A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifiers

Valentin Barriere
Universidad de Chile – DCC, CENIA
Santiago, Chile
vbarriere@dcc.uchile.cl

Sebastian Cifuentes
CENIA
Santiago, Chile
sebstian.cifuentes@cenia.cl

Abstract

In this paper, we apply a method to quantify biases associated with named entities from various countries. We create counterfactual examples with small perturbations on target-domain data instead of relying on templates or specific datasets for bias detection. On widely used classifiers for subjectivity analysis, including sentiment, emotion, hate speech, and offensive text using Twitter data, our results demonstrate positive biases related to the language spoken in a country across all classifiers studied. Notably, the presence of certain country names in a sentence can strongly influence predictions, up to a 23% change in hate speech detection and up to a 60% change in the prediction of negative emotions such as anger. We hypothesize that these biases stem from the training data of pre-trained language models (PLMs) and find correlations between affect predictions and PLMs likelihood in English and unknown languages like Basque and Maori, revealing distinct patterns with exacerbate correlations. Further, we followed these correlations in-between counterfactual examples from a same sentence to remove the syntactical component, uncovering interesting results suggesting the impact of the pre-training data was more important for English-speaking-country names. Our anonymized code is available here.

1 Introduction

Recent trend in Natural Language Processing research, like in works published at conference such as ACL (Rogers et al., 2023), is to provide open-source data and models (Scao et al., 2022). This practice not only enhances its value for general research purposes but also facilitates the deployment of these models in diverse operational settings by companies or stakeholders. Applications such as customer experience, CV screening, Social Media analyses and moderation are example of applications that will directly impact the users in different ways. For this reason, the models applied at large scale should be scrutinized in order to understand their behavior and should tend to be fair by passing successfully a series of test to reduce their biases toward various target groups. Past study (Ladhak et al., 2023) showed that PLMs are impacted by names, and Barriere and Cifuentes (2024) proposed a method to quantify this to detect biases of the model toward specific countries, using the country most common names as a proxy. We are showing in this paper that this bias is systematic in several widely-used off-the-shelf classifiers on English data, and propose a method to directly link the bias level with the perplexity of the PLM

Contributions We propose an investigation into biases related to country-specific names in widely used off-the-shelf models (Barbieri et al., 2020, 2022), commonly deployed in production environments for Twitter data.[1] Our analysis reveals distinct biases in sentiment, emotion, and hate speech classifiers, showing a propensity to favor names from certain countries while markedly disfavoring those from less Westernized nations, often by a large margin. Furthermore, we establish a global-level correlation between the perplexity of associated PLMs and model predictions across both known and unknown (i.e., Out-of-Distribution; OOD) languages, demonstrated through examples in English, Basque, and Maori. At a local level, we mitigate the influence of syntax on perplexity by examining the correlation among counterfactual examples generated through minor perturbations. Notably, our findings suggest that the frequency of a name's occurrence during the training phase directly impacts the sentiment model's tendency to produce positive outputs, which highly disadvantage the non-English (i.e., OOD) persons in a world

[1] Regarding the number of monthly downloads of `cardiffnlp` models from Barbieri et al. (2020, 2022) in the Huggingface Model Hub at the time of writing (>4m for sentiment).

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 569–579

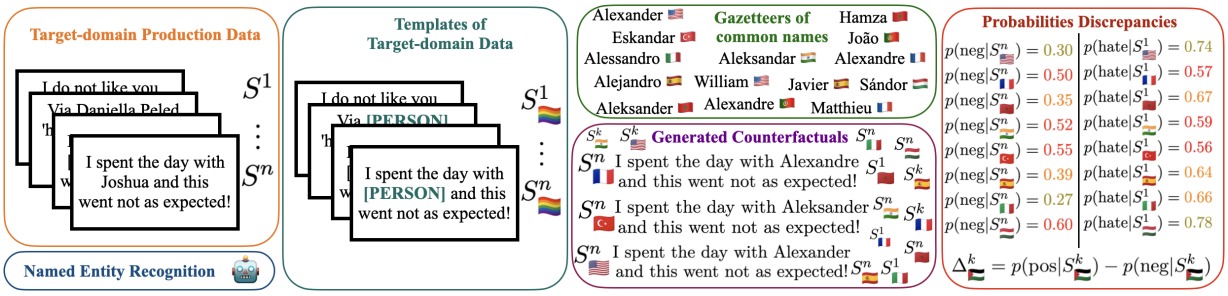

Figure 1: Overview of the counterfactual example creations. We show examples with sentiment and hate speech for variation of the name "*Alexander*" and two sentences S^1 and S^n. S^1 : "*I do not like you [PER] you fucking bitch*". The **NER** is applied to the **production data** to create **templates**, which are then filled randomly with most common names from **gazeeters** of different countries to create a pool of **counterfactuals**. The **discrepancies** in probabilities is quantified using metrics such as Δ.

where English is widely utilized as pivot language. ***Our method is unsupervised, moreover it can be applied to any classifier and any dataset***.

2 Related Work

As it is known that models still learn bias when fine-tuned on downstream tasks and that the correlation is low between the intrinsic bias scores of the initial model and its extrinsic bias scores after fine-tuning (Kaneko et al., 2022a, 2024), we use a method to evaluate an already trained classifier and not the pre-trained language model. Some works propose such thing as general "unit-test" for NLP models (Ribeiro et al., 2020) or even applying a battery of fairness tests (Nozza et al., 2022). However, extrinsic methods mainly relies on template or datasets (Czarnowska et al., 2021; Kurita et al., 2019; Guo and Caliskan, 2021), which have been proven to influence considerably the bias estimation and conclusion across template modification (Seshadri et al., 2022). A potential solution is to apply perturbation on the test data. Perturbations can be used for attribution methods (Fel et al., 2023), but also for testing a model's robustness (Ribeiro et al., 2020). They allow getting rid of the aforementioned template issue and data collection methodology: directly used on the target domain data, it prevents for not properly evaluating the intended notion of bias (Blodgett et al., 2020).

The origin of the bias generally comes from the training data (Caliskan et al., 2017), as a lot of information can be stored in the network (Petroni et al., 2019; Carlini et al., 2021, 2018) due to repetitions of the same sentences or concepts. This type of over-representation in the training data involve a representation bias, such as the one demonstrated by Kaneko and Bollegala (2022) regarding the gender as masculine was over-represented. This was found out to be correlated with the likelihood of the model. For example, Barikeri et al. (2021) propose a perplexity-based bias measure meant to quantify the amount of bias in generative language models along several bias dimensions. For this reason, Kaneko et al. (2022b) propose to use the likelihood as a proxy to estimate the bias on gender. In our case, we validate that the bias is already present in the PLM, by calculating the correlation between the likelihood and different classes for country-name. This technique is even more efficient with generative models (Ouyang et al., 2022; Jiang et al., 2024) as one can apply it directly on production model.

Although names are not inherently linked to a specific nationality, research has revealed the presence of nationality biases within them. Delving into this underexplored domain, Venkit et al. (2023) shed light on the influence of demographic on biases associated with countries in language models. An and Rudinger (2023) offer insights into the intricate relationship between demographic attributes and tokenization length, particularly focusing on biases related to first names. Zhu et al. (2023) propose to mitigate name bias by disentangling it from its semantic context in machine reading comprehension tasks. Ladhak et al. (2023) investigate the propagation of name-nationality bias, demonstrating through intrinsic evaluation with templates how names and nationalities are intrinsically linked and how biases manifest as hallucinations. Lastly, Barriere and Cifuentes (2024) showed that using names as proxy works to detect country-related biases depends on the sentence's language, in multilingual sentiment and stance recognition models (Barriere and Balahur, 2023; Barriere and Jacquet, 2022; Barriere et al., 2022).

3 Method

We first rely on Named Entity Recognition (NER) to create counterfactual examples from the target-domain, specific of target groups, following the methodology of Barriere and Cifuentes (2024). The bias is assessed by quantifying the differences in the model outputs. Second, we ran a series of experiences studying the correlation between the output variations and the perplexity. Figure 1 shows an overview of the bias detection.

3.1 Perturbation-based Counterfactuals

Counterfactual Generation A set of counterfactual examples are constructed from the target-domain data using a NER system combined with a list of most common names from different countries. Each named entity automatically tagged as person is substituted by a random common name from a specific country. Note that the original entity is conserved, by looking in our gazeeters its corresponding gender. More details are found in the Appendix A.

Bias Calculation In order to assess the bias, we calculate the percentage of change in terms of tagged examples, using the confusion matrices. For sentiment, we also computed the change in difference in probability between positive predictions and negative predictions Δ.

3.2 Perplexity and Likelihood

General and Pseudo-Perplexity The perplexity of a language model measures the likelihood of data sequences and represents how fluent it is (Carlini et al., 2018). In simpler terms, perplexity reflects how unexpected a particular sequence is to the model. A higher perplexity suggests that the model finds the sequence more surprising, while a lower perplexity indicates that the sequence is more likely to occur. We refer to the definition of pseudo-log-likelihood introduced by Salazar et al. (2020), the pseudo-perplexity being the opposite of it. For a sentence $S = w_1, w_2, ..., w_{|S|}$, the pseudo-log-likelihood (PLL) score given by Eq. 1, can be used for evaluating the preference expressed by an MLM for the sentence S.

$$PLL(S) = -\sum_{i=1}^{|S|} \log P_{MLM}(w_i | S_{\setminus wi}; \theta) \quad (1)$$

The Log-Perplexity as defined in Carlini et al. (2018) is the negative log likelihood, hence **we use pseudo-log-perplexity as simply the opposite of the PLL.**[2] More details are provided in Appendix B. In the following, we will not use the term pseudo- when talking about the pseudo- perplexity or likelihood.

Bias quantification We calculated the Pearson correlation between the probabilities output and likelihood in two ways. First, what we call *global* correlation, i.e., between all the examples of the dataset, in order to shed lights on a general pattern between perplexity and subjectivity. Second, what we call *local* correlations, i.e., between elements coming from the same original sentence, before averaging them. In this way, we can disentangle the syntactic aspect of the sentences that have an impact in the likelihood calculation. This is similar to normalizing the perplexity and likelihood of every examples coming from the same sentence before calculating the Pearson correlation.

4 Experiments and Results

4.1 Experiments

Bias Detection Our first experiment focuses on quantifying the country names bias for different off-the-shelf models previously learned on tasks that are related to affects, looking at the probability of positiveness and the percentage of change in number of predicted examples per class.

Global Perplexity The second experiment aims to show that the model predictions are in general intricately linked with the perplexity even for unknown languages. We first create datasets in these unknown languages using Machine Translation (MT) in order to preserve the semantic content in-between the different languages, as they did in in Balahur and Turchi (2013). We then calculate the "global" correlation between perplexity and output probabilities in English and unknown languages such as Maori and Basque, which we obtain using Google Translate.[3] More details in Appendix C.

Local Perplexity To remove the syntactic aspect influencing both perplexity and predictions, we conduct experiments focusing on what we call "local" correlation, which is between the relative probabilities of each class among counterfactual examples

[2] Contrary to the definition of Salazar et al. (2020) defining it on a complete corpus, summing between all the sentences before passing it to exponential.

[3] Google MT is based on the LLM PaLM 2 (Google, 2023), which should work reasonably well for these two languages already used in production.

Country	Sentiment				Emotion				Hate		Offensive	
	Δ	−	≈	+	Joy	Opt.	Anger	Sad.	Non-hate	Hate	Non-off.	Off.
United Kingdom	-1.43	5.4	1.3	-4.6	-2.1	0.6	2.7	6.4	-0.2	23.5	-0.4	4.8
United States	-1.35	5.0	1.7	-4.9	-2.3	-0.5	4.0	6.5	-0.2	22.0	-0.5	6.1
Canada	-1.43	5.5	1.5	-5.0	-1.6	-0.2	2.3	5.0	-0.2	21.0	-0.4	4.5
Australia	-1.37	5.7	1.2	-4.7	-2.3	0.9	3.2	6.6	-0.2	23.0	-0.3	4.3
South Africa	-1.58	5.9	1.2	-4.8	-1.5	0.4	1.0	6.1	-0.2	22.5	-0.3	3.9
India	-2.70	7.9	-0.1	-4.4	-2.5	-6.1	8.7	5.0	-0.1	10.0	0.1	-1.6
Germany	-2.14	6.4	1.3	-5.3	-0.0	-4.8	-0.2	4.7	-0.1	19.0	-0.3	3.3
France	-1.58	7.7	-0.2	-4.0	0.9	-5.1	-2.5	3.8	-0.1	10.5	-0.0	0.1
Spain	-2.46	6.0	2.6	-6.5	1.7	-13.0	-0.4	2.7	-0.0	6.0	-0.2	2.7
Italy	-1.98	7.1	1.1	-5.4	2.5	-15.5	-0.9	1.5	-0.1	12.5	-0.2	2.5
Portugal	-2.30	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5	-0.1	1.8
Hungary	-2.26	4.9	2.7	-6.1	2.4	-17.2	-1.4	4.0	-0.1	6.5	0.2	-2.1
Poland	-2.02	3.4	3.6	-6.3	2.0	-13.7	-2.4	5.1	-0.1	9.5	0.1	-1.3
Turkey	-2.33	6.8	0.7	-4.7	0.2	-11.9	4.8	1.7	-0.1	7.5	0.0	-0.3
Morocco	-2.04	4.2	2.4	-5.2	-9.0	-33.2	60.3	-17.4	-0.0	2.0	0.4	-4.9

Table 1: Changes in probability output (Δ) and in percentage of examples in each of the predicted classes, both relative to the original unmodified sentence to compare with the model's likely real-world production settings.

(i.e., generated with minor perturbations) and their associated relative perplexity.

4.2 Experimental Protocol

Gazeeters We used the dataset collected from Wikidata Query Service.[4] by the authors of `Checklist`, composed of common first and last names as well as the associated cities from several countries. This makes a total of 16,771 male first names, 12,737 female first names, 14,797 last names from 194 countries.

NER We use a multilingual off-the-shelf NER system available on the Spacy library (AI, 2023) and created for social media (named `xx_ent_wiki_sm`) to identify entities for removal in target-domain data, aligning with the data used during model deployment.

Perturbation For every sentence x, we create 50 random perturbations of this sentence for each of the target countries.

Dataset In order to apply our method to data similar to production data, we collected 8,891 ***random tweets in English*** by using the IDs from the Eurotweets dataset (Mozetič et al., 2016). The 8,891 tweets used in the experiment correspond to a random selection of 10% of the English tweets of the EuroTweets dataset (Mozetič et al., 2016) downloaded in June 2020.[5]

Tested Classifiers The models used were the ones of (Barbieri et al., 2020, 2022) for multilingual sentiment analysis, monolingual hate speech, emotion recognition and offensive text detection: `cardiffnlp/twitter-xlm-roberta-base-sentiment`, `cardiffnlp/twitter-roberta-base-hate`, `cardiffnlp/twitter-roberta-base-emotion`, and `cardiffnlp/twitter-roberta-base-offensive`. 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.

4.3 Results

Bias Detection Table 1 provides a comprehensive overview of the impact of country-specific named entities on sentiment, emotion, hate speech, and offensive text classifications across diverse classifiers. Notably, it reveals significant variations in model predictions based on the presence of different country names within textual data. For sentiment analysis, it is striking to observe substantial shifts in sentiment probabilities (Δ)[6] across countries. For instance, countries like India, Turkey or Spain exhibit noteworthy deviations in sentiment probabilities, indicating potential biases in classifier outputs concerning specific national contexts.[7] The percentages of predicted negative, neutral, and positive sentiments further underscore the nuanced nature of these biases, with certain countries con-

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

[5] No label were used.

[6] Δ's standard deviations are proportional to its values.

[7] This is interesting as Spanish (resp. Indian dialects) are the main foreign languages of migrants in US (resp. UK).

Task	Label	English	Basque	Maori
Hate		3.17	23.07	22.31
Sentiment	−	-11.39	25.48	35.33
	≈	19.27	-19.98	-36.23
	+	-5.41	-3.04	5.86

Table 2: Global correlations between PPL and classes for different languages, tasks or pre-trainings.

sistently receiving more positive or negative sentiment classifications compared to others. Emotion analysis reveals intriguing patterns in the distribution of predicted emotions across countries. Optimism shows an interesting pattern where the non-English names highly decrease this prediction, up to -33% for Moroccan. It is also notable that Moroccan names provoke a very high increase (60%) of anger predictions at the expense of the other classes. Finally, a similar pattern can be seen for the hate speech and offensive text classifiers. English-speaking countries names highly favor hate speech detection, even as a false positive, compared to other countries. For offensive text detection, there is an increase of 6.1% with counterfactuals using US names and a decrease of 4.9% and 2.1% using Moroccan and Hungarian names.

Global Subjectivity-Perplexity Correlation Table 2 shows the correlations between the perplexity and the labels for Sentiment and Hate speech tasks using tweets from different languages, obtained using Machine Translation. For the hate speech model, the global correlation between the hate speech class and the perplexity is almost close to zero for English data, which is good since showing no spurious pattern between perplexity and hate speech prediction. However, the correlations are higher for the unknown language such as Basque and Maori, where it reaches more than 22%. The model tends to classify as hate speech more easily texts having a higher perplexities, i.e., that are outside the training distribution. For the Sentiment model, the pattern for Basque and Maori language is the same, high positive/negative correlation for the negative/positive class, which means that the less the sentence is similar to the train distribution, the more negative it would be. Additional experiments using other languages are confirming the results, and are available in Appendix D.

Local Subjectivity-Perplexity Correlation Table 3 shows correlations between the relative perplexity of the model and the probabilities of different classes. The results are very different from global correlations. Notably, there is a negative correlation between perplexity and positiveness of the sentiment, which implies that names that are more similar to what was seen during the PLM pre-training will imply a more positive output of the sentiment classifier. This trend is particularly pronounced among English-speaking countries. Due to lack of space, more details and results can be found in Appendix E.

Country	Sentiment		
	−	≈	+
United Kingdom	15.03	5.89	-18.26
United States	14.70	6.63	-18.41
Canada	15.18	4.91	-17.68
Australia	15.68	5.46	-18.52
South Africa	13.12	5.87	-16.67
India	7.64	5.18	-11.75
Germany	13.62	4.50	-16.34
France	8.18	4.42	-11.47
Spain	11.37	4.16	-14.23
Italy	11.09	3.79	-13.57
Portugal	9.45	2.93	-11.97
Hungary	8.37	2.89	-10.79
Poland	9.88	3.22	-12.32
Turkey	9.62	2.79	-11.86
Morocco	9.07	-0.16	-8.25
Overall	11.17	4.63	-14.40

Table 3: Correlations between the relative perplexity of the model and the relative output probabilities.

5 Conclusion

Bias at the nationality level can also occur with the most common entities of the country such as names. We show its occurrence in this paper for a set of tasks that are related with affect and subjectivity classification, using several transformer models widely used on Twitter data. Motivated by prior research, we studied the link between this bias and the perplexity of the PLM showing *(i)* exacerbate correlations in unknown languages, and *(ii)* verify that correlation can be related to names using counterfactual sentences. We found out interesting patterns using the Pearson correlations between the classes and perplexity, revealing higher correlations for English-speaking country names, meaning that the exposition bias on names impacts the predictions also in-between a country.

6 Limitations

First, 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, meaning that the classifier behavior is biased. Second, even if our method slightly perturbates the data from the target distribution, it does not explicitly keep it inside, creating examples that might be a bit outside the distribution of the production data. We think that is the reason why we see a general shift toward a more negative sentiment when comparing perturbated examples and true examples (negative predictions always augment while positive predictions always decrease). It would be more natural to use target-data-specific lexicons, or use a generative model to do the job. However, we think that this is a fair comparison toward all the countries and it can drive a pertinent conclusion on a relative bias between the different countries. Another bias induction can also come from the fact that some names can be non gendered in some context, such as Claude as a first-name or Jane as a surname (for a man) that would be tagged as feminine. Co-reference resolution could mitigate this issue, even though we believe it is uncommon. Finally, we compare a masked language model, but further experiments are left for future workusing generative models such as flan-T5 (Chung et al., 2022) or Mixtral (Jiang et al., 2024) where the same model computes both label and perplexity, for example using label tokens probabilities to estimate the probabilities (Hegselmann et al., 2023).

Acknowledgements

The authors thank the reviewers for the various comments that helped to improve the manuscript. This work has been partially funded by National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

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.

Alexandra Balahur and Marco Turchi. 2013. Improving sentiment analysis in twitter using multilingual machine translated data. *International Conference Recent Advances in Natural Language Processing, RANLP*, (September):49–55.

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*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 1644–1650.

Soumya Barikeri, Anne Lauscher, Ivan Vulic, and Goran Glavaš. 2021. REDDITBIAS: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955.

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.

Valentin Barriere and Sebastian Cifuentes. 2024. Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method with Least Confounding Variables. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1511–1518, Torino, Italia. ELRA and ICCL.

Valentin Barriere and Guillaume Jacquet. 2022. CoFE : A New Dataset of Intra-Multilingual Multi-target Stance Classification from an Online European Participatory Democracy Platform. *AACL-IJCNLP*.

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.

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

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2018. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H Chi, Jeff Dean, Jacob Devlin, Adam Robert, Denny Zhou, Quoc V Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models.

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.

Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. 2023. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. In *CVPR*.

Google. 2023. PaLM 2 Technical Report. (May).

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.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, pages 5549–5581.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the Mask - Evaluating Social Biases in Masked Language Models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 11954–11962.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. The Gaps between Pre-train and Downstream Settings in Bias Evaluation and Debiasing.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022a. Debiasing isn't enough! – On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks. In *Proceedings - International Conference on Computational Linguistics, COLING*, volume 29, pages 1299–1310.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022b. 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.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLoS ONE*, 11(5):1–26.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for Social Bias Testing of Large Language Models. *2022 Challenges and Perspectives in Creating Large Language Models, Proceedings of the Workshop*, pages 68–74.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Ameida, Carroll L. Wainwright, Pamela Mishkin, C L Mar, Jacob Hilton, Amanda Askell, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv*, https://op.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases?

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

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Figure 1, pages 2699–2712.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin

Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

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

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.

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.

A Counterfactual Examples Creation

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,[8] and $|T|$ the number of target groups that we consider. The set of source examples $\mathcal{S} = \{S^1, S^2, ..., S^{|\mathcal{S}|}\}$ contains the sentences from our target-domain data with at least one named entity (such as a person or a location), and $\mathcal{S}' = \{S'_1, ..., S'_{|\mathcal{S}|}\}$ the set of sets of perturbated examples, $S'_i = \{S^i_{t,j}, j = 1..E\}$ the set of perturbated examples of the sentence i for the target group t, with E the number of counterfactual examples. We use Φ as the score functions, and d as the distance metrics used on top of the score functions.

In the example in Figure 1, for simplicity reasons we show only one example of name per country, which means $j = 1$ in $S^i_{t,j}$ and t is represented as the flag of the country.

Country-Specific Entities Gazeeters 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 last names $\mathcal{L}_{\text{France}} = \{\text{Lepennec}, \text{Fourniol}, \text{Denis}, ...\}$. The proposed method relies on gazeeters that are country-specific, that can be for different type of named entities.

Data Perturbation The detected entities, in combination with attributes $\mathcal{A}$, form a dataset for generating contrastive examples $\mathcal{S}' = \{S'_1, ..., S'_{|\mathcal{S}|}\}$ related to specific target groups. The random subtraction process follows Ribeiro et al. (2020) method using simple patterns and the Spacy library (AI, 2023). Even though the model utilized is robust and widely employed in the industry, given the noisy nature of tweets, it may occasionally miss a name but is more likely to rightfully detect one (with lower recall but higher precision on noisy data). We manually examined 100 examples where a Person (PER) entity was detected in our downloaded data, and found a satisfying precision of the NER to be 88%. Subsequently, our method utilizes as templates examples with detected names (which are pertinent templates if precision is high).

B Pseudo-Likelihood

It noteworthy that it is possible to use other metrics such as the All Unmasked Likelihood (AUL) or AUL with Attention weights of Kaneko and Bollegala (2022). Nevertheless, in our case we use

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

Label	English	Dutch	Spanish	Hindi	Malayalam	Turkish	Basque	Maori
−	-11.39	-13.87	-6.28	-10.89	-7.03	-6.02	25.48	35.33
≈	19.27	21.61	19.00	25.54	9.12	16.54	-19.98	-36.23
+	-5.41	-7.13	-11.10	-13.50	-1.94	-10.32	-3.04	5.86

Table 4: Global correlations between PPL and classes for different languages using the multilingual sentiment model

Country	Sentiment			Emotion				Hate	Offensive
	−	≈	+	Anger	Joy	Opt.	Sadness		
United Kingdom	15.03	5.89	-18.26	2.02	6.82	-16.46	14.87	3.96	2.75
Ireland	11.69	5.78	-15.72	0.21	8.77	-15.30	11.78	2.67	5.20
United States	14.70	6.63	-18.41	1.99	8.23	-19.01	17.09	4.44	4.90
Canada	15.18	4.91	-17.68	1.62	7.10	-16.73	15.22	2.97	4.31
Australia	15.68	5.46	-18.52	2.06	7.70	-17.55	15.50	4.10	3.03
New Zealand	15.17	4.80	-17.65	3.29	5.95	-17.53	16.48	3.23	2.21
South Africa	13.12	5.87	-16.67	1.47	6.79	-16.26	14.97	3.67	3.50
India	7.64	5.18	-11.75	-0.37	-12.23	10.32	1.84	2.50	12.03
Germany	13.62	4.50	-16.34	2.66	4.37	-12.99	11.61	2.12	4.15
France	8.18	4.42	-11.47	1.66	5.37	-10.79	7.51	2.59	10.19
Spain	11.37	4.16	-14.23	1.97	4.47	-9.59	6.10	-1.16	2.36
Italy	11.09	3.79	-13.57	0.39	1.69	-5.67	6.14	-1.92	0.76
Portugal	9.45	2.93	-11.97	0.51	3.29	-7.23	6.09	-1.15	2.73
Hungary	8.37	2.89	-10.79	2.02	-0.57	-5.71	7.08	-3.95	0.73
Poland	9.88	3.22	-12.32	-0.99	5.47	-6.72	3.67	-4.45	6.66
Turkey	9.62	2.79	-11.86	1.25	-1.25	-5.50	9.02	-2.74	0.73
Morocco	9.07	-0.16	-8.25	2.07	-25.60	21.88	8.76	1.53	-4.44
Overall	11.17	4.63	-14.40	2.77	-3.66	-5.05	10.61	1.69	2.38

Table 5: Correlations between the relative perplexity of the model and the relative probabilities of the different classes. We only use hate and offensive speech detection as it is binary classification.

examples from the target domain, hence we do want to take into account the bias introduced by the other unmasked token words in the context. Indeed, the models studied in this work are likely be deployed on data following the same distribution.

C Machine Translation

Google Translate was employed as MT, known for its up-to-date machine translation capabilities, although originally intended for general text rather than tweets. However, we do not see this as crucial. We did not check if the label is conserved because it is not the purpose as our method does not even use the original labels: the method in the 2nd experiments measures the correlation between output labels and tweet perplexity, whether it is in English, Maori or Basque. Our aim in utilizing MT was to maintain tweet content while creating our tweets in low-resource languages, as Balahur and Turchi (2013) did.

D Global Subjectivity-Perplexity Correlation

We extend the experiments of Table 2, using the exact same setting, but with other languges: Dutch, Spanish, Hindi, Malayalam and Turkish. We show the results in Table 4. It is possible to see that the sentiment model is behaving for these "known languages" the same way it behaves with English, with a negative correlations on the negative and positive sentiment and a positive correlation with the neutral sentiment. The behavior that we see for out-of-distribution languages such as Maori or Basque is very different.

E Local Subjectivity-Perplexity Correlation

Table 5 show the local correlations between the perplexity and probability outputs for all the classifiers. Regarding emotions, optimism and sadness show the same patterns than positive and negative sentiments. Surprising reverse trends are observed for

Indian and Moroccan names in the positive emotion, which means the more (resp. less) stereotype is the name, the more it tend to classify joy (resp. optimism). Regarding hate speech and offensive text, the correlation are low. However, for hate speech we can notice that the trend is almost reverse between English-speaking and non-English-speaking countries.