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

Valentin Barriere, Sebastian Cifuentes

Universidad de Chile – DCC — CENIA

LREC-COLING 24, Torino

Introduction

Intro

- Fairness in IA

Intro

- Fairness in IA

- Generally coarse, based on GDP

Intro

- Fairness in IA

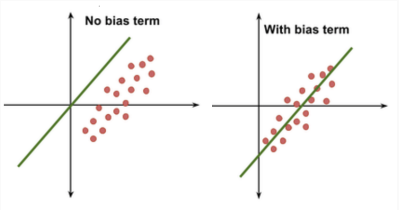
- Generally coarse, based on GDP

- The world has high diversity of languages, cultures, due to internal/external migrations

What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

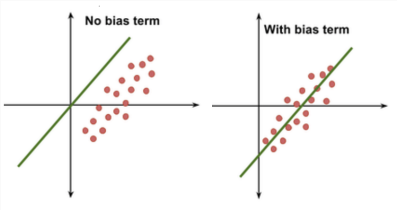
- A bias in a linear model to fit data



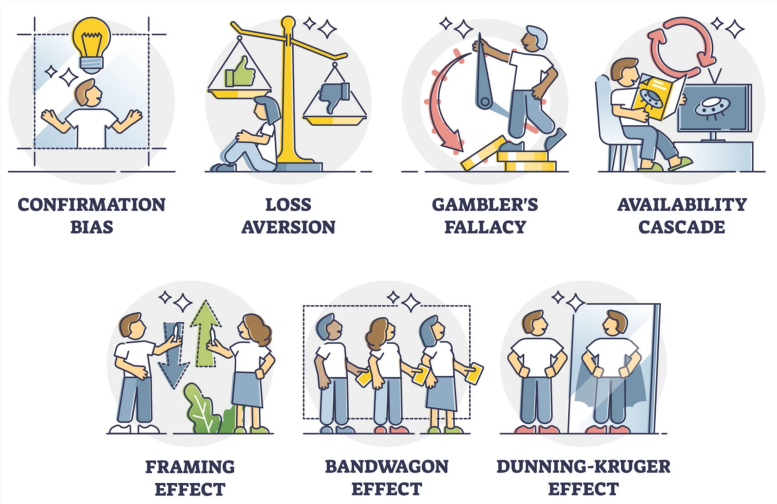
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



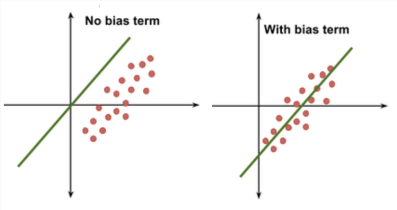
- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...



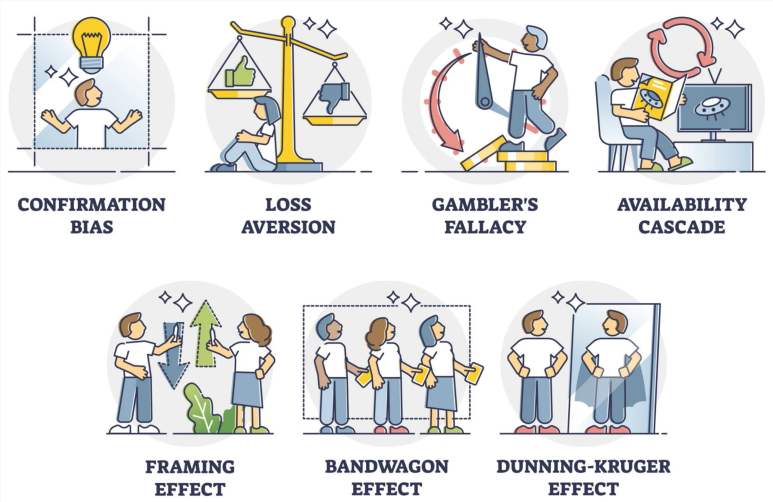
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...

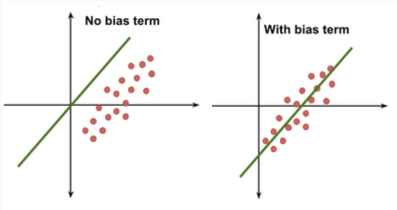


- A social bias like a cultural bias, people have different norms

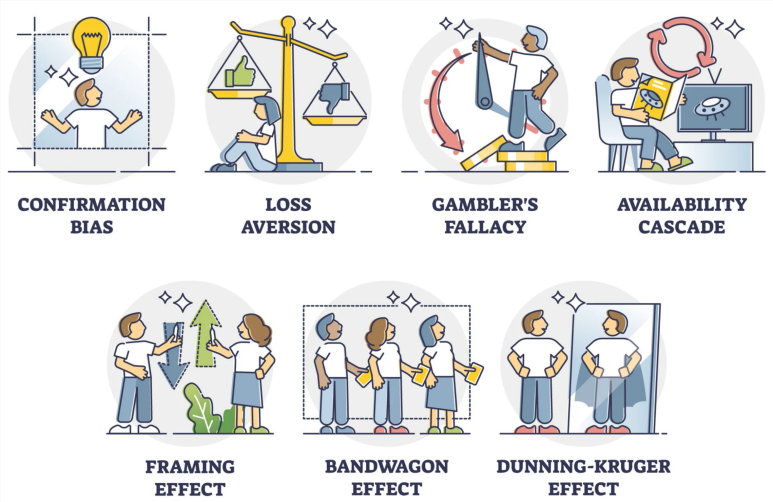
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...



- A social bias like a cultural bias, people have different norms

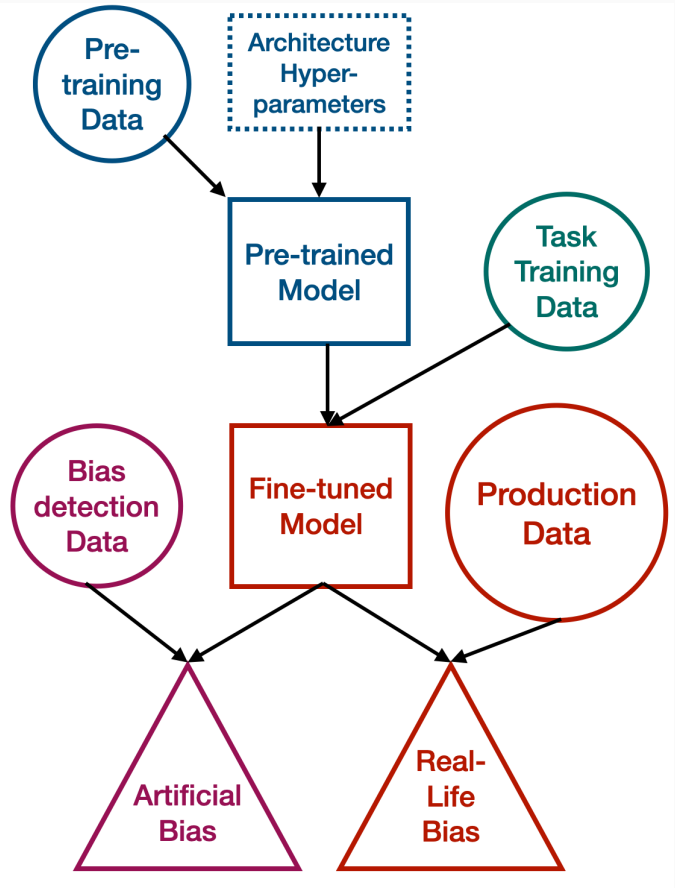
In a decision-making process, a bias can be seen as a change of decision actioned by a non-causal variable.

Confounding variables problem

General Issue

All the bias measurement process is biased itself by different variables such as the bias detection dataset or the fine-tuning dataset. Let's propose a method applied to classifiers using real-world target data.

- **Fine-tuning a model inducts biases** because of the task training data
- Bias detection on pre-trained LM, not on the **final classifier**
- Bias assessment methods relies bias-detection datasets, not **target data distribution**



Method's General idea

Our solution to these problems:

- We propose a method to use the target data, by perturbating any real-life examples
- Our method at the difference of outputs between the perturbated examples, without the need for label
- We use names as a proxy to estimate the bias
- We look at country-related bias and hence be more geographically fine-grained

We found out biases in multilingual models in English and non-English toward several countries, depending on the target language

Related works I

Intrinsic methods

More general but their correlation to downstream tasks is questionable

- Relation between intrinsic metrics and actual deviant behavior is opaque [9, 6]
- Methods based on embeddings lack of transparency and interpretability [19]

Extrinsic methods

More interpretable but

- depends on the choice of variables [1]
- dataset used for evaluation [16]

Even intrinsic methods relying on templates [7, 12, 10]

Related works II

Data

- Considerable variations in bias values and conclusions across template modifications [18]
- Different works propose a multilingual dataset [8, 5]
- A few resources for non-English languages, especially out of a non-Western context [20]

Nationality bias

- [21] shows influence of demographic attributes on country biases
- Names have been shown to contains nationality biases [13]
- [7] dividing the nationalities in 6 groups based on their GDP

[17] proposes `Checklist`, using a perturbation method in order to assess the robustness of a model

Our method: Perturbation of target-distribution examples

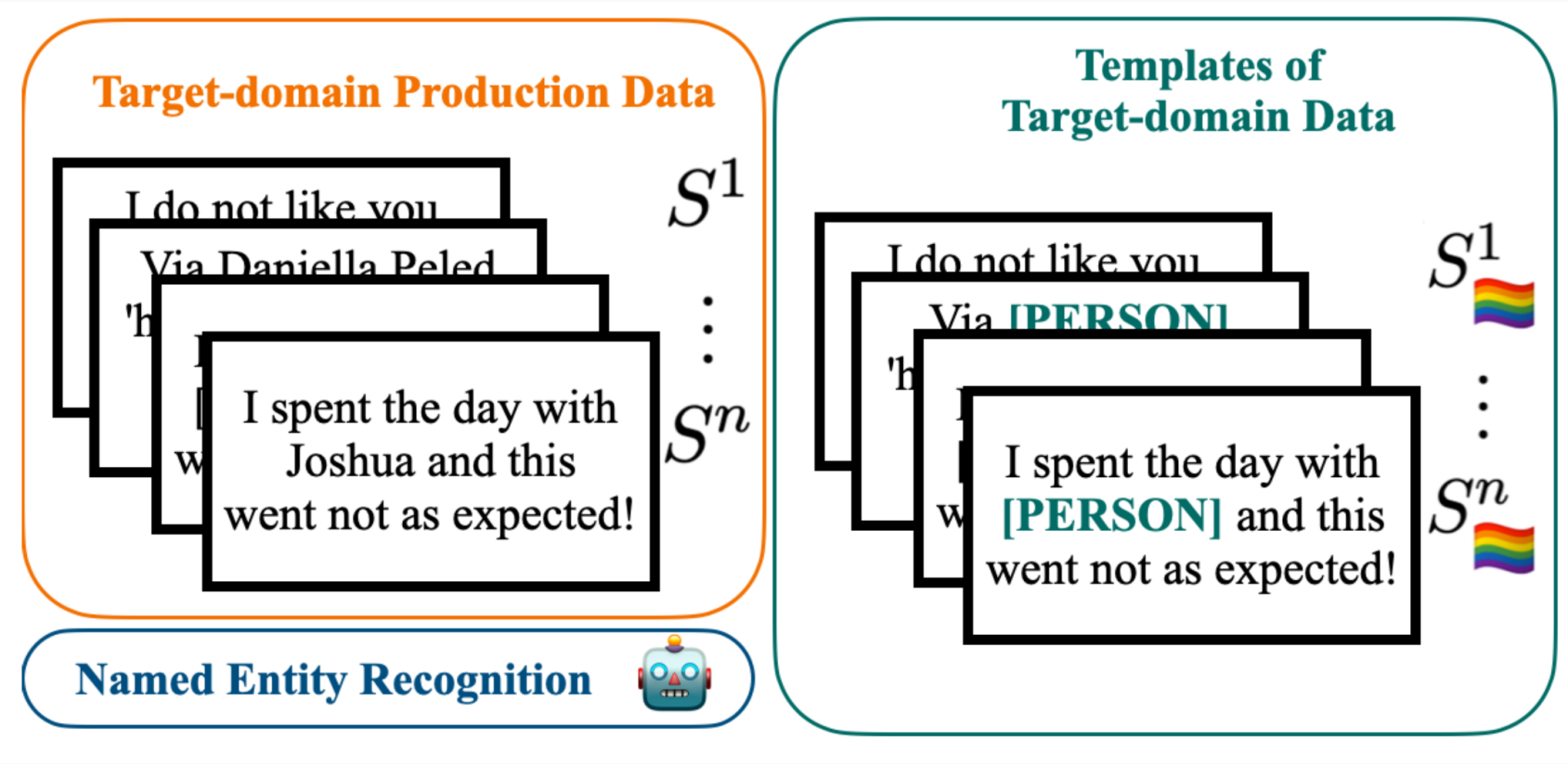


Figure 1: Overview of the counterfactual example creations

Our method: Perturbation of target-distribution examples

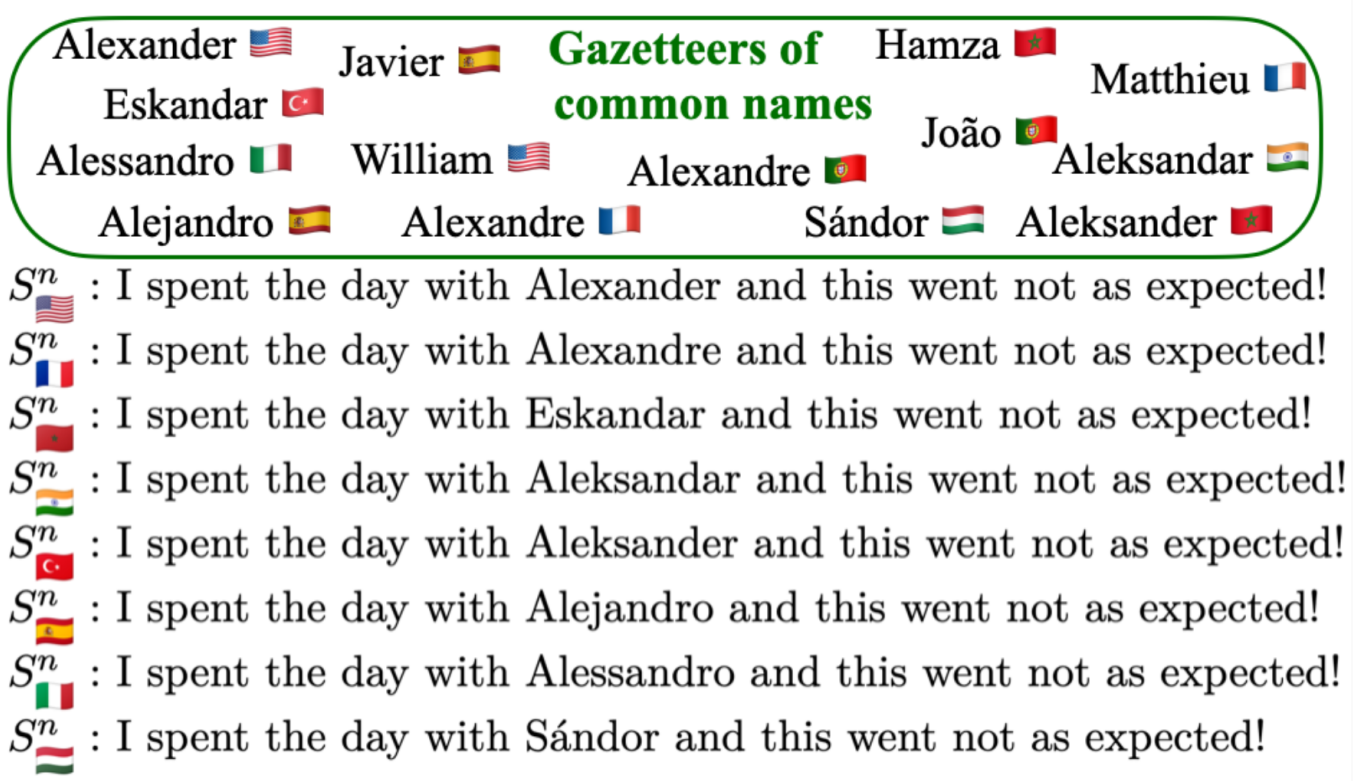


Figure 1: Overview of the counterfactual example creations

Our method: Perturbation of target-distribution examples

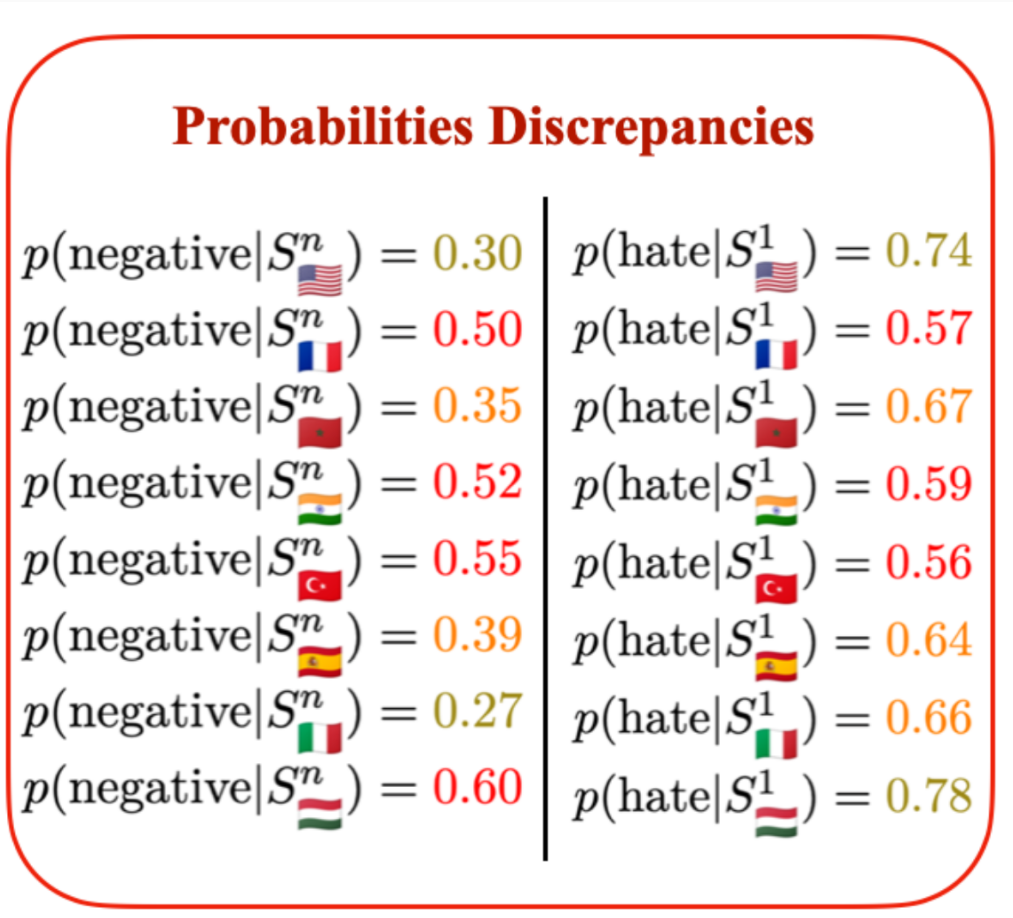


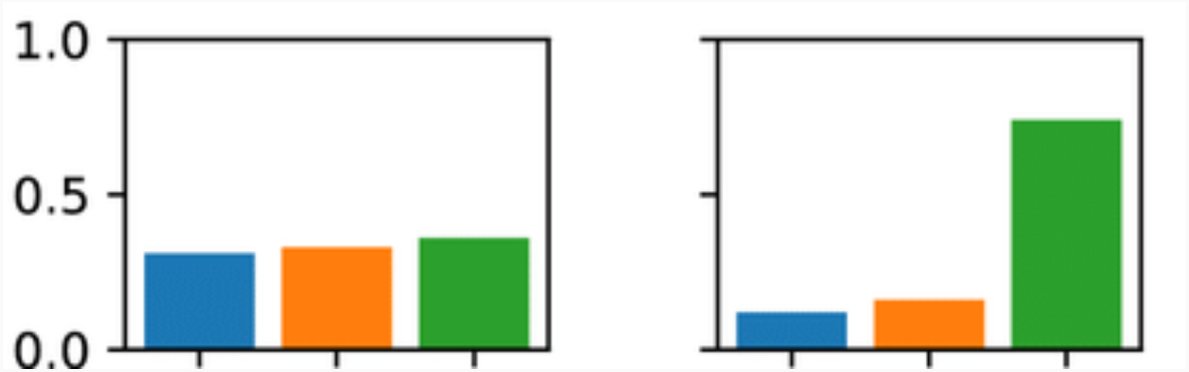
Figure 1: Overview of the counterfactual example creations

Problem: Sentences with names from certain countries will more likely to be classified as negative when it's not, and less likely to be classified as hate speech when it is!

How do we detect a bias?

In a decision-making process, a bias can be seen as a change of decision actioned by a non-causal variable:

- Look at the change in distribution when perturbating the input data with a non-causal change
- A bias is non necessary negative: a change of a Language Model's distribution might reflects the world[1]
- For some models, when the labels have an explicit valence, it is possible to quantify the positiveness of the bias



[1] In their paper "*A Natural Bias for Language Generation Models*" [14], the authors introduce a way to initialize the bias of a LM in order to fasten the learning phase

We used several metrics

A general one

- Distribution distance (Jensen–Shannon divergence, Wasserstein distance, Sinkhorn distance).
- Can be used to say that a bias exists.

A label-oriented one

- Percentage of augmentation/diminution of the predicted examples in each of the classes.
- Can be used to interpret the type of bias regarding the class and target groups.

A valence-oriented one

- $\Delta = \sum_{pos} p_{pos} - \sum_{neg} p_{neg}$.
- Can be used to detect if a bias is harmful or not toward a target group.

Experiments

Experimental Protocol

Models

- Widely used off-the-shelf Twitter multilingual sentiment classifiers based on `XLM-T` [2],[2] which had $>$ 1M monthly download
- Multilingual stance classifier from [4]

Datasets

- Datasets from the `TweetEval` [3] benchmark (AR, EN, ES, DE, FR, IT, PT) and downloaded Tweets frm [15] (PL, HU) and [11] (TK).
- Zero-shot stance recognition dataset `CoFE` from [4]
- Gazeeters of most common names and surnames for each country (from Wikidata, like [17]): $\approx$ 15k names from from 194 countries.

Others

We used the KL divergence, we created 50 random perturbations per sentence, and for stance recognition we used the classes *In Favor* and *Against* as positive and negative.

English Language using Stance Classifier

Gender	Male					Female				
Metric	Δ	Other	Against	In Favor	KL	Δ	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. Δ 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.

English Language using Stance Classifier

Gender	Male					Female				
Metric	Δ	Other	Against	In Favor	KL	Δ	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. Δ 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.

English Language using Stance Classifier

Gender	Male					Female				
Metric	Δ	Other	Against	In Favor	KL	Δ	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. Δ 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.

English Language using Stance Classifier

Gender	Male					Female				
Metric	Δ	Other	Against	In Favor	KL	Δ	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. Δ 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.

English Language using Stance Classifier

Gender	Male					Female				
Metric	Δ	Other	Against	In Favor	KL	Δ	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. Δ 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.

Multilingual Texts

Model tends to prefer the names coming from the sentence's language.
Impulsing for the name **AI Xenophobia**, the fear of the stranger.

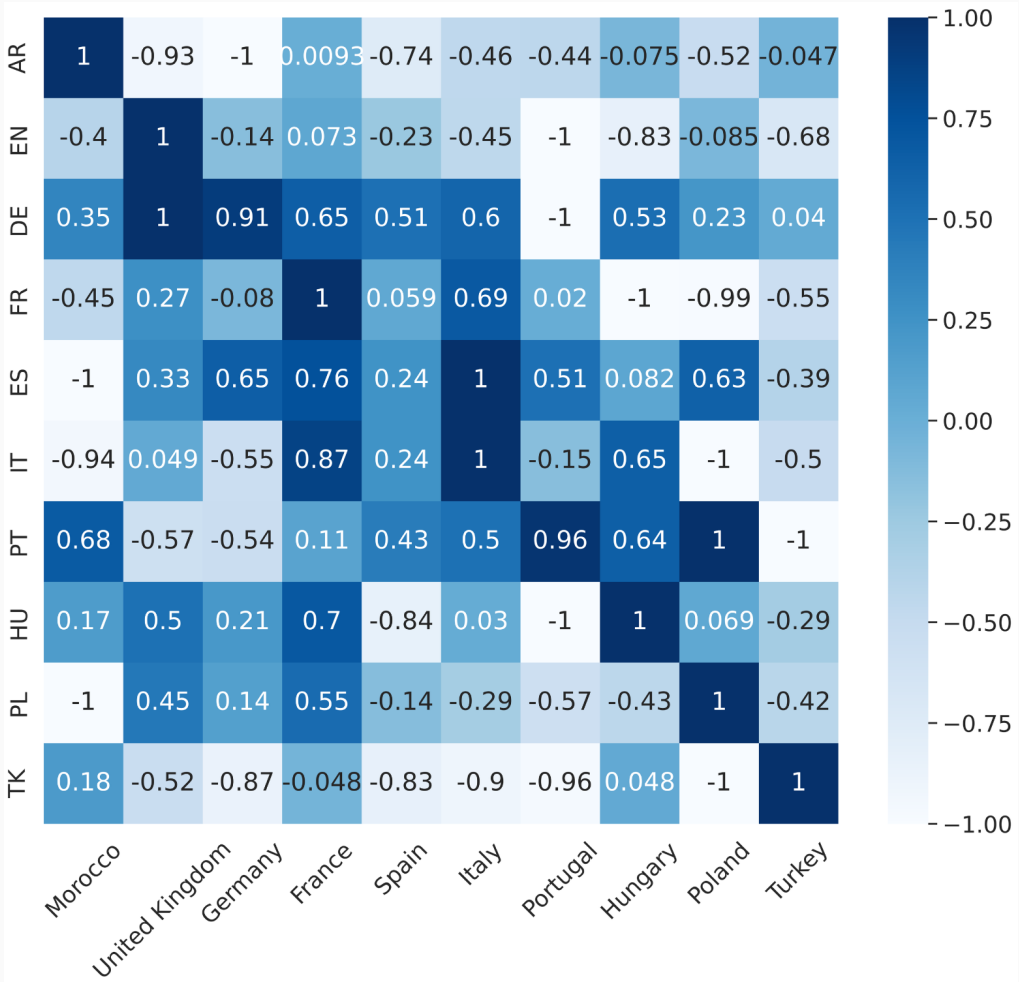


Figure 2: Matrix of Δ normalized per language from multilingual sentiment

Conclusion

- New technique to detect **country-related bias** minimizing confounding variables
- Detection of the bias in **broadly used off-the-shelf affect-related classifiers**
- Xenophobia: Bias change w.r.t. **the language of the sentence**

Thanks for listening!

Questions?

References i


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


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


F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke.
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, 2020.

References ii

V. Barriere and A. Balahur.
Multilingual Multi-target Stance Recognition in Online Public Consultations.
accepted to MDPI Mathematics, 2023.

A. Câmara, N. Taneja, T. Azad, E. Allaway, and R. Zemel.
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, 2022.

References iii

Y. T. Cao, Y. Pruksachatkun, K. W. Chang, R. Gupta, V. Kumar, J. Dhamala, and A. Galstyan.
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, 2022.

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

References iv

S. Goldfarb-tarrant, A. Lopez, R. Blanco, and D. Marcheggiani.
Bias Beyond English : Counterfactual Tests for Bias in Sentiment Analysis in Four Languages.
In *Findings of ACL: ACL 2023*, pages 4458–4468, 2023.

S. Goldfarb-Tarrant, R. Marchant, R. M. Sánchez, M. Pandya, and A. Lopez.
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, 2021.

References

W. Guo and A. Caliskan.
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, 2021.

A. Koksal and A. Ozgur.
Twitter Dataset and Evaluation of Transformers for Turkish Sentiment Analysis.
In *29th Signal Processing and Communications Applications Conference (SIU)*, 2021.

K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov.
Measuring Bias in Contextualized Word Representations.
pages 166–172, 2019.


F. Ladhak, E. Durmus, M. Suzgun, T. Zhang, D. Jurafsky, K. McKeown, and T. Hashimoto.
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, 2023.


C. Meister, W. Stokowiec, T. Pimentel, L. Yu, L. Rimell, and A. Kuncoro.
A Natural Bias for Language Generation Models.
In *ACL*, volume 2, pages 243–255, 2022.


I. Mozetič, M. Grčar, and J. Smailović.
Multilingual twitter sentiment classification: The role of human annotators.
PLoS ONE, 11(5):1–26, 2016.


H. Orgad and Y. Belinkov.
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, 2022.


M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh.
Beyond Accuracy: Behavioral Testing of NLP Models.
ACL, 2020.


P. Seshadri, P. Pezeshkpour, and S. Singh.
Quantifying Social Biases Using Templates is Unreliable.
(Tsrml), 2022.

References viii


F. Valentini, G. Rosati, D. Blasi, D. F. Slezak, and E. Altszyler.
On the interpretation and significance of bias metrics in texts: a PMI-based approach.
In *ACL*, volume 2, pages 509–520, 2023.


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


P. N. Venkit, S. Gautam, R. Panchanadikar, T. H. Huang, and S. Wilson.
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, 2023.