An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models

Victor Steinborn and **Philipp Dufter*** and **Haris Jabbar** and **Hinrich Schütze**
Center for Information and Language Processing (CIS)
LMU Munich, Germany
{vsteinborn, philipp, jabbar}@cis.lmu.de

Abstract

Warning: *This work deals with statements of a stereotypical nature that may be upsetting.*

Bias research in NLP is a rapidly growing and developing field. Similar to CrowS-Pairs (Nangia et al., 2020), we assess gender bias in masked-language models (MLMs) by studying pairs of sentences that are identical except that the individuals referred to have different gender. Most bias research focuses on and often is specific to English. Using a novel methodology for creating sentence pairs that is applicable across languages, we create, based on CrowS-Pairs, a multilingual dataset for English, Finnish, German, Indonesian and Thai. Additionally, we propose S_{JSD}, a new bias measure based on Jensen–Shannon divergence, which we argue retains more information from the model output probabilities than other previously proposed bias measures for MLMs. Using multilingual MLMs, we find that S_{JSD} diagnoses the same systematic biased behavior for non-English that previous studies have found for monolingual English pre-trained MLMs. S_{JSD} outperforms the CrowS-Pairs measure, which struggles to find such biases for smaller non-English datasets.

1 Introduction

Pretrained language models (PLMs) have greatly benefited NLP (Raffel et al., 2020; Peters et al., 2018; Devlin et al., 2019; Zhuang et al., 2021). However, commonly used PLMs such as BERT have been shown to encapsulate social biases, including those relating to gender and race (Kurita et al., 2019; Nadeem et al., 2021; Nangia et al., 2020). The general consensus is that these biases are learned from the statistical distributional co-occurrence of words relating to a group (such as terms relating to men or women) with a context in which that group is often mentioned in corpora (Bolukbasi et al., 2016; Webster et al., 2021). For example, "doctor" may co-occur with "man" more often than with "woman", leading to an internal representation in the model where a gender-neutral concept, such as being a doctor, is more closely associated with male-related terms than with female-related terms (Bolukbasi et al., 2016).

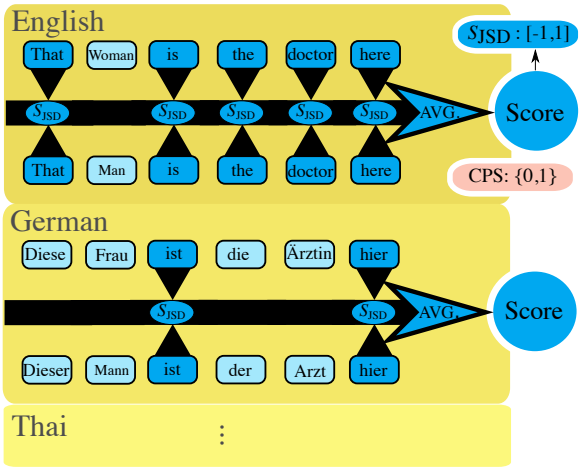

Figure 1: Following Nangia et al. (2020), we assess multilingual gender bias in MLMs by matching gender-specific tokens (light blue) in the context of non-gender-specific tokens (dark blue) in sentence pairs. We develop a methodology for creating sentence pairs that we argue is applicable across languages in contrast to prior work. We mask unchanged tokens one at a time and calculate S_{JSD}, a novel information-theoretic bias measure whose sentence-level average we show to be better behaved than competing measures.

In this work we tackle this type of binary stereotypical representational gender bias (henceforth simply "gender bias") in MLMs in a multilingual setting. We propose a multilingual approach to study gender bias in MLMs, outlined in Figure 1, which, to the best of our knowledge, can in principle be extended to any language.[1]

The importance of developing AI systems that

*Now at Apple

[1]Code and dataset with additional languages available at: https://github.com/VSteinborn/s_jsd-multilingual-bias

Findings of the Association for Computational Linguistics: NAACL 2022, pages 921 - 932

are mindful of different societal groups, such as people of different genders, is a topic much discussed in the area of fairness research in NLP (Blodgett et al., 2020). However, a shortfall of this area is its almost exclusive focus on English. As far as we are aware, ours is the first study to attempt to create a truly multilingual approach to study gender bias in language models. Previous multilingual approaches were largely limited to sentences with fixed templates and grammar structures, which heavily constrains the range of languages that may be studied with a given template (González et al., 2020). Our approach builds on Nangia et al. (2020) and attempts to study natural sentences by comparing a pair of sentences that differ only by the gender of persons mentioned, a process which we will refer to as *gender swapping*.

To illustrate the problem of using templates, consider the following sentence pair and its German translation.

(1) a. **He** is the doctor here.
b. **She** is the doctor here.

(2) a. **Er** ist **der Arzt** hier.
b. **Sie** ist **die Ärztin** hier.

In German the only parts that remain the same are "ist" and "hier" under gender swapping, as the German word for the profession "doctor" and its associated definite article change form depending on the gender of the person. Thus, template structures developed for English of the form

(3) [*person*] is the [*profession*] here.

have to be heavily modified and constrained to create grammatically correct sentences in German. The problem is exacerbated in multilingual studies, where appropriate templates need to be defined for each language.

We take inspiration from CrowS-Pairs (CPS) (Nangia et al., 2020), which studies pairs of crowdsourced sentences, for a range of social biases. It includes gender-swapped pairs for the diagnosis of gender bias. However, we found that we cannot simply translate CPS into other languages. The main problem is that English pronouns are clear indicators of gender – at least of binary gender, which we focus on in this paper. But this clear indication gets lost in translation for languages that have gender-neutral pronouns like Finnish and those that predominantly use null pronouns like Thai.[2] We could mandate that only words with "gender-inherent" meaning like "mother", "wife" and "sister" are used, but that would exclude many topics that we need to cover in a good diagnostic dataset, e.g., work life and sports.

The solution we propose is to simply use names to indicate gender. Our assumption here is that all languages have words for names and that there are two subsets of names that can only have female and male referents. Note that there are certainly "unisex" names, i.e., names that can refer to both men and women, even in English ("Jess", "Leslie"). But as far as we know there is no language that has no "monosex" names, i.e., names that can refer to only one gender. We rely on such monosex names to construct sentence pairs.

In English, we select a few frequent male and female names; we only use them for English. Before translating the sentence pairs into another language, we first identify corresponding frequent male and female names in the target language. The translators are then instructed to only use those names. This methodology should be applicable universally, so that we can construct a multilingual gender bias resource for any set of languages. In this paper, we translate the CPS dataset into German, Indonesian, Thai and Finnish. We edit the original CPS dataset before translation to heed the recommendations of Blodgett et al. (2021). A more detailed description of dataset creation will be given in §3.1.[3]

The second contribution of this paper is S_{JSD}, a novel measure based on the Jensen–Shannon divergence (Lin, 1991), to test MLMs for social biases by using sentence pairs that capture a binary contrast between two groups. The measure used in CPS (see §3.2) makes use of a binary decision process, which has the effect of removing information of the probability values from the MLM, which we show reduces the measure's predictive power. Our motivation for introducing S_{JSD} is to retain as much information from the MLM output probabilities as possible in our final reported score in order to make effective use of the limited amount of human-translated sentences that are available.

Thus, our contributions are (1) developing a method for creating multilingual datasets for diagnosing gender bias in language models that is ap-

[2]The English sentence "she ate it" is simply expressed as "ate" in many "pro-drop" languages as long as subject and object of "ate" are clear from context.

[3]Blodgett et al. (2021) argue against using names for race. Their arguments do not apply to gender in our setup. See §3.1.

plicable across the diverse set of human languages, (2) applying this method, taking the CPS dataset (Nangia et al., 2020) as a starting point, and creating a multilingual gender bias diagnosis dataset for English, German, Thai, Indonesian and Finnish, (3) proposing the S_{JSD} measure, which retains information regarding the numeric output probabilities of MLMs.

2 Related Work

Given this work focuses on multilingual methods to measure gender bias in MLMs, this discussion will focus on evaluation measures and techniques; a thorough discussion of debiasing methods is beyond the scope of this paper.

Bias Measures in MLMs. Recently, pretrained masked language models, such as BERT (Devlin et al., 2019), have significantly gained in popularity, which in turn has led to numerous studies analyzing their behavior, including their encapsulation and reproduction of social bias. Prior to the emergence of these models however, it was already well known that NLP models can learn social biases from corpora, as exemplified in work by Bolukbasi et al. (2016) who demonstrated that word embeddings contain societal gender biases. Subsequently, further tests, such as the word embedding association test (WEAT) by Caliskan et al. (2017), demonstrated that word embeddings also have other biases, including racial biases. May et al. (2019) extended WEAT to sentence encoders, including BERT, with the sentence encoder association test (SEAT), to study sentence-level social biases in these models using template constructed sentences. However, the results of this study were inconclusive, and Kurita et al. (2019) showed that the cosine-based methods used in WEAT and SEAT are not appropriate for contextualized embeddings, and instead use a scoring method based on the prediction probability of an attribute given a target in template sentences.

The evaluation method used in StereoSet (Nadeem et al., 2021) was inspired by SEAT while CPS (Nangia et al., 2020) uses pseudo-log-likelihood MLM scoring (Salazar et al., 2020). A contribution of CPS and StereoSet is to provide techniques that evaluate natural sentences instead of simple templates. One disadvantage of template approaches is that they have been shown to be highly dependent on the template chosen, as well as on the terms that are chosen to substitute into the template (Delobelle et al., 2021; Antoniak and Mimno, 2021). Nonetheless, Kaneko and Bollegala (2021) criticize CPS and StereoSet for their evaluation measures, arguing that the act of masking tokens results in a systematic overestimate in measured biases. However, they also describe this effect as systematic, and thus we would expect systematic trends in bias scores between models to remain conserved when masking tokens.

Multilingual Studies of Bias in MLMs. As far as we are aware, there are no studies that have attempted to develop a multilingual method to test for gender bias in MLMs without template structures. However, there are several multilingual studies. For example, González et al. (2020) constructed sentence templates for languages with type B reflexivization (including Swedish and Russian), which can be used to construct challenge datasets to measure gender bias. Similarly, Câmara et al. (2022) used template structures to test MLMs for intersectional biases in English, Spanish and Arabic. Bartl et al. (2020) also constructed templates to study biases in German and English BERT models, but sometimes a different form of a template has to be used depending on the gender of a mentioned person. Liang et al. (2020) examined the case of English and Chinese using templates while focusing on the cross-lingual transfer of removing biases in Chinese using English training data.

Counter Factual Data Augmentation (CFA). Our work generally falls under the category of CFA. CFA has been used to train a model on an augmented training corpus by swapping target terms, which has been shown to be effective for debiasing in multilingual settings via zero-shot transfer learning (Lauscher et al., 2021). However, simple substitution methods employed in CFA fail at producing grammatical sentences in languages with gender agreement rules. For such languages, other strategies, such as machine translation (Jain et al., 2021), have to be employed.

Barikeri et al. (2021) create templates from real-world conversational text that can be used to evaluate language models for social biases. These templates then produce so-called "counterfactual pairs" Zhao et al. (2018) by substituting terms representing different social groups, resulting in sentence pairs similar to those in CrowS-Pairs (Nangia et al., 2020). Again, as we discussed in the introduction, templates are difficult to use for many languages.

In contrast to most work on CFA, we do not use

templates, we target non-English, we create data by crowdsourcing and our focus is measuring bias cross-lingually, as opposed to debiasing.

Bias From a Social Science Perspective. A critical survey of 146 NLP papers by Blodgett et al. (2020) outlines common pitfalls in NLP research, including the CPS study, when attempting to study social bias. We attempt to take into account their recommendations in this work.

3 Methodology

3.1 Dataset

A major obstacle in transferring existing techniques to measure gender bias in languages beyond English is that we need to adapt methods to the target language's gender agreement system. Methods for measuring gender bias in MLMs often rely on fixed sentence templates, where predefined words are inserted that test some aspect of bias, such as occupational gender bias (e.g., (Kurita et al., 2019; Webster et al., 2021)). While these template structures can be modified and applied to a range of languages, once a template is chosen, the range of languages that can be studied is restricted (González et al., 2020).

Thus, to design a multilingual approach to gender bias, we want to move beyond the rigid artificial sentence structures that result from using templates. We also speculate that moving away from rigid sentence structures allows us to probe the language model more deeply for biases. It may be possible that superficially debiased language models can perform well on certain bias evaluation tasks that use templates, similar to the situation for linearly debiased word embeddings that perform well on some bias measures but still encapsulate significant distributional biases (Gonen and Goldberg, 2019).

Two evaluation datasets that go beyond templates are StereoSet (Nadeem et al., 2021) and CPS (Nangia et al., 2020). One important difference between them is the masking pattern. While StereoSet's context association test masks words that may be gendered in a different language (e.g., adjectives in Spanish), CPS consists of pairs of sentences and only masks tokens that are shared by the two sentences. Here we will only consider the CPS dataset, which also marks which of the two sentences is more stereotypical (Nangia et al., 2020).

For our dataset, we consider sentence pairs where people of the male and female gender are being contrasted, for example:

(4) a. **He** is a pilot.
b. **She** is a pilot.

For this example, we assume each word is a separate token. The unmodified tokens common to both sentences are: "is", "a", "pilot". For each sentence, the unmodified tokens form a set U and the remaining modified tokens a set M ("**He**" for (4)a, for example). Thus, for each sentence, the set of all tokens is the union of U and M.

We will make the assumption that, for sufficiently long and complex sentences, when swapping the gender of a person reference in a sentence there remain sections of the sentence that remain unchanged and that this is true for all languages. From this observation, we found the masking pattern CPS implements to be appropriate for multiple languages and thus the sentences labeled with the "gender" tag in the CPS dataset were selected as the basis for subsequent translations.

The CPS dataset was recently criticized for lacking clear explanations of what types of social biases are being measured (Blodgett et al., 2021). For this reason the selected CPS sentences have been minimally modified to be mindful of the pitfalls outlined in (Blodgett et al., 2021). For example, some sentences were omitted because the contrasted groups were unrelated to the stated "gender" label, such as for sentences that contrasted two racial groups instead.

We will now outline the modifications of the CPS dataset for this study.

First, we ensured each sentence only compares binary gender. Non-binary gender adds a level of complexity in the multilingual context that we leave for future work. We also removed sentences that compare clothing items, most likely intended as a proxy for gender. Clothing items and their significance differ across cultures, so such sentences are difficult to translate.

Second, for sentences that only used a pronoun to identify gender, we exchanged the pronoun with a common name that is stereotypically associated with one gender in the English dataset. Subsequently, when translating the English dataset into other languages, the names were exchanged for others that are common gendered names in the target language. We limited the number of names in the English dataset to four to simplify the subsequent translation process. Names were introduced

because many languages do not have gendered pronouns, and thus information relating to gender may be lost in translation. For example, a typical translation of (4) into Indonesian results in two identical sentences, which makes the sentence pair useless for Indonesian. Using names as a proxy for identifying a social group is discouraged in (Blodgett et al., 2021) for race bias, but using stereotypically gendered names as a proxy for binary gender seems unproblematic to us. For example, whereas names only indirectly and ambiguously identify race (at least in English), we can easily find names that are "monosex", i.e., names that can only have either male or female referents. Thus, we would modify example (4) as follows for our dataset:

(5) a. **Robert** is a pilot.
b. **Olivia** is a pilot.

Finally, we removed sentences that did not correctly isolate a stereotype, an issue noted in the original paper (Nangia et al., 2020).

In this work we investigate binary gender stereotypes as a representational harm across languages, to use the terminology of Blodgett et al. (2020). The CPS dataset was created by US crowdworkers (Nangia et al., 2020). We make the assumption that most aspects of gender bias should be part of a diagnostic test across languages and cultures. For example, the associations of "doctor" with "male" or of "childcare" with "female" are biases that most cultures are at risk for. So we should test whether our language models exhibit these biases for all cultures. There probably are aspects of gender bias that are relevant to only a small subset of cultures (e.g., the association of "being *eligible* to drive a car" with "male"). We stress the importance of investigating gender bias multilingually. Given that our study is the first to do this, we feel justified to leave the issue of how to comprehensively test for all aspects of bias in gender diagnosis to future work.

Note that we do not make the assumption that gender bias is the same across languages! If "childcare" is strongly associated with "female" in (the training corpus of) language A, but not in (the training corpus of) language B, then (assuming we use models that pick up bias from their training corpora) our methodology will find less gender bias for language B – and this would be the intended result of our work.

For the translations, we hired translators to translate the modified English dataset into their native language. Translators were paid an agreed upon amount above the minimum wage in their respective country of residence and were informed of the intended use of their translations. Each translator was provided an instruction sheet, which exemplifies the translation process of CPS sentence pairs from English to German. The translation instructions can be found in the supplementary material and the target languages of the translations were German (De), Finnish (Fi), Indonesian (Id) and Thai (Th). We chose these languages to cover different language families and because translators for them were easily available to us.

	De	En	Fi	Id	Th
#w	5470	5548	4151	4790	6693
#w/s	13	13	10	11	16

Table 1: Our multilingual bias diagnosis dataset consists of 212 sentence pairs in five languages. The table gives total number of words (#w) and words per sentences (#w/s) for each language. Thai was tokenized with Deepcut (Kittinaradorn et al., 2019).

An overview of the metadata of the edited and translated dataset is given in Table 1.

3.2 Bias Measure

Our aim is to create a bias measure that can retain meaningful information from the model output that is relevant for detecting multilingual gender bias. Before introducing our proposed measure, we will go over the CPS measure (Nangia et al., 2020).

CrowS-Pairs Measure. Given is a pair of gender swapped sentences. One sentence is judged to be socially more stereotypical than the other by the annotators in the CPS study (Nangia et al., 2020). We refer to the two sentences as "more" and "less".

The set of tokens that are shared (resp. are not shared, i.e., modified) between the two sentences is denoted as U (resp. M) – see §3.1. For each sentence the tokens in U are masked one at a time. Each time a token is masked, the sentence is passed through the model and the model output probabilities are obtained. Following Nangia et al. (2020)'s notation, we denote the output probability of the model for the i^{th} correct token under the mask $u_{G,i} \in U$ in the more stereotypical sentence as $P_{\text{more}}(u_{G,i}) \equiv P(u_i|U_{\backslash u_i}, M, \theta)$, where M are the unique tokens in the more stereotypical sentence and θ are the model parameters. The output probability for the other sentence P_{less} is defined analo-

gously.

The score for a sentence in the pair is its pseudo-log-likelihood, calculated as the sum of $\log P(u_{G,i})$ over all u in U where P is either P_{more} or P_{less}. The sentence pair is assigned a binary score of 1 (resp. 0) if the more stereotypical sentence has a larger (resp. smaller) score. A possible advantage of this binarization is that the numerical value of the pseudo-log-likelihood cannot be interpreted (hence "pseudo") (Nangia et al., 2020; Salazar et al., 2020), so one can only rely on the comparison of the scores, not on their absolute values. The final score is the percentage of sentences that have been assigned a score of 1.

According to Nangia et al. (2020), an ideal unbiased model would achieve a score of 50 on a dataset. However, it is important to keep in mind that each sentence pair contributes with equal weight to the final score, due to binarization. Consider as an example a language in which a small part of the sentence pairs are diagnosed as extremely biased, but most sentence pairs do not show bias, so their final score will be randomly 0 or 1. In such a case, CPS does not distinguish strong bias from weak bias and sentence pairs that are not biased contribute noise to the final measure. Hence, unusually biased behavior of the model may not be effectively captured by the measure, and in order to obtain meaningful results a large number of human-annotated sentence pairs is required.

The following simulated scenario will illustrate the effect of dataset size. Let us ignore the internal mechanisms of the model and for simplicity assume that a biased model has a fixed probability of $p = 0.55$ to assign a binary score of 1. This may be modeled as a Bernoulli process (Papoulis and Pillai, 2002). For such a model and for a set of $n = 200$ sentence pairs, roughly the number of sentences we consider in our study, the expected dataset score is 55 and the standard error 3.5 (since the standard error is $\sim \frac{1}{\sqrt{n}}$ for Bernoulli). Thus, the CPS measure must rely on a large number of sentence pairs to obtain statistically meaningful results because of the binary decision process that disregards information regarding the extent of the discrepancy between S_{more} and S_{less}. The measures of Nadeem et al. (2021) in StereoSet and of Kaneko and Bollegala (2021) also employ binarization and therefore do not make efficient use of the available data to measure bias.

The Proposed S_{JSD} measure. Our goal in developing the S_{JSD} measure was to create a theoretically well founded measure that retains information regarding MLM output probabilities, avoiding the binary decision process in CPS. This is especially important for our study, where we had limited resources to create the translated dataset.

The S_{JSD} measure is based on the Jensen-Shannon divergence (Lin, 1991), a quantity bounded to the range $[0, 1]$, that measures the similarity between two probability distributions, P and Q, defined as follows:

$$\text{JSD}(P||Q) = H\left(\frac{P+Q}{2}\right) - \frac{H(P) + H(Q)}{2} \tag{1}$$

where H is entropy. If P and Q are unrelated and share no overlap $\text{JSD}(P||Q) = 1$ and if they are the same distribution (maximum overlap) $\text{JSD}(P||Q) = 0$. The square root of the Jensen-Shannon divergence, the Jensen–Shannon distance, is a metric, i.e., it satisfies a range of properties intuitive to measures of distance, including the triangle inequality (Endres and Schindelin, 2003).

Define the *gold distribution* as a one-hot distribution G that identifies the correct token under the mask. We then define our measure S_{JSD} as the difference of two distances: the Jensen–Shannon distance between P_{more} (resp. P_{less}) and the gold distribution:

$$S_{\text{JSD}} = \sqrt{\text{JSD}(P_{\text{more}}||G)} - \sqrt{\text{JSD}(P_{\text{less}}||G)} \tag{2}$$

This definition may also be expressed purely in terms of the model output probability for the token under the mask $P_{\text{more/less}}(u_G)$, as $\text{JSD}(P||G)$ may be expressed in the form shown in Eq. 3 for any distribution P. Thus only human annotated text is evaluated.

$$\begin{aligned}\text{JSD}(P||G) &= \frac{1}{2}(P_G \log_2(P_G) \\ &- (P_G+1)\log_2(P_G+1)+2), \quad P(u_G) \equiv P_G\end{aligned} \tag{3}$$

The quantity S_{JSD} is also bound to the range $[-1, 1]$, which limits the effect of outliers. The theoretically ideal non-biased model should yield a value of 0 for S_{JSD} when the distance of P_{more} to G is equal to the distance of P_{less} to G. When P_{more} is closer to G than P_{less}, we take this as a sign of bias for the stereotypical sentence, thus we expect biased models to systematically generate negative S_{JSD} scores.

Model	Lang.	$S_{\text{JSD}} \times 10^{-3}$	CPS	B.S_{JSD}
mBERT	En	$-0.05_{\pm 1}$	$57_{\pm 3}$	$50_{\pm 3}$
xlmR	En	$-1_{\pm 2}$	$62_{\pm 3}$	$54_{\pm 3}$
mBERT	De	$-1_{\pm 2}$	$57_{\pm 3}$	$55_{\pm 3}$
xlmR	De	$-2_{\pm 2}$	$51_{\pm 3}$	$50_{\pm 3}$
mBERT	Id	$-3_{\pm 1}$	$46_{\pm 3}$	$51_{\pm 3}$
xlmR	Id	$-4_{\pm 2}$	$51_{\pm 3}$	$54_{\pm 3}$
mBERT	Th	$-4_{\pm 2}$	$60_{\pm 3}$	$60_{\pm 3}$
xlmR	Th	$-4_{\pm 2}$	$57_{\pm 3}$	$57_{\pm 3}$
mBERT	Fi	$-0.2_{\pm 2}$	$44_{\pm 3}$	$50_{\pm 3}$
xlmR	Fi	$-3_{\pm 2}$	$51_{\pm 3}$	$53_{\pm 3}$

Table 2: CPS and S_{JSD} scores and standard errors on our multilingual bias diagnosis dataset. The S_{JSD} scores systematically identify the stereotypical sentence as indicated by the negative scores. Some CPS scores are below 50, indicating the measure cannot capture the stereotypical behavior of the model for this dataset. The binarized version of S_{JSD} (B.S_{JSD}) also illustrates the effect of binarization. B.S_{JSD} has scores of 50 in three cases where S_{JSD} is negative, suggesting that binarization reduces the predictive power of the measure.

	Unperturbed		Perturbed	
Model	S'_{JSD}	CPS	S'_{JSD}	CPS
BERT	$-6_{\pm 1}$	$60.5_{\pm 1.3}$	$-6_{\pm 1}$	$58.6_{\pm 1.3}$
RoBERTa	$-10_{\pm 1}$	$65.5_{\pm 1.2}$	$-10_{\pm 1}$	$63.5_{\pm 1.2}$
ALBERT	$-13_{\pm 1}$	$67.0_{\pm 1.2}$	$-11_{\pm 1}$	$64.5_{\pm 1.2}$
mBERT	$-4_{\pm 1}$	$53.6_{\pm 1.3}$	$-3_{\pm 1}$	$55.6_{\pm 1.3}$
xlmR	$-4_{\pm 1}$	$57.1_{\pm 1.3}$	$-4_{\pm 1}$	$56.6_{\pm 1.3}$

Table 3: Scores and standard errors on the original CPS dataset (Nangia et al., 2020), for which BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021) and ALBERT (Lan et al., 2020) were used. $S'_{\text{JSD}} = S_{\text{JSD}} \times 10^{-3}$. Unperturbed and perturbed conditions where a sentence is perturbed by removing the final character. For this larger dataset both S_{JSD} and CPS show the same systematic trends in bias scores between the models, in agreement with the results of Nangia et al. (2020). Under the effect of the perturbation, the dataset is of sufficient size that both measures are robust and retain their systematic trends. The number of significant figures for CPS was chosen to match the results of the original CPS study.

To generate a score for a sentence pair, we take the average of S_{JSD} scores. For the score of the entire dataset, we take the average of the sentence scores.

Error Analysis. For an analysis of the error of the reported score on the dataset, we bootstrap the sentence scores to determine an estimate for the standard error using SciPy (Efron and Tibshirani, 1993; Virtanen et al., 2020). For CPS we achieve this by bootstrapping the binary sentence scores.

4 Experiments

For our experiments we make use of the Transformers library (Wolf et al., 2020). We use two multilingual models, multilingual BERT (mBERT) (Devlin et al., 2019), trained on Wikipedia, and base xlm-RoBERTa (xlmR) (Conneau et al., 2020), trained on Wikipedia and filtered CommonCrawl data from the internet (Wenzek et al., 2020). We choose xlmR as it has been shown to significantly outperform mBERT on numerous cross-lingual tasks (Conneau et al., 2020). As of this writing, xlmR seems to be the best performing multilingual model in the Transformers library (Wolf et al., 2020; Conneau et al., 2020). The two models differ in training data by the CommonCrawl, which we assume to be more of a source of bias than Wikipedia, based on the results of the CPS study. Nangia et al. (2020) found RoBERTa, trained on Wikipedia and the CommonCrawl, among other datasets (Zhuang et al., 2021), to generally have higher bias scores, compared to BERT (Devlin et al., 2019), although this was not true for gender bias.

We run the two models on our translated datasets and calculate CPS and S_{JSD} scores. Running a model on a single language using an Intel Xeon Processor E5-2680 v2 takes roughly 15 minutes.

We also test S_{JSD} on the models and dataset used in the CPS study (Nangia et al., 2020).

Finally, we test the effect of model size on the scores by comparing the large and base xlm-RoBERTa models. See the appendix for a list of all models used.

5 Results and Analysis

Table 2 shows results for CPS and S_{JSD} on the multilingual dataset. We observe that the CPS measure reports scores well under 50 for multiple languages. This goes against the intuition that MLMs learn stereotypical associations from data: it wrongly suggests that male stereotypes are associated with women and female stereotypes with men. We suspect this behavior of CPS comes from the binary decision problem outlined in §3.2, which is especially relevant for smaller datasets.

A first indication to suspect that we might be in this regime is that the CPS standard errors are close in value to the estimated standard errors assuming a Bernoulli process, as discussed in §3.2. Thus we cannot make a reliable inference regarding model bias. We can also observe a clustering of CPS sentence scores, before binarization, around

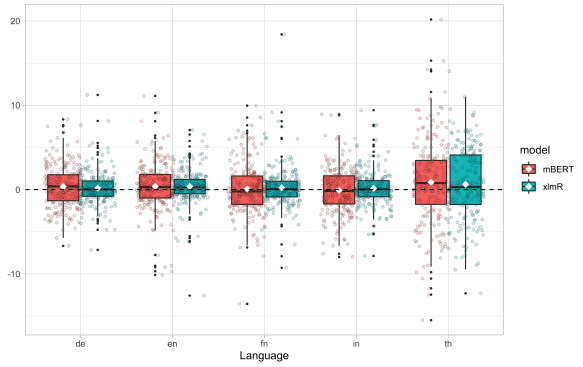

Figure 2: The difference $S_{\text{more}} - S_{\text{less}}$ for our multilingual bias diagnosis dataset. The white points mark the averages and the box and whiskers plots mark the quartiles. Most of the scores cluster around the decision boundary denoted by the horizontal dotted line.

the decision boundary in Figure 2, indicating that slight variations in bias scores can substantially change the CPS score. Furthermore, the effect of binarizing S_{JSD} (i.e., following the CPS method but replacing $\log P_{\text{more}}(u_{G,i})$ with the JSD distance to the gold token) is shown in Table 2. These binarized S_{JSD} scores fail to detect bias by yielding scores of 50 in three cases – whereas the S_{JSD} score predicts bias as expected. All this, coupled with the discussion in §3.2, reinforces our argument that binarization harms measure performance and that S_{JSD} is numerically more suitable and theoretically justified as a measure compared to CPS, especially on smaller datasets. Note that we did not unbinarize CPS scores as they have no clear statistical interpretation (Nangia et al., 2020; Salazar et al., 2020); see discussion in §3.2.

Table 2 shows that S_{JSD} has negative values, i.e., indicates bias consistently across all languages and models. Interestingly, xlmR consistently yields equal or more negative S_{JSD} scores than mBERT; this supports our hypothesis that xlmR encapsulates more gender bias. However, we also note that some of the standard errors are large compared to their associated average. This, we suspect, is largely due to the limited number of sentences that are available in the analysis.

To test our measure on a larger dataset, we applied S_{JSD} to the original CPS dataset (which is much larger since it covers many different types of bias) in Table 3. We also tested the stability of the measures by perturbing the dataset by removing the final character of each sentence, which most commonly resulted in the removal of a period. From this we observe that both the CPS and the S_{JSD} standard errors decrease with dataset size, as expected.

Ideally, future MLM bias datasets would have a statistical justification that the calculated sample error is adequate to make judgments about a model's bias as probed by a particular measure.

Finally, we also tested for model size by comparing the base and large xlmR models (Conneau et al., 2020). While both models were found to be biased for the more stereotypical sentence, the scores between the two models were too similar and no pattern distinguishing them could be found using either bias measure. We suspect this is due to the small size of the dataset. More complex factors could be at play as well, such as multilingual models benefiting from more multilingual data, in terms of gender bias mitigation, a hypothesis supported by the more negative S_{JSD} scores for monolingual models in Table 3.

Qualitative Analysis. In Table 2 the English CPS scores reproduce the systematic behavior from the CPS study (Nangia et al., 2020); however, this is not true for all languages, such as German. This may simply be a result of the much larger amount of English training data available to the models (Conneau et al., 2020). The nature of this additional data does not need to be biased in nature. A contributing factor may be that this additional data forces the tokenizer to adjust the tokens in its vocabulary to be whole words for English whereas for other languages tokens are more often subwords, which cannot be exchanged for others without forming grammatically incorrect sentences. To make this point, consider the following sentence pair from the dataset:

(6) a. Women are terrible at handiwork.
b. Men are terrible at handiwork.

Using mBERT, three tokens in the set U together form the composite word "handiwork" ("hand", "##i", "##work"); the remaining four are separate words and the period. However, it is almost trivial to predict any one of the subword tokens from the composite word, thus differences in MLM prediction probabilities may not be informative for detecting gender bias. In this case CPS assigns a sentence score of 1 and S_{JSD} $-.0075$. The value of $S_{\text{more}} - S_{\text{less}}$ for CPS is 1.29, placing it close to the decision boundary in Figure 2 and thus making CPS prone to noise.

For the German translation of the sentence, three

tokens in U are individual words or the period, while the remaining five form composite words. CPS assigns a sentence score of 0 and S_{JSD} $-.0135$. In this case $S_{\text{more}} - S_{\text{less}}$ for CPS is $-.98$, once again placing it close to the decision boundary in Figure 2.

Over the whole dataset, for German, 57% and 75% of tokens in the more stereotypical sentence were correctly predicted using mBERT and xlmR, respectively, whereas for English the prediction accuracy was lower at 56% and 68%, despite having more training data. Thus, compared to German, the CPS measure may be better suited for English, where individual tokens are not as trivial to predict and the CPS measure is not as prone to being influenced by noise from subword tokens.

6 Summary of Limitations

Our results indicate that S_{JSD} is superior to the original CrowS-Pairs measure. But like the CrowS-Pairs measure, S_{JSD} does not provide reliable measurements consistently. The most noticeable case of this is that for many models, "reverse bias" is well within the confidence interval of the bias measures, i.e., values below 50 are within the confidence interval for the CrowS-Pairs measure and positive values for S_{JSD}. We use reverse bias to refer to bias that is the opposite from the stereotype. Examples would include that the model favors women to be doctors and men to cry easily. While we did not confirm this experimentally, it seems not possible that a language model would learn a (spurious) stereotype even though the reverse of the stereotype dominates in the training corpus. Thus, this finding suggests that the measures must be interpreted with caution.

One of our original goals was a cross-lingual quantitative comparison of subtypes of gender bias. For example, maybe the "doctors are men" subtype of gender bias is less prevalent in Sweden than in Germany. Or the subtype "childcare is women's business" is stronger in Russia than in Canada. However, the two measures are not reliable on a sentence-pair by sentence-pair basis, so that one would need hundreds of examples of a subtype to make such inferences. This would require a dataset two orders of magnitude larger than the one we created.

We hypothesize that the main reason for the unreliability of the measures for individual sentence pairs is that predicting subwords is easy and not strongly linked to the difficulty of predicting a word; see " Qualitative Analysis" in the last section. Since most non-English languages will contain words broken into subwords in a given sentence pair, unrealistically high prediction accuracy and a lack of comparability of scores of a sentence pair across languages are the result.

7 Conclusion

In this paper, we developed a method for creating a multilingual gender bias diagnosis dataset that can be used across languages. Based on CrowS-Pairs (Nangia et al., 2020), we used this method to construct a multilingual gender bias diagnosis dataset for English, Finnish, German, Indonesian and Thai. Additionally, we proposed a new measure based on the Jensen–Shannon divergence from information theory, S_{JSD}, to study bias in MLMs using sentence pairs that contrast two groups. Using this measure we found that all studied models showed signs of gender bias for more stereotypical sentences across all five languages. Our hope is that our methods can be used for better evaluation of bias and debiasing in MLMs. We also hope that our work will foster more multilingual work on bias in language models.

In the future, since most recent bias research focused on PLMs and word embeddings, we plan to develop measures for downstream tasks as recommended by Blodgett et al. (2020) and Delobelle et al. (2021), which may be incorporated in a development pipeline when releasing models (Nozza et al., 2022).

8 Ethical Considerations

The dataset presented in this paper aims to make progress in the evaluation of multilingual gender bias in MLMs, however we argue that it should not be used to train such models. As the presented dataset is intended as a test set, training on it would defeat its purpose as a test of gender bias in MLMs. The presented dataset is based on the CPS dataset, an English crowdsourced dataset aimed at evaluating social biases in the United States (Nangia et al., 2020). For the purpose of this study we made the assumption that the biases in the CPS dataset relating to gender can be extended to the other languages studied and are relevant in cultures where the languages are spoken, however we caution against the blind implementation of such systems without an understanding of the target culture.

This work also focused exclusively on binary gender. The non-trivial nature of representing non-binary people in languages with strong gender agreement rules, such as German, substantially complicates the process of creating natural sentences that could be used for evaluation. For this reason, and because it is an important area with its own challenges, we leave the representation of non-binary people in multilingual settings to future work, where it can be studied with care as a topic in its own right.

Additionally, we caution against concluding that models are completely bias free when they generate scores that theoretically unbiased models are expected to generate. It may be that these models still encode biases that cannot be captured using the proposed measure or dataset, which may later manifest once a model is implemented.

Acknowledgements. We thank Sheng Liang and David Lowell for their thoughtful feedback and suggestions and the translators for their valuable insights into their native languages. This project was funded by a *2020 Award for Inclusion Research* from Google Research, the European Research Council (grant #740516) and the German Federal Ministry of Education and Research (BMBF, grant #01IS18036A).

References

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904, Online. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational 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 1941–1955, Online. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

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 Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in english, spanish, and arabic. *ArXiv*.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *ArXiv*.

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.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*. Number 57 in Monographs on statistics and applied probability. Chapman & Hall.

D.M. Endres and J.E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Unmasking the mask – evaluating social biases in masked language models. *ArXiv*.

Rakpong Kittinaradorn, Korakot Chaovavanich, Titipat Achakulvisut, Kittinan Srithaworn, Pattarawat Chormai, Chanwit Kaewkasi, Tulakan Ruangrong, and Krichkorn Oparad. 2019. DeepCut: A Thai word tokenization library using Deep Neural Network. *Zenodo*.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. 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.

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

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Challenges & Perspectives in Creating Large Language Models*.

Athanasios Papoulis and S. Unnikrishna Pillai. 2002. *Probability, random variables, and stochastic processes*, 4th ed edition. McGraw-Hill.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. *ArXiv*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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. 2018. 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.

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.

A Language Models

Model	Multilingual	Parameters
mBERT	yes	178M
xlmR (base)	yes	278M
xlmR (large)	yes	560M
BERT (uncased)	no	110M
RoBERTa	no	355M
ALBERT	no	206M

Table 4: Details of models used in this study.