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

Valentin Barriere, Sebastian Cifuentes

Universidad de Chile - CENIA

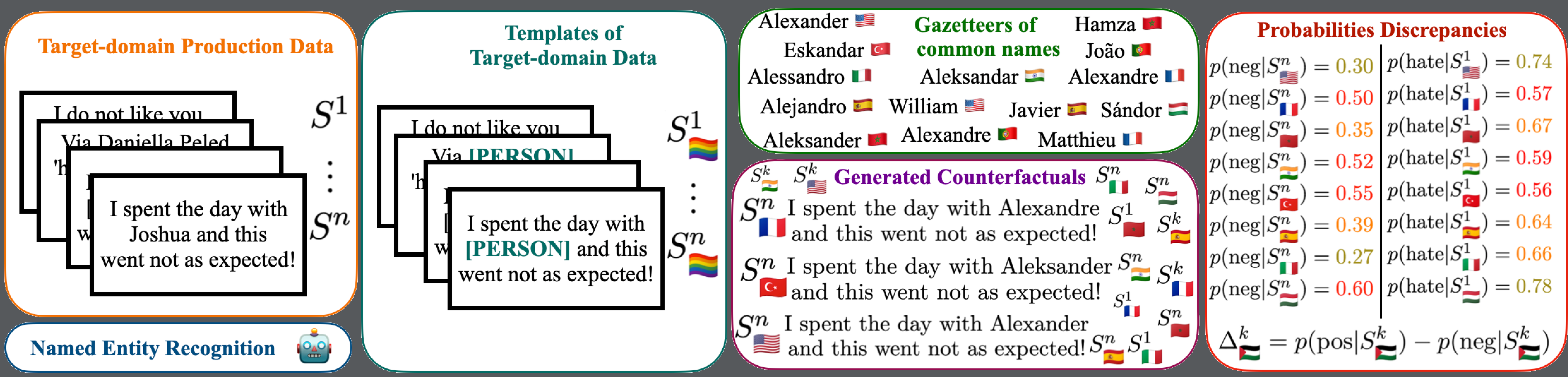
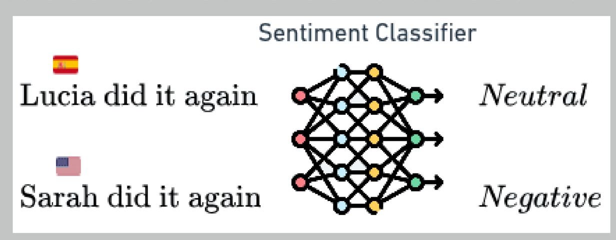


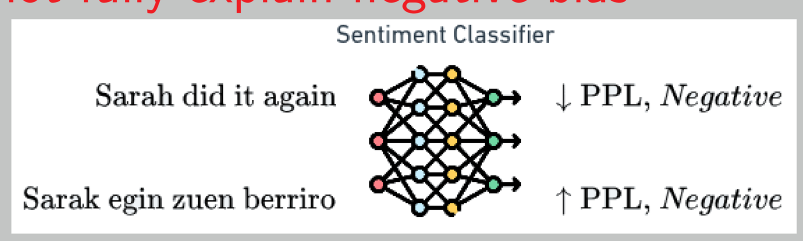
Figure 1: Overview of the counterfactual examples creation. We show examples with sentiment and hate speech for variation of the name "*Alexander*" and two sentences.

In a nutshell

- **Country-related Names:** Using names as a proxy allows detecting country-related bias.
  ⇨ We found negative biases towards several countries in several classifiers



- **Global and Local Perplexity:** Studying the link between OOD words, perplexity, and sentiment predictions.
  ⇨ Perplexity does not fully explain negative bias



Experiments overview

- **Experiment 1: Bias Detection**
  - **Motivation**: Quantify country-name biases of widely used classifiers.
  - **Results**: Significant variations in model predictions based on the presence of different country-names.
- **Experiment 2: Global Perplexity Correlations**
  - **Motivation**: Show the influence of the origin language on the correlation of model predictions and perplexity.
  - **Results**: Model predictions tend to be more negative for unfamiliar languages.
- **Experiment 3: Local Perplexity Correlations**
  - **Motivation**: Show the influence of country-name groups on the correlation of model predictions and perplexity.
  - **Results**: Country-names that are more similar to pre-training data imply a more positive prediction.

Bias quantification

We look at the change in classifiers behavior using different perturbation techniques and quantifying bias using two approaches:

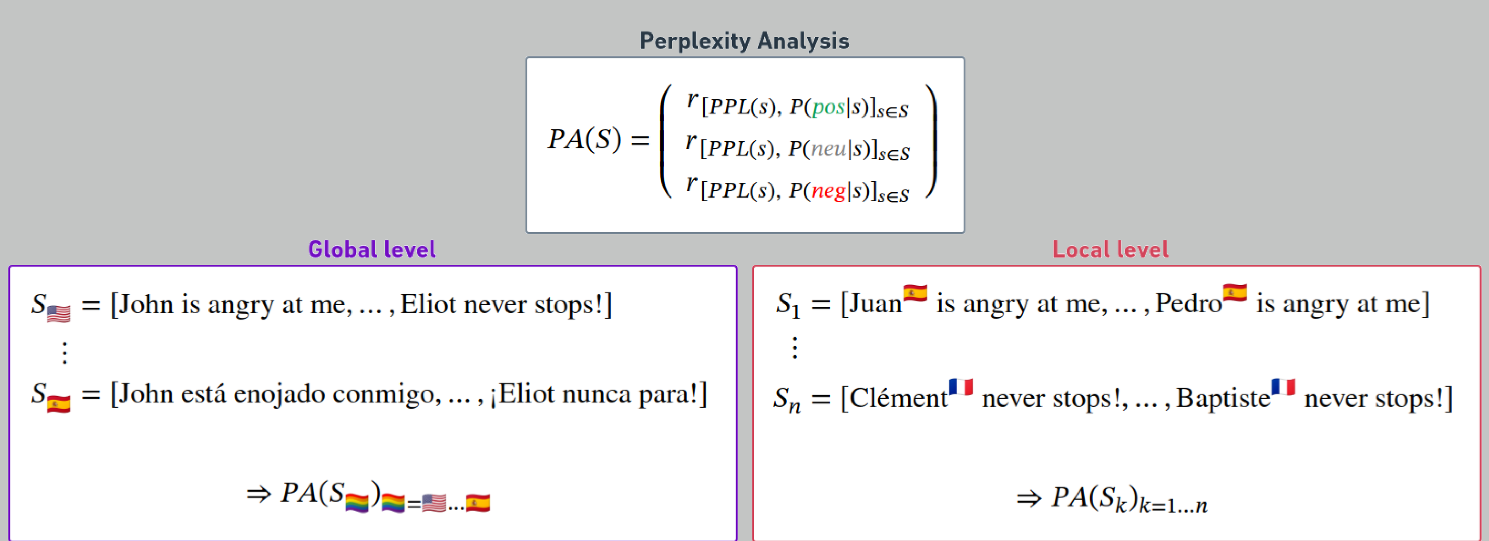
- **Output discrepancy**: Using counterfactual examples (Fig. 1), we analyze changes in output proportions and shifts in output probability ($\boldsymbol{\Delta}$).

$$\boldsymbol{\Delta} = \sum_{pos} p_{pos} - \sum_{neg} p_{neg}$$

- **Perplexity Analysis**: We measure the correlation between perplexity (PPL) and sentiment label probabilities.
  - We use the opposite of pseudo-log-likelihood (PLL) to measure perplexity.

$$PLL(s) = -\sum_{i=1}^{|s|} \log P_{MLM}(w_i | s_{\setminus w_i}; \theta)$$

  - We measure this correlation at two levels: **global** and **local**.



Experimental setup

- A dataset of **8,891 English-language tweets** from Eurotweets Dataset.
- **Gazetteers containing common first and last names from 194 countries**, sourced from Wikidata by the authors of `Checklist`.
- **A multilingual off-the-shelf NER system.**
- **Widely used Affect-related Off-the-shelf Classifiers**: Multilingual sentiment, Monolingual hate speech, emotion recognition and offensive text detection.

Exp. 1: Bias varies between countries

Country	Sentiment				Emotion				Hate	
	$\boldsymbol{\Delta}$	$-$	$\approx$	$+$	Joy	Opt.	Anger	Sad.	Non-hate	Hate
United Kingdom	**-1.43**	5.4	1.3	-4.6	-2.1	0.6	2.7	6.4	-0.2	**23.5**
United States	**-1.35**	5.0	1.7	-4.9	-2.3	-0.5	4.0	6.5	-0.2	**22.0**
Canada	**-1.43**	5.5	1.5	-5.0	-1.6	-0.2	2.3	5.0	-0.2	**21.0**
South Africa	**-1.58**	5.9	1.2	-4.8	-1.5	0.4	1.0	6.1	-0.2	**22.5**
India	**-2.70**	7.9	-0.1	-4.4	-2.5	-6.1	8.7	5.0	-0.1	10.0
Germany	**-2.14**	6.4	1.3	-5.3	-0.0	-4.8	-0.2	4.7	-0.1	19.0
France	**-1.58**	7.7	-0.2	-4.0	0.9	-5.1	-2.5	3.8	-0.1	10.5
Spain	**-2.46**	6.0	2.6	-6.5	1.7	-13.0	-0.4	2.7	-0.0	6.0
Portugal	**-2.30**	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5
Turkey	**-2.33**	6.8	0.7	-4.7	0.2	-11.9	4.8	1.7	-0.1	7.5
Morocco	**-2.04**	4.2	2.4	-5.2	-9.0	**-33.2**	**60.3**	**-17.4**	-0.0	2.0

Table 1: Changes in probability output ($\boldsymbol{\Delta}$) and in percentage of examples per predicted class.

Exp. 2/3: PPL-Prediction patterns changes for OOD languages

Label	English	Dutch	Spanish	Hindi	Turkish	Basque	Maori
$-$	-11.39	-13.87	-6.28	-10.89	-6.02	25.48	35.33
$\approx$	19.27	21.61	19.00	25.54	16.54	-19.98	-36.23
$+$	-5.41	-7.13	-11.10	-13.50	-10.32	-3.04	5.86

Table 2: Global Perplexity-Prediction correlations: switch for unknown languages.

- **Well-known languages**: model tends to classify OOD (high PPL) as neutral.
- **Unknown languages**: it tends to classify OOD as negative.
- **Correlation for Names** is like for unknown languages: the more OOD the more negative.
- But also the less OOD the more positive!

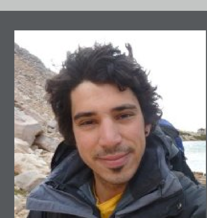
Country	Sentiment		
	$-$	$\approx$	$+$
United Kingdom	**15.03**	5.89	**-18.26**
United States	**14.70**	6.63	**-18.41**
Canada	**15.18**	4.91	**-17.68**
South Africa	**13.12**	5.87	**-16.67**
India	**7.64**	5.18	**-11.75**
Germany	**13.62**	4.50	**-16.34**
France	**8.18**	4.42	**-11.47**
Spain	**11.37**	4.16	**-14.23**
Portugal	**9.45**	2.93	**-11.97**
Turkey	**9.62**	2.79	**-11.86**
Morocco	**9.07**	-0.16	**-8.25**
Overall	11.17	4.63	-14.40

Table 3: Local Perplexity-Prediction correlations.

Conclusion

- Nationality bias in widely used affect-related tweet classifiers.
- Bias is linked to the perplexity of the underlying PLM, suggesting a connection to the data used for pre-training.
- Relation between changes in the model perplexity and it's corresponding classification.

Contact Information:



https://github.com/valbarriere/biases_ppl/

vbarriere@dcc.uchile.cl // sebastian.cifuentes@cenia.cl