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

Valentin Barriere, Sebastian Cifuentes

Universidad de Chile - CENIA

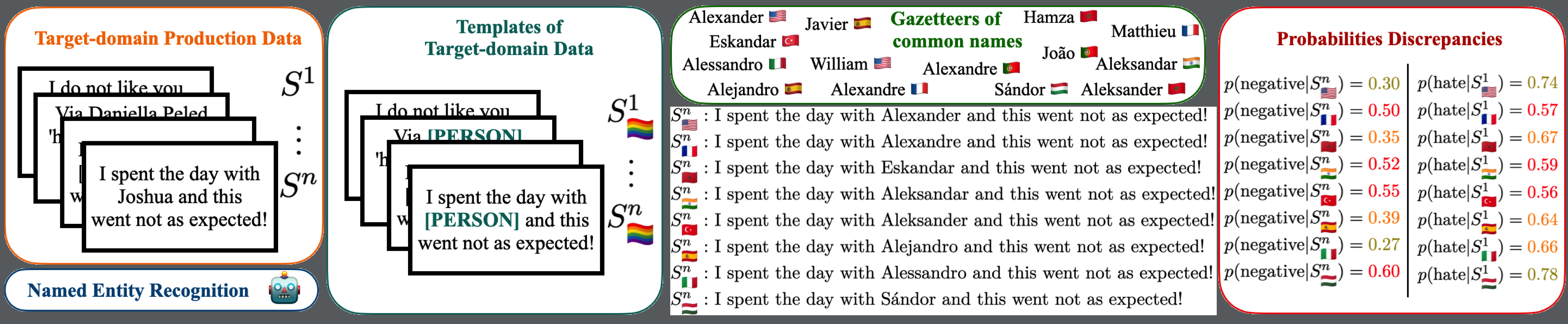
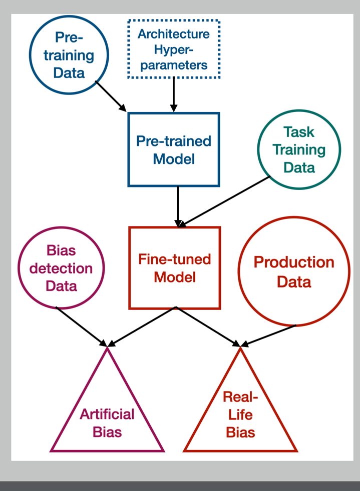


Figure 1: Overview of the counterfactual example creations. We show examples with sentiment and hate speech for variation of the name "*Alexander*" and two sentences.

General issues

Coarse-grain: Classical bias detection methods regarding geography are usually restrained to coarse-grained scales

Confounding variables problem: All the bias measurement process is biased itself by different variables such as the bias detection dataset or the fine-tuning dataset. Our method applies 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

How do we detect a bias?

We look at the change in distribution when perturbating the input data with a non-causal change

- **A general one**: Can be used to say that a bias exists
  - Distribution distance (Jensen–Shannon divergence, Wasserstein distance, Sinkhorn distance).
- **A label-oriented one**: expert knowledge helps understand
  - 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**: when the labels have an explicit valence, it is possible to quantify the bias' harmfulness toward a target group
  - $\boldsymbol{\Delta} = \sum_{pos} p_{pos} - \sum_{neg} p_{neg}$.

Related works

- Intrinsic methods: General but correlation to downstream tasks is questionable: opaque relation between intrinsic non-interpretable metrics and model behavior
- Extrinsic methods: Interpretable but depends on choice of variables/dataset
- Data: A few resources for non-English languages out of a non-Western context, and considerable variations in bias values and conclusions across template modifications
- Nationality bias: studies showed influence of demographic attributes at the country-level, or name-nationality using templates and generative models
- `Checklist` [1] uses a perturbation method in order to assess the robustness of a model

References

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

[2] Valentin Barriere and Alexandra Balahur. Multilingual Multi-target Stance Recognition in Online Public Consultations. *accepted to MDPI Mathematics*, 2023.

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

[4] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis 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.

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

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

In a nutshell

- Bias assessment using the production model on 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 to 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.

Experimental Protocol

- One experiment using Stance Recognition `CoFE` dataset and model [2]
- One experiment using widely used Twitter multilingual sentiment classifier based on `XLM-T` [3] and Tweets data from `TweetEval` + Others [4, 5, 6] (10 languages; AR, EN, ES, DE, FR, IT, PT, PL, HU, TK)
- Gazeteers of most common names and surnames from each country (from Wikidata, like [1]): ≈ 15k names from from 194 countries.
- We created 50 random perturbations per sentence using most common names. For stance recognition we used the classes *In Favor* and *Against* as positive and negative.

English Stance Recognition

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
United States	-0.61	0.0	12.0	-4.0	3.99	-0.46	0.0	8.0	-5.0	3.77
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
Morocco	-1.44	2.0	6.0	-11.0	5.48	-1.41	3.0	2.0	-13.0	5.42
Turkey	-1.58	2.0	5.0	-12.0	5.13	-1.34	2.0	5.0	-12.0	4.78

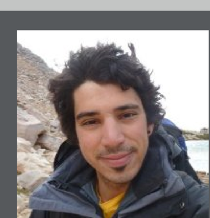
Table 1: **Δ**: difference of probability of the positive class and the negative class. The other values by class and gender are percentages of change in the classification output.

English-speaking country names exhibit highest **Δ** (i.e., more positive outcome). Female names more positive, except for India.

Multilingual Sentiment Classification

- Matrix of **Δ** normalized per language from multilingual sentiment
- Model prefers names from the sentence's language
- Strong implications with the global use of English, or for people with foreign names due to immigration

Contact Information:



https://github.com/valbarriere/Bias_COLING24/

vbarriere@dcc.uchile.cl