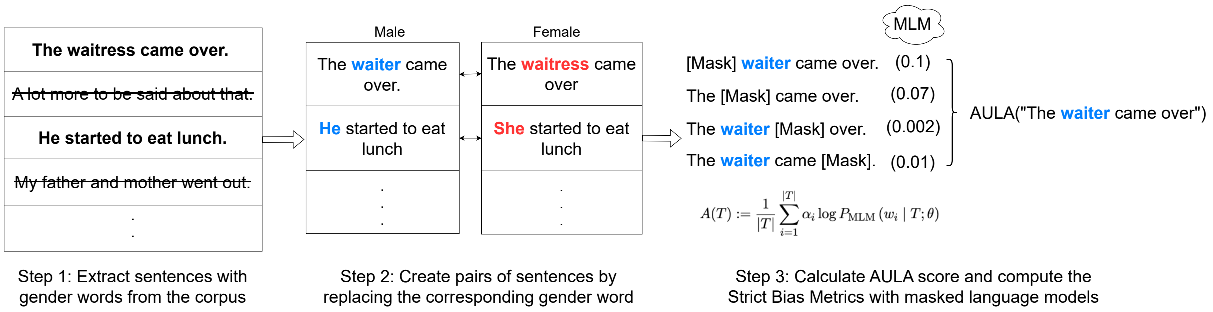
The waitress came over.
~~A lot more to be said about that.~~
He started to eat lunch.
~~My father and mother went out.~~
Male
Female
The waiter came over.
The waitress came over
He started to eat lunch
She started to eat lunch
MLM
[Mask] waiter came over. (0.1)
The [Mask] came over. (0.07)
The waiter [Mask] over. (0.002)
The waiter came [Mask]. (0.01)
AULA("The waiter came over")
$A(T) := \frac{1}{|T|}\sum_{i=1}^{|T|} \alpha_i \log P_{\mathrm{MLM}}(w_i \mid T;\theta)$
Step 1: Extract sentences with gender words from the corpus
Step 2: Create pairs of sentences by replacing the corresponding gender word
Step 3: Calculate AULA score and compute the Strict Bias Metrics with masked language models

Figure 1: Lexicon-Based Sentence Extraction

Second, MBE calculates the All Unmasked Likelihood with Attention weights (AULA; (Kaneko and Bollegala, 2022)) for each sentence in T_f and T_m. With T_j being a sentence in the language of interest, T_j can be decomposed to $T_j = [w_0, w_1, ..., w_{|T_j-1|}]$ where w_i denotes each token within the sentence. $P_{\mathrm{MLM}}(w_i \mid T_j;\theta)$ measures the probability of MLMs predicting token w_i given all the tokens of T_j and the pre-trained parameters θ. AULA is the summation of the log-likelihood multiplied by the average of multi-head attention α_i associated with w_i, which emphasizes relatively important words within the sentence as shown in Equation 1.

$$A(T) := \frac{1}{|T|}\sum_{i=1}^{|T|} \alpha_i \log P_{\mathrm{MLM}}(w_i \mid T;\theta) \quad (1)$$

Finally, MBE exhaustively compares the AULA likelihood of sentences in T_f and T_m. MBE utilizes an indicator function in order to pick out which male sentences have higher AULA scores when compared to those from female sentences. $C(T_f, T_m)$ measures the cosine similarity of sentence embedding for T_f and T_m. The MBE score shows the percentage of T_m preferred by the MLMs over T_f within the parallel corpus:

$$\frac{\sum_{T_m \in \mathcal{T}_m} \sum_{T_f \in \mathcal{T}_f} C(T_m, T_f)\mathbb{I}\big(A(T_m) > A(T_f)\big)}{\sum_{T_m \in \mathcal{T}_m} \sum_{T_f \in \mathcal{T}_f} C(T_m, T_f)} \quad (2)$$

3.2 Strict Bias Metrics

One of the limitations of the MBE scoring method is that the comparisons are not rigorous enough. When the AULA computes the likelihood of sentences, it sums up the likelihood of each token and calculates the score, which leaves too much freedom for other tokens, unrelated to gender, to affect the score. This especially is problematic when the AULA likelihoods are compared even when the sentences are distinctively dissimilar.

This margin of error can be reduced in measuring the bias of MLMs if likelihoods are only compared on the pairs of sentences with the only difference being the gender word using Strict Bias Metrics (SBM). Unlike MBE, SBM requires information about which word is the gender word within the sentence. Therefore, this study creates a list of gender words in different languages in §4.1. While the number of comparisons in the AULA scores decreases, SBM only makes meaningful comparisons by capturing the difference in the likelihood incurred by the difference in gendered words. The equation of SBM is as follows:

$$\frac{\sum_{(T_m, T_f) \in \mathcal{T}_m \times \mathcal{T}_f} \mathbb{I}\big(A(T_m) > A(T_f)\big)}{|\mathcal{T}_m|} \quad (3)$$

3.3 Lexicon-based Sentence Extraction

SBM takes pairs of sentences with gender words as the input. This study proposes two methods for making pairs of sentences with the only difference between the sentences being the gender words.

The first approach named the Lexicon-based method finds the sentences with one gender word. Then, it constructs another sentence with the opposite gender word replacing the original gender word. For example, if the corpus contains the sentence "The waitress came over", this method creates a new sentence "The waiter came over" as shown in Figure 1. Between these two sentences, this method computes the SBM to compare the likelihood of the MLMs predicting these two sentences, considering the relative importance of gender words. The corpus in evaluation might have a skewed number of sentences toward male-gendered words or female-gendered words. To mitigate this, the Lexicon-based approach matches the number of sentences that contains female words to the number of sentences with male words.

The sentences that have more than one gender word are excluded because they cannot be classified as male or female sentences with conflicting information. If the sentence "His answer is more accurate than hers" is used as both male and female sentences, the unmasked gender word will provide unwanted clues for the MLMs in generating AULA scores when either ’His’ or ’hers’ is masked, making the SBM score not reliable.

3.4 Model-based Sentence Extraction

The Model-based approach also makes pairs of sentences with distinctive gender words, shown in Figure 2. First, sentences that contain one gender word are extracted and simultaneously the gender word is masked. Similar to the Lexicon-based approach, only sentences with one gendered word are included. The key difference from Lexicon-based sentence extraction is that Model-based approach uses the highest-likely male or female word predictions made by language-specific BERT-based models: German (Chan et al., 2020), Spanish (Cañete et al., 2020), Portuguese (Souza et al., 2020), and Chinese (Cui et al., 2020).

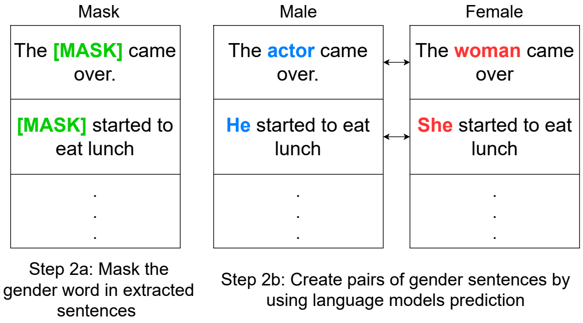
Mask
Male
Female
The [MASK] came over.
The actor came over.
The woman came over
[MASK] started to eat lunch
He started to eat lunch
She started to eat lunch
Step 2a: Mask the gender word in extracted sentences
Step 2b: Create pairs of gender sentences by using language models prediction

Figure 2: Model-Based Sentence Extraction

In this process, if a language model does not predict the [Mask] as a female or male word, this study excludes those sentences. If it predicts the [Mask] as a female word, but not as a male word, this work adapts the Lexicon-based method and uses the corresponding male pair word of the female predicted word to create a sentence with the male-gendered word, and vice-versa, as shown in the second row of Figure 2. Optimally, MLMs make predictions for both male and female words, in which the Model-based method utilizes the both highest-likely male and female predictions. Finally, SBM is applied to the pairs of sentences to quantify whether the MLMs are biased towards males or females.

MLMs can produce a probability for words within their list of vocabulary, so this study restricts a threshold in which the likelihood is significant. This study assumes that any predictions made by the language models with a likelihood less than 0.01 are not significant. In turn, this study inspects which top k value successfully captures the pool of predictions that have a likelihood greater than 0.01. After extracting all the sentences that have a male or female word, predictions are made by the MLMs. Then, appropriate k is determined by looking into which pool of top k predictions covers most of a male or female prediction with its likelihood greater than 0.01 when iterating k from 1 to 15. This study concludes using the top 10 predictions is appropriate in that the model has only marginal changes in coverage rate after k equals to 10, as shown in Figure 3.

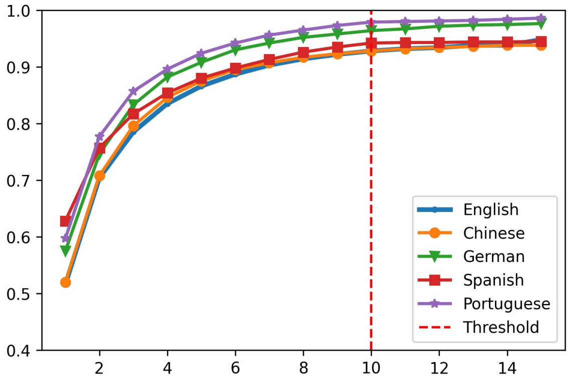

Figure 3: Sentences with gender words within the top 10 predictions by Masked Language Models

3.5 Direct Comparison Bias Metrics

By utilizing AULA score, SBM shows how much MLMs prefer male sentences compared to female sentences. Because MLMs produce the likelihood of a word when filling in the [Mask], a comparison at the word level is also conducted in this study. The likelihoods of a male word to a female word being predicted within a sentence are compared, given that at least one of the predictions, either male or female, produces a significant probability (greater than 0.01). Direct Comparison Bias Metrics is formulated to compare the likelihoods of the words from different gender sentences:

$$\frac{\sum_{(T_m,T_f)\in\mathcal{T}_m\times\mathcal{T}_f}\mathbb{I}(P(w_m;T_m)>(P(w_f;T_f))}{|\mathcal{T}_m|} \tag{4}$$

4 Data Preparation

This paper makes the most of parallel corpus when creating the data for gender bias evaluation for multilingual language models. The corpus used in this study is TED corpus. This parallel corpus is comprised of approximately 4,000 TED talks that make up 427,436 English sentences within the corpus. The transcripts were manually translated by certified translators in more than 100 languages and later reviewed before being made public (Reimers and Gurevych, 2020). We conducted all our experiments on an M1 Pro Chip with a 14-core GPU using the transformers implementation (Wolf et al., 2020).

4.1 Multi-lingual Gender Lexicon

This work takes the list of male and female words in Bolukbasi et al. (2016) to produce an English Gender Lexicon. It is ensured that each word within the list has a counterpart word of the opposite gender, which makes this list applicable to the Lexicon-based and the Model-based method. Unlike Kaneko et al. (2022), common first names from the CrowS-Pairs data set are disregarded in making Multilingual Gender Lexicon (MGL) (Nangia et al., 2020). This is due to the fact that when these first names are translated, they are often transliterated meaning that the closest pronunciation is reproduced in the translated language. This in turn lessens the interpretability of the first names leading to the loss of gender information.

The words in the English gender lexicon are translated into 8 languages by using three machine translation systems: Bing translator, DeepL translator, and Google translator. If machine translation systems yield different outputs, these translations are reviewed by 3 native speakers for all languages. The native speakers examine if the translation sounds natural. When the majority of reviewers determine that the translation does not sound natural, the translation is replaced with their translation. Additionally, if the majority of reviewers believe that the English word is translated into a gender-neutral word or if the translation has multiple meanings including a gender-neutral word, the translation and its counterpart are excluded from the MGL.

For example, the pronoun ’sie’ in German has two meanings relating to pronouns: she and you (honorific). Even though ’sie’ is the most widely used pronoun to refer to a female, it was taken out in order to leave no room for ambiguity on whether ’sie’ connotes gender information or not.

4.2 MGL validation

Before using MGL to extract sentences with gender words across 8 languages, this work observes if the words in MGL exist in the translated sentences that originally contain gender words. This study makes the most of parallel corpora by referring to the original sentence and the corresponding target sentence within TED corpus.

11,000 English sentences are randomly sampled from the TED corpus. Then, the sentences with gender words in English are extracted. Next, the extracted English sentences are mapped to the target language $\mathcal{T}_g$ to check if the translation exists. Within the translated sentences $\mathcal{T}_t$, we analyze how many sentences contain words from the MGL. Table 1 shows the results from this validation process, including the coverage percentage of MGL.

Language	$\|T_t\|$	$\|T_t \cap T_g\|$	Coverage Percent
German	1226	1124	**91.7**
Japanese	1288	466	36.6
Arabic	1327	252	19.0
Spanish	1380	1125	**81.5**
Portuguese	1206	928	**76.9**
Russian	1289	583	45.2
Indonesian	671	312	46.5
Chinese	1325	997	**75.2**

Table 1: The total percent of sentences that contain words from MGL across 8 different languages.

gender words from MGL do not successfully cover the majority portion of the sentences in four languages, even though the corresponding English sentence includes gender information. In Indonesian, a lot of gender-neutral words replace what used to be gender-specific words in English because it would sound unnatural to translate to a gender-specific translation (Dwiastuti, 2019). This resulted in using less than 50 percent of the English lexicons, resulting to a low coverage percentage. Both Russian and Arabic are morphologically rich languages, making our MGL cover-

Lang	Kaneko_sub	Kaneko_all	Lexicon-based	Model-based	Total
English	—	39,040	25,993	28,112	34,970
German	4,700	26,639	32,436	29,667	33,154
Spanish	7,100	37,808	76,972	96,995	114,168
Portuguese	5,700	29,975	24,608	31,670	36,072
Chinese	6,800	36,270	22,196	22,616	30,547

Table 2: The number of sentences used for different methods.

age rate low (Al-Haj and Lavie, 2010; Rozovskaya and Roth, 2019). Japanese is a pro-drop language meaning that it allows for omitting the subject of a sentence. This is crucial in that the subject usually accounts for the part of a sentence that contains gender words, which in turn explains the low coverage percent for Japanese.

Previous Kaneko et al. (2022)'s method depends on a sweeping assumption that the gender information from an English sentence with a gender word is retained in the corresponding sentence in other languages, even if gender-specific words do not exist in the translated sentence. This assumption does not hold for languages that are mentioned above unless the sentence contains a very gender-specific context. For example, if the Japanese translation omitted the subject of the sentence, it becomes extremely difficult to classify the [Omitted Subject] as female or male (e.g., *[Omitted Subject] is a student*). Therefore, this study only uses the MGL on languages that have relatively high coverage: German, Spanish, Portuguese, and Chinese.

4.3 Sentence pair generation

This paper utilizes MGL in English, German, Spanish, Portuguese, and Chinese to make two different data sets with an equal number of male and female sentences. Table 2 shows the number of sentences used in evaluating gender bias for MLMs in different languages.

Even though this paper utilizes the same MGL in extracting sentences with only one gender word from the same corpus, the number of sentences used for the Model-based approach is different compared to the Lexicon-based approach. The loss of sentences in the Lexicon-based approach occurs when matching the number of sentences with male and female-gender words.

Data truncation occurs in the Model-based approach when MLMs do not predict the masked part as either a male or a female-gender word. Table 3 reveals the percentage of sentences that had both female and male, just one, or zero predictions across the five languages.

Lang	Comparisons	None	One-sided	Both
English	28,112	19.6%	16.6%	63.8%
German	29,667	10.5%	58.8%	30.7%
Spanish	96,995	15.0%	33.1%	51.9%
Portuguese	31,670	12.2%	44.8%	43.0%
Chinese	22,616	26.0%	14.9%	59.1%

Table 3: Distribution of gender word prediction by Masked Language Models in Model-based Approach.

Some languages like German have gendered articles, adjectives, demonstrative, possessive, and attributive pronouns. When these components are related to the gender word within MGL, they need to be altered as well to ensure grammatical soundness when extracting the sentence pairs. Of all extracted gender indicative sentences in German using the Lexicon-based approach (§3.3), 19.85% of these sentences contained articles, adjectives, or specific pronouns that are related to the gender nouns. This study thus changed those relevant words with the corresponding gendered word. For example, if 'Mann' was the gender word in 'ein guter Mann,' it was adjusted to 'eine gute Frau.'

5 Experiments

5.1 MBE on the Entire Corpus

When replicating his work, we notice that Kaneko et al. (2022) only makes use of about one-fourth of the sentences within the TED corpus as seen in Table 2. This study compares the MBE score presented in Kaneko et al. (2022)'s work to the MBE score computed on the entire corpus for English, German, Spanish, Portuguese, and Chinese Masked Language Models (MLMs). From Table 4, the second column is the MBE score from the sub-sample of sentences presented in Kaneko et al. (2022)'s work, and the third column is the MBE score on the entire TED corpus using the same extraction method and metrics. Surprisingly, the MBE score evaluated on the entire corpus show

Lang	Kaneko_sub	Kaneko_all	Lexicon-based	Model-based	Direct	Male / Female
English	—	52.07(±1.34)	50.39(±0.28)	45.49	75.18	62.83 / 37.17
German	54.69	45.78(±1.72)	52.31(±0.64)	55.43	44.72	48.92 / 51.08
Spanish	51.44	48.52(±1.04)	41.68(±0.76)	50.74	72.34	66.29 / 33.71
Portuguese	53.07	46.70(±0.81)	51.77(±0.44)	61.04	73.36	65.89 / 34.11
Chinese	52.86	46.67(±0.55)	46.42(±0.68)	53.15	89.62	63.67 / 36.33

Table 4: Left: Comparison of MBE scores evaluated between the sub-sampled sentences and the sentences extracted over the entire corpus (confidence level at 99%). Middle: Lexicon-based (confidence level at 99%), Model-based, and Direct Comparison Bias Metrics score. Right: Distribution of male and female sentences from the original corpus.

contradictory results compared to the MBE score from sub-sampled sentences.

5.2 Lexicon-based and Model-based SBM

This study evaluates the gender bias of MLMs with the sentence created by the Lexicon-based approach. Table 2 indicates the number of sentences used for detecting gender bias when using Lexicon-based and Model-based sentence extraction. From Table 4, the scores obtained from Lexicon-based Strict Bias Metrics (SBM) are greater than 0.5 for English, German, and Portuguese, showing that these MLMs are biased toward males. In contrast, Chinese and Spanish MLMs show bias towards females.

The Model-based SBM represents the gender bias score when evaluated on sentences created by language models. The English language model shows bias towards females, whereas the rest of the language models in German, Spanish, Portuguese, and Chinese show bias toward males.

5.3 Direct Comparison Bias Metrics Score

This section looks into the gender bias score evaluated on the word level by using Direct Comparison Bias Metrics (DCBM). In doing so, this study uses sentences extracted from Model-based sentence extraction that have at least a one-sided prediction. Therefore, the number of sentences used in computing Direct Comparison Bias Metrics (DCBM) is the same as the number used for obtaining Model-based SBM.

The scores from DCBM are much more extreme compared to MBE and SBM. All the scores follow the distribution of male and female sentences from the original corpus. This implies that using DCBM in a heavily gender-skewed data set to quantify the gender bias of the MLMs can lead to a faulty result. DCBM is thus more suitable for quantifying the bias of the evaluation corpus, rather than the bias of MLMs.

6 Analysis

6.1 Performance Analysis

Kaneko et al. (2022)'s scores, when compared to the results on the entire corpus, contradict each other. This is due to the fact that only a subsample of the entire corpus is utilized in the evaluation. Because the sentences are randomly removed in the process of evening out the number of male and female sentences for MBE and Lexicon-based approach, MBE score and Lexicon-based SBM are not consistent. It is noteworthy that the standard deviation of SBM scores obtained from the Lexicon-based method after running the experiments five times is smaller than that of MBE scores for all languages except for Chinese. However, when the bias scores are close to the threshold, this inconsistency from the randomness of the data set can incur contradicting results.

Unlike MBE or Lexicon-based approach, the Model-based approach does not even out the number of male and female sentences. The model-based approach does not require equating the number of male and female sentences due to the assumption that masking the gendered words effectively decorrelates gender information within the sentence. This assumption holds because all the sentences that were extracted only have one gender word. Ultimately, Table 5 reaffirms that the Model-based approach will retain most of the sentences in all languages except German.

Lang	MBE	Lexicon-based	Model-based
English	31.98%	25.67%	**19.60%**
German	35.35%	**2.17%**	10.51%
Spanish	31.99%	32.58%	**15.04%**
Portuguese	30.90%	31.78%	**12.20%**
Chinese	32.13%	27.34%	**25.96%**

Table 5: Percentage of data truncation for MBE, Lexicon-based, and Model-based approach.

Given that the Model-based approach is consistent and least prone to truncation, some might question, "Why adapt the Lexicon-based approach at all within the Model-based approach?" If the evaluation corpus size becomes larger, the number of sentences with both male and female predictions indeed becomes large enough.

However, the downside of only utilizing the Model-based approach, meaning only using the sentences that have both male and female predictions within the top 10 predictions, is that predictions made by the Model-based approach are less diverse. Figure 4 shows that the numbers of vocabulary used from MGL, across all five languages in evaluation, are greater in sentences extracted with the Lexicon-based approach compared to the sentences extracted with the Model-based approach. The numbers of unique words used to fill the [Mask] do not exceed 50% for all the languages except for Chinese. The Chinese MLM is peculiar in that it fills the [MASK] of over 18,000 sentences with only 10 male nouns and 6 female nouns. By incorporating the diverse set of vocabulary from one-sided prediction when implementing the Model-based method, the SBM quantifies MLMs' preference even on sentences that they are not inclined to generate.

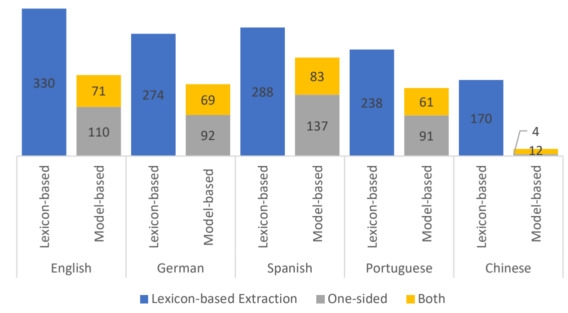

Figure 4: Number of gender words used for Lexicon-based and Model-based sentence extraction

6.2 Discussion

In constructing MGL, this work depended on machine translators and cross-validation from 3 native speakers. MGL can be improved upon by using a word alignment tool on a verified parallel corpus which will in turn allow the Lexicon-based and Model-based approaches to be expanded to other languages. Additionally, morphological analysis is essential for languages that have grammatically gendered words not just in nouns, but also in articles, adjectives, and verbs.

With this MGL of English, German, Spanish, Portuguese, and Chinese, future works can be conducted to evaluate gender bias in a plethora of MLMs not explored in this paper. In addition, future work is not dependent on using a parallel corpus and therefore can evaluate gender bias in any corpus. Rather than measuring the gender bias of MLMs with translations from parallel corpus, utilizing a corpus with that specific language can lead to more interesting results. We also defer the work of testing the limit of BERT-based language models to probe if increasing the length of the sentence length leads to different gender bias scores for the same language models.

7 Conclusion

We present rigorous methods of evaluating gender bias in English, German, Spanish, Portuguese, and Chinese Masked Language Models. With the three newly presented methods, a comparative analysis of the results is conducted to conclude that the Model-based approach is the most generalizable and consistent method. Despite the great achievements, our work is only limited to 5 languages. We want to further our methods to explore gender bias in more language models and use evaluation corpora that are not a parallel corpus.

Bias evaluation is an untapped field with many new studies coming up with new methodologies and metrics to quantify bias. While there is no consensus on how to detect, evaluate, and mitigate bias, we believe that a collective effort to investigate bias in NLP from diverse perspectives is imperative for this line of research. The bias evaluation system should not be biased, and the best way to ensure this is to tackle bias evaluation in a multi-faceted way. All our resources including the MGL and source codes for evaluation are available through our open-source project [1].

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hassan Al-Haj and Alon Lavie. 2010. The impact of Arabic morphological segmentation on

[1] https://github.com/anonymous

broad-coverage English-to-Arabic statistical machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.

Meisyarah Dwiastuti. 2019. English-Indonesian neural machine translation for spoken language domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 309–314, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. *Unmasking the Mask* - Evaluating Social Biases in Masked Language Models. In *Proc. of the 36st AAAI Conference on Artificial Intelligence*, pages 11954–11962.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pre-trained 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.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Fabio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

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

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.