Cross-lingual Transfer Can Worsen Bias in Low-Resource Sentiment Analysis

Anonymous ACL submission

Abstract

Sentiment analysis (SA) systems are ubiquitous in natural language processing, used in many products and hundreds of languages. SA systems in high-resource languages are known to have gender and racial biases reflecting those that appear in their training data. In low-resource languages, scant training data is often supplemented by the use of pre-trained models, including multilingual models trained on other languages, and sometimes with even supervision coming entirely from other languages. What biases do systems import from these pre-trained models? To answer this question, we use a counterfactual evaluation procedure to test whether biases are imported (1) from monolingual pre-training; and (2) from other languages via pre-training or supervision data in those languages. Across five languages, we find that bias *does* change: in general, systems built with monolingual pre-trained models become *less* biased, while those built with multi-lingual pre-trained models become *more* biased. We also find racial biases to be much more prevalent than gender biases. To further research on this topic, we release a new evaluation corpus for racial and gender bias in four languages, our code, and 1,525 sentiment models.[1]

1 Introduction

Sentiment Analysis (SA) systems are among the most widely deployed NLP systems, used in hundreds of languages (Chen and Skiena, 2014). Since SA requires supervision but has substantial training resources in only a handful of languages, much research has been devoted to expanding it to low-resource languages. Creating supervised training data in a new language is expensive, so two strategies are commonly used to avoid this cost. The first, which we call **monolingual transfer**, is to pre-train on a large corpus in the target language, fine-tune on a small amount of supervision data in that language, and apply the model in that language (Gururangan et al., 2020). The second, which we call **zero-shot cross-lingual transfer**, is to pre-train on a large corpus in *many* languages, fine-tune on supervision data in a high-resource language, and use the model directly in the target language (Eisenschlos et al., 2019; Ranasinghe and Zampieri, 2020).

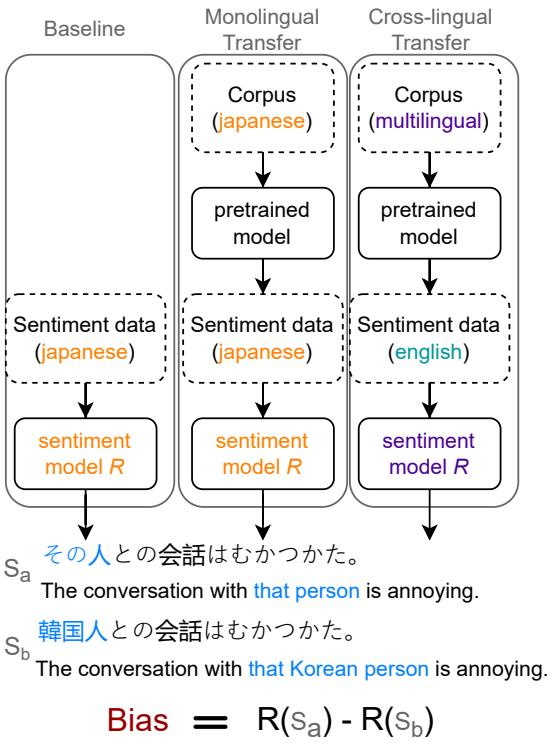


Figure 1: We use counterfactual evaluation to evaluate how bias is transferred from training data. Counterfactual pairs (e.g. sentences *a*, *b*) vary a single demographic variable (e.g. race). We measure bias as the difference in scores for the pair. An unbiased model should be invariant to the counterfactual, with a difference of zero.

It is well-known that high-resource SA models exhibit gender and racial biases (Kiritchenko and Mohammad, 2018; Thelwall, 2018; Sweeney and Najafian, 2020). Yet, it is unclear what biases exist in SA models for low-resource settings. What biases do models import from pre-training (**RQ1**)?

[1] URL to code and models withheld for anonymity.

The question of bias in low-resource settings is especially pertinent for zero-shot cross-lingual transfer, since different cultures and languages express different biases. Does bias transfer across languages (**RQ2**)? Specific cultural meanings, multiple word senses, and dialect differences often contribute to errors in multilingual SA systems (Mohammad et al., 2016; Troiano et al., 2020), and are also sources of bias (Sap et al., 2019). For example, the word *foreigner* is translated from English to Japanese as *gaijin* (外人) which has approximately the same meaning, but more negative sentiment. Bias may arise from differences in what is explicitly expressed. For example, there is evidence that syntactic gender agreement increases gender information in representations (Gonen et al., 2019; McCurdy and Serbetci, 2017), and there is also evidence that gender information in representations correlates with gender bias (Orgad et al., 2022). Could multilingual pre-training on languages with gender agreement produce more gender bias in target languages without gender agreement?

We investigate these questions via counterfactual evaluation, in which test examples are edited to change a single variable of interest—such as the race of the subject—so that any change in model behaviour can be attributed to that edit. Expanding on the methodology of Kiritchenko and Mohammad (2018), we create counterfactual evaluation datasets for gender and racial bias in five languages: Japanese (ja), simplified Chinese (zh), Spanish (es), German (de), and English (en). Though not low-resource languages, they enable us to simulate the task of transfer across languages. We find that:

1. Monolingual pre-training makes SA systems less biased *in aggregate*, though in surprising ways: our non-pretrained models exhibit extreme changes in behaviour on counterfactual examples, whereas pre-trained models exhibit many small nuanced changes.

2. Gender bias is influenced by multilingual pre-training, in directions that are predictable by the presence or absence of syntactic gender agreement in the target language.

3. Bias can also transfer across languages via zero-shot multilingual transfer. Racial bias in particular can change dramatically.

4. The common strategy of compressing models via distillation can often reduce biases, though not always.

We recommend the community use our evaluation procedure in expanding SA to low-resource languages and tasks. We release our counterfactual evaluation corpus, and all models and code used for our experiments, to facilitate further research.[1]

2 Background

We focus on sentiment **polarity detection** (Pang and Lee, 2007): a text classification task where our output label indicates the overall sentiment, with five possible results, each reported as an ordinal **score** shown in parentheses: very negative (1), negative (2), neutral (3), positive (4), or very positive (5).[2]

2.1 Transfer and Zero-Shot Learning

The aim of transfer learning is to leverage a plentiful resource to bootstrap learning for a task with few resources. For example, we can train a Japanese language model on a web crawl (plentiful resource) and use its learned parameters to initialise a polarity detection model that we fine-tune on a few Japanese reviews (scarce resource). Cross-lingual transfer learning (Ruder et al., 2019; Pires et al., 2019; Wu and Dredze, 2019) extends this idea to transferring across *languages*. The idea is to train a model on text in many languages, including both our target language and a language with substantial resources in the target task. For example, we pre-train a model on a multilingual web crawl containing both English and Japanese, and fine-tune on many English reviews (plentiful resource). We then assume that since the model knows about both Japanese and polarity detection, it can be applied to the task even though it has never seen examples of polarity detection in Japanese. We call this zero-shot cross-lingual transfer (**ZS-XLT**). An alternative approach is few-shot transfer, where we also use a very small amount of target-language supervision. We focus on zero-shot transfer because it makes clear any causal link between multilingual training and bias transfer.

2.2 Counterfactual Evaluation

Counterfactual (or contrastive) evaluation establishes causal attribution by modifying a single input variable, so that any changes in output can be attributed to that intervention (Pearl, 2009). When

[2]This is the most common approach for sentiment systems trained on user reviews, i.e. IMDB, RottenTomatoes, Yelp, Amazon products (Poria et al., 2020).

evaluating model fairness, we assert that model predictions should be invariant to intervention on a demographic or protected variable such as race or gender (Kusner et al., 2017).[3] For example, if our variable of interest is gender, and our original sentence is *The conversation with that boy was irritating*, then our intervention creates the counterfactual sentence *The conversation with that girl was irritating*. Importantly, we change no other variables, such as age (*boy* → *woman*), register (*boy* → *lady*), or relationship (*boy* → *sister*). We then evaluate the behavior of our model across many such pairs of original and counterfactual sentences. In a model with no gender bias, sentiment should not change under this intervention. If it does, and does so *systematically* over many counterfactuals, we conclude that our model is biased. Biased models for SA are likely to propagate representational harm (Crawford, 2017) by systematically associating minoritised groups with more negative sentiment. They also can propagate allocational harm by being less stable at sentiment prediction in the presence of certain demographic information, as, definitionally, variance under the counterfactual is incorrect.

Counterfactual evaluation is frequently used in bias research on classification tasks (Garg et al., 2019), and sometimes even on generation tasks (Huang et al., 2020). There have also been works exposing weaknesses of counterfactual evaluation sets and pitfalls in their design (Blodgett et al., 2021; Zhang et al., 2021; Krishna et al., 2022). Anyone expanding or replicating our counterfactual evaluation work should consult these as prerequisites.

3 Methodology

Our general approach follows four steps (Figure 1):

1. **Create counterfactual corpora** of paired sentences varying gender or racial variables.

2. **Evaluate models** on the counterfactual corpus, separately for each demographic.

3. **Compare models** via summary statistics for an overall measure of bias.

4. **Visualise differences** between the models with confusion matrices, to analyze sources of difference.

In Step 1, we use the Equity Evaluation Corpus (Kiritchenko and Mohammad, 2018) for English. For our other languages, no counterfactual evaluation corpora existed, so we created them using the method of Kiritchenko and Mohammad (2018).

While we built this procedure for SA models, it can be made task independent, and applied to any classification task, such as Dixon et al. (2018) for toxicity detection. Be aware that the invariance assertions that need to be made for counterfactuals will vary by task.

3.1 New Counterfactual Evaluation Corpora

To create counterfactual examples for non-English languages we use template sentences (Table 1). In all languages we have native speakers[4] translate the template structure of Kiritchenko and Mohammad (2018), often modifying them to preference naturalness in the target language while preserving sentiment. Templates have placeholders for demographic words for counterfactuals, and emotion words to test different levels of sentiment polarity.

The templates of Kiritchenko and Mohammad (2018) covered only the weak agreement and inflectional morphology of English, so we extend their methodology to handle a variety of grammatical phenomena in other languages. For example, in German we add gender agreement (masculine, feminine, neuter) and noun declension; in Spanish we add gender agreement (masculine, feminine, plural of both) and idiomatic verb usage;[5] in Japanese we add a distinction between active and passive forms. Chinese requires no special handling since it lacks gender agreement or inflectional morphology.

In all languages, we create a gender bias test set with paired terms expressing the male/female demographic axis. In German and Japanese we also create pairs of terms for racial and anti-immigrant bias, which we derive from NGOs, sociology and anthropology resources, and government census data (Buckley, 2006; Weiner, 2009; Muigai, 2010; , FADA). For racial and anti-immigrant bias, we usually leave the privileged group unmarked to avoid

[3] There are tasks where invariance to demographics doesn't make sense, such as hate speech classification. We design our evaluation data so that all examples should be invariant, since this is an important part of counterfactual evaluation design.

[4] Details of speakers in Appendix C.

[5] Many emotions in Spanish can idiomatically only be expressed with 'to be' or 'to have', but not both. Some take both, e.g., estoy enfadado vs. tengo un enfado — I am angry vs. I have an anger, but some emotions can use only one, or as in that example, the form changes.

	Template	Counterfactual sentences
en	The conversation with <person object> was <emotional situation word>.	The conversation with [him\her] was irritating.
ja	<person> との会は <emotion word passive>た	[彼\彼女] との会は イライラさた。
zh	跟 <person> 的谈话很 <emotional situation word>.	跟 [他\她] 的谈话很 令人生气.
de	Das Gespräch mit <person dat. object> war <emotional situation word>.	Das Gespräch mit [ihm\ihr] war irritierend.
es	La conversación con <person> fue <emotional situation word female>.	La conversación con [él\ella] fue irritante.

Table 1: Example sentence templates for each language and their counterfactual words that, when filled in, create a contrastive pair; in this case, for gender bias. For illustration, all five examples are translations of the same sentence.

	Gender	Race/Immigrant
Japanese	3340	2004
Chinese	4928	-
German	3200	5236
Spanish	4240	6360
English	2880	5760

Table 2: Counterfactual pairs in each evaluation set, including original English. Differences in corpus size are due to differing number of grammatical variants and demographic words across languages.

the unnaturalness of markedness (Blodgett et al., 2021). For example, we contrast *a person* (German *er*) with *a Turkish person* (German *der Türke*)[6]. For Spanish anti-immigrant bias we use name lists from Goldfarb-Tarrant et al. (2021), based on social science research (Salamanca and Pereira, 2013). We lacked equivalent resources for Chinese, so we test only gender bias. The resulting corpora (Table 2) are comparable to or larger than other common contrastive evaluation benchmarks (Blodgett et al., 2021).

3.2 Metrics

We need an aggregate measure of overall bias and a way to look at results in more detail. For our aggregate metric, we measure the difference in sentiment score between each pair of counterfactual sentences, and then analyse the mean and variance over all pairs. Formally, each corpus consists of n sentences, $S = \{s_i ... s_n\}$, and a demographic variable $A = \{a, b\}$ where a is the privileged class (*male* or *privileged / unmarked race*) and b is the minoritised class (*female* or *racial minority*). The sentiment classifier produces a score R for each sentence, and our aggregate measure of bias is:

$$\frac{1}{N}\sum_{i=0}^{n} R(s_i \mid A = a) - R(s_i \mid A = b)$$

In this formulation, values greater than zero indicate bias against the minoritised group, values less than zero indicate bias against the privileged group, and zero indicates no bias. Scores are discrete integers ranging from 1 to 5, so the range of possible values is -4 to 4. For example, if a sentence received a score of 4 with the male demographic term, and a score of 1 with the female demographic term, the score gap for that example is 3.

To put our results in context, Kiritchenko and Mohammad (2018) found the average bias of a system to be $\leq 3\%$ of the output score range, which corresponds to a gap of 0.12 on our scale. In practice, this is equivalent to reducing the sentiment score by one for twelve out of every hundred reviews mentioning a minoritised group, or to flipping the score from maximally positive to maximally negative for three out of every hundred.

For more granular analysis we examine confusion matrices of privileged vs. minoritised scores for each example. This enables us to distinguish between many minor changes in sentiment or fewer large changes, which are otherwise obscured by aggregate metrics as described above.

4 Experimental Setup

For our pretrained models, we train both standard and distilled models, since the latter are often preferred in practice for efficiency. As we are doing a comparison of transfer learning conditions that would happen in practice, we use pretrained models from huggingface (Wolf et al., 2020) that have been standard as per sentiment benchmarks and previous work on our data.[7] Models have as similar numbers of parameters and fine-tuning procedures as is possible (Appendix A), and are trained until convergence using early stopping on the development set. Pretrained models converge to equivalent performance on the fine-tuning data as previous work Keung et al. (2020). F1 scores and steps to convergence are included in Appendix B.

[6] Default or privileged groups are usually *unmarked*. *White person* is usually only used in racially charged contexts: a *person* is assumed to be white unless specified otherwise.

[7] https://paperswithcode.com/task/sentiment-analysis#benchmarks

In all experiments, we shuffle all fine-tuning datasets with the same, fixed random seed so that ordering is random but fixed across models. Our *pretraining* data is from from Wikipedia and CommonCrawl, Paracrawl, or the target language equivalent. Hence, there is a domain shift between the pretraining data and our fine-tuning data.

We train each model five times with different random seeds (or five separate runs for the baseline) and then ensemble by taking their majority vote, a standard procedure to reduce variance. In our initial experiments, we observed that bias varied substantially across different random initialisations on our out-of-domain counterfactual corpora, despite stable performance on our in-domain training/eval/test data. Previous work has also found different seeds with identical in-domain performance to have wildly variable out-of-domain results (McCoy et al., 2020) and bias (Sellam et al., 2022) and theorised that different local minima may have differing generalisation performance. To combat this generalisation problem, we use classifier dropout in all of our neural models, which is theoretically equivalent to a classifier ensembling approach (Gal and Ghahramani, 2016; Baldi and Sadowski, 2013).

5 RQ1: How Does Monolingual Transfer Influence Bias?

What is the effect of switching from a baseline machine learning model, trained on only a small amount of supervision data, to a pre-trained model trained on much more data? What biases are imported from pre-training?

5.1 Setup

Models. Our *baseline (no pretraining) models* are bag-of-words linear kernel support vector machines (SVMs) trained using `scikit-learn` (Pedregosa et al., 2011) on monolingual supervision data, using hinge loss, 1000 maximum iterations with early stopping, and L2 regularization. Our *Monolingual transfer (mono-T) models* are pretrained `bert-base` (Devlin et al., 2018) for each language. We randomly initialise a linear classification layer and simultaneously train the classifier and fine-tune the language model on monolingual supervision data. We also include a distilled monolingual transformer, which is identical save that the pretrained model is `distilbert-base` (Sanh et al., 2019).

Fine-tuning data. For each model, we use the language appropriate subset of the Multilingual Amazon Reviews Corpus (MARC; Keung et al., 2020), which contains 200 word reviews in English, Japanese, German, French, Chinese and Spanish, with discrete sentiment labels ranging from 1-5, balanced across labels. We use the standard train/dev/test splits of 200k, 5k, 5k examples (the same for all languages).

Like many bias tests, these experiments have positive predictive power: If bias in mono-T models is higher than baseline, then we conclude that bias has been transferred from pretraining. If not, then we conclude that bias is not transferred or that other differences, such as increased modelling power, reduce bias enough to neutralise it.

5.2 Results

The baseline models are the most biased for both gender and race in all languages (Figure 2), though not always against minoritised groups: systems are often biased against the male demographic, consistent with previous work on SA (Thelwall, 2018).

Analyzing the granular differences (Figure 3) reveals interesting behavior that is not captured by aggregate metrics: much of the bias exhibited by the baselines arises from consistently flipping specific labels in the counterfactual, while bias exhibited by the pre-trained models is subtler and more varied.[8] For example, the Japanese baseline exhibits racial bias by frequently changing neutral labels to very negative labels, whereas in the mono-T model the change under the counterfactual is expressed as many less extreme changes. The model is still biased overall: though the changes are more varied, in aggregate they associate racial minorities with more negative sentiment. The German baseline model is more extreme: when the demographic variable changes from privileged to minoritised, the model changes its prediction from very positive to very negative. The German mono-T model also makes biased choices, though more moderately (neutral to negative) and there is more 'counter-bias' in the upper triangle, which greatly minimises overall bias.

Summary Any biases introduced by pretraining are in most cases outweighed by the other benefits of using a larger, more expressive model, so overall using pretrained models produces less biased mod-

[8] We show Japanese and German for illustration; the trend is present in all languages. All graphs are in Appendix E.

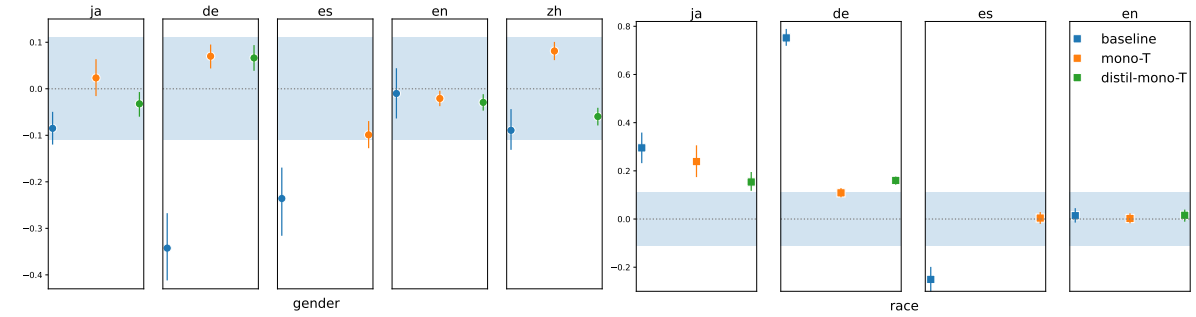


Figure 2: Aggregate bias metrics for RQ1: Comparison of baseline (blue), mono-T (orange), and distil mono-T (green) models. Mean and variance of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.2). Spanish (es) distilled model is intentionally missing for lack of comparable pretrained model.

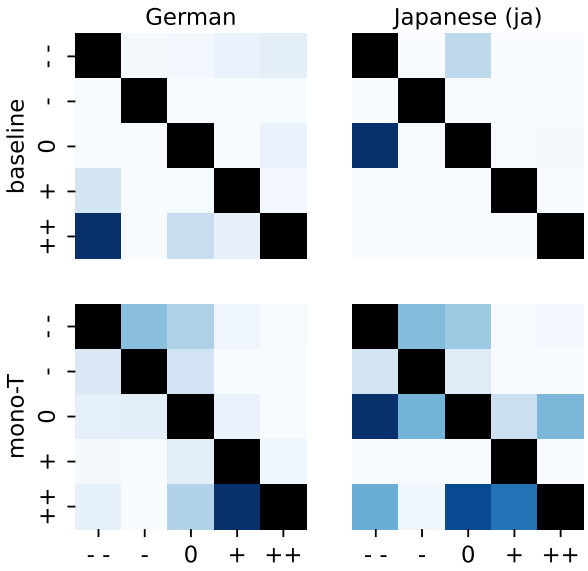


Figure 3: Confusion matrices for racial demographic counterfactual pairs for Japanese and German, comparing the difference between baseline and monolingual transfer (mono-T) models (RQ1). Higher colour saturation in the lower triangle is bias against the minoritised group, in the upper triangle is bias against the privileged group. Other languages and bias types follow the same pattern between baseline and mono-T models (regardless of absolute level of bias).

els than a simple bag-of-words approach. However, be aware that changing from a baseline to mono-T model can change the direction of gender bias.

6 RQ2: Does bias transfer across languages, from pretraining or supervision?

We examine whether a decision to use zero-shot cross-lingual transfer (ZS-XLT), instead of monolingual transfer, affects bias. Both are reasonable choices: monolingual transfer can make the most of a small amount of higher quality monolingual supervision data, while ZS-XLT leverages large amounts of supervision data in English. But there are thus two potential sources of bias in ZS-XLT: from the multilingual *pretraining*, or from the English *supervision*. Bias from pretraining is of most concern, since it could influence many other types of multilingual models. We conduct two experiments to control each variable separately and identify the source of changes in bias.

6.1 Setup

We compare three models: the mono-T models from RQ1, a cross-lingual transfer (ZS-XLT) model, and a multilingual model trained on the RQ1 monolingual data (mono-XLT). The mono-XLT model is an experimental ablation on ZS-XLT, allowing us to disentangle the effects of pre-training data and supervision, and thus attribute changes in model behaviour to one or the other.

Models. Monolingual models are the same as in RQ1. ZS-XLT models are `mbert-base` in English only, to emulate standard cross-lingual transfer. We also include a distilled ZS-XLT transformer, identical save that the pretrained model is `distilmbert-base`. Mono-XLT models are the same pretrained multilingual models, where fine-tuning and classification layer training are done on the monolingual data from RQ1.

Fine-tuning data. We fine-tune on the US segment of the Amazon Customer reviewss corpus.[9] This dataset is not balanced across labels,[10] so we balance it in order to match the monolingual setup. After balancing we use 2 million reviews,

[9] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[10] As is common in user-generated review data, the distribution is skewed towards extreme labels, and in the original review data 1 and 5 are 73% of data.

ten times more than the monolingual training data. We create our own train/evaluation/test split of 0.95/0.025/0.025, and fix this across models.

If the mono-XLT model has higher bias than the mono-T model, this bias comes from multilingual pretraining. If the ZS-XLT model has higher bias than the mono-XLT model, this bias comes from cross-lingual supervision. As with RQ1, these experiments have positive predictive power.

6.2 Results

Can bias transfer across languages from pretraining data? Yes. The difference between mono-T and mono-XLT is generally small for race, except in German, and large for gender (Figure 4). This demonstrates that bias from a language included in pretraining can come appear in a model targeted to a different target language.

The larger effect on gender than on race is as we expected; both because gender biases are less culturally specific than racial biases, and because some languages have stronger syntactic gender signal than others. We hypothesised that the increase in gender information from grammatical agreement seen in previous work might manifest in *increased* gender bias when using cross-lingual transfer for languages with weak gender agreement, and *decreased* gender bias when using transfer for languages with strong gender agreement. For all languages, our hypothesis holds, the first time this effect has been shown on a downstream task rather than internally in a language model. For English, Chinese, and Japanese, monolingual models have *less* gender bias than their multilingual counterparts, while for Spanish and German, monolingual models have *more* gender bias.[11]

Summary Bias can transfer across languages from pretraining data, and gender bias is increased if pretraining languages have more gender agreement than the target language, or decreased if they have less. It is not clear whether racial bias can be transferred from pretraining, and this may be because racial biases tend to be less common across languages and cultures.

Can bias transfer from English supervision data to other languages? To answer this, we compare ZS-XLT to mono-XLT and mono-T. For gender bias, ZS-XLT aggregate bias is the same or very similar to mono-T bias in German, and Spanish, though of higher variance. In Japanese and Chinese gender bias is increased in the ZS-XLT model, though for different reasons: Japanese has anti-male bias from the change to an XLT model (since it appears in both the mono-XLT and the ZS-XLT models) whereas Chinese cross-lingual transfer has anti-male bias from the supervision data (as it appears only in ZS-XLT and not mono-XLT). This is strange, as it does not appear in all models: gender bias in English mono-XLT and ZS-XLT models is the same, so the supervision data does not universally change bias across languages. This must be mediated by some other effects, which would be valuable to investigate in future work.

For racial bias, the trend is less systematic. Sometimes the ZS-XLT model bias is unchanged, as with Japanese and English, or sometimes increased, as with German and Spanish.

The presence of cross-lingual racial bias is surprising. Racial bias tends to be culturally specific, so we did not expect it to transfer across language data the way gender bias might; we expected ZS-XLT to have either equivalent or less racial bias than mono-T. A possible source of cross-lingual bias supported by the patterns in our results is that there may be more overlap between German and Spanish racial biases and cross-lingual pretraining data or English fine-tuning data. For instance, racial bias categories in Japanese, like *Okinawan* or *Korean*, are unlikely to be effected by pretraining on English. Racial bias categories in German, though German-specific, may be shared by other high resource Western languages, such as *Muslim* or *Arab*. Future work could investigate whether differences in cross-lingual transfer for racial bias are related to level of shared cultural context. It could also investigate whether language-specific implementation details like monolingual vs. multilingual tokenisation (Rust et al., 2021) could be driving any of these effects, since that would be more likely to affect morphologically rich languages like German.

Summary In answer to RQ2, both gender and racial bias can transfer across languages via supervision data.

7 Recommendations & Conclusion

This broad set of experiments has shown that bias can change drastically as a result of any of the stan-

[11] As a further controlled experiment on agreement, we 'scrubbed' gender information from English (De-Arteaga et al., 2019), to compare standard English with a less gendered version of the same data. Our results showed that scrubbing does not work as intended on all data domains, so we have left them out of this analysis. They are included in Appendix D.

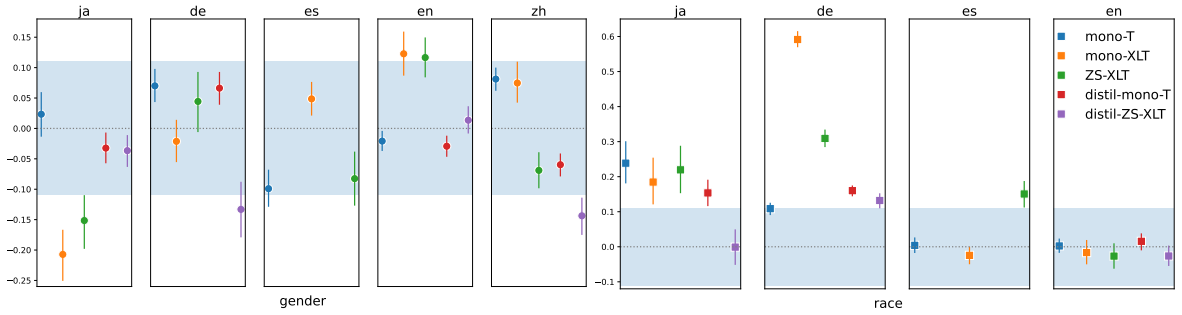


Figure 4: Aggregate bias metrics for RQ2: Comparison of mono-T (blue), and mono-XLT (orange), ZS-XLT (green), distil mono-T (red), and distil ZS-XLT (purple) models. Mean and variance of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.2). Spanish (es) distilled model is intentionally missing for lack of comparable pretrained model.

dard engineering choices for making an SA system in a lower resourced language. In light of these results, we make the following recommendations:

Use our bias evaluation procedure. Assess bias of all model *and* data choices. Use granular bias by sentiment label, as well as aggregate bias, to make decisions that best suit the intended application.

Don't rely solely on aggregate measures. Our results highlight how summary statistics can make different underlying distributions appear identical, a point made by Matejka and Fitzmaurice (2017) in general, and by Zhao and Chang (2020) for specifically bias, but still frequently overlooked in most bias research. Though both are problematic, model that consistently associates slightly more negative sentiment to a minoritised group is qualitatively different from a model that sometimes flips very positive sentiment to very negative sentiment.

Use monolingual transfer if available. For our tasks, monolingual transfer models (mono-T) are universally less biased than baseline models, even though pretraining data can affect bias, and pretraining data contains undesirable content (Luccioni and Viviano, 2021). The more consistent behaviour from a better trained model is still better.

Beware of bias introduced cross-lingually. Bias can transfer across languages from pretraining or from supervision data, which means that cross-lingual transfer has the opportunity to introduce non-local biases. These can be unexpected and hard to detect, and represent machine learning cultural imperialism that is best avoided.

Be particularly aware of racial biases. Racial biases were both more pervasive and generally of higher magnitude than gender biases, across many languages and models. Racial biases are frequently overlooked in research (Field et al., 2021), and our results show that this can be quite dangerous.

Consider compressing models. Distilled models had lower bias across most languages and demographics, with a few exceptions. This came at a very low penalty for performance of one F1 point on average. Previous work had contradictory conclusions regarding model compression, with some vision models showing worse bias in compressed models (Hooker et al., 2020) and some NLP generation models showing less bias under compression (Vig et al., 2020). Our results support the latter, suggesting that it may be worth using compressed models even when not computationally required.

We have laid the groundwork for investigating bias in low-resource and cross-lingual sentiment analysis. We created resources, presented an evaluation procedure, and used it to do the first analysis of bias in SA in a simulated low-resource setting across multiple languages. We have also raised many open questions. Do these results hold for in-domain sentiment analysis? Out-of-domain, which we tested, is the most common use case for sentiment analysis since training data is plentiful only in narrow domains (e.g. reviews), but in-domain SA has also many applications. What are the key mechanisms of cross-lingual transfer? A causal analysis (Vig et al., 2020), or saliency and attribution methods, could enable us to understand, and perhaps control, whether to transfer information when it ameliorates bias (as may be the case with cross-lingual or cross-task information attenuating bias signal) or block transfer when it makes bias worse. We invite the NLP community to use the data and models from this work to answer these and other questions.

References

Pierre Baldi and Peter J. Sadowski. 2013. Understanding dropout. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2814–2822.

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.

Sandra Buckley. 2006. *Encyclopedia of contemporary Japanese culture*. Routledge.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, Sahin Cem Geyik, K. Kenthapadi, and A. Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

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.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AIES '18*.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

The Federal Anti-Discrimination Agency (FADA). 2020. Equal rights, equal opportunities: Annual report of the federal anti-discrimination agency.

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.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Sara Hooker, Nyalleng Moorosi, G. Clark, S. Bengio, and Emily L. Denton. 2020. Characterising bias in compressed models. *ArXiv*, abs/2010.03058.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 4563–4568, Online. Association for Computational Linguistics.

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.

Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5830–5842, Dublin, Ireland. Association for Computational Linguistics.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Justin Matejka and George Fitzmaurice. 2017. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1290–1294, New York, NY, USA. Association for Computing Machinery. [link].

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

K. McCurdy and Oguz Serbetci. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *ArXiv*, abs/2005.08864.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Intell. Res.*, 55:95–130.

Githu Muigai. 2010. Report of the special rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, githu muigai, on his mission to germany (22 june - 1 july 2009).

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. *ArXiv*, abs/2204.06827.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *CoRR*, abs/2005.00357.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual 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 3118–3135, Online. Association for Computational Linguistics.

Gastã Salamanca and Lidia Pereira. 2013. PRESTIGIO Y ESTIGMATIZACIÃ"N DE 60 NOMBRES PROPIOS EN 40 SUJETOS DE NIVEL EDUCACIONAL SUPERIOR. *Universum (Talca)*, 28:35 – 57.

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.

Maarten Sap, D. Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 359–368, New York, NY, USA. Association for Computing Machinery.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Michael Weiner. 2009. *Japan's minorities: the illusion of homogeneity*, volume 38. Taylor & Francis.

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double perturbation: On the robustness of robustness and counterfactual bias evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3899–3916, Online. Association for Computational Linguistics.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, 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.

A Model Implementation Details

Monolingual transformer models have 110 million parameters ($\pm$ 1 million) and vocabularies of 30-32k with 768D embeddings. Multilingual models have 179 million parameters, a vocabulary of 120k, with 768D embeddings. We train the monolingual models with the same training settings as preferred in Keung et al. (2020), and allow the pretrained weights to fine-tune along with the newly initialised classification layer. The multilingual models are trained identically, save that they have a 100x larger learning rate, and learning rate annealing.

All models were trained for 5 seeds, models trained on monolingual data (mono-T, mono-XLT, and distil-mono-T) were checkpointed 15 times. ZS-XLT models were checkpointed 6 times. In total we train 1525 models: 3 monolingual (non-baseline) model types with 5 seeds across 5 languages and 15 checkpoints (1,225 models) and 2 multilingual model types (ZS-XLT, distil-XLT) with 5 seeds and 5 languages and 6 checkpoints (300) models.

This study was done on only the converged models, but all models are released for further study.

Computational Resources. Each model was trained on 4 NVIDIA Tesla V100 GPUs with 16GB memory. mono-T and mono-XLT models took 6-8 hours to converge, ZS-XLT and distil-ZS-XLT took 15 hours. This is a total of 620 total hours, or 2,480 GPU hours on our resource.

B Model Performance

	Standard		**Distilled**		**Baseline**
	F1	Steps	F1	Steps	F1
ja	**0.62**	44370	0.61	60436	0.38
zh	**0.56**	35190	0.53	43750	0.42
de	**0.63**	36720	0.63	52621	0.51
es	**0.61**	41310		-	0.48
en	**0.65**	27050	**0.65**	44285	0.53
en_s	**0.65**	26520	0.64	17193	0.52
ZS-XLT	**0.69**	75000	0.68	33336	-

Table 3: F1 at convergence and steps at convergence for standard size, distilled, and baseline models. Monolingual model performance is measured on the MARC data, ZS-XLT model performance on the US reviews data. en_s is english scrubbed of gender information (see D)

C Dataset Creation

We worked alongside native speakers in Japanese, German, Spanish, and Chinese, to create the dataset we release as a new bias evaluation corpus. The Chinese, German, and Spanish speakers had training in linguistics, the Japanese speaker did not but had professional translation experience. These were collaborators, not crowdworkers. It took the authors a week to create and refine the procedure for dataset creation, but once it was created it took each native speaker about 4 hours to complete.

The Japanese and German speakers are from Japan and Germany, respectively. The Spanish speaker is from Spain and the Chinese speaker is from mainland China. This may cause some limitations in coverage of other types of Spanish (e.g. South American) or Chinese (Traditional). They match the supervision data, which is Spain-Spanish and simplified Chinese, but may cause limitations in future work.

D Experiments with Scrubbing Gender from English

We additionally analysed removing gender information from English (via the method in De-Arteaga et al. (2019)). A comparison of all the English monolingual models for with scrubbed English models can be seen in Figure 5a, with a breakdown of the source of the differences in the baseline model in Figure 5b.

Previous work has experimented with removing gender signal (Gonen et al., 2019; Zhao et al., 2018; De-Arteaga et al., 2019) and found it to decrease gender bias. However, our results differ from this, for the monolingual transformer and monolingual distilled models there is little and negligible difference (respectively) and in the baseline English has very little gender bias, but scrubbed English has far more. Why is this? Is it just random noise, or is there a true effect?

A very small percentage of tokens were removed from the data overall ($< 1\%$), so the effect is not attributable to a change in training data size.

From examination of the data, it is very different from the previous data of De-Arteaga et al. (2019). They applied scrubbing to biographies, where the task was to classify the occupation (from a closed set). Biographies have more words that explicitly encode gender (he, she, husband, etc) than do Amazon reviews. Review data *does* include a lot of gender information, but less explicitly mentioned;

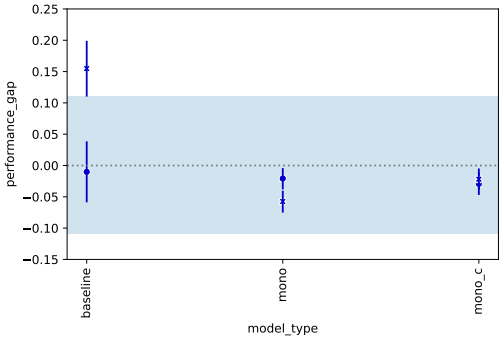
it will instead be in style of speaking and gendered associations of objects. So in “My wife loved wearing this purse.”, gender information is not as diluted by the deletion of “wife” as it is in ‘My wife went to law school.’. Based on this analysis, we experimented with using a logistic regression classifier (in `scikit-learn` (Pedregosa et al., 2011)) and found that it got far better than chance accuracy at determining the gender of the removed tokens (67% acc) when applied to the scrubbed review data. So we can conclude that even after scrubbing, there is still gender information in the text.

Why would this make the model more biased? It is not clear, but it is possible that the deletion of those tokens has made the model rely on deeper gender cues and relationships. This would be in line with the findings of (Mendelson and Belinkov, 2021; Orgad et al., 2022) who use information theoretic probes and find it makes information *more* available in many debiasing conditions.

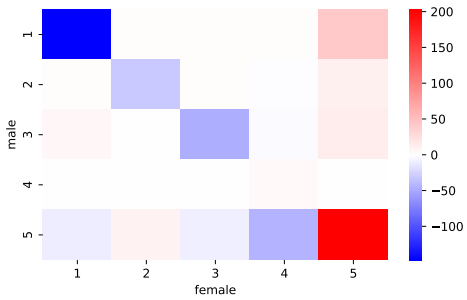
Further work could use a similar information theoretic approach to verify whether that is the reason for the results here. This could have important implications for future work, since scrubbing is a common debiasing technique, so it is notable if it is actually a data-dependent debiasing technique, and is contra-indicated in some cases.

E Full set of confusion matrices comparing baseline and monolingual models.

Figure 6 contains all confusion matrices for all languages, of which we displayed a subset in the body of this work.

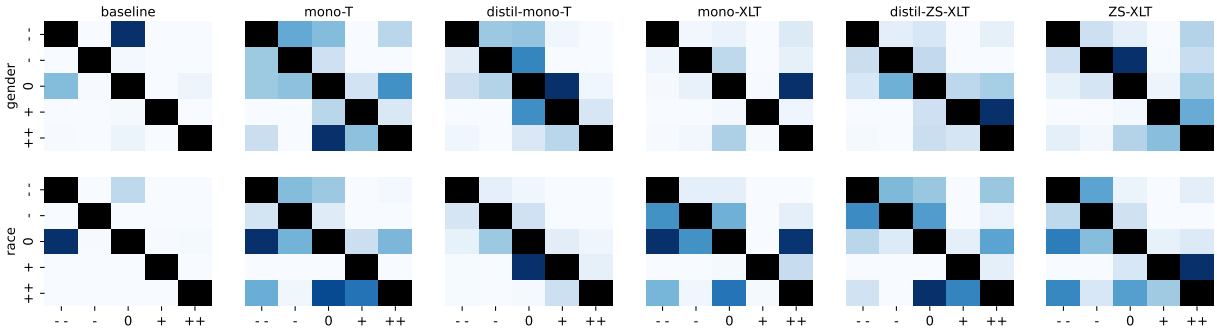


(a) Mean and variance of the performance gaps for the gender counterfactuals in monolingual models of English (O) and scrubbed English (X). Positive numbers are bias against the minoritised group (in this case, female), 0 is no bias, and negative is bias against the privileged group. Shaded region represents 3% of total range (see 3.2).

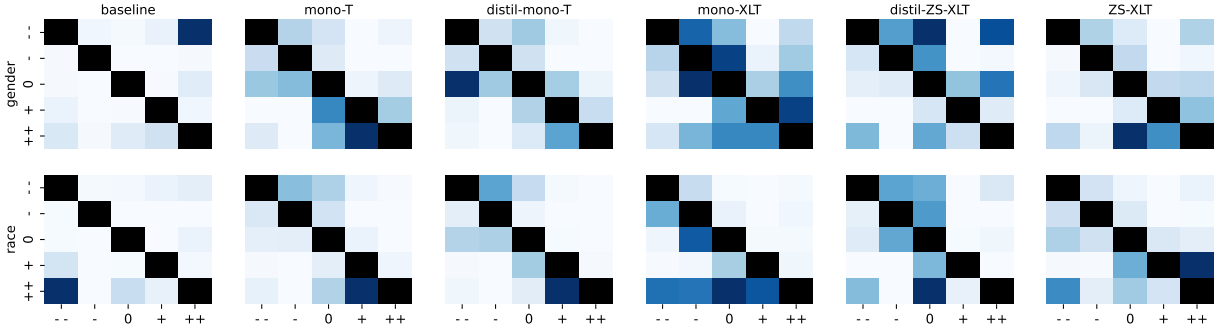


(b) Difference between confusion matrices of the gender counterfactual for English and scrubbed English baseline model (male demographic labels on x axis, female demographic counterfactual on y). The diagonal represents no bias in either direction, the upper right is bias against men, the lower left is bias against women. Red cells mean the English model has more, blue mean the scrubbed English model has more, and white mean their numbers are identical.

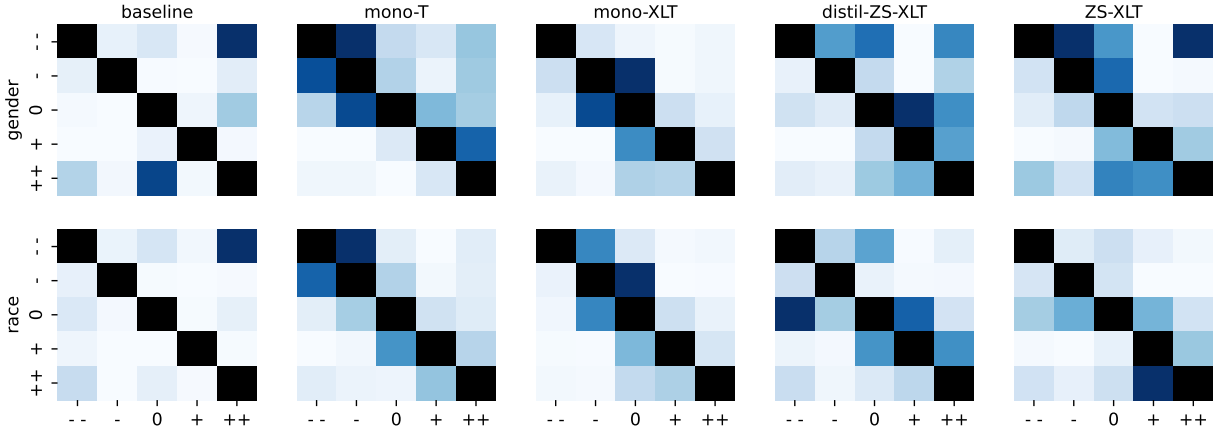
Figure 5: Comparisons of gender bias for English and English with explicit gender identifiers removed (aka "scrubbed" English). Scrubbing increases bias in the baseline, but does not much change either neural model.



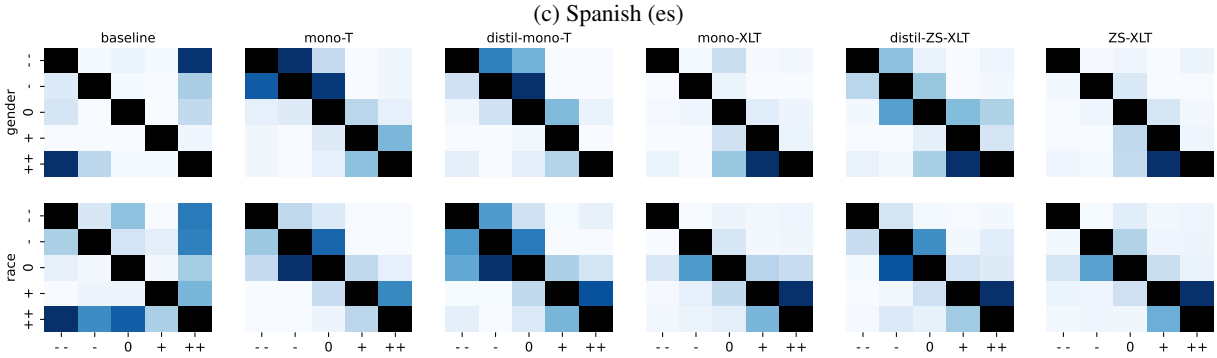
(a) Japanese (ja)



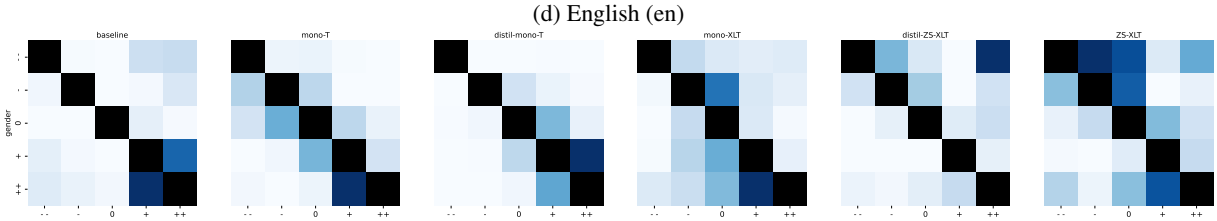
(b) German (de)



(c) Spanish (es)



(d) English (en)



(e) Chinese (zh_cn)

Figure 6: All confusion matrices for experiments in this paper. Higher colour saturation in the lower triangle is bias against the minoritised group, in the upper triangle is bias against the privileged group. Saturations are not normalised across all languages and models; this is not a proxy for aggregate comparative bias, it shows the pattern across sentiment scores.