Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages

Seraphina Goldfarb-Tarrant[*†] **Adam Lopez**[†]
Roi Blanco[‡] **Diego Marcheggiani**[‡]
[†]University of Edinburgh, [‡]Amazon
s.tarrant@ed.ac.uk roiblan@amazon.com
alopez@inf.ed.ac.uk marchegg@amazon.com

Abstract

Sentiment analysis (SA) systems are used in many products and hundreds of languages. Gender and racial biases are well-studied in English SA systems, but understudied in other languages, with few resources for such studies. To remedy this, we build a counterfactual evaluation corpus for gender and racial/migrant bias in four languages. We demonstrate its usefulness by answering a simple but important question that an engineer might need to answer when deploying a system: What biases do systems import from pre-trained models when compared to a baseline with no pre-training? Our evaluation corpus, by virtue of being counterfactual, not only reveals which models have less bias, but also pinpoints changes in model bias behaviour, which enables more targeted mitigation strategies. We release our code and evaluation corpora to facilitate future research.[1]

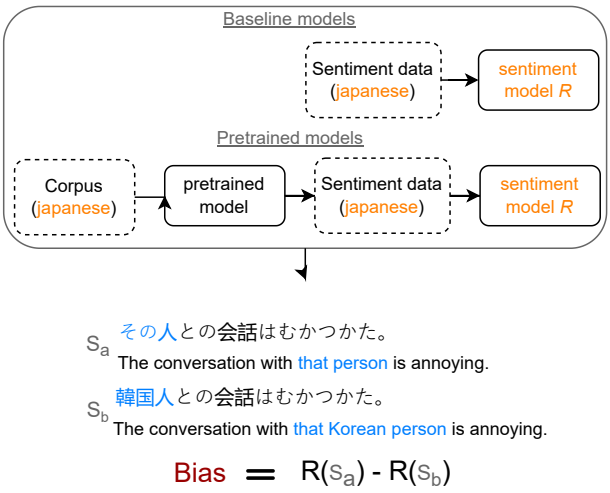


Figure 1: We create corpora and then do counterfactual evaluation to evaluate how bias is transferred from training data. Counterfactual pairs (e.g. sentences *a*, *b*) vary a single demographic variable (e.g. race). We measure bias as the difference in scores for the pair. An unbiased model should be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

1 Introduction

Sentiment Analysis (SA) systems are among the most widely deployed NLP systems, used in hundreds of languages (Chen and Skiena, 2014). It is well-known that English SA models exhibit gender and racial biases (Kiritchenko and Mohammad, 2018; Thelwall, 2018; Sweeney and Najafian, 2020), which are acquired from their training data, training objective, and other system choices (Suresh and Guttag, 2019). Other languages are understudied; though many papers study SA bias in English, few study SA bias in other languages. This may be partly attributable to resource constraints: there are fewer corpora available to audit systems for bias in non-English languages. To remedy this, we create evaluation datasets to evaluate gender and racial bias in four languages: Japanese (ja), simplified Chinese (zh), Spanish (es), German (de). Each of these four languages has publicly available data for training SA systems (Keung et al., 2020b), and together they represent three distinct language families. To complement their existing resources with a new resource that measures bias, we use counterfactual evaluation (Figure 1), in which test examples are edited to change a single variable of interest—such as the race of the subject—extending previous work done in English (Kiritchenko and Mohammad, 2018). We release the evaluation dataset to facilitate further research.[1]

We demonstrate the value of these evaluation resources by answering the following research questions: (RQ1) What biases do we find in other languages, compared to in English? (RQ2) How does the use of pre-trained models affect bias in SA systems? While pre-trained models are common in NLP, they may import biases not present in task

[*] Correspondence to s.tarrant@ed.ac.uk. Work completed while at an internship at Amazon.

[1]All code, evaluation data, and links to models and raw data can be found here: https://github.com/seraphinatarrant/multilingual_sentiment_analysis

Findings of the Association for Computational Linguistics: ACL 2023, pages 4458–4468
July 9-14, 2023

supervision data, since a large pre-training corpus may embody biases not present in the supervision corpus. On the other hand, pre-training might diminish biases that arise from the small sample sizes typical of SA training corpora.

Our experiments show that both gender and racial bias are present in SA systems for all four languages: when model architecture, data quantity, and domain are held constant, SA systems in other languages display quantitatively more bias than SA systems in English. For RQ2, we find that pre-training also makes SA systems less biased for all languages, *in aggregate*, though in surprising ways: our non-pre-trained models exhibit extreme changes in behaviour on counterfactual examples, whereas pre-trained models exhibit many small nuanced changes.

2 New Counterfactual Evaluation Corpus

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). 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 on 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.

To create counterfactual examples for non-English languages we use template sentences, illustrated in Table 1. Each template has a placeholder for a demographic word, in order to represent the counterfactual; and an emotion word, in order to represent different levels of sentiment polarity.

The templates of Kiritchenko and Mohammad (2018) only needed to handle 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;[2] 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 by providing contrasting pairs of male/female terms that can fill the placeholder for demographic variable. In German and Japanese we also provide 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). We usually leave the privileged group unmarked to avoid the unnaturalness of markedness (Blodgett et al., 2021).[3] For Spanish anti-immigrant bias, we create pairs of names by using name lists that are strongly associated with migrants or with non-migrants, sourced from Goldfarb-Tarrant et al. (2021), which are 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).

To produce the templates, we worked alongside native speakers in Japanese, German, Spanish, and Chinese to translate the English templates of Kiritchenko and Mohammad (2018), often modifying them to prefer naturalness in the target language while preserving sentiment. Our Japanese translator had professional translation experience, while our German, Spanish, and Chinese translators had training in linguistics. While collaborative development and refinement of the translation process required about a week, actual translation took about four hours for each dataset. Further details in A.

3 Methodology

For our SA task, we focus on sentiment **polarity detection** (Pang and Lee, 2007), where the output label represents the sentiment of a text as an ordinal **score** (shown in parentheses): very negative

[2] 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.

[3] For example, for anti-Turkish bias in German, we replace `person dative object` in Table 1 by contrasting *dem Türken* (Turkish person (male gender)) with the unmarked *ihm* (him).

	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 reference English. Differences in corpus size are due to differing number of grammatical variants and demographic words across languages.

(1), negative (2), neutral (3), positive (4), or very positive (5).[4]

3.1 Metrics

We measure the mean and variance of the differences in sentiment score between each pair of counterfactual sentences. 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*) 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:

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

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.

Our counterfactual evaluation process enables us to examine bias behaviour more granularly as well. We generate confusion matrices of privileged vs. minoritised scores such that an unbiased model would have all scores along the diagonal. 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.

In results we shade 3% of total range for easier visual inspection. This is an arbitrary choice: 'no bias' differs by application and values within the shaded range may still be unacceptable. Intuitively, this corresponds to models being maximally biased for three of every hundred examples, or making minor biased errors for twelve of every hundred.

4 Experiments

We want to answer the questions: what biases arise in SA systems in each of these languages (RQ1)? Does pre-training improve or worsen biases (RQ2)? To answer these questions, we measure the bias of a baseline SVM classification model to a model based on a pre-trained transformer model. We compare standard and distilled transformer models; distilled models are often used in practice since they are better suited to the computational constraints of real-world systems.

Our *baseline (no pre-training) models* are bag-of-words linear kernel support vector machines (SVMs) trained on the supervision data in each language. Our *pre-trained (mono-T) models* are pre-trained `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 the same supervision data. Our *distilled (distil-mono-T) models* are identical, but based on `distilbert-base` (Sanh et al., 2019).

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. All models converge to performance on par with SotA on this task and data. Training details and F1 scores on the SA task are reported in Appendix B and C.

Training data For each model, we use the language appropriate subset of the Multilingual Amazon Reviews Corpus (MARC; Keung et al., 2020a), which contains 200 word reviews in English,

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

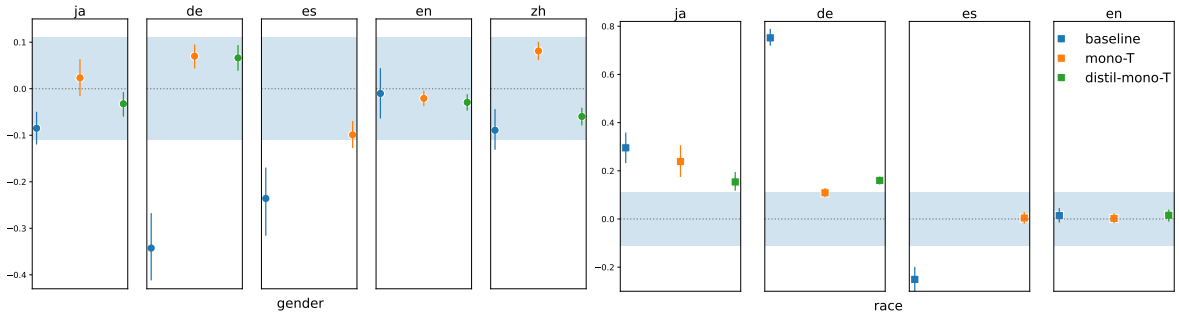


Figure 2: Aggregate bias metrics for baseline (blue), pretrained mono-T (orange), and pretrained distil mono-T (green) models. Mean and variance of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.1). Spanish (es) distilled model is intentionally missing for lack of comparable pretrained model.

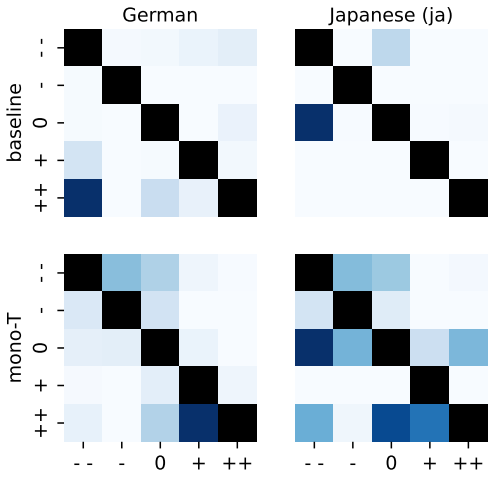


Figure 3: Confusion matrices for racial counterfactual pairs for Japanese and German, comparing baseline and pretrained models. Higher colour saturation in the lower triangle is bias against the minoritised group, against the privileged group in the upper triangle.

Japanese, German, French, Chinese and Spanish, with discrete sentiment labels ranging from 1-5, balanced across labels.

5 Results

The baseline models are 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). [5]

Figure 2 also shows that English models tend to be less biased than the other languages.

Analyzing the granular differences (Figure 3) reveals interesting behaviour 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 pre-trained models is more varied.[6] 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 lessens overall bias.

6 Related Work and Conclusion

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 common pitfalls in the design of counterfactuals (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. The contemporary work of Seshadri et al. (2022) find many ways that other

[5] Because this task is sentiment analysis, it is more possible to get bias against a male demographic than if the task were, say, biography classification. For the latter, the male demographic is associated with prestige roles (and thus generally bias is anti-female), but for sentiment analysis, male demographics can be associated with negative characteristics (violence, aggression, if a model is stereotyping) as well as with competence, so a few works have found female subjects to sometimes have more positive sentiment, depending on context.

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

templates for bias evaluation can be brittle, so future work should take this into account and take measures to ensure robustness, such as testing with multiple paraphrases of the templates.

We have laid the groundwork for investigating bias in sentiment analysis beyond English. 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 showed that using pre-trained models produces *much less* biased models than using baseline SVMs. We also showed that pre-trained models have very different *patterns* of bias; a type of analysis that is enabled by the counterfactual design of our corpus. We invite the NLP community to use the data and methods from this work to continue analysis of languages beyond English.

7 Limitations

Like all bias tests, these experiments have *positive* predictive power: they can find the biases they test for, but they cannot eliminate the possibility of there being biases that the tests overlook.

Our Japanese, German, Spanish, and Chinese translators were from Japan, Germany, Spain, and mainland China, respectively. Hence, their translations may reflect their native dialects of these languages. While these dialects are consistent with the corresponding training datasets in these languages, this fact may limit conclusions that we or others can draw about SA in other dialects of these languages, such as Central and South American dialects of Spanish, or Chinese (Traditional).

8 Ethics Statement

Because of the aforementioned limitation regarding positive predictive power, there is always a risk with research on social biases that it can give practitioners a false sense of security. It is absolutely possible to evaluate on our corpus and get no bias, and still end up causing harm to racial or gender demographics, since they do not cover all biases or all domains. This should be kept in mind whenever applying this research.

Acknowledgements

We thank Björn Ross for many comments and helping shape the draft, Lluís Màrquez for helping manage the project at Amazon, and the Amazon Barcelona Search team for their enthusiastic support of the project.

References

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.

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.

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

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.

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. 2020a. 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.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020b. Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

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.

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).

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.

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.

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.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

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*.

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

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.

A Benchmark Dataset Creation

We followed the recommendations of Blodgett et al. (2021) to ensure the validity of our datasets. Many of the pitfalls enumerated in their work do not apply to our dataset, as we are measuring sentiment, rather than stereotypes, but we took care to avoid those that do apply. These are:

Markedness. In most cases we contrast the minority group, e.g. *Turkish people* with the unmarked group, e.g. *people*. Using a marked privileged group—white people, straight people, etc—is in most cases uncommon and occurs in only particular settings, which threatens the validity of the contrastive test (Blodgett et al., 2021). We do make a few exceptions and mark privileged groups. We do mark them for gender bias, since gender is explicitly marked in language more than other demographic traits (e.g. we contrast *woman* with *man*, not with *person*). We also sometimes use first names as proxies for demographics such as race, class, and immigration status (in Spanish and English) and in these cases the privileged group is another name.

Naturalistic Text. Some of the sentences in the original Kiritchenko and Mohammad (2018) would be valid grammatical sentences if translated directly into other languages, but would not sound natural. For example, reflexive pronouns (himself, herself) aren't used the same way in Chinese as in English, so in translating the English template `<person subject> found himself/herself in a/an <emotional situation word> situation.` we instead used the Chinese template `<person subject> 经历了一件<emotional situation word>的事.`, which means `<person subject> was in a <emotional situation word> situation.` These small changes preserve the same rough semantics, and more importantly preserve naturalness.

Indirect Demographic Identification. Blodgett et al. (2021) caution against the use of proper names or other proxies as a stand in for a demographic group, because their reliability for this use is untested. We would add that names are difficult to use in a contrastive pair where we need to change only *one* demographic variable, because names indicate many bits of demographic information at once: race, gender, class, place of birth, period of birth, etc. We intentionally avoid this by using identity terms (Turk, Korean, etc) most of the time, which do sometimes conflate race and country of origin, but are otherwise the most precise option. We use proper names only in Spanish based on the work of Goldfarb-Tarrant et al. (2021) and Salamanca and Pereira (2013), who show that there is data backing up the migrant vs. non-migrant names. Even so, there is some conflation between migrant status and socioeconomic class in that set of names: we consider that acceptable for our purposes. There are also names as a proxy for African-Americans in English, as the dataset is from Kiritchenko and Mohammad (2018) and that is what they use.

Basic Consistency A few other applicable pitfalls, which Blodgett et al. (2021) capture under the heading 'Basic Control and Consistency' we avoid organically by our template based construction, e.g. differences in sentence length between sentences A and B, are a possible confound, but by construction we contrast only one word in a pair and the sentence is otherwise unperturbed.

Once we had designed our translation process, we did a multi-step qualitative evaluation. After we had settled on the first version of the three sets of templates, demographic terms, and emotion words in each language, we worked with the native speaker to iterate and make sure there were no accidental unnatural sentences or grammatical errors. We generated a few examples for each template + emotion + demographic combination, manually reviewed 200 examples per language, and then made corrections to the templates, words and the rules for combining them. We then repeated this exact process a second time after the adjustments.

B Model Implementation Details

Monolingual transformer models have 110 million parameters (± 1 million) and vocabularies of 30-32k with 768D embeddings. We train the monolingual models with the same training settings as preferred in Keung et al. (2020a), and allow the pretrained weights to fine-tune along with the newly initialised classification layer.

C Model Performance

Performance at convergence for models in each language is given in Table 3.

We determined convergence by examining loss curves and selecting the model where training loss was flat, and validation had not yet increased. We

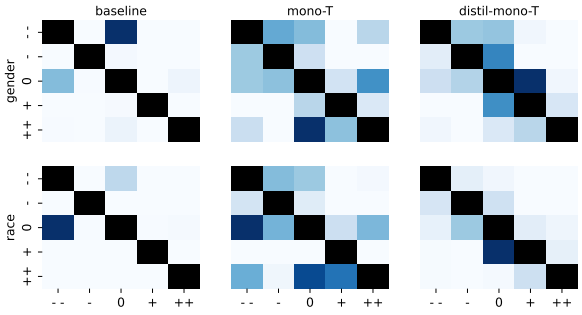
did not use early-stopping, as we wanted to save many model checkpoints in order to study the training dynamics of bias, including *after* convergence when the model was overtrained. However, we found no clear trends in how bias changed over the course of training, so for this study we used only one model, at convergence, per language. We hope that by releasing all model checkpoints (15 per language), other researchers may be able to expand our work into the training dynamics of bias.

	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

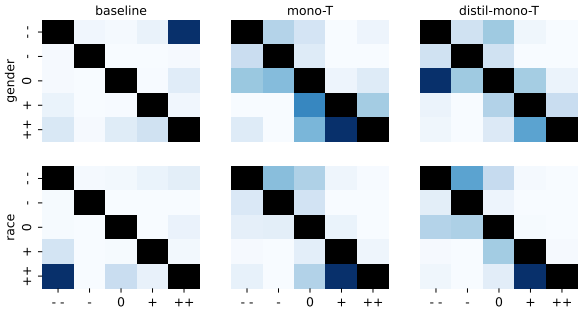
Table 3: F1 at convergence and steps at convergence for standard size, distilled, and baseline models. Performance is measured on the MARC data.

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

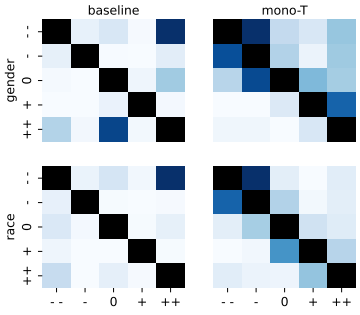
Figure 4 contains all confusion matrices for all languages, of which we displayed a subset in the body of this work.



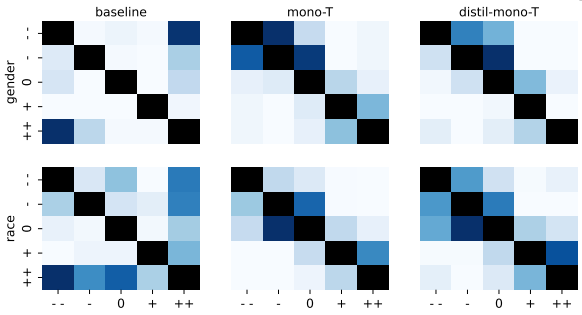
(a) Japanese (ja)



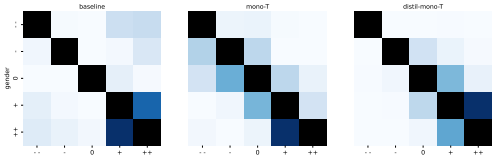
(b) German (de)



(c) Spanish (es)



(d) English (en)



(e) Chinese (zh_cn)

Figure 4: 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.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2

☑ B1. Did you cite the creators of artifacts you used?
2

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ 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)?
3

☐ 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.?
2

☑ 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.
3 methodology

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A

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

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ 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?
4.1

☑ 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.)?
4

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.