# Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method With Least Confounding Variables

**Valentin Barriere, Sebastian Cifuentes**
Universidad de Chile | CENIA
Computer Science Department
Santiago, Chile
vbarriere@dcc.uchile.cl, sebastian.cifuentes@cenia.cl

**Abstract**

Classical bias detection methods used in Machine Learning are themselves biased because of the different confounding variables implied in the assessment of the initial biases. First they are using templates that are syntactically simple and distant from the target data on which the model will be applied. Second, current methods are assessing biases in pre-trained language models or in dataset, but not directly on the fine-tuned classifier that can actually produce harms. We propose a simple method to detect the biases of a specific fine-tuned classifier on any type of unlabeled data. The idea is to study the classifier behavior by creating counterfactual examples directly on the target data distribution and quantify the amount of changes. In this work, we focus on named entity perturbations by applying a Named Entity Recognition on target-domain data and modifying them accordingly to most common names or location of a target group (gender and country), and this for several morphosynctactically different languages spoken in relation with the countries of the target groups. We used our method on two models available open-source that are likely to be deployed by industry, and on two tasks and domains. We first assess the bias of a multilingual sentiment analysis model trained over multiple-languages tweets and available open-source, and then a multilingual stance recognition model trained over several languages and assessed over English language. Finally we propose to link the perplexity of each example with the bias of the model, by looking at the change in label distribution with respect to the language of the target group. Our work offers a fine-grained analysis of the interactions between names and languages, revealing significant biases in multilingual models.

**Keywords:** Country-specific Bias, Machine Learning Classifiers, Perturbation Method, Multilingual Bias

## 1. Introduction

Biases in natural language processing (NLP) are everywhere, starting by the data (Wiegand et al., 2019), the annotations (Santy et al., 2023; Sap et al., 2022) and even the annotation campaign instructions (Parmar et al., 2023). Among other things, NLP models can drag moral (Hämmerl et al., 2022), social (Sap et al., 2020) or political biases (Feng et al., 2023).

The quantification of social bias is a prominent theme in recent research. It can be in multimodal data like image captioning (Hirota et al., 2022) or in general text (Czarnowska et al., 2021). This can be done using intrinsic methods that are evaluating the model's internal representation in different ways, or using extrinsic methods that measure how a model's performance on some task is sensitive to some attributes of a target group (Blodgett et al., 2020). The intrinsic methods are more general but their correlation to downstream tasks is questionable (Goldfarb-Tarrant et al., 2021; Cao et al., 2022) since the relation between intrinsic metrics and actual deviant behavior of the model that could be observed with extrinsic metrics is very opaque. Moreover, intrinsic metrics based on word embedding remains opaque because of the lack of transparency and interpretability (Valentini et al., 2023).

Extrinsic methods are based on the model performances (if they are lower for a target group) and predictions (if they change when the target group change). They are more straightfoward in the assessment of the model bias, however these approaches themselves are not immune to bias as they highly depends on the choice of variables (Badilla et al., 2020) and dataset used for evaluation (Orgad and Belinkov, 2022).

More generally, it is difficult to assess the impact of various variables on the bias of a deployed model, such as the target data and fine-tuning data. Indeed, *when assessing the bias of the pre-trained model, we ignore their final impact albeit they are potential confounding variables* (cf. Figure 1). First, the biases are assessed on a certain data distribution (i.e. a domain), and even intrinsic methods relying on templates (Czarnowska et al., 2021; Kurita et al., 2019; Guo and Caliskan, 2021) have been proven sensitive to template choice, revealing considerable variations in bias values and conclusions across template modifications (Seshadri et al., 2022). Then, existing techniques to assess biases in text-based models often fall short in providing comprehensive insights into the behavior of these models in production settings: the classifier models used afterward are not the same when fine-tuned over new data. By studying the production



*LREC-COLING 2024*, pages 1511–1518

model itself, we reduce the number of confounding variables that can impact the bias thereafter.

Even though names are not inherently associated with a particular nationality, they have been shown to contains nationality biases (Ladhak et al., 2023). Venkit et al. (2023) delve into the underexplored domain of nationality bias in language models, spotlighting the influence of demographic attributes on country biases. An and Rudinger (2023) provide insights into the interplay between demographic attributes and tokenization length, with a focus on first name biases. Zhu et al. (2023) present a novel approach for mitigating name bias by disentangling it from its semantics in machine reading comprehension. Lastly, Ladhak et al. (2023) investigate the propagation of name-nationality bias, showing that names and nationalities are binded using a intrinsic evaluation with templates, and how biases manifest themselves as hallucinations.

Because a set of different nationalities generally implies a set of different languages, multilingualism should be embedded in the bias evaluation method. The studies on multilingual bias assessment offer insights into detecting and mitigating biases in low-resource and non-English language contexts, but there are few resources for non-English languages, especially out of a non-Western context (Vashishtha et al., 2023). Kaneko et al. (2022) introduced the Multilingual Bias Evaluation score to bridge the gap in bias assessment for non-English languages using Machine Translation, however they created another bias by using Machine Translation on non-Western context. It is difficult to create a dataset for bias detection at multilingual scale because of the difference in cultures and religions. Template can seem like the easiest options (Das et al., 2023), because otherwise annotation is very costly (Sahoo et al., 2023). Finally, notable is the work of Câmara et al. (2022) who study intersectional bias for multiple languages, using simple templates. These studies collectively enhance our understanding of biases in multilingual settings, emphasizing the need for culturally relevant assessments. In the same way it is important to assess the bias of one model in different languages (Goldfarb-tarrant et al., 2023; Goldfarb-Tarrant et al., 2023), we argue that it should also be tested for different data domains.

Finally, social bias can also be annotated in order to explicitly detect them in a sentence, whether they are explicit or implicit (Sahoo et al., 2023), but this annotation part is costly and very language- and culture- dependent. The fine nuances in source language make machine translation hardly usable for this kind of task (Kaneko et al., 2022), making it impossible to use methods based on this like the one proposed by Barriere and Balahur (2020).

This paper addresses these challenges by proposing a novel method for detecting biases in

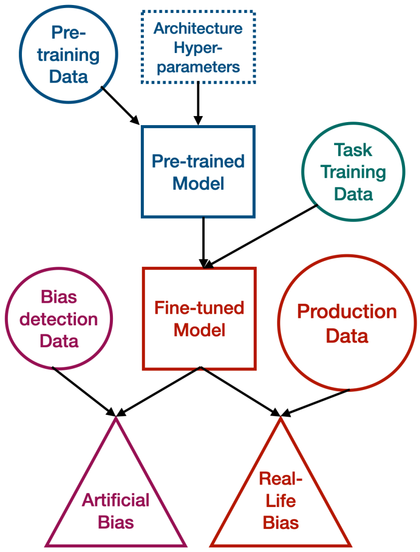


Figure 1: Different variables can alter the bias of the production model on production data. Assessing the bias on a dataset that does not follow the production data distribution adds a new confounding variable. Training phase is in blue while real-world model application is in red. Bias-detection datasets impact the bias estimation.

fine-tuned classifiers applied to unlabeled data. Unlike existing techniques that rely on syntactically simple templates or assess biases in pre-trained models or datasets, our dataset-agnostic approach directly evaluates the impact of a classifier on the target data distribution, allowing to uncouple datasets and metrics. We achieve this by conducting invariance tests by creating counterfactual examples using Named Entity Recognition (NER) and country-specific lexicons, before quantifying changes introduced by the classifier.

The work that is closer to our work is the one of Goldfarb-tarrant et al. (2023), where the authors are proposing a multilingual dataset in order to assess different biases. In our case, we are not relying on a particular dataset since this would implicitly add a new confounding variable in the bias assessment, as we are using automatic Named Entity Recognition (NER) on any sentence in order to create counterfactual data. Another close work is the one of Ribeiro et al. (2020) that propose invariance tests that consist of replacing named entities with others and look at the shift in the model's output. The difference in our work is that we are analyzing the problem at country-level, looking at the interactions between names and languages, with more fine-grained metrics.

As far as the authors know, no current method proposes to concretely assess the bias directly on production models that are deployed in our society, with explicit extrinsic metrics (Orgad and Belinkov, 2022; Orgad et al., 2022). Moreover, nobody proposed a multilingual study over names in order to assess the bias that they are dragging in a multilingual model. In their general framework, Czarnowska et al. (2021) are dividing the nationalities in 6 groups based on their GDP, but we argue that this division should be even more fine-grained and related to the country language. This paper shows that for at least two models based on one of the most used multilingual transformers (Conneau et al., 2020), there are strong biases towards names that are changing with respect to the language used. As we show patterns of aversion for names coming from countries not speaking the language used in the sentence, we name this phenomena '*AI model xenophobia*'.[1]

## 2. Method

The proposed method relies on NER to create counterfactual examples from the target-domain and specific of target groups, and to assess the bias quantifying the differences in the model outputs.[2]

**Notation** We decide to slightly change the notations of Czarnowska et al. (2021) because our target groups are country-related which can be defined by different attributes such as names of persons or locations. We use $\mathcal{A}$ as a set of target words sets such that $\mathcal{A} = \{A_1, A_2, ..., A_{|T|}\}$ where $A_t$ represents the target words set of the target group $t$ for the attribute $A$,[3] and $|T|$ the number of target groups that we consider. The set of source examples $X = \{x_1, x_2, ..., x_{|X|}\}$ contains the sentences from our target-domain data with at least one named entity (such as a person or a location), and $S' = \{S'_1, ..., S'_{|X|}\}$ the set of sets of perturbated examples, $S_j^{\prime t_i}$ the set of perturbated examples of the sentence j for the target group $i$. We use $\Phi$ as the score functions, and $d$ as the distance metrics used on top of the score functions.

**Country-Specific Entities Gazeteers** Our method is relying on country-specific gazeeters, that can be for different type of named entities: one gazeeter of a specific attribute $A$ from a given country $t$ will contain words related to this country. For example, if the name is the attribute and the country is France, we will obtain the set of the most common French names for man or woman $\mathcal{N}_{\text{France}} = \{\text{Matthieu}, \text{Jean}, \text{Sophie}, ...\}$ or surnames $\mathcal{S}_{\text{France}} = \{\text{Lepennec}, \text{Fourniol}, \text{Dubois}, ...\}$. The proposed method relies on gazeeters that are country-specific, that can be for different type of named entities. The authors of Ribeiro et al. (2020) collected common first and last names, but also the associated cities from several countries. This makes a total of 16771 male first names, 12737 female first names, 14797 last names and 5445 cities from 194 countries. For more information, the reader is referred to Appendix A.

**Data Perturbation** We use a multilingual NER system to identify entities for removal in target-domain data, aligning with the data used during model deployment. These entities, in combination with attributes $\mathcal{A}$, form a dataset for generating contrastive examples $S' = \{S'_1, ..., S'_{|X|}\}$ related to specific target groups. The random subtraction process follows Ribeiro et al. (2020) method using simple patterns and the Spacy library (AI, 2023).

**Bias Quantification** We use different methods to quantify biases. The most naive is the shift in output distribution caused by a non-causal perturbation of the input, assessed here with a distance $d$ between the distributions of the original and counterfactual examples. Even if this value means there is a bias, analyzing the class-level predictions is necessary to define it. We propose to compute a class-specific distance for models predicting classes related to positive or negative outcomes and infer a general valence. We compute the difference in positive and negative probabilities between the original and counterfactual examples, which we call $\Delta$ (see Eq. 1) and that represents how more positive the counterfactual example is. Finally, we also look at the augmentation/diminution of the predicted examples in each of the classes.

$$\Delta = \sum_{pos} p_{pos} - \sum_{neg} p_{neg} \tag{1}$$

## 3. Experiments

In a series of experiments using various datasets, we first investigate the impact of perturbations on a multilingual stance recognition system, focusing on English data in Section 3. This analysis aims to uncover how different countries influence English language and to gauge the gender-related impact across various languages. Subsequently, in Section 3, we extend this analysis to a multilingual context, utilizing a sentiment analysis model trained on Twitter data. We evaluate biases within this widely used model across 11 morphosyntactically diverse languages, all from the same domain.

**English Stance Recognition** We are focusing on the multilingual stance recognition dataset CoFE (Barriere et al., 2022), with the baseline model of

[1]Xenophobia is the fear of the strangers

[2]Our code is available online: `https://github.com/valbarriere/Bias_COLING24/`

[3]It can be name regarding the gender, surname, location,...

| Gender Metric | Male $\Delta$ | Other | Against | In favor | KL | Female $\Delta$ | Other | Against | In favor | KL |
|---|---|---|---|---|---|---|---|---|---|---|
| United Kingdom | -0.55 | 0.0 | 13.0 | -3.0 | 4.01 | -0.46 | 0.0 | 8.0 | -4.0 | 3.83 |
| Ireland | -0.62 | 0.0 | 12.0 | -4.0 | 4.23 | -0.57 | 0.0 | 10.0 | -5.0 | 4.18 |
| United States | -0.61 | 0.0 | 12.0 | -4.0 | 3.99 | -0.46 | 0.0 | 8.0 | -5.0 | 3.77 |
| Australia | -0.58 | 0.0 | 13.0 | -3.0 | 4.16 | -0.49 | 0.0 | 9.0 | -4.0 | 3.91 |
| New Zealand | -0.55 | 0.0 | 12.0 | -4.0 | 4.12 | -0.43 | 0.0 | 9.0 | -4.0 | 3.84 |
| Canada | -0.68 | 0.0 | 11.0 | -4.0 | 4.14 | -0.64 | 0.0 | 7.0 | -5.0 | 3.92 |
| South Africa | -0.66 | 0.0 | 10.0 | -4.0 | 4.07 | -0.59 | 1.0 | 7.0 | -6.0 | 3.80 |
| India | -0.81 | 0.0 | 6.0 | -5.0 | 4.72 | -1.17 | 1.0 | 8.0 | -9.0 | 4.73 |
| Germany | -0.98 | 0.0 | 10.0 | -6.0 | 4.26 | -0.77 | 1.0 | 8.0 | -6.0 | 3.94 |
| France | -1.03 | 1.0 | 8.0 | -7.0 | 4.29 | -0.91 | 2.0 | 3.0 | -9.0 | 4.13 |
| Spain | -1.70 | 2.0 | 7.0 | -11.0 | 4.80 | -1.52 | 2.0 | 6.0 | -11.0 | 4.52 |
| Italy | -1.82 | 2.0 | 8.0 | -12.0 | 4.74 | -1.47 | 2.0 | 5.0 | -12.0 | 4.31 |
| Portugal | -1.66 | 2.0 | 8.0 | -11.0 | 5.08 | -1.43 | 2.0 | 6.0 | -11.0 | 4.45 |
| Morocco | -1.44 | 2.0 | 6.0 | -11.0 | 5.48 | -1.41 | 3.0 | 2.0 | -13.0 | 5.42 |
| Hungary | -1.43 | 2.0 | 8.0 | -11.0 | 4.64 | -1.46 | 2.0 | 7.0 | -11.0 | 4.68 |
| Poland | -1.52 | 1.0 | 11.0 | -10.0 | 4.69 | -1.41 | 2.0 | 7.0 | -11.0 | 4.49 |
| Turkey | -1.58 | 2.0 | 5.0 | -12.0 | 5.13 | -1.34 | 2.0 | 5.0 | -12.0 | 4.78 |

Table 1: Metrics on the stance recognition model. $\Delta$ represents the difference of probability of the positive class and the negative class. The other values by class and by gender are the percentage of change in the classification output of the model.

a recent shared-task (Bondarenko et al., 2023) on this dataset. The data contains proposals from the online participatory democracy platform called the "*Conference for the Future of Europe*" which took place between 2021 and 2022. Any participant can write a proposals with a title and associated description, and comment over other participants proposals. In our study, we only focus on English comments.[4]

**Multilingual Sentiment Analysis** The second experiment is focusing on multilingual sentiment analysis models trained over tweets. We focused on the widely recognized XLM-T model from Barbieri et al. (2022), as it is a very frequently (>1M monthly downloads) employed model for multilingual sentiment analysis over tweets. As for the data, we focused on the associated datasets in Arabic, English, German, French, Spanish, Italian, and Portuguese from the same paper,[5] and added three other datasets to extend the variety of languages. For this reason, we collected tweets in languages from family that were initially missing, by using the Eurotweets (Mozetič et al., 2016) and the Bounti Turkish (Köksal and Özgür, 2021) datasets. Tweets from Polish, Hungarian and Turkish were added as languages from slavic, uralic and altaic families were not present.[6]

[4]We tried more languages but there was not enough entities detected in the other languages.

[5]We removed Hindi as we wanted to focus on near-Europe languages

[6]Note that we never used the sentiment label as our method only relies on the model's output distribution.

**Metrics** We set $\Phi$ as the output layer of the neural network, which consists of the distribution of class probabilities. For the class-agnostic metric we chose to set $d$ as the symmetrical Kullback–Leibler divergence on the probability distributions. Since we are focusing on a tri-class Sentiment Analysis and Stance Recognition, and because of the inherent nature of these tasks, the valence of the bias can be easily inferred using the three groups of classes *Positive*/*In favor*, *Negative*/*Against* and *Neutral*. We use this grouping in order to compute $\Delta$, which also serves as $d$.

**Experimental Protocol** For every sentence $x$, we create $50$ random perturbations of this sentence for each of the target countries. Other details can be found in Appendix B.

## 4. Results

**Multilingual Sentiment Analysis** The results presented in Figure 2 reveal a significant correlation between the language of the text and the language of the named entity concerning the difference in output probabilities between the positive and negative classes ($\Delta$) that we normalized per language to obtain a number between -1 and 1. In Arabic, English, French, Italian, Hungarian, Polish, and Turkish, entities from these languages yield the most positive results. Following closely, Portuguese and German names exhibit the second-highest positivity in their respective languages. Notably, Spanish names do not receive a positive sentiment score in Spanish text. These findings also highlight interesting connections between closely related languages: Italian

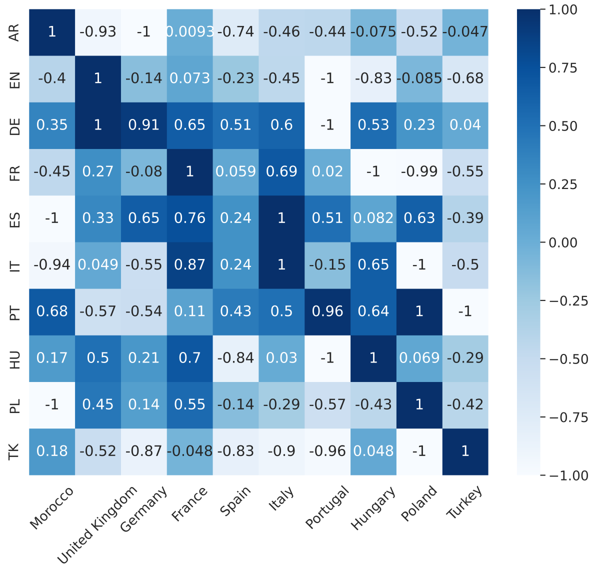


Figure 2: Matrix of $\Delta$ normalized per language.

names are perceived very positively in Spanish text, while names from the United Kingdom are rated highly positively in German. Less intuitive observations include English names having a more positive impact in German, but the reverse is not true. Surprisingly, Polish names are viewed very positively in Portuguese text.

**English Stance Recognition** Table 1 displays metrics related to the names of different countries. Notably, English-speaking country names, including those from countries with different primary languages like India and South Africa, consistently exhibit the lowest $\Delta$ values, indicating a more positive outcome. Specifically, names from the United States exhibit the lowest KL divergence, with values of 4.01 for males and 3.83 for females. However, it's important to highlight that Indian female names differ in their $\Delta$ compared to female names from other English-speaking countries. Concerning the gender, in general names from female are more positive than the ones from males, moreover the augmentation of *Against* prediction on the counterfacutal examples is lower than for the males (4%). Finally, it is also worth noting that the perturbed examples are less positive than the original ones, which might be due to distribution perturbation. And surprisingly when analyzing the predicted classes, the overall more positive countries show a higher augmentation of negative classifications than for other countries, and inversely the diminution of positive outcomes is far less.

## 5. Conclusion

We introduce an approach aimed at quantifying classifier biases with respect to named entities originating from various countries. Our method leverages counterfactual examples generated from data within the target domain, thereby mitigating the influence of confounding variables when assessing model biases deployed in practical applications. Furthermore, our investigation reveals a consistent phenomenon across two distinct multilingual tasks, namely stance recognition and sentiment analysis. In these two tasks, we first show that a bias can be detected by looking at the probability distribution, and second, that this bias can be defined more precisely. The models exhibit a propensity to assign more positive output to sentences containing named entities from countries where the language of the sentences is spoken, impulsing for the name '*AI model xenophobia*'.

## 6. Limitations

Our method only relies on Named Entities, so it does miss all the implicit hate speech. Nevertheless, it is a system with low recall but high precision as when it detects a change, it means that the classifier behavior is biased.

The use of lexicons implies another bias, even though they are the most frequent names of people and places. First, Paris in French is Parigi in Italian. we do not take this into account. Second, the script of the lexicon is always Latin, which is not true for every languages: Arabic is in Arabic script but Moroccan names were added in Latin script.[7]

Looking at the change in distribution is complex to interpret, for example in the case of a language model the distribution of words might change, because of the data distribution that indeed influence the prediction, without the possibility to explicitly find whether or not this mathematical bias (Meister et al., 2022) is a social bias.

## 7. Ethical Considerations

Our research on detecting and mitigating biases in fine-tuned NLP models places specific ethical considerations at the forefront. We are committed to the elimination of biases that could perpetuate discrimination or harm marginalized groups, prioritizing non-discrimination and fairness. We have made our code open-source to facilitate the accessibility and utilization of our method by anybody on their models and datasets.

[7] Interestingly, we still detect the same pattern than for other languages/names

## 8. Acknowledgements

Valentin thanks both Alexandra Balahur and Felipe Bravo for the early discussions on this work. This research has been funded by National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

## 9. Bibliographical References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, pages 265–283.

Explosion AI. 2023. spaCy: Industrial-strength Natural Language Processing in Python.

Haozhe An and Rachel Rudinger. 2023. Nichelle and Nancy : The Influence of Demographic Attributes and Tokenization Length on First Name Biases. In *ACL*, volume 2, pages 388–401.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: The word embeddings fairness evaluation framework. *IJCAI International Joint Conference on Artificial Intelligence*, 2021-Janua:430–436.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: A Multilingual Language Model Toolkit for Twitter. In *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis @ ACL*.

Valentin Barriere and Alexandra Balahur. 2020. Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation. In *COLING*.

Valentin Barriere and Alexandra Balahur. 2023. Multilingual Multi-target Stance Recognition in Online Public Consultations. *MDPI Mathematics – Special issue on Human Language Technollogy*, 11(9):2161.

Valentin Barriere, Alexandra Balahur, and Brian Ravenet. 2022. Debating Europe : A Multilingual Multi-Target Stance Classification Dataset of Online Debates. In *Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), LREC*, June, pages 16–21, Marseille, France. European Language Resources Association.

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

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan-Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché,2023: Argument and Causal Retrieval. In *ECIR*.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic. In *LTEDI 2022 - 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, Proceedings of the Workshop*, pages 90–106.

Yang Trista Cao, Yada Pruksachatkun, Kai Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2:561–570.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-Lingual Representation Learning at Scale. pages 31–38.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward Cultural Bias Evaluation Datasets: The Case of {B}engali Gender, Religious, and National Identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair

NLP Models. In *ACL*, volume 1, pages 11737–11762.

Seraphina Goldfarb-tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias Beyond English : Counterfactual Tests for Bias in Sentiment Analysis in Four Languages. In *Findings of ACL: ACL 2023*, pages 4458–4468.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 1926–1940.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models. In *Findings of ACL: ACL 2023*, pages 2209–2225.

Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin A. Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *Findings of ACL: ACL 2023*, pages 2137–2156.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying Societal Bias Amplification in Image Captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:13440–13449.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender Bias in Masked Language Models for Multiple Languages. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 2740–2750.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. pages 166–172.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 3198–3211.

Clara Meister, Wojciech Stokowiec, Tiago Pimentel, Lei Yu, Laura Rimell, and Adhiguna Kuncoro. 2022. A Natural Bias for Language Generation Models. In *ACL*, volume 2, pages 243–255.

Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. *GeBNLP 2022 - 4th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop*, pages 151–167.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How Gender Debiasing Affects Internal Model Representations, and Why It Matters. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 2602–2628.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1771–1781.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models. *ACL*.

Nihar Ranjan Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With Prejudice to None : A Few-Shot , Multilingual Transfer Learning Approach to Detect Social Bias in Low Resource Languages. In *Findings of ACL: ACL 2023*, pages 13316–13330.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. 1:9080–9102.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics*, pages 5477–5490.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5884–5906.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying Social Biases Using Templates is Unreliable. (Tsrml).

Francisco Valentini, Germán Rosati, Damián Blasi, Diego Fernandez Slezak, and Edgar Altszyler. 2023. On the interpretation and significance of bias metrics in texts: a PMI-based approach. In *ACL*, volume 2, pages 509–520.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On Evaluating and Mitigating Gender Biases in Multilingual Settings. In *Findings of ACL: ACL 2023*, pages 307–318.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 116–122.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: The problem of biased datasets. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:602–608.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. Causal Intervention for Mitigating Name Bias in Machine Reading Comprehension. In *Findings of ACL: ACL 2023*, 2021, pages 12837–12852.

## 10. Language Resource References

Barriere, Valentin and Jacquet, Guillaume Guillaume and Hemamou, Leo. 2022. *CoFE: a new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform*.

Köksal, Abdullatif and Özgür, Arzucan. 2021. *Twitter dataset and evaluation of transformers for Turkish sentiment analysis*. IEEE.

Mozetič, Igor and Grčar, Miha and Smailović, Jasmina. 2016. *Multilingual Twitter sentiment classification: The role of human annotators*. Public Library of Science San Francisco, CA USA.

## A. Gazeteers

These lexicons were obtained from the Wikidata Query Service.[8] As we noticed incoherences in the cities per country lexicons (France did not have big cities like Toulouse), we decided to enhance these lexicons by running our own requests and added the biggest cities.[9] This makes a total of 16771 male first names, 12737 female first names, 14797 last names and 5445 cities from 194 countries.

## B. Experimental Protocol

In all our experiments we avoided examples seen during the training phase of the model. The test sets partition from Barbieri et al. (2022) were used for the XLM-T dataset, and CF$_{E-T}$ and CF$_{U}$ were used as test sets on CoFE (Barriere and Balahur, 2023). We use the best model, pre-trained over Debating Europe (Barriere et al., 2022). Experiments were run using Tensorflow 2.4.1 (Abadi et al., 2016), transformers 3.5.1 (Wolf et al., 2019), a GPU Nvidia RTX-8000 and CUDA 12.0.

[8] https://query.wikidata.org/

[9] We also collected the most common names by scrapping Wikipedia pages and added this resource to our code, but we found out it was more straightforward to use the Checklist toolbox.