Tackling Biases In or Using Generative AI

Valentin Barriere

Universidad de Chile – DCC — CENIA

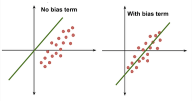
Journées Scientifiques Inria Chile 2024

Biases and Fairness

What can be biases?

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

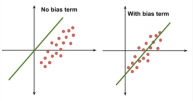
- A bias in a linear model to fit data



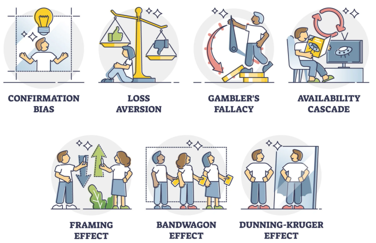
What can be biases?

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

- A bias in a linear model to fit data



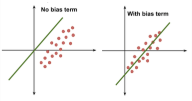
- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...



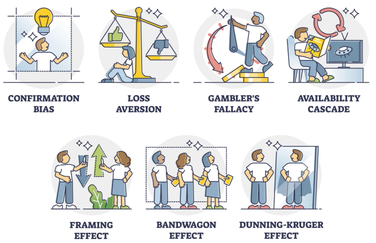
What can be biases?

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, ...

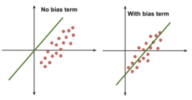


- A social bias like a cultural bias, people have different norms

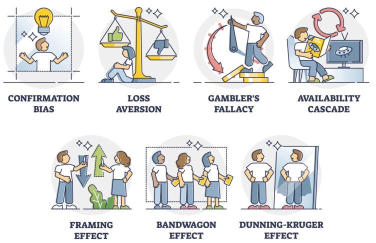
What can be biases?

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.

Cognitive Biases and Fast/Slow Thinking [12]

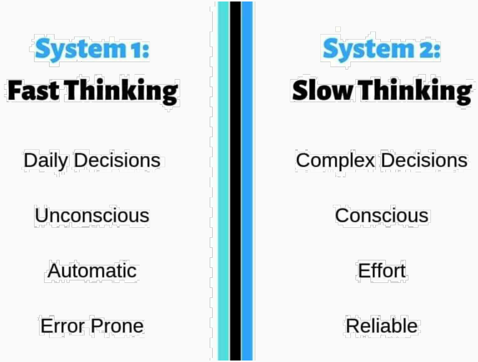


Figure 1: Nobel-winning Daniel Kahneman's book " *Thinking fast and slow*"

- **System 1** is the brain's automatic, fast, and intuitive mode of thinking. It relies on heuristics (mental shortcuts) to make quick judgments and decisions, often based on past experiences or stereotypes
- **System 2** is slower, more deliberative, and analytical. It kicks in when we need to process complex problems, weigh evidence carefully, and revise our beliefs based on reasoning.

Cognitive Biases

Closing this analogy part

- **ML models are trained on biased data can develop biased priors** leading to unfair or skewed prediction
- Similar to how **individuals may develop and act on biased stereotypes**.

Cognitive Biases

Closing this analogy part

- **ML models are trained on biased data can develop biased priors** leading to unfair or skewed prediction
- Similar to how **individuals may develop and act on biased stereotypes**.

In conclusion:

- Data can be biased because of spurious correlations due to hazard or confounding variables,
- The model will take advantage of this bias like a human would do

Fairness

Fairness

Generally, when talking about unfair models, we are looking for negative biases toward certain target groups.

This can happen in different ways:

- **Heterogeneous valence over target groups**: sentiment more negative for arabic names, recidivism prediction higher for black people, lower salary for women or minorities...
- **Heterogeneous performances over target groups**: face recognition system that works badly for Asian users, ASR only works for Castilian or Mexican Spanish, ...
- **Stereotypes**: Co-reference model thinks women is the nurse while the man the doctor
- **Lack of knowledge**: LLM is less knowledgeable when talking about Oriental than Occidental Culture

Structure of the Talk

I. **Tackling Biases in Generative AI**: Biases related to names from different countries in LLMs

II. **Tackling Biases using Generative AI**: Targeted Image Data Augmentation reducing biases related to low-frequency relations between entities

LLM Bias Detection through Names

Motivations

- Fairness in IA

Motivations

- Fairness in IA

- Generally coarse, based on GDP

Motivations

- Fairness in IA

- Generally coarse, based on GDP

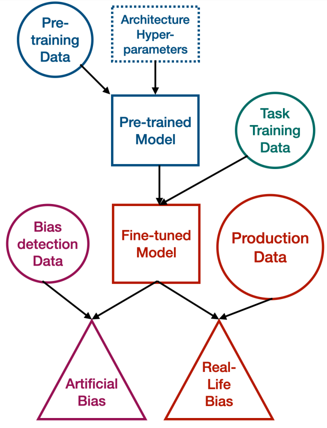
- The world has high diversity of languages, cultures, due to internal/external migrations

Confounding variables problem

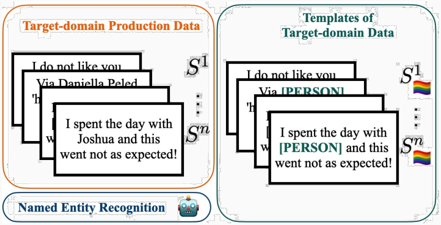
General Issue

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**

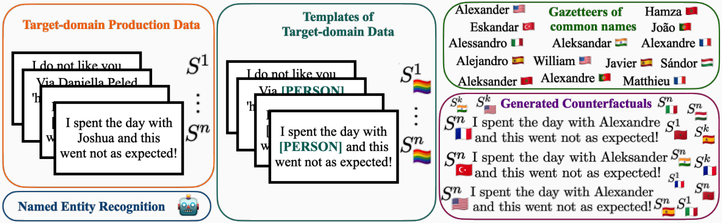


Counterfactual Example Generation & Bias Calculation



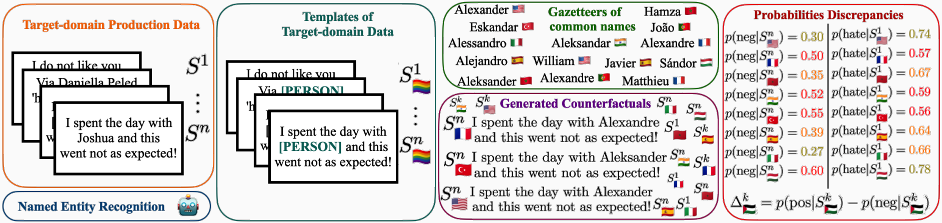
- NER creates **target-domain templates**

Counterfactual Example Generation & Bias Calculation



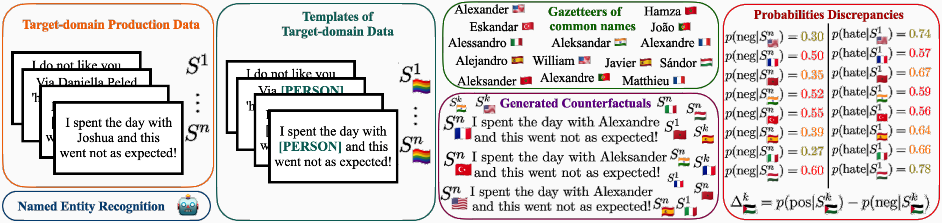
- NER creates **target-domain templates**
- Templates filling using **most common country names**

Counterfactual Example Generation & Bias Calculation



- NER creates **target-domain templates**
- Templates filling using **most common country names**
- **Output discrepancy quantification** between perturbed examples

Counterfactual Example Generation & Bias Calculation



- NER creates **target-domain templates**
- Templates filling using **most common country names**
- **Output discrepancy quantification** between perturbed examples

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

Counterfactual Example Generation & Bias Calculation

Probabilities Discrepancies

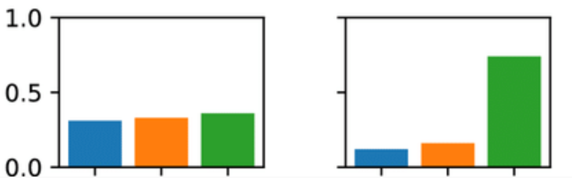
$p(\text{negative} \mid S^n_{\text{🇺🇸}}) = 0.30$	$p(\text{hate} \mid S^1_{\text{🇺🇸}}) = 0.74$
$p(\text{negative} \mid S^n_{\text{🇫🇷}}) = 0.50$	$p(\text{hate} \mid S^1_{\text{🇫🇷}}) = 0.57$
$p(\text{negative} \mid S^n_{\text{🇲🇦}}) = 0.35$	$p(\text{hate} \mid S^1_{\text{🇲🇦}}) = 0.67$
$p(\text{negative} \mid S^n_{\text{🇮🇳}}) = 0.52$	$p(\text{hate} \mid S^1_{\text{🇮🇳}}) = 0.59$
$p(\text{negative} \mid S^n_{\text{🇹🇷}}) = 0.55$	$p(\text{hate} \mid S^1_{\text{🇹🇷}}) = 0.56$
$p(\text{negative} \mid S^n_{\text{🇪🇸}}) = 0.39$	$p(\text{hate} \mid S^1_{\text{🇪🇸}}) = 0.64$
$p(\text{negative} \mid S^n_{\text{🇮🇹}}) = 0.27$	$p(\text{hate} \mid S^1_{\text{🇮🇹}}) = 0.66$
$p(\text{negative} \mid S^n_{\text{🇭🇺}}) = 0.60$	$p(\text{hate} \mid S^1_{\text{🇭🇺}}) = 0.78$

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*" [16], 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

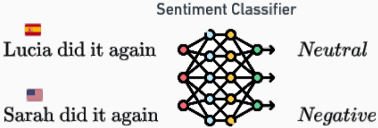
- $\Delta = \sum_{pos} p_{pos} - \sum_{neg} p_{neg}$.
- Can be used to detect if a bias is harmful or not toward a target group.

Our Key findings

- Using names as a proxy allows detecting **country-related bias**
- Bias of a multilingual model depends on the **sentence language**
- Studying the link between OOD words, perplexity, and sentiment predictions

Our Key findings

- Using names as a proxy allows detecting **country-related bias** ⇒ Negative biases towards several countries in several classifiers

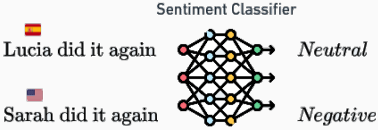


- Bias of a multilingual model depends on the **sentence language**

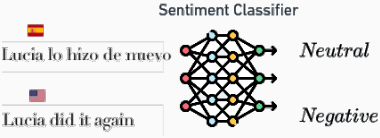
- Studying the link between OOD words, perplexity, and sentiment predictions

Our Key findings

- Using names as a proxy allows detecting **country-related bias** ⇒ Negative biases towards several countries in several classifiers



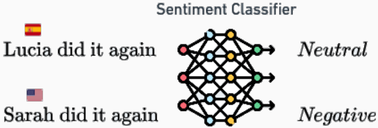
- Bias of a multilingual model depends on the **sentence language** ⇒ Model favor names from the countries speaking the language



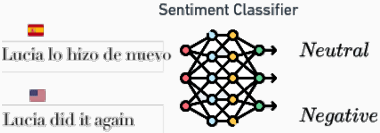
- Studying the link between OOD words, perplexity, and sentiment predictions

Our Key findings

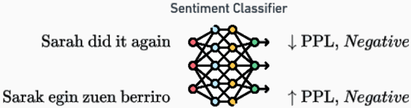
- Using names as a proxy allows detecting **country-related bias** ⇒ Negative biases towards several countries in several classifiers



- Bias of a multilingual model depends on the **sentence language** ⇒ Model favor names from the countries speaking the language



- Studying the link between OOD words, perplexity, and sentiment predictions ⇒ Perplexity does not fully explain negative bias



Related Works

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 [21]

Extrinsic methods

More interpretable but

- depends on the choice of variables [1]
- dataset used for evaluation [18]

Even intrinsic methods relying on templates [7, 14, 10]

Related Works II

Data

- Considerable variations in bias values and conclusions across template modifications [20]
- Different works propose a multilingual dataset [8, 5]
- A few resources for non-English languages, especially out of a non-Western context [22]

Nationality bias

- [23] shows influence of demographic attributes on country biases
- Names have been shown to contains nationality biases [15]
- [7] dividing the nationalities in 6 groups based on their GDP

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

Experiments

Experimental Protocol

Models

- Widely used[2] off-the-shelf Twitter multilingual and English classifiers based on XLM-T [2]: sentiment, emotion, hate speech, ...
- Multilingual stance classifier from [4]

Datasets

- Datasets from the TweetEval [3] benchmark (AR, EN, ES, DE, FR, IT, PT) and/or downloaded Tweets [17, 13] (EN, PL, HU, TK)
- Zero-shot stance recognition dataset CoFE from [4]
- Gazeeters of most common names and surnames for each country (from Wikidata, like [19]): $\approx$ 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.

[2] cardiffnlp/twitter-xlm-roberta-base-sentiment had $>$ 1M monthly download

Experiments Overview

Experiment 1/2: Bias Detection

- **Motivation**: Quantify country-name biases of widely used classifiers.
- **Results**: There are significant variations in model predictions based on the presence of different country-names, showing pattern for negative bias.

Experiment 3: AI Xenophobia

- **Motivation**: Show the influence of the origin language on the bias
- **Results**: Model tends to favor locals' names

Experiment 4/5: Perplexity Correlations

- **Motivation**: Show the influence of country-name groups on the correlation of model predictions and perplexity.
- **Results**:
    - Model predictions tend to be more negative for unfamiliar languages
    - Country-names that are more similar to pre-training data imply a more positive prediction

Experiment 1: English Language using Stance Classifier

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

Experiment 1: English Language using Stance Classifier

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

Experiment 1: English Language using Stance Classifier

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

Experiment 1: English Language using Stance Classifier

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

Experiment 1: English Language using Stance Classifier

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

Experiment 2: English Language using Multilingual Sentiment

Country	Sentiment				Emotion				Hate	
	Δ	−	≈	+	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
Australia	-1.37	5.7	1.2	-4.7	-2.3	0.9	3.2	6.6	-0.2	23.0
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
Italy	-1.98	7.1	1.1	-5.4	2.5	-15.5	-0.9	1.5	-0.1	12.5
Portugal	-2.30	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5
Hungary	-2.26	4.9	2.7	-6.1	2.4	-17.2	-1.4	4.0	-0.1	6.5
Poland	-2.02	3.4	3.6	-6.3	2.0	-13.7	-2.4	5.1	-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 2: Changes in probability output (Δ) and in percentage of examples in each of the predicted classes.

Experiment 2: English Language using Multilingual Sentiment

Country	Sentiment				Emotion				Hate	
	Δ	−	≈	+	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
Australia	**-1.37**	5.7	1.2	-4.7	-2.3	0.9	3.2	6.6	-0.2	23.0
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
Italy	**-1.98**	7.1	1.1	-5.4	2.5	-15.5	-0.9	1.5	-0.1	12.5
Portugal	**-2.30**	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5
Hungary	**-2.26**	4.9	2.7	-6.1	2.4	-17.2	-1.4	4.0	-0.1	6.5
Poland	**-2.02**	3.4	3.6	-6.3	2.0	-13.7	-2.4	5.1	-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 2: Changes in probability output (Δ) and in percentage of examples in each of the predicted classes.

Experiment 2: English Language using Multilingual Sentiment

Country	Sentiment				Emotion				Hate	
	Δ	$-$	$\approx$	$+$	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**
Australia	-1.37	5.7	1.2	-4.7	-2.3	0.9	3.2	6.6	-0.2	**23.0**
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
Italy	-1.98	7.1	1.1	-5.4	2.5	-15.5	-0.9	1.5	-0.1	12.5
Portugal	-2.30	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5
Hungary	-2.26	4.9	2.7	-6.1	2.4	-17.2	-1.4	4.0	-0.1	6.5
Poland	-2.02	3.4	3.6	-6.3	2.0	-13.7	-2.4	5.1	-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 2: Changes in probability output (Δ) and in percentage of examples in each of the predicted classes.

Experiment 2: English Language using Multilingual Sentiment

Country	Sentiment				Emotion				Hate	
	Δ	$-$	$\approx$	$+$	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
Australia	-1.37	5.7	1.2	-4.7	-2.3	0.9	3.2	6.6	-0.2	23.0
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
Italy	-1.98	7.1	1.1	-5.4	2.5	-15.5	-0.9	1.5	-0.1	12.5
Portugal	-2.30	6.9	1.6	-5.9	1.9	-12.9	1.1	-0.4	-0.1	9.5
Hungary	-2.26	4.9	2.7	-6.1	2.4	-17.2	-1.4	4.0	-0.1	6.5
Poland	-2.02	3.4	3.6	-6.3	2.0	-13.7	-2.4	5.1	-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 2: Changes in probability output (Δ) and in percentage of examples in each of the predicted classes.

Experiment 3: 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.

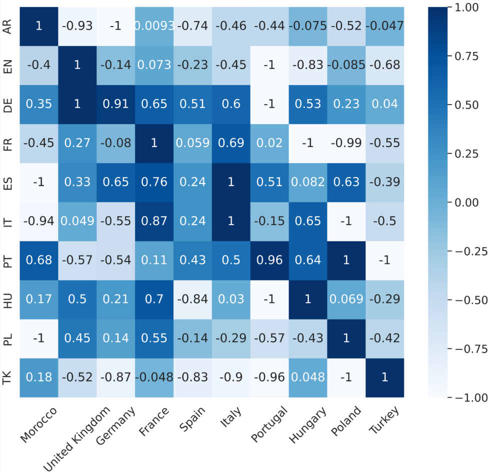


Figure 2: Matrix of Δ normalized per language from multilingual sentiment

Experiment 4/5: Perplexity Analysis

- We conducted a **perplexity analysis** to explore the model's confidence given certain changes

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

Experiment 4/5: Perplexity Analysis

- We conducted a **perplexity analysis** to explore the model's confidence given certain changes

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

Perplexity Analysis

$$PA(S) = \begin{pmatrix} r[PPL(s),\ P(pos|s)]_{s \in S} \\ r[PPL(s),\ P(neu|s)]_{s \in S} \\ r[PPL(s),\ P(neg|s)]_{s \in S} \end{pmatrix}$$

Global level

$S_{🇺🇸}$ = [John is angry at me, … , Eliot never stops!]

⋮

$S_{🇪🇸}$ = [John está enojado conmigo, … , ¡Eliot nunca para!]

$\Rightarrow PA(S_{🏳️‍🌈})_{🏳️‍🌈=🇺🇸…🇪🇸}$

Local level

S_1 = [Juan🇪🇸 is angry at me, … , Pedro🇪🇸 is angry at me]

⋮

S_n = [Clément🇫🇷 never stops!, … , Baptiste🇫🇷 never stops!]

$\Rightarrow PA(S_k)_{k=1…n}$

Exp. 4/5: PPL Prediction patterns changes for OOD

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 3: 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

Exp. 4/5: PPL Prediction patterns changes for OOD

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 3: 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

Exp. 4/5: PPL Prediction patterns changes for OOD

Label	English	Dutch	Spanish	Hindi	Turkish	Basque	Maori
−	-11.39	-13.87	-6.28	-10.89	-6.02	25.48	35.33
≈	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 3: 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

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 4: Local Perplexity-Prediction correlations.

Exp. 4/5: PPL Prediction patterns changes for OOD

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 3: Global Perplexity-Prediction correlations: switch for unknown languages.

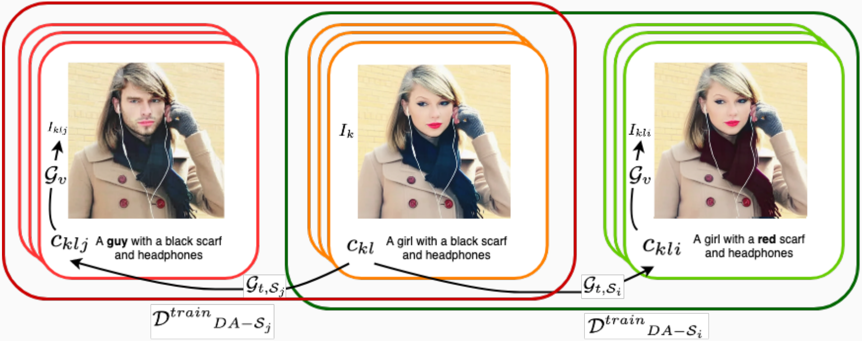
- **Well-known languages**: model tends to classify OOD (high PPL) as neutral
- **Unknown languages**: it tends to classify OOD as negative
- **Correlation for Names** is like for unknown languages: the more OOD the more negative
- But also the less OOD the more positive!

Country	Sentiment $-$	Sentiment $\approx$	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 4: Local Perplexity-Prediction correlations.

LMM Bias Removal using Image Generators

Targeted Image Data Augmentation



- Some situations are less seen in the data: a red tree, a football match with several balloons, or a woman snowboarding [11]
- How to augment data so the model can adapt to a new situation?
- In Image Captioning: generating new content with a text2image via perturbations on the caption data

Results

Test / Train	#DA	BLEU@1-4				RefCLIPScore			
		$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$
$\mathcal{D}^{train}$ (Vanilla)	0	51.8	44.0	49.9	49.7	79.9	79.3	79.8	80.3
$\mathcal{D}^{train}{}_{SD-rnd}$	60k	51.3	44.1	49.2	49.6	80.0	79.5	79.7	80.2
$\mathcal{D}^{train}{}_{SD-clr}$	20k	*51.7*	44.0	*49.3*	49.5	79.8	79.4	79.6	80.1
$\mathcal{D}^{train}{}_{SD-ctg}$	20k	*51.7*	*44.4*	49.2	49.7	79.9	*79.5*	79.7	80.2
$\mathcal{D}^{train}{}_{SD-gdr}$	20k	51.2	43.4	48.5	48.8	*80.0*	79.2	*79.9*	80.3
$\mathcal{D}^{train}{}_{SD-all}$	60k	51.8	44.9	**50.1**	**50.5**	80.1	**79.7**	80.1	**80.5**

What do we have using BLIP2 for Image Captioning?

- We focused on 3 basic human skills: **Gender, Color and Counting**

Results

Test / Train	#DA	BLEU@1-4				RefCLIPScore			
		$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$	$\mathcal{D}^{test}{}_{clr}$	$\mathcal{D}^{test}{}_{ctg}$	$\mathcal{D}^{test}{}_{gdr}$	$\mathcal{D}^{test}$
$\mathcal{D}^{train}$ (Vanilla)	0	51.8	44.0	49.9	49.7	79.9	79.3	79.8	80.3
$\mathcal{D}^{train}{}_{SD-rnd}$	60k	51.3	44.1	49.2	49.6	80.0	79.5	79.7	80.2
$\mathcal{D}^{train}{}_{SD-clr}$	20k	*51.7*	44.0	*49.3*	49.5	79.8	79.4	79.6	80.1
$\mathcal{D}^{train}{}_{SD-ctg}$	20k	*51.7*	*44.4*	49.2	49.7	79.9	*79.5*	79.7	80.2
$\mathcal{D}^{train}{}_{SD-gdr}$	20k	51.2	43.4	48.5	48.8	*80.0*	79.2	*79.9*	80.3
$\mathcal{D}^{train}{}_{SD-all}$	60k	51.8	44.9	**50.1**	**50.5**	80.1	**79.7**	80.1	**80.5**

What do we have using BLIP2 for Image Captioning?

- We focused on 3 basic human skills: **Gender, Color and Counting**
- TIDA helps the model to **get better on specific subsets related to these skills**, and on the general test set
- The model use skill-associated words **more often when the caption should contain one and less when it should not**
- It works better than random (non-targeted DA)

Conclusion

Conclusion

Tackling bias in Generative IA:

- 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**
- **Bias is linked to the perplexity of the underlying PLM**, suggesting a connection to the data used for pre-training
- However, this relation is not that simple!

Tackling bias with Generative IA:

- Biases can be due to low correlation relations
- Generative model are useful to create data to reduce them

References

Valentin Barriere, Felipe del Rio, Andres Carvallo, Carlos Aspillaga, Eugenio Herrera and Cristian Buc, Targeted Image Data Augmentation Increases Basic Skills Captioning Robustness, **GEM Workshop @ EMNLP 2023**

Valentin Barriere and Sebastian Cifuentes, Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method With Least Confounding Variables, **COLING 2024**

Valentin Barriere and Sebastian Cifuentes, A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifier, **EMNLP 2024**

Valentin Barriere, Fantastic Biases (What are They) and Where to Find Them, **Bites de Ciencias** *26 (2024), 02-13*

Thank you for your attention!

Contact:

vbarriere@dcc.uchile.cl

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.

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

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 v

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.

L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach.
Women Also Snowboard: Overcoming Bias in Captioning Models.
Proceedings of the European Conference on Computer Vision (ECCV),, pages 771–787, 2018.

D. Kahneman.
Thinking, Fast and Slow.
2011.

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.

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.