A Comparative Study on the Impact of Model Compression Techniques on Fairness in Language Models

Krithika Ramesh
Microsoft Research
kramesh.tlw@gmail.com

Arnav Chavan*
Indian Institute of Technology, Dhanbad
arnavchavan04@gmail.com

Shrey Pandit*
BITS, Pilani
pandit.shrey.01@gmail.com

Sunayana Sitaram
Microsoft Research
sunayana.sitaram@microsoft.com

Abstract

Compression techniques for deep learning have become increasingly popular, particularly in settings where latency and memory constraints are imposed. Several methods, such as pruning, distillation, and quantization, have been adopted for compressing models, each providing distinct advantages. However, existing literature demonstrates that compressing deep learning models could affect their fairness. Our analysis involves a comprehensive evaluation of pruned, distilled, and quantized language models, which we benchmark across a range of intrinsic and extrinsic metrics for measuring bias in text classification. We also investigate the impact of using multilingual models and evaluation measures. Our findings highlight the significance of considering both the pre-trained model and the chosen compression strategy in developing equitable language technologies. The results also indicate that compression strategies can have an adverse effect on fairness measures.

1 Introduction

Despite their increasing popularity, machine learning models have been known to exhibit biases in their outputs, present privacy risks, and have potentially negative environmental consequences from their training and deployment. (Bender et al., 2021; Talat et al., 2022). Language models suffer from biases that result in unequal resource distributions (**allocational harms**), in addition to the undesired tendency to reproduce biases and stereotypes in content that is reflective of hegemonic worldviews (**representational harms**). Although measures have been proposed in tasks such as text classification (Czarnowska et al., 2021) to investigate the disparate allocational treatment of different classes, much of the research on fairness in language models centers on addressing representational harms (Blodgett et al., 2020). The potential of these models to further stigmatize marginalized communities is demonstrated in (Dressel and Farid, 2018), which illustrates how recidivism prediction systems are biased against black defendants, who have a higher baseline risk for repeat offences. Biases are also prevalent in computer vision applications such as facial recognition technologies. Within NLP, (Bolukbasi et al., 2016), one of the first forays that studied this phenomenon in language, noted that word embeddings contained stereotypical associations with respect to gender. Language models can exhibit biases toward different dialects for tasks like toxicity and hate speech detection (Garg et al., 2022; Sap et al., 2019), generate stereotypical representations and narratives (Lucy and Bamman, 2021), and are capable of the outright erasure of underrepresented identities (Dev et al., 2021). Compressed models that are biased may have detrimental consequences in the real world, as they are typically deployed on edge devices, which can further disadvantage communities without access to other forms of technology. Consequently, these issues have compelled a shift towards developing more inclusive systems.

Hooker et al. (2020) demonstrates how compression techniques, when applied to models that deal with tabular data, lead to the disparate treatment of less-represented classes. However, equivalent studies in NLP (Tal et al., 2022; Ahn et al., 2022; Silva et al., 2021) do not provide a conclusive observation as to whether compression methods are effective for reducing bias in NLP, and are centered mainly solely around model distillation being the compression technique of choice. This paper aims to resolve the following questions by benchmarking a wide range of metrics and datasets to study bias in text classification systems.

- How does model compression using pruning, quantization, or distillation impact bias in language models, and to what extent?

*Equal contribution

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15762–15782
July 9-14, 2023 ©2023 Association for Computational Linguistics

- To what extent are these observations influenced by variables such as the utilization of different techniques within a specific compression method or a change in model architecture or size?
- How does multilinguality affect these observations in compressed models?

2 Related Work

Compression techniques such as pruning, distillation, and quantization have proven effective at reducing the size of models while maintaining their performance. **Pruning** can be done in two ways, via structured and unstructured pruning. While structured pruning involves removing groups of neurons, unstructured pruning removes individual neurons by zeroing out their values. Structured pruning methods generally achieve faster inference speeds, along with a reduction in parameter size. **Knowledge distillation** techniques are another alternative that have been demonstrated to effectively transfer knowledge from a teacher model to a smaller student model, using a loss function designed to minimize the distance between the features or the outputs of the student and teacher models. We also incorporate a third form of model compression - **quantization**, where model weights and/or activations are represented using lower-bit precisions. There are two main approaches to quantization: post-training quantization, which is applied to a pre-trained model, and quantization-aware training (Zafrir et al., 2019a), which incorporates quantization into the training process in order to mitigate the loss of accuracy that can occur with post-training quantization. Although several techniques for pruning and quantization have been developed, we acknowledge that our work consists only of models compressed using post-training dynamic quantization and the pruning method proposed in Zafrir et al. (2021).

Whilst there has been research at the confluence of fairness and efficiency in natural language processing (NLP), the results from these studies can be inconclusive, limited in their research design, and at times, contradict the results from previous analyses. Talat et al. (2022); Orgad and Belinkov (2022); Field et al. (2021); Blodgett et al. (2020) provide critical insights into the current state of fairness in NLP and delve into the details of what research studies must consider when conducting work in this area. The discussion thus far concerning fairness, in general, has mainly been Anglo-centric, but recent forays (Kaneko et al., 2022; Huang et al., 2020b; Gonen et al., 2019; Zhao et al., 2020) have explored bias in multilingual spaces and languages beyond English.

In the context of model compression, Tal et al. (2022) show that while larger models produce fewer gendered errors, they produce a *greater proportion* of gendered errors in coreference resolution whilst Xu and Hu (2022) suggest that distillation and pruning have a regularizing effect that mitigates bias in text classification. On the other hand Silva et al. (2021); Ahn et al. (2022); Hessenthaler et al. (2022) all demonstrate how distillation can have an adverse impact on model fairness.

Hessenthaler et al. (2022) strongly casts doubt on the results from Xu and Hu (2022) by showing that knowledge distillation decreases model fairness. Additionally, the findings from Mohammadshahi et al. (2022) point toward the fact that pruning can amplify bias in multilingual machine translation models. It must also be noted that with the exception of Hessenthaler et al. (2022); Tal et al. (2022); Xu and Hu (2022); Mohammadshahi et al. (2022), many of these studies do not validate the fairness of these models over downstream tasks. This is essential as bias measurements over a model's pre-trained representations cannot be used as a proxy to assess bias in its downstream outputs (Goldfarb-Tarrant et al., 2021). Lauscher et al. (2021); Gupta et al. (2022) explore the efficient debiasing of models via the use of adapters and an adapted form of distillation, respectively.

To our knowledge, our work is the first comprehensive study on fairness in NLP with respect to pruning, distillation and quantization, in addition to which it addresses both monolingual and multilingual models.

3 Methodology and Setup

3.1 Pruning, Quantization, Distillation

Our pruning approach uses the Prune Once For All (Prune OFA) (Zafrir et al., 2021) method on the base models. The Prune OFA method is a state-of-the-art pruning strategy that prunes models during the pre-training phase, eliminating the need for additional pruning on downstream tasks.

We employ dynamic quantization (Zafrir et al., 2019b) as a post-training quantization method for fairness evaluation. This approach converts model weights to INT8 format post-training and dynamically quantizes activations during runtime based on the range of data. This method has the advantage of minimal hyperparameter tuning and additional flexibility in the model, which minimizes any potential performance loss.

For knowledge distillation, we consider models compressed using the techniques employed in (Sanh et al., 2019; Wang et al., 2020a), with the primary difference in these methods being the type of feature representations that the student is encouraged to mimic. We utilize pre-trained distilled models that are publicly available[1][2] for all of our experiments. The complete list of models we considered for these experiments is in the appendix (Table 9).

3.2 Fairness Evaluation in Language Models

To examine bias in LMs, we rely on a combination of **intrinsic** and **extrinsic** measures. Intrinsic measures primarily evaluate bias in the pre-trained representations of language models, such as in the static and contextualized embedding spaces. On the other hand, extrinsic measures estimate bias in the outputs produced by the LLM in the downstream task it is fine-tuned for. Extrinsic evaluation measures are capable of identifying both allocational and representational harms, while intrinsic measures only address the latter. The inconsistencies and lack of correlation between these two kinds of metrics (Goldfarb-Tarrant et al., 2021; Cao et al., 2022) has led to calls for better evaluation practices that prioritize extrinsic evaluation. We have included detailed explanations of the metrics and datasets in the next section and provided a broad overview and additional details in the appendix in Table 11.

4 Intrinsic measures

StereoSet (Nadeem et al., 2021) is an English dataset used for analyzing's a model's proclivity for stereotypical and anti-stereotypical data across the axes of gender, race, religion, and profession. We consider only the intrasentence samples from StereoSet and evaluate the test set split. The ICAT (Idealized Context Association Test) score combines both the language model score (LMS) and the stereotype score (SS) such that it is maximized when the model is unbiased and simultaneously proficient at language modeling as shown in Equation 1.

$$ICAT = LMS * \frac{\min(SS, 100 - SS)}{50} \quad (1)$$

Similar to StereoSet, **CrowS-Pairs** (Nangia et al., 2020) is a crowdsourced dataset that allows us to observe bias along the dimensions of gender, race, and religion. The distance between the stereotype and anti-stereotype pairs is kept to a minimum, and the metric involves the pseudo-log likelihood scoring mechanism from Salazar et al. (2020). However, both StereoSet and CrowS-Pair have been subject to critique for the inconsistencies in their datasets (Blodgett et al., 2021).

5 Extrinsic measures

For extrinsic measurement over downstream tasks, we have used multiple datasets with different fairness definitions (details in Table 11 in the appendix). The Jigsaw dataset is used to evaluate bias in toxicity detection systems across multiple demographic identities. We do this by assessing the difference in False Positive Rates (FPR) across subgroups to ensure that text from one group is not unfairly flagged as toxic. We report ROC-AUC as a metric on three specific subsets:

- **Subgroup AUC** : The test set is restricted to samples that mention the specific identity subgroup. A low value suggests that the model is ineffective at differentiating between toxic and non-toxic remarks that mention the identity.
- **BPSN AUC** (Background Positive, Subgroup Negative) : The test set is restricted to the non-toxic examples that mention the identity and the toxic examples that do not mention the identity. A low value suggests that the model predicts a higher toxicity score than it should for a non-toxic example mentioning the identity.
- **BNSP AUC** (Background Negative, Subgroup Positive) : The test set is restricted to the toxic examples that mention the identity and the non-toxic examples that do not mention the identity. A low value here indicates that the model predicts lower toxicity scores than it should for toxic examples mentioning the identity.

[1] https://huggingface.co
[2] https://github.com/microsoft/unilm/tree/master/minilm

The other monolingual extrinsic measure includes the **African American Vernacular English-Standard American English (AAVE-SAE)** dataset (Groenwold et al., 2020a), which consists of intent-equivalent SAE and AAVE sentence pairs. Sap et al. (2019) has shown that AAVE language is more likely to be identified as hate speech compared to the standardized form of American English. A fair, unbiased model on this data would produce similar sentiment scores for both AAVE and SAE. We have also included the results for the Equity Evaluation Corpus (EEC), a template-based dataset that evaluates the emotional intensity of sentiment classification systems over four categories of data- anger, fear, sadness, and joy, in the appendix (Section C.1).

5.1 Multilingual Datasets

To test if these observations are consistent with results across multilingual models, we use a binarized **hate speech detection** dataset, originally sourced from Huang et al. (2020a). It consists of online data accumulated from Twitter along with labels containing information pertinent to the user's age, gender, country, and race/ethnicity, and the details regarding the distribution of data and labels across languages are provided in the Appendix in Table 17. The fairness evaluation objective for the hate speech detection task involves measuring the **equality differences (ED)** metric across each of the groups corresponding to the aforementioned demographic factors. The **ED** is defined as the difference between the true positive/negative and false positive/negative rates for each demographic factor. For instance, the ED for false positive rates (FPED) is defined below, where d is representative of each demographic group within a demographic factor D (for example, gender is a demographic factor, and male is a corresponding representative demographic group).

$$FPED = \sum_{d \epsilon D} \|FPR_d - FPR\| \tag{2}$$

We also make use of **reviews datasets** sourced from Trustpilot, Yelp, and Amazon, with a rating (1-5) for each review (Hovy et al., 2015; Huang and Paul, 2019). The data includes user information, such as age, gender, and country (our analysis is constrained to gender). For this specific task, the dataset has been transformed into a binary sentiment analysis classification task, where reviews with a rating above 3 are classified as positive, and those with a rating below 3 are classified as negative. Reviews with a rating of 3 are discarded. As with the hate speech dataset, the **equality difference** metric is used to evaluate group fairness over this task along a given dimension.

6 Analysis of Results

6.1 StereoSet

The findings of the StereoSet evaluation are presented in Table 1, wherein a higher ICAT score implies a lesser biased model [3]. According to the results, the monolingual models' distilled and pruned versions exhibit more bias than their original counterparts. However, this trend does not necessarily apply to the multilingual or quantized versions of these models (Table 13). There is also an indication that the extent of pruning is potentially proportional to the negative impact on fairness in these models for this metric. Additionally, the MiniLM models, which employ a different distillation technique than the one used for DistilBERT, show a significant decrease in the ICAT score. However, it is worth noting that they are relatively smaller (MiniLMv2 being approximately one-third the size of DistilBERT). Among the three techniques, quantization appears to be the rank the lowest in terms of bias according to the intrinsic StereoSet measure. That said, these results may not accurately predict the model's performance in downstream tasks (Goldfarb-Tarrant et al., 2021). Based on the ICAT score measurement, the models distilled using MiniLMv2 exhibit the highest level of bias, while the quantized models demonstrate the best performance in this metric.

DistilBERT emerges as the least biased among the distilled models, while the quantized version of BERT-base shows the least bias among the quantized model sets. We highlight that while quantization results in a higher ICAT score for BERT, this is not the case for RoBERTa. Furthermore, although we have aggregated the scores for the dimensions of gender, race, and religion, these trends do not persist uniformly across individual dimensions. This observation is also reflected in our evaluation of the CrowS-Pair dataset.

[3] A green arrow indicates that the model is less biased in comparison to the parent model (in bold), while a red arrow indicates the opposite.

Model	Overall ICAT Score
bert-base-uncased	70.30
distilbert-base-uncased	69.52 ↓-0.78
miniLMv2-L6-H384-uncased	53.94 ↓-16.36
bert-base-uncased-90%-pruned	69.44 ↓-0.86
bert-base-uncased-85%-pruned	68.50 ↓-1.8
bert-base-uncased-quantized	72.06 ↑1.76
bert-base-multilingual-cased	64.94
distilbert-base-multilingual-cased	67.99 ↑3.05
xlm-roberta-large	71.29
multilingual-MiniLM-L12-H384	52.47 ↓-18.82
roberta-base	67.18
distilroberta	66.68 ↓-0.5
roberta-base-quantized	65.81 ↓-1.37
bert-large-uncased	69.50
miniLMv2-L6-H384-uncased	49.74 ↓-19.76
bert-large-uncased-90%-pruned	68.91 ↓-0.59
bert-large-uncased-quantized	70.20 ↑0.7

Table 1: We report the overall ICAT score for the model evaluations over the StereoSet dataset. The higher the ICAT score, the less biased the model.

Model	Gender	Race	Religion
bert-base-uncased	57.25 +7.25	62.33 +12.33	62.86 +12.86
distilbert-base-uncased	56.87 +6.87	60.97 +10.97	66.67 +16.67
miniLMv2-L6-H384-uncased	50.76 +0.76	50.68 +0.68	72.38 +22.38
bert-base-uncased-90%-pruned	51.91 +1.91	59.61 +9.61	60.95 +10.95
bert-base-uncased-85%-pruned	51.91 +1.91	53.01 +3.01	58.10 +8.10
bert-base-uncased-quantized	57.25 +7.25	62.14 +12.14	46.67 -3.33
bert-base-multilingual-cased	47.71 -2.29	44.66 -5.34	53.33 +3.33
distilbert-base-multilingual-cased	50.38 +0.38	41.94 -8.06	53.33 +3.33
xlm-roberta-large	54.41 +4.41	51.65 +1.65	69.52 +19.52
multilingual-MiniLM-L12-H384	39.85 -10.15	60.39 +10.39	47.62 -2.38
roberta-base	60.15 +10.15	63.57 +13.57	60.00 +10.00
distilroberta	52.87 +2.87	60.08 +10.08	63.81 +13.81
roberta-base-quantized	53.64 +3.64	58.53 +8.53	49.52 -0.48
bert-large-uncased	55.73 +5.73	60.39 +10.39	67.62 +17.62
miniLMv2-L6-H384-uncased	43.13 -6.87	50.1 +0.1	57.14 +7.14
bert-large-uncased-90%-pruned	54.20 +4.20	60.19 +10.19	69.52 +19.52
bert-large-uncased-quantized	50.38 +0.38	63.11 +13.11	55.24 +5.24

Table 2: The results for the CrowS-Pairs metric for different model families have been reported, with values closer to 50 indicating less biased models according to this metric.

6.2 CrowS-Pair

In Table 2, the results for CrowS-Pair have been presented for gender, race, and religion, along with the deviation from the ideal baseline score of 50. According to this metric, a higher magnitude of deviation indicates more bias in the model. Our findings reveal inconsistent disparities in the scores across different compression methods and their base and large counterparts. For example, while the results suggest that DistilBERT is less biased than BERT-base in terms of gender and race, this does not hold true for religion. While this may also be in due part to the relatively smaller sample size of the data for each dimension (Meade et al., 2022), it would be essential to understand if a model demonstrating lower bias in one dimension generalizes to other dimensions or data that incorporates intersectional identities. However, it is important to acknowledge that intrinsic and extrinsic measures do not necessarily correlate with each other. Additionally, Aribandi et al. (2021) highlights the substantial variance in likelihood-based and representation-based diagnostics during empirical evaluations, emphasizing the need for caution when interpreting findings from intrinsic measures.

6.3 Jigsaw

To evaluate the potential harm caused by these models, it is essential to assess bias in the context of downstream tasks. We fine-tuned the models on the Jigsaw dataset and examined how well they performed on various forms of protected identity mentions. Table 3 presents the aggregated scores for all subgroups across the metrics discussed in Section 5.[4]

The overall trend suggests that compression methods can have a negative impact on fairness. Distilled models generally appear to demonstrate a higher level of bias compared to their pruned and quantized counterparts. In contrast to the findings from intrinsic measurements, quantization does lead to a decrease in performance in these models, and this drop is also observed in the multilingual models. However, the pruned and quantized models generally exhibit a lower magnitude of bias compared to the distilled models.

Among all the compressed models evaluated, the base form of DistilBERT exhibits the highest degree of bias. These findings may vary at different training stages, and they warrant further probing to see if training the models further to improve the performance of these compressed models could also significantly contribute to reducing bias.

6.4 AAVE-SAE

Given the proclivity of hate speech detection systems to flag AAVE language as hate speech (Sap et al., 2019; Groenwold et al., 2020b), we aimed to assess whether SST-2 fine-tuned models also tend to classify AAVE language as negative. The underlying fairness objective in this context is to evaluate the robustness of sentiment analysis models to data from diverse dialects. We make use

[4]Results for the pruned version of BERT-large excluded due to low performance on Jigsaw and AAVE-SAE.

Model	Subgroup AUC	BPSN AUC	BNSP AUC
bert-base-uncased	0.918	0.934	0.975
distilbert-base-uncased	0.878 ↓-0.04	0.892 ↓-0.042	0.972 ↓-0.003
miniLM-L12-H384-uncased	0.917 ↓-0.001	0.943 ↑0.009	0.970 ↓-0.005
bert-base-uncased-90%-pruned	0.915 ↓-0.003	0.932 ↓-0.002	0.973 ↓-0.002
bert-base-uncased-85%-pruned	0.917 ↓-0.001	0.933 ↓-0.001	0.974 ↓-0.001
bert-base-uncased-quantized	0.917 ↓-0.001	0.933 ↓-0.001	0.974 ↓-0.001
bert-base-multilingual-cased	0.914	0.936	0.971
distilbert-base-multilingual-cased	0.895 ↓-0.019	0.913 ↓-0.023	0.969 ↓-0.002
xlm-roberta-base	0.914	0.942	0.969
multilingual-MiniLM-L12-H384	0.904 ↓-0.01	0.926 ↓-0.016	0.968 ↓-0.001
roberta-base	0.920	0.947	0.971
distilroberta	0.901 ↓-0.019	0.921 ↓-0.026	0.971 0
roberta-base-quantized	0.918 ↓-0.002	0.943 ↓-0.004	0.971 0
bert-large-uncased	0.913	0.922	0.975
bert-large-uncased-quantized	0.909 ↓-0.004	0.922 0	0.971 ↓-0.004

Table 3: We report the results for the Jigsaw dataset. The higher the AUC, the less biased the model. The scores for the identity subgroups have been aggregated and presented in this table.

Model	Negative to Positive	Positive to Negative	Total Changes
bert-base-uncased	238	89	327
distilbert-base-uncased	326 ↑88	76 ↓-13	402 ↑75
bert-base-uncased-90%-pruned	205 ↓-33	128 ↑39	333 ↑6
bert-base-uncased-85%-pruned	340 ↑102	147 ↑58	487 ↑160
bert-base-uncased-quantized	281 ↑43	93 ↑4	374 ↑47
xlm-roberta-base	247	56	303
multilingual-MiniLM-L12-H384	294 ↑47	73 ↑17	367 ↑64
roberta-base	241	102	343
distilroberta	238 ↓-3	108 ↑6	346 ↑3
roberta-base-quantized	207 ↓-34	115 ↑13	322 ↓-21
roberta-large	178	110	288
miniLM-L12-H384-uncased	265 ↑87	64 ↑-46	329 ↑41
bert-large-uncased	230	72	302
bert-large-uncased-quantized	175 ↓-55	156 ↑84	331 ↑29

Table 4: The results depict the count of non-concurrent predictions for the SST-2 fine-tuned models tested over the AAVE-SAE dataset.

of well-optimized, pre-trained models that were fine-tuned on the Stanford Sentiment Bank (SST-2) dataset (Socher et al., 2013), and we fine-tuned the pruned pre-trained models over SST-2. Additionally, we applied quantization techniques to the existing models and compared the outcomes of dynamically quantized models with other compressed variations. We examined the change in predictions when considering the AAVE intent-equivalent counterpart of the SAE language. We term the contradictory predictions of the classifier on AAVE-SAE sentence pairs as *non-concurrent predictions*, and our results are presented in Table 4.

A consistent pattern is observed where distilled models demonstrate a significantly higher degree of bias in this particular task than their base models. While the BERT-base pruned models also show a decline in performance, the 90% pruned version appears to be more robust than the 85% pruned version. Across all cases, except for the dynamically quantized form of RoBERTA-base, the quantized models show an increase in these non-concurrent predictions. Another interesting point of note is that several of these models seem to record positive to negative non-concurrent predictions when considering AAVE language instead of its SAE intent-equivalent counterpart.

7 Multilingual Datasets

To investigate whether the observed trends in a monolingual setting extend to a multilingual scenario, we conducted experiments using a separate set of models, with information about their size provided in Table 10 in the appendix. For these experiments, we employed the same techniques of pruning, distillation, and quantization as used in the monolingual experiments.

7.1 Hate Speech Detection

The hate speech dataset evaluation results are presented in Table 5 and Table 7. In contrast to the trends observed in the monolingual evaluations conducted for English, the impact on fairness, as measured by the equality differences (**ED**) metric, is not as consistently evident among the compressed models in the multilingual setup. In the quantized and distilled models, the trends with respect to English remain consistently negative.

The training for all these models was constrained to 5 epochs, and the F1 and AUC scores for the base models are lower than their compressed counterparts. The compressed models demonstrate greater performance gains within the same training duration as compared to their base forms, and this observed improvement in performance could contribute to enhanced fairness outcomes as well.

Furthermore, it is worth considering that in previous monolingual tasks and even in the multilingual evaluation of Trustpilot reviews (Table 8), the compressed models were more likely to experience a drop in the ED metric. However, it is essential to highlight that the magnitude of this drop observed in the current results is considerably less pronounced. Additionally, the F1 and AUC performance of these models over these datasets is significantly higher.

Across nearly all the experiments conducted and languages documented in Tables 5, 7, and 8,

Model	Language	AUC	F1-macro	Age	Gender
bert-base-multilingual-cased	English	0.743	0.645	0.110	0.043
	Italian	0.662	0.509	0.064	0.070
	Polish	0.735	0.648	0.302	0.266
	Portuguese	0.616	0.539	0.194	0.181
	Spanish	0.676	0.618	0.177	0.179
distilbert-base-multilingual-cased	English	0.790	0.702	0.199 ↑+0.089	0.084 ↑+0.041
	Italian	0.673	0.551	0.123 ↑+0.059	0.102 ↑+0.032
	Polish	0.706	0.638	0.264 ↓-0.038	0.249 ↓-0.017
	Portugese	0.651	0.513	0.031 ↓-0.163	0.173 ↓-0.008
	Spanish	0.695	0.617	0.134 ↓-0.043	0.135 ↓-0.044
bert-base-multilingual -cased-quantized	English	0.750	0.641	0.141 ↑+0.031	0.080 ↑+0.037
	Italian	0.675	0.509	0.089 ↑+0.025	0.078 ↑+0.008
	Polish	0.735	0.628	0.314 ↑+0.012	0.242 ↓-0.024
	Portuguese	0.602	0.493	0.191 ↓-0.003	0.026 ↓-0.155
	Spanish	0.670	0.613	0.217 ↑+0.040	0.173 ↓-0.006
bert-base-multilingual-cased-90%-pruned	English	0.813	0.708	0.135 ↑+0.025	0.075 ↑+0.032
	Italian	0.666	0.537	0.150 ↑+0.086	0.238 ↑+0.168
	Polish	0.698	0.580	0.221 ↓-0.081	0.230 ↓-0.036
	Portuguese	0.697	0.540	0.209 ↑+0.015	0.054 ↓-0.127
	Spanish	0.659	0.616	0.185 ↑+0.008	0.150 ↓-0.029
bert-base-multilingual-cased-50%-pruned	English	0.764	0.657	0.078 ↓-0.032	0.048 ↑+0.005
	Italian	0.648	0.553	0.168 ↑+0.104	0.178 ↑+0.108
	Polish	0.711	0.622	0.245 ↓-0.057	0.233 ↓-0.033
	Portuguese	0.644	0.505	0.115 ↓-0.079	0.108 ↓-0.073
	Spanish	0.684	0.625	0.246 ↑+0.069	0.085 ↓-0.094
bert-base-multilingual-cased-10%-pruned	English	0.745	0.644	0.089 ↓-0.021	0.051 ↑+0.008
	Italian	0.670	0.565	0.210 ↑+0.146	0.260 ↑+0.190
	Polish	0.670	0.597	0.160 ↓-0.142	0.167 ↓-0.099
	Portuguese	0.590	0.480	0.142 ↓-0.052	0.048 ↓-0.133
	Spanish	0.681	0.620	0.347 ↑+0.170	0.188 ↑+0.009
xlm-roberta-large	English	0.529	0.218	0.005	0.004
	Italian	0.629	0.549	0.246	0.119
	Polish	0.580	0.520	0.080	0.067
	Portuguese	0.447	0.398	0.126	0.045
	Spanish	0.590	0.556	0.251	0.088
multilingual-MiniLM-L12-H384	English	0.701	0.605	0.060 ↑+0.055	0.032 ↑+0.028
	Italian	0.622	0.571	0.337 ↑+0.091	0.191 ↑+0.072
	Polish	0.643	0.587	0.138 ↑+0.058	0.098 ↑+0.031
	Portuguese	0.606	0.559	0.336 ↑+0.210	0.237 ↑+0.192
	Spanish	0.624	0.570	0.270 ↑+0.019	0.096 ↑+0.008

Table 5: The results for the age and gender categories of the Hate Speech dataset. The lower the ED, the less biased the model.

Language	Race	Country	Age	Gender
English	distilbert-base-multilingual-cased	distilbert-base-multilingual-cased	distilbert-base-multilingual-cased	distilbert-base-multilingual-cased
Italian	-	-	bert-base-multilingual-cased-10%-pruned	bert-base-multilingual-cased-10%-pruned
Spanish	multilingual-MiniLM-L12-H384	bert-base-multilingual-cased-90%-pruned	bert-base-multilingual-cased-10%-pruned	bert-base-multilingual-cased-10%-pruned
Portuguese	multilingual-MiniLM-L12-H384	multilingual-MiniLM-L12-H384	multilingual-MiniLM-L12-H384	multilingual-MiniLM-L12-H384
Polish	-	-	multilingual-MiniLM-L12-H384	multilingual-MiniLM-L12-H384

Table 6: The list of compressed models which demonstrate the sharpest increase in the ED metric relative to their base model.

Model	Language	AUC	F1-macro	Race	Country
bert-base-multilingual-cased	English	0.743	0.645	0.059	0.031
	Portuguese	0.616	0.539	0.200	0.109
	Spanish	0.676	0.618	0.087	0.130
distilbert-base-multilingual-cased	English	0.790	0.702	0.086 ↑+0.027	0.077 ↑+0.046
	Portuguese	0.651	0.513	0.105 ↓-0.095	0.089 ↓-0.020
	Spanish	0.695	0.617	0.089 ↑+0.002	0.127 ↓-0.003
bert-base-multilingual-cased-quantized	English	0.750	0.641	0.066 ↑+0.007	0.043 ↑+0.012
	Portuguese	0.602	0.493	0.069 ↓-0.131	0.037 ↓-0.072
	Spanish	0.670	0.613	0.039 ↓-0.048	0.149 ↑+0.019
bert-base-multilingual-cased-90%-pruned	English	0.813	0.708	0.041 ↓-0.018	0.026 ↓-0.005
	Portuguese	0.697	0.540	0.151 ↓-0.049	0.106 ↓-0.003
	Spanish	0.659	0.616	0.033 ↓-0.054	0.289 ↑+0.159
bert-base-multilingual-cased-50%-pruned	English	0.764	0.657	0.038 ↓-0.019	0.020 ↓-0.011
	Portugese	0.644	0.505	0.086 ↓-0.114	0.118 ↑+0.009
	Spanish	0.684	0.625	0.092 ↑+0.005	0.217 ↑+0.087
bert-base-multilingual-cased-10%-pruned	English	0.745	0.644	0.024 ↓-0.025	0.009 ↓-0.022
	Portuguese	0.590	0.480	0.193 ↓-0.007	0.024 ↓-0.085
	Spanish	0.681	0.620	0.130 ↑+0.043	0.249 ↑+0.119
xlm-roberta-large	English	0.529	0.218	0.005	0.003
	Portuguese	0.447	0.398	0.121	0.175
	Spanish	0.590	0.556	0.030	0.376
multilingual-MiniLM-L12-H384	English	0.701	0.605	0.011 ↑+0.006	0.027 ↑+0.024
	Portuguese	0.606	0.559	0.263 ↑+0.142	0.232 ↑+0.057
	Spanish	0.624	0.570	0.097 ↑+0.067	0.383 ↑+0.007

Table 7: The results for the race/ethnicity and country categories of the Hate Speech dataset. The lower the ED, the less biased the model.

the MiniLM model distilled from XLM-R Large demonstrates higher levels of bias compared to the base model. These results also exhibit variations across languages and dimensions under consideration. A model may produce fairer outcomes for data in one language but not necessarily generalize to another language or dimension. Additionally, the trends observed in the ED values for pruning the multilingual BERT-base model are not consistently monotonic. We have included the results for the most significant decrease in magnitude across each dimension and language for these experiments in Table 6. Our benchmarking of these compressed models indicates that various elements in the experimental setup, such as the selection of techniques within a given compression method or the choice of pre-trained model architecture, are likely to have consequences in the measurements we observe.

7.2 Trustpilot Reviews Dataset

We also fine-tuned these models using a dataset comprising Trustpilot reviews from four different languages. The results for the equality difference (ED) for gender are presented in Table 8. Although the compressed models generally exhibit poorer performance in terms of their overall equality difference, the magnitude of the difference in ED between the compressed models and their base forms is considerably smaller compared to the values observed in the previous task. However, it is worth noting that the results for the English reviews dataset (Table 12 in the appendix) contradict this pattern. In that case, the compressed versions of BERT demonstrate less bias, whereas the opposite is true for XLM-R Large.

8 How Does Model Compression Affect Fairness?

8.1 Distillation, Pruning and Quantization

The claim that distillation tends to amplify biases in models aligns with our findings in monolingual evaluation experiments. However, the impact on fairness metrics can vary, and this pattern does not necessarily hold true in multilingual settings, as evidenced by our evaluation of multilingual fairness datasets. Similar observations can be made regarding pruned models, although further investigation is warranted to understand how different pruning strategies and levels of pruning may influence these effects.

In contrast, our approach of post-training quantization has yielded more diverse outcomes. While its impact on fairness may be relatively less pronounced, it can sometimes lead to impractical models for downstream tasks due to their low perfor-

Model	Language	F1-W Avg	AUC-W Avg	Total ED
bert-base-multilingual-cased	English	0.981	0.987	0.026
	French	0.976	0.990	0.022
	German	0.979	0.985	0.014
	Danish	0.971	0.992	0.015
distilbert-base-multilingual-cased	English	0.975	0.987	0.026 0.000
	French	0.971	0.984	0.037 ↑+0.015
	German	0.976	0.977	0.043 ↑+0.029
	Danish	0.964	0.992	0.020 ↑+0.005
bert-base-multilingual-cased-quantized	English	0.978	0.984	0.047 ↑+0.021
	French	0.969	0.984	0.048 ↑+0.026
	German	0.976	0.980	0.005 ↓-0.009
	Danish	0.970	0.991	0.021 ↑+0.006
bert-base-multilingual-cased-90%-pruned	English	0.976	0.988	0.029 ↑+0.003
	French	0.973	0.986	0.036 ↑+0.014
	German	0.975	0.982	0.025 ↑+0.011
	Danish	0.963	0.991	0.024 ↑+0.009
bert-base-multilingual-cased-50%-pruned	English	0.980	0.989	0.020 ↓-0.006
	French	0.975	0.989	0.038 ↑+0.016
	German	0.977	0.988	0.025 ↑+0.011
	Danish	0.970	0.991	0.019 ↑+0.004
bert-base-multilingual-cased-10%-pruned	English	0.979	0.988	0.026 0.000
	French	0.976	0.988	0.028 ↑+0.006
	German	0.976	0.981	0.017 ↑+0.003
	Danish	0.969	0.993	0.017 ↑+0.002
xlm-roberta-large	English	0.987	0.993	0.018
	French	0.984	0.991	0.024
	German	0.985	0.992	0.031
	Danish	0.985	0.994	0.008
multilingual-MiniLM-L12-H384	English	0.976	0.991	0.042 ↑+0.024
	French	0.972	0.989	0.041 ↑+0.017
	German	0.975	0.986	0.023 ↓-0.008
	Danish	0.970	0.993	0.017 ↑+0.009

Table 8: The results for the gender category of the Trustpilot Reviews dataset. The lower the ED, the less biased the model.

mance. Therefore, careful consideration is required when employing post-training quantization to strike a balance between fairness and task effectiveness.

8.2 Multilingual vs Monolingual Models

While monolingual evaluation generally negatively impacts fairness, the same cannot be said for multilingual evaluation, which varies across languages and dimensions. It would be valuable to investigate the underlying causes for the decrease in fairness during compression and explore its relationship with the multilingual and monolingual aspects of the model. It also remains to be seen whether well-optimized models for a specific task are more prone to demonstrating increased bias in their compressed versions, thereby possibly relying on unfair associations to make predictions.

8.3 Additional Considerations

There are still lingering questions regarding the influence of various elements, such as model size, architecture choices, different variants of compression techniques, and their impact on our evaluations. While our results seem to indicate otherwise for some of these parameters (such as size), it is essential to explore whether these observations translate across different tasks. As evinced by Tal et al. (2022), the size of a model does not necessarily correlate with reduced biases, a notion that is further supported by our own findings. It would be worthwhile to extensively examine how these models are affected when different compression methods are combined or constrained to the same parameter count.

9 Conclusion

In this work, we conduct a comprehensive evaluation of fairness in compressed language models, covering multiple base models, compression techniques, and various fairness metrics. While prior studies have evaluated the fairness of compressed models, the results have not always been conclusive. In contrast, our extensive benchmarking provides evidence that challenges recent research suggesting that model compression can effectively reduce bias through regularization, and we demonstrate that this is the case for both multilingual and monolingual models across different datasets.

The compression of language models through distillation, quantization, and pruning is crucial for the practical use of language technologies in various real-world applications. While it is essential to preserve performance during compression, it is equally imperative that the compressed versions of language models maintain or even enhance fairness measures to avoid potential harm.

10 Ethics Statement

Our results indicate that compression does harm fairness, particularly in the monolingual setting. The potential harm that the system may cause and the application it will be used for should be considered when selecting a model compression technique, in addition to factors like accuracy, latency, and size. Although we have not observed absolute trends across models, datasets, and compression techniques, it is especially crucial to evaluate compressed models for fairness and accuracy before deployment and, on a broader note, to understand why compressed models might exhibit issues with respect to fairness.

In our paper, we conducted evaluations of multilingual language models using fairness metrics for various languages, including English. We observed varying trends regarding their performance on fairness metrics across different languages. However, it is vital to consider the potential influence of the lack of well-optimized models for these specific tasks, which may mitigate some of these issues. Additionally, evaluation datasets are scarce for assessing bias in languages other than English and for different fairness definitions. We also acknowledge that fairness trends identified in English evaluations may not necessarily be true for all languages.

While our benchmarking encompassed multiple intrinsic and extrinsic metrics, it is important to acknowledge their limitations in capturing all dimensions of fairness. Further research is needed to develop comprehensive extrinsic metrics across diverse tasks. Although our work has been centered around fairness in allocation-based (classification) applications, addressing fairness concerns in other types of language models, such as natural language generation models, is necessary. In generative tasks, the measurement of unfair outcomes would be distinct from the methods we have used. Another area of potential future work could involve benchmarking debiasing methods for compressed models and developing new compression-aware methods.

Limitations

The primary motivation behind this paper was to provide a comprehensive benchmarking study that explores the impact of model compression techniques on bias in large language models. While our work is among the first efforts to address fairness in compressed language models across multiple compression methods, including exploring multilingual settings, we are aware of the inherent limitations associated with our benchmarking study. Some of the limitations and potential directions for future work that builds on our study include the following:

- Our study primarily focused on benchmarking pre-trained models and evaluating their performance in the downstream text classification task. Expanding our investigation to encompass other tasks, particularly those involving generative models or large language models (LLMs), would be a valuable contribution to the research community. Examining the impact of model compression techniques on fairness in these domains would provide further insights and contribute to a more comprehensive understanding of bias in different types of language models.

- While our work includes a multilingual evaluation component, we acknowledge that there is room for further improvement and comprehensiveness in our benchmarking study, particularly with regard to quantization and pruning techniques. Apart from this, we did not provide a comparative analysis of monolingual and multilingual models using the same extrinsic data, which could provide valuable insights into the disparate impact of compression on the bias across languages. These are potential areas for future research that could contribute to a more thorough understanding of bias in compressed language models.

- Despite showing results for state-of-the-art pruning methods, further benchmarking is necessary to observe how bias varies across different pruning techniques. Similarly, whilst our method serves as a proxy to estimate bias trends in quantized models, a thorough quantization-specific study is needed.

- Different compression strategies yield varied benefits in terms of latency, memory, and so forth. Investigating the tradeoffs between these elements and fairness and accuracy would yield valuable insights for obtaining realistic estimations in real-world scenarios. Additionally, conducting case-study analyses would give practitioners in the field a deeper understanding of the potential harm these methods may introduce.

References

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, 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.

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.

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. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4:eaao5580.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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 1905–1925, Online. Association for Computational Linguistics.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling bias in toxic speech detection: A survey.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020a. Investigating african-american vernacular english in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020b. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723, Seattle, United States. Association for Computational Linguistics.

Xiaolei Huang, Xing Linzi, Franck Dernoncourt, and Michael J. Paul. 2020a. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelveth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

Xiaolei Huang and Michael J. Paul. 2019. Neural user factor adaptation for text classification: Learning to generalize across author demographics. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 136–146, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020b. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Jigsaw.

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.

Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2021. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. 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.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. What do compressed multilingual machine translation models forget?

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.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020a. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2020b. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6151–6162.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, 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, Rémi 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 M. 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.

Guangxuan Xu and Qingyuan Hu. 2022. Can model compression improve nlp fairness. *ArXiv*, abs/2201.08542.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019a. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 36–39. IEEE.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019b. Q8BERT: quantized 8bit BERT. *CoRR*, abs/1910.06188.

Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. *arXiv preprint arXiv:2111.05754*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

A Appendix

A.1 Methodology and Setup

A.1.1 Pruning

We adopt the **Prune Once For All** or **Prune OFA** method (Zafrir et al., 2021) as our central pruning strategy. Prune OFA has demonstrated state-of-the-art performance in terms of compression-to-accuracy ratio for BERT-based models, and it also eliminates the need to conduct task-specific pruning, as the sparse pre-trained language model can be directly fine-tuned on the target task. This simplifies our comparisons, as the same pruned model can be fine-tuned on different datasets.

A.1.2 Distillation

We use the pre-trained distilled variants of base models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2019), namely DistilBERT (Sanh et al., 2019), DistilRoBERTa, and multilingual MiniLM (Wang et al., 2020a), which are publicly available through the HuggingFace API (Wolf et al., 2020) for our experiments. DistilBERT selects one layer from each pair of alternate layers in the teacher architecture (BERT-base), lowering the number of layers in the distilled model by half. MiniLM is distilled from the final attention layer of the teacher model, thus making this knowledge distillation method task-independent. In addition to evaluating bias in these pre-trained models using intrinsic metrics, we fine-tuned some distilled models on the SAE-AAVE, Jigsaw, and Equity Evaluation Corpus (EEC) datasets for evaluation using extrinsic metrics.

A.1.3 Quantization

Dynamic quantization is particularly effective when the time required to execute a model is dominated by loading weights from memory rather than computing matrix multiplications, as with transformer models. Therefore, we adopt dynamic quantization in all of our experiments. With this approach, model parameters are converted to INT-8 format post-training, and the scale factor for activations is dynamically determined based on the range of the data observed at runtime, which helps to maintain flexibility in the model and minimize any loss in performance. Additionally, dynamic quantization requires minimal hyperparameter tuning and is easy to deploy in production.

A.2 Further Details on Pruning, Quantization and Distillation

A.2.1 Pruning

Neural architecture pruning aims at eliminating redundant parts of neural networks while maintaining model performance. Unstructured pruning removes individual neurons by setting the value of these parameters to zero, whereas structured pruning removes groups of neurons such as layers, attention heads, and so forth. (Sanh et al., 2020) presents a form of unstructured weight pruning in which individual weights can be eliminated to create a sparse network. Although massive reductions in the parameter count are observed, the inference speeds show no such improvement. On the other hand, structured pruning methods (Wang et al., 2020b) achieve faster inference speeds along with a reduction in parameter size. (Lagunas et al., 2021) extend the work of movement pruning to the structured and semi-structured domains. Recently, (Zafrir et al., 2021) showed that integrating pruning during the pre-training of language models gives high-performing sparse pre-trained models, thus removing the burden of pruning for a specific downstream task.

A.2.2 Distillation

Knowledge distillation (KD) (Hinton et al., 2015) has been shown to effectively transfer knowledge from a teacher model to a smaller student model, with a loss function designed to minimize the distance between the features or the outputs of the student and teacher models. Numerous alterations can be made to the KD setup, such as choosing intermediate layers of the teacher model for initializing the student architecture (Sanh et al., 2019), distilling the final attention layer of the teacher transformer architecture (Wang et al., 2020a), introducing bottlenecks for distillation (Sun et al., 2020). However, biases in the teacher model could potentially propagate into the distilled models making it more biased compared to the original teacher model (Silva et al., 2021).

A.2.3 Quantization

Quantization compresses models by representing model weights and/or activations with lower bit precisions. It can also make it possible to carry out inference using integer-only operations, as demonstrated by Kim et al. (2021). There are two main approaches to quantization: post-training quantization, which is applied to a pre-trained model, and quantization-aware training (Zafrir et al., 2019a), which incorporates quantization into the training process in order to mitigate the loss of accuracy that can occur with post-training quantization.

B Additional Results

We have included the results and a brief description for certain monolingual and multilingual measures below. Our decision to include the Equity Evaluation Corpus (EEC) and Log Probability Bias Score (LPBS) metric measures in the appendix is motivated by the fact that both these metrics consist of template-based data lacking concrete fairness objectives, and are therefore not a reflection of harms that can be caused in real-world applications. Recent research (Alnegheimish et al., 2022) has effectively highlighted the sensitivity of template-based evaluations to the selection and design of templates, which can bias the results. Furthermore, the LPBS is an intrinsic measure, and Aribandi et al. (2021) addresses the instability of likelihood and representation-based model diagnostic measures. Therefore, we advise readers to exercise caution when drawing conclusions from these results.

B.1 Multilingual Datasets

The findings of our multilingual evaluation on the English reviews dataset, comprising reviews obtained from platforms such as Amazon and Yelp, have been presented in Table 12. Additionally, we have included the results of the evaluation of multilingual models on the English version of StereoSet in Table 13, as well as the evaluation of these models for the Crows-Pair dataset for English in Table 14. Ideally, we intended to study the performance of models on datasets with comparable fairness notions or objectives in both monolingual and multilingual contexts. Unfortunately, we encountered limitations in sourcing such datasets, and therefore, we leave this as an avenue for future research.

C Fine-tuning Setup

For the extrinsic measures, the models are fine-tuned over a specific training dataset before the fairness evaluation is carried out over the test set. Most of our fine-tuning setups have been derived from previous work (Huang et al., 2020a; Huang, 2022; Câmara et al., 2022). The intrinsic measures do not require a hyperparameter search, as they are evaluated over pre-trained model representations. For the extrinsic measures, we relied on pre-trained SST-2 fine-tuned models available on HuggingFace. We performed fine-tuning solely for all the models trained on the Jigsaw Toxicity Classification dataset, the final details of which are as follows:

- Batch Size : 16
- Learning Rate : 1e-4
- Weight decay: 0.01
- Warmup Ratio: 0.06
- Epochs: 5
- Optimizer: AdamW

C.1 Equity Evaluation Corpus

The Equity Evaluation Corpus (Kiritchenko and Mohammad, 2018) is a template-based corpus for evaluating sentiment analysis systems for emotional intensity across four categories (joy, sadness, anger and joy). In this particular task, we measure the Pearson Correlation Coefficient (PCC) of the predictions of these models against the gold label. the It must be noted that previous research (Alnegheimish et al., 2022) indicates that bias evaluation is sensitive to design choices in template-based data, and that evaluating our models over natural sentence-based datasets would be a better alternative to gauge the impact these models can have. The fairness objective here looks to address the disparity in terms of the PCC across all the models across the different categories of template data. The results have been reported in Table 15.

C.2 Log Probability Bias Score

The Log Probability Bias Score (LPBS) (Kurita et al., 2019) was proposed as a modification to the DisCo metric (Webster et al., 2020). LPBS operates similarly to WEAT, using template sentences (e.g., '[TGT] likes to [ATT]') in which TGT represents a list of target words and ATT represents a list of attributes for which we aim to measure biased associations. The test also accounts for the prior probability of the target attribute, allowing us to evaluate bias solely based on the attributes without being influenced by the prior probability of the target token. The attribute categories that we have taken into consideration are a list of professions, positive words, and negative words (Bartl et al., 2020) (Kurita et al., 2019). The results have been reported in Table 16.

Model Name	Parameter #	Jigsaw	EEC	AAVE-SAE	StereoSet	CrowS-Pair	LPBS
bert-base-uncased	110 M	✓	✓	✓	✓	✓	✓
distilbert-base-uncased	66 M	✓	✓	✓	✓	✓	✓
miniLM-L12-H384-uncased	33 M	✓	✓	✓	✗	✗	✗
bert-base-uncased-85%-pruned	16.5 M	✓	✓	✓	✓	✓	✓
bert-base-uncased-90%-pruned	11 M	✓	✓	✓	✓	✓	✓
bert-base-uncased-quantized	110 M	✓	✓	✓	✓	✓	✓
bert-large-uncased	340 M	✓	✓	✓	✓	✓	✓
bert-large-uncased-90%-pruned	34 M	✗	✓	✗	✓	✓	✓
bert-large-uncased-quantized	340 M	✓	✓	✓	✓	✓	✓
bert-base-multilingual-cased	178 M	✓	✓	✗	✓	✓	✓
distilbert-base-multilingual-cased	135 M	✓	✓	✗	✓	✓	✓
xlm-roberta-large	560 M	✗	✗	✗	✓	✓	✓
multilingual-MiniLM-L12-H384 [xlm-roberta-large]	117 M	✗	✗	✗	✓	✓	✓
xlm-roberta-base	278 M	✓	✓	✓	✗	✗	✗
multilingual-MiniLM-L12-H384 [xlm-roberta-base]	117 M	✓	✓	✓	✗	✗	✗
roberta-base	125 M	✓	✓	✓	✓	✓	✓
distilroberta	82 M	✓	✓	✓	✓	✓	✓
roberta-base-quantized	125 M	✓	✓	✓	✓	✓	✓

Table 9: Details about the models and which metrics they were evaluated for in the monolingual fairness experiments. The parameter counts for the pruned models indicates the total number of non-sparse parameters. Some of the models could not be evaluated for the intrinsic measures due to their architectural setup.

Model Name	Parameter #
bert-base-multilingual-cased	178 M
distilbert-base-multilingual-cased	135 M
bert-base-multilingual-cased-10%-pruned	160 M
bert-base-multilingual-cased-50%-pruned	89 M
bert-base-multilingual-cased-90%-pruned	17 M
bert-base-multilingual-cased-quantized	178 M
xlm-roberta-large	560 M
multilingual-MiniLM-L12-H384	117 M

Table 10: Parameter count for all the models used for the multilingual fairness evaluation experiments. The parameter counts for the pruned models indicates the total number of non-sparse parameters. These models have been used uniformly for all the multilingual datasets.

Metric	Type of Metric	Downstream Task	Template-Based	Fairness Objective	Dimensions
Monolingual					
Jigsaw Toxicity Unintended Bias	Extrinsic	Toxicity Detection	No	Increased likelihood of being classifying comment as toxic based on identity group mentions	Multiple [Gender, Religion, Race/Ethnicity, Sexual Orientation, Disability, etc]
AAVE-SAE	Extrinsic	Sentiment Classification	No	Increased likelihood of being classifying comment as negative based on dialect used	Dialect
EEC	Extrinsic	Sentiment Classification	Yes	Difference in emotion categories for emotional intensity prediction	Emotional Intensity
StereoSet	Intrinsic	N/A	No	Evaluation of model preference for stereotypical sentences	Gender, Race/Ethnicity, Religion, Profession
CrowS-Pair	Intrinsic	N/A	No	Evaluation of model preference for stereotypical sentences	Gender, Race/Ethnicity, Religion
LPBS	Intrinsic	N/A	Yes	Evaluation of model preference for stereotypical associations	Gender
Multilingual					
Hate Speech	Extrinsic	Hate Speech Detection	No	Measuring performance across data based on the demographic groups they are sourced from	Age, Gender, Country, Race/Ethnicity
Reviews Dataset	Extrinsic	Sentiment Classification	No	Measuring performance across data based on the demographic groups they are sourced from	Gender

Table 11: List of all the details pertaining to the fairness metrics used.

Model	F1-W Avg	AUC-W Avg	Total ED
bert-base-multilingual-cased	0.872	0.916	0.499
distilbert-base-multilingual-cased	0.868	0.914	0.350 ↑-0.149
bert-base-multilingual-cased-quantized	0.854	0.892	0.317 ↑-0.182
bert-base-multilingual-cased-10%-pruned	0.869	0.921	0.258 ↑-0.241
bert-base-multilingual-cased-50%-pruned	0.865	0.918	0.313 ↑-0.186
bert-base-multilingual-cased-90%-pruned	0.862	0.910	0.442 ↑-0.057
xlm-roberta-large	0.908	0.947	0.290
multilingual-MiniLM-L12-H384-distilled-XLMR-Large	0.839	0.898	0.402 ↓+0.112
xlm-roberta-large-quantized	0.865	0.928	0.474 ↓+0.184
xlm-roberta-base	0.787	0.900	0.349 ↓+0.059

Table 12: We report the performance of multilingual models and the ED (equality differences) fairness estimate over a set of English reviews sourced from websites such as Amazon, Yelp, etc. The higher the ED, the less fair the model.

Model	Overall ICAT Score
bert-base-multilingual-cased	64.94
distilbert-base-multilingual-cased	67.99 ↑+3.05
bert-base-multilingual-cased-quantized	64.78 ↓-0.16
bert-base-multilingual-cased-10%-pruned	67.82 ↑+2.88
bert-base-multilingual-cased-50%-pruned	66.67 ↑+1.73
bert-base-multilingual-cased-90%-pruned	67.00 ↑+2.06
xlm-roberta-large	71.29
multilingual-MiniLM-L12-H384	52.47 ↓-18.82
xlm-roberta-large-quantized	69.63 ↓-1.66

Table 13: The overall ICAT score for the multilingual models for the StereoSet (English) dataset. The higher the ICAT score, the less biased the model.

Model	Gender	Race	Religion
bert-base-multilingual-cased	47.71 -2.29	44.66 -5.34	53.33 +3.33
distilbert-base-multilingual-cased	50.38 +0.38	41.94 -8.06	53.33 +3.33
bert-base-multilingual-cased-quantized	52.29 +2.29	42.72 -7.28	52.38 +2.38
bert-base-multilingual-cased-10%-pruned	47.71 -2.29	47.57 -2.43	58.1 +8.1
bert-base-multilingual-cased-50%-pruned	49.24 -0.76	48.54 -1.46	56.19 +6.19
bert-base-multilingual-cased-90%-pruned	50.0 0	57.48 +7.48	53.33 +3.33
xlm-roberta-large	54.41 +4.41	51.65 +1.65	69.52 +19.52
multilingual-MiniLM-L12-H384	39.85 -10.15	60.39 +10.39	47.62 -2.38
xlm-roberta-large-quantized	52.87 2.87	57.28 +7.28	71.43 +21.43

Table 14: The results for the CrowS-Pairs metric for multilingual models have been reported, with values closer to 50 indicating less biased models according to this metric.

Model	Joy	Sadness	Anger	Fear
bert-base-uncased	0.600	0.533	0.557	0.552
distilbert-base-uncased	0.623	0.587	0.623	0.565
distilbert-base-uncased-60%-pruned	0.586	0.551	0.585	0.540
miniLM-L12-H384-uncased	0.352	0.195	0.230	0.245
miniLM-L12-H384-uncased-70%-pruned	0.600	0.539	0.573	0.547
bert-base-uncased-85%-pruned	0.550	0.432	0.464	0.478
bert-base-uncased-90%-pruned	0.523	0.418	0.450	0.472
bert-base-uncased-quantized	0.455	0.382	0.383	0.410
bert-base-multilingual-cased	0.506	0.386	0.364	0.408
distilbert-base-multilingual-cased	0.478	0.380	0.328	0.410
xlm-roberta-base	0.491	0.476	0.039	0.354
multilingual-miniLM-L12-H384	0.336	0.012	0.019	0.046
roberta-base	0.495	0.305	0.393	0.450
distilroberta-base	0.540	0.503	0.508	0.557
roberta-base-quantized	0.177	0.230	0.360	0.108
bert-large-uncased	0.545	0.450	0.549	0.503
bert-large-uncased-90%-pruned	0.614	0.476	0.519	0.547
bert-large-uncased-quantized	0.375	0.314	0.356	0.364

Table 15: The results for the emotionality intensity regression task over the EEC corpus. The results represent the Pearson Correlation Coefficient of the model for each emotion and indicates how the model performs on that particular category's template data.

	Profession	Positive	Negative
bert-base-uncased	0.694	0.040	0.111
distilbert-base-uncased	1.113	0.279	0.218
distilbert-base-uncased-60%	0.206	0.422	0.361
bert-base-uncased-85%-pruned	1.393	0.090	0.135
bert-base-uncased-90%-pruned	1.943	0.070	0.048
bert-base-uncased-quantized	1.116	0.102	0.006
bert-base-multilingual-cased	1.326	0.322	0.052
distilbert-base-multilingual-cased	0.660	0.005	0.053
xlm-roberta-large	2.007	0.031	0.073
multilingual-MiniLM-L12-H384	0.667	1.739	0.028
roberta-base	4.704	0.016	0.014
distilroberta	6.218	0.287	0.271
roberta-base-quantized	3.657	0.019	0.014
bert-large-uncased	0.155	0.359	0.343
bert-large-uncased-90%-pruned	0.899	0.293	0.269
bert-large-uncased-quantized	1.861	0.115	0.082

Table 16: The results for the effect size from the LPBS metric. The higher the effect size (calculated using Cohen's d), the higher the magnitude of bias in the model.

Dataset	Languages
StereoSet	en
Crows-Pair	en
Reviews Dataset	en, fr, de, dk
Hate Speech Detection	en, pt, es, it, po

Table 17: List of the multilingual datasets and their corresponding languages.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the final section of the paper.

☒ A2. Did you discuss any potential risks of your work?
No, as our work deals with pointing out the potential risks of certain strategies used to optimize for efficiency that could inadvertently cause models to make more biased decision. We have extensively discussed the limitations and other potential implications of our findings, however.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, Section 1. Section 8 does this as well.

☑ A4. Have you used AI writing assistants when working on this paper?
Yes, Grammarly, for ensuring that the content was grammatically correct and coherent.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Yes, we cited the relevant papers/authors that developed these datasets in whichever sections they were brought up.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we used are open access and not proprietary.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We use the datasets for their intended use i.e to evaluate models.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The details for the datasets are listed in tables in the Appendix.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 6 onwards

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, Section 4 onward. We utilized hyperparameters that have been used in previous related research and did not specifically perform a hyperparameter search to find the most optimized model, due to the computational expense required given the extent of our benchmarking. The point of this paper was to evaluate and benchmark fairness across a series of compression methods to provide a conclusive answer as to how compression may affect fairness, and to what extent.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No, our experiments are based on fairness evaluation and do not require multiple reruns.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We have referenced previous research that proposes evaluation metrics using which we have carried out these evaluations.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.