On the Sources of Bias in NLP Models

Origins, Impact, Challenges, and the Ways Forward.

Fatma Elsafoury

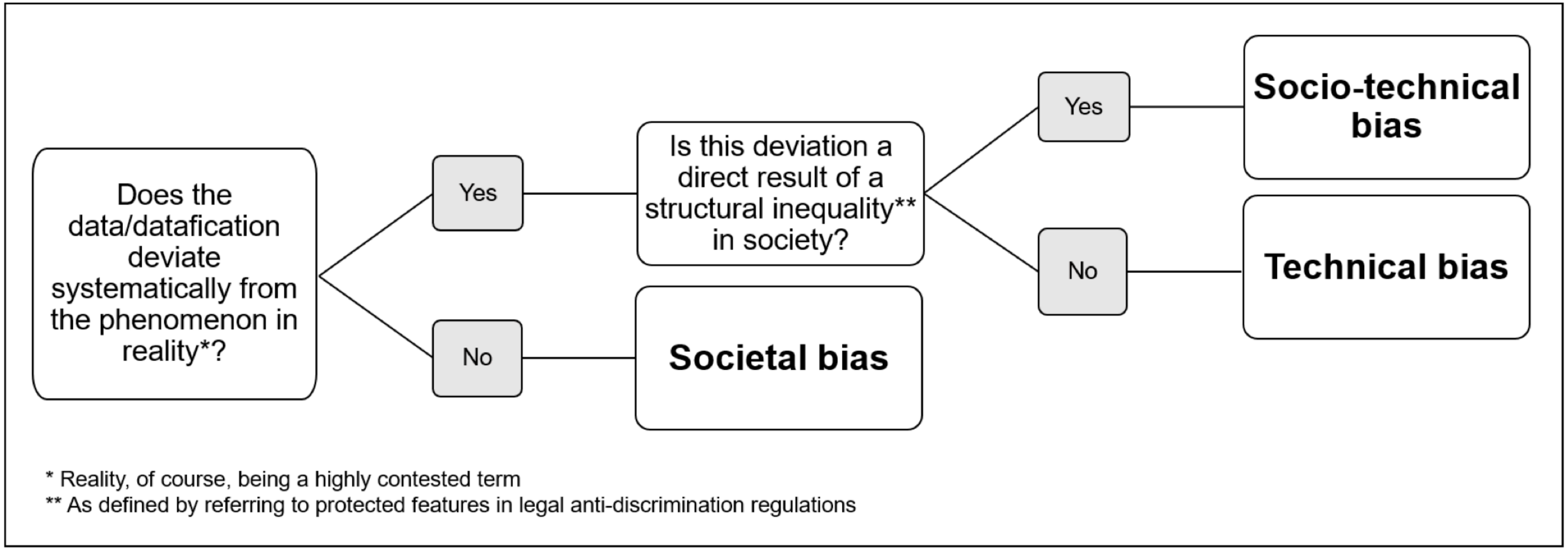
The impact of **Bias** on the **Fairness** of **Toxicity detection.**

Fatma Elsafoury, and Stamos Katsigiannis. ”On Bias and Fairness in NLP: Investigating the Impact of Bias and Debiasing in Language Models on the Fairness of Toxicity Detection”. A long paper **under-submission at the Computational Linguistics journal.**

What is **Bias?**

Bias Definition

Based on Legal anti-discrimination regulations, Paola Lopez distinguishes between 3 types of bias[1]:

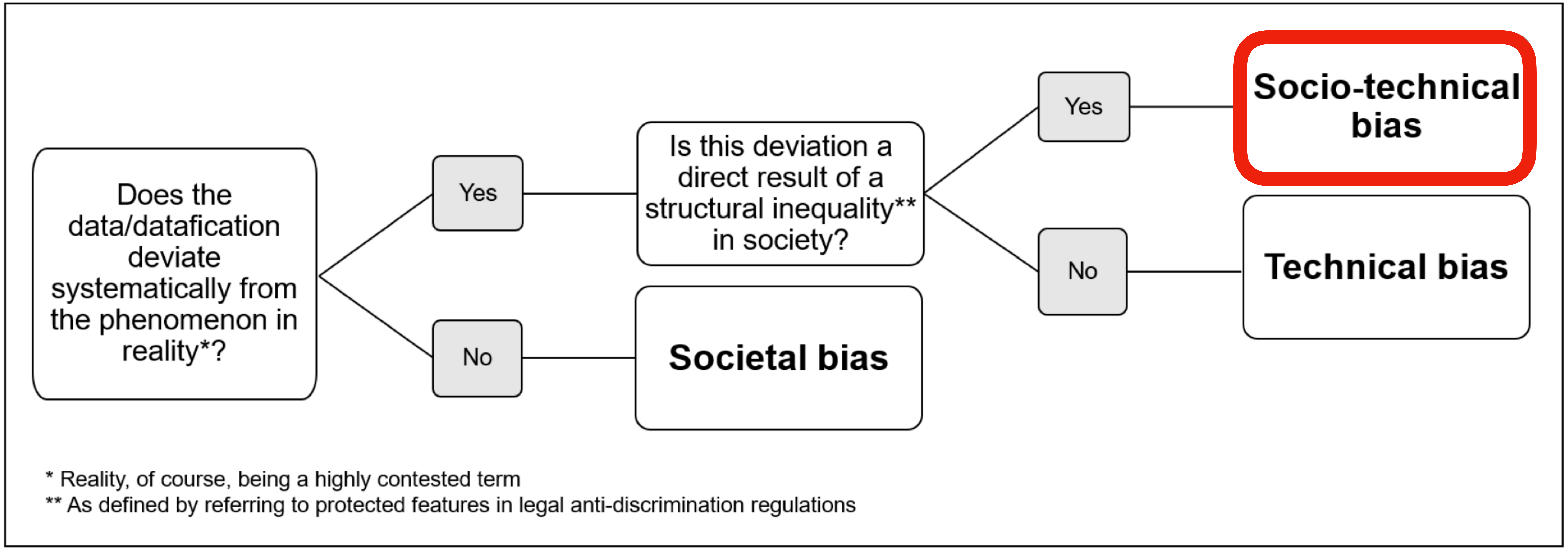


Bias Scheme [1]

[1] Lopez, Paola. 2021. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29.

Bias Definition

Based on Legal anti-discrimination regulations, Paola Lopez distinguishes between 3 types of bias[1]:



Bias Scheme [1]

[1] Lopez, Paola. 2021. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29.

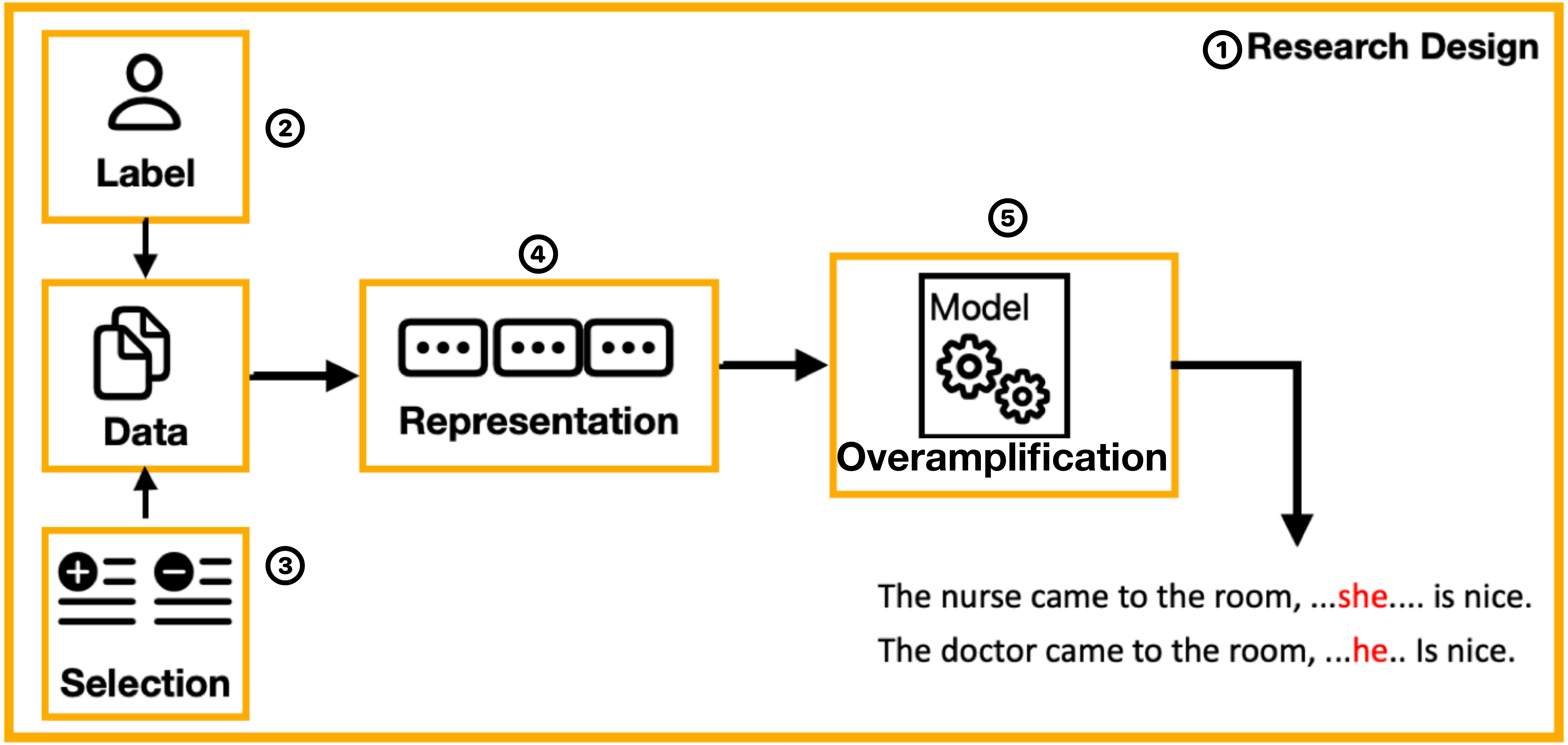
Bias Definition

Socio-technical Bias	Statistical definition of Bias
"A systematic deviation due to structural inequalities"[1]	*"A systematic distortion in the sampled data that compromises its representatives"[2]*

How these definition are related?
Is data the only form of inequalities in the NLP process?

[1] Lopez, Paola. 2021. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29.
[2] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Sources of Bias in NLP

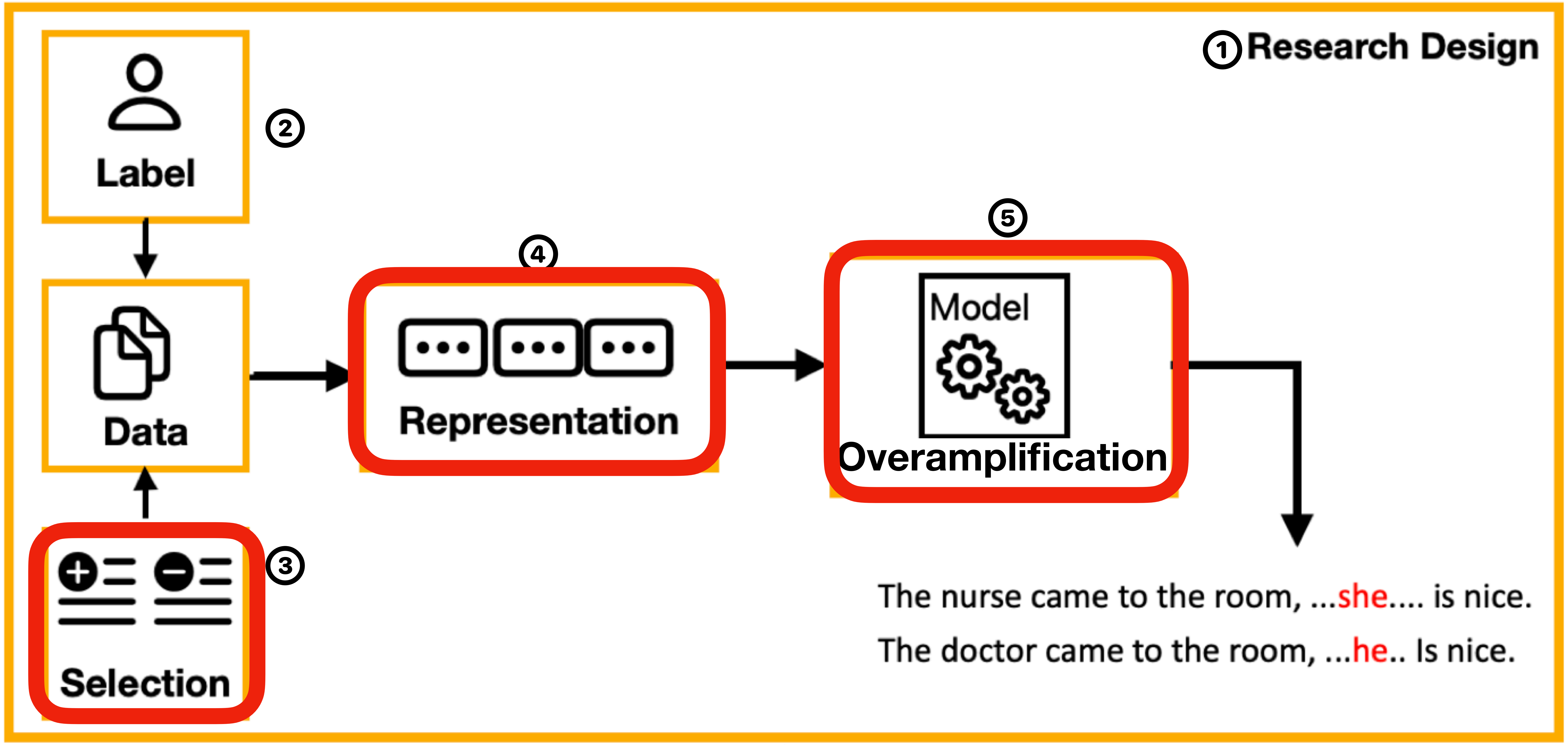


Conceptual framework of five sources bias in NLP models [1,2]

[1] Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

[2] Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

Sources of Bias in NLP



Conceptual framework of five sources bias in NLP models [1,2]

[1] Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

[2] Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

What is Fairness?

Fairness Definition

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A clarification of the nuances in the fairness metrics landscape

Alessandro Castelnovo[1,2,3], Riccardo Crupi[1,3], Greta Greco[1,2,3], Daniele Regoli[1,3]✉, Ilaria Giuseppina Penco[1] & Andrea Claudio Cosentini[1]

In recent years, the problem of addressing fairness in machine learning (ML) and automatic decision making has attracted a lot of attention in the scientific communities dealing with artificial intelligence. A plethora of different definitions of fairness in ML have been proposed, that consider different notions of what is a "fair decision" in situations impacting individuals in the population. The precise differences, implications and "orthogonality" between these notions have not yet been fully analyzed in the literature. In this work, we try to make some order out of this zoo of definitions.

Fairness Definition

Group Fairness Metrics

"*Compare the outcome of the classification algorithm for two or more groups*"[1].

$$FPR\_gap_{g,\hat{g}} = |FPR_g - FPR_{\hat{g}}|$$

$$TPR\_gap_{g,\hat{g}} = |TPR_g - TPR_{\hat{g}}|$$

$$AUC\_gap_{g,\hat{g}} = |AUC_g - AUC_{\hat{g}}|$$

Where g and gˆ, are different groups of people based on sensitive attributes like gender, race, etc.

[1] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. ACM Comput. Surv. 56, 7, Article 166 (July 2024), 38 pages. https://doi.org/10.1145/3616865

[2] Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In WWW '19: Companion Proceedings of The 2019 World Wide Web Conference, pages 491–500.

What is Toxicity detection?

Bias Definition

A toxic comment is

"rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion"[1]

Subjective definition which is hard to quantify and to label.

[1]Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. pages 67–73, 12 2018. .

Toxicity detection
Dataset

- Jigsaw Unintended bias dataset[1]
    - Civil Comments Platform.
    - ~ 2 Million comments.
    - Toxicity and Identity labels.
- Models: Bert-base-uncased, RoBERTa-base, AlBERT-base.

Sensitive attribute	Marginalized	Non-marginalized
Gender	Female	Male
Race	Black, Asian	White
Religion	Jewish, Muslim	Christian

Table 1: The examined sensitive attributes and identity groups.

	AUC scores		
Dataset	**BERT**	**RoBERTa**	**AlBERT**
Jigsaw-unintended	0.902	0.908	**0.911**

Table 2: Performance of different Models

[1] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 491–500. https://doi.org/10.1145/3308560.3317593

Fairness of Toxicity detection

Fairness dataset

- **<u>Original fairness dataset:</u>** Subset of the the test set .
    - Imbalance between the different identity groups:
        - size and ratio of toxic sentences.

This poses a challenge on the measured fairness score.

Fairness of Toxicity detection

Fairness dataset

- **<u>Original fairness dataset:</u>** Subset of the the test set .
    - Imbalance between the different identity groups:
        - size and ratio of toxic sentences.

This poses a challenge on the measured fairness score.

We create data perturbations to balance the dataset (toxic and non-toxic) comments.

We use lexical word replacement to create the perturbations with race and religion.

For gender with the different pronouns, we use the AugL tool to swap gender information[1].

[1] Papakipos, Zoe and Joanna Bitton. 2022. Augly: Data augmentations for robustness.

Fairness of Toxicity detection

Fairness dataset

For example

Muslims are terrorists	Christians are terrorists	Jews are terrorists
Black people are violent	White people are violent	Asian people are violent
Women belong to the kitchen	Men belong to the kitchen	

Fairness of Toxicity detection

Fairness dataset

For example

Muslims are terrorists	Christians are terrorists	Jews are terrorists
Black people are violent	White people are violent	Asian people are violent
Women belong to the kitchen	Men belong to the kitchen	

What about *Asymmetric Counterfactuals*?

Fairness of Toxicity detection

Fairness dataset

Asymmetric Counterfactuals[1]:

Happens when the created counterfactual makes the toxicity label invalid.

For example:

N****ers came to me (Toxic)

Whites came to me (~~Toxic~~)

[1] Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 219–226, ACM.

Fairness of Toxicity detection

Fairness dataset

Two assumptions of Asymmetric Counterfactuals[1]:

1. **Identity attacks:** When toxicity targets a marginalised group, it is based on identity only with **no other toxicity signals**.

2. **Stereotyping comments**: are more likely to occur in a **toxic comment** attacking marginalised groups.

[1] Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 219–226, ACM.

Fairness of Toxicity detection

Fairness dataset

- No offensive identity terms in toxic comments.
- Stereotyping expressions found in toxic and non-toxic comments.
    - e.g., *"Police"* which stereotype Black people used in **toxic** and **non-toxic.**
    - *"supremacist"* which stereotypes White people used in **toxic** and **non-toxic.**

The Asymmetric counterfactual is not a problem with the Jigsaw dataset.

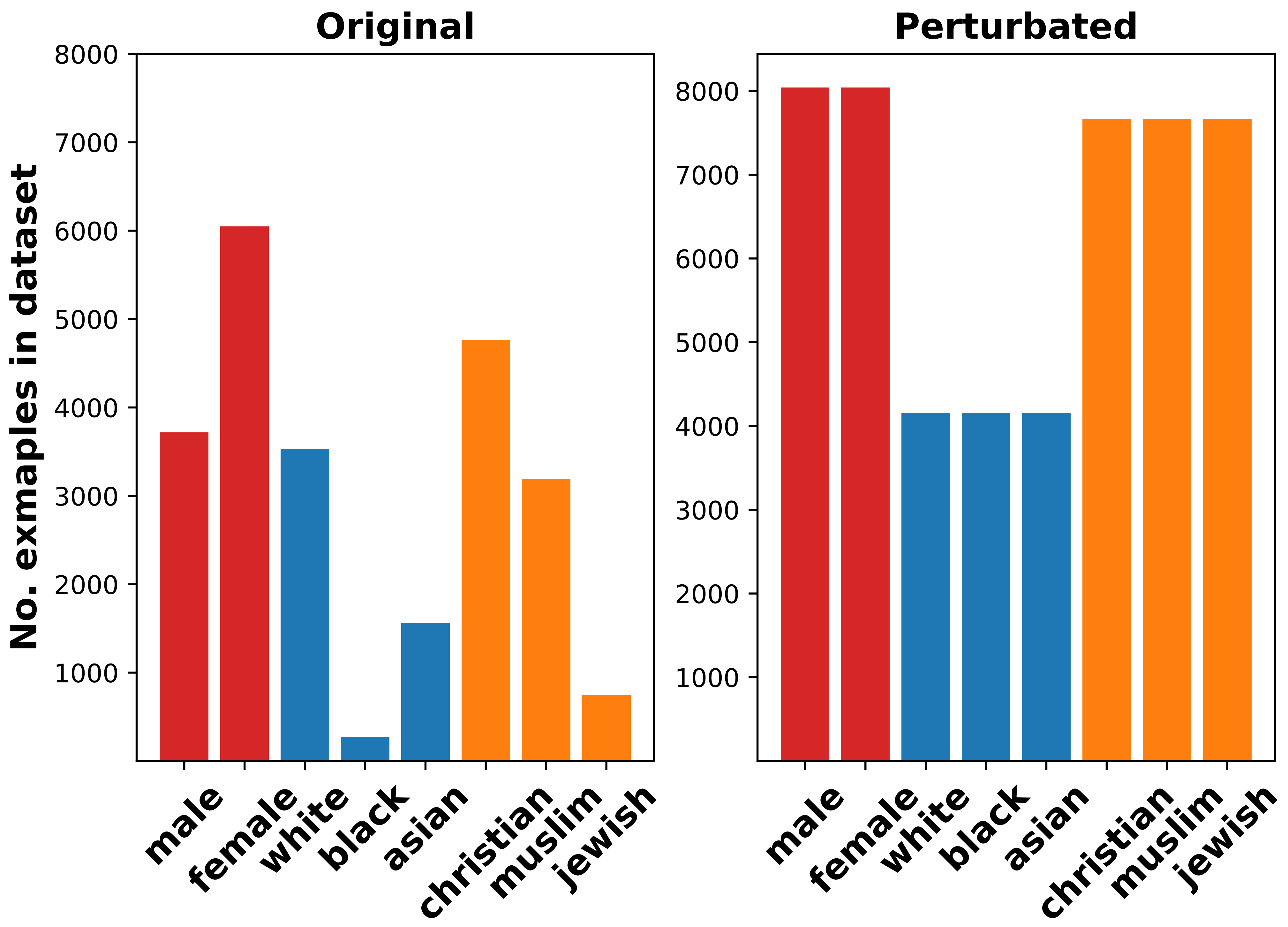
Identity	Toxic sentences	Non-toxic sentences
Black	black, people, blacks, racist, police,Black, other, man, white, men	black, people, blacks, man, police,other, Black, white, many, men
Asian	people, Asian, many, repair, chef, country, racist, real, citizens, Korean	Asian, other, Chinese, people, many, countries, years, women, more, country
White	white, people, racist, men, supremacists, man, racism, right, supremacist, White	white, people, men, racist, right, other, man, many, supremacists, male
Female	women, woman, people, white, other, many, sexual, time, life, sex	women, woman, people, many, other, more, time, right, life, abortion
Male	man, men, white, black, people, male, women, stupid, racist, males	man, men, white, people, male, other, many, right, time, way
Muslim	Muslim, people, women, white, many, other, muslim, terrorists, religion, muslims	Muslim, people, countries, women, other, many, country, ban, world, muslim
Jewish	Jewish, people, anti, black, hate, women, good, other, white, man	Jewish, people, anti, other, white, right, way, state, many, world
Christian	people, white, Christian, women, right, other, many, sex, Catholic,life	Catholic, people, Christian, church, many, women, other, right, time, good

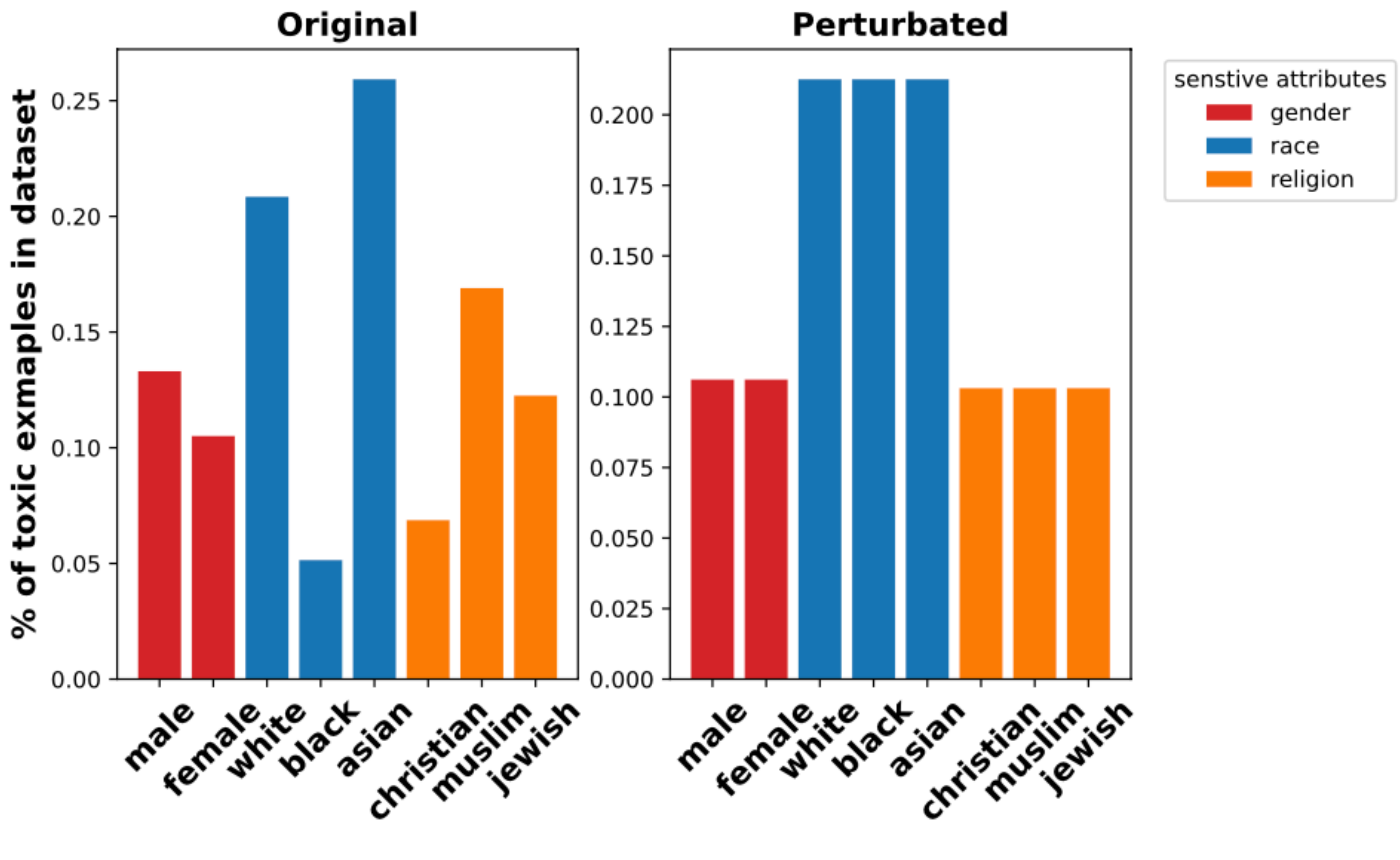
The most common nouns and adjectives in the Jigsaw dataset

Fairness of Toxicity detection

Fairness dataset

After perturbaiton, we have balanced fairness dataset.





Fairness of Toxicity detection

Balanced vs. Original Fairness dataset

- Different fairness metrics give different results.

- With the balanced fairness dataset, we get more reliable fairness results.

Attribute	Model	Dataset	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	Original	0.001	0.081	0.025
		Balanced	↑ 0.006	↓ 0.038	↓ 0.003
	BERT	Original	0.002	0.111	0.026
		Balanced	↑ 0.008	↓ 0.036	↓ 0.009
	RoBERTa	Original	0.007	0.084	0.017
		Balanced	↓ 0.004	↓ 0.031	↓ 0.011
Race	ALBERT	Original	0.007	0.044	0.003
		Balanced	↑ 0.008	↓ 0.0016	↑ 0.018
	BERT	Original	0.008	0.017	0.048
		Balanced	↑ 0.015	↓ 0.002	↓ 0.025
	RoBERTa	Original	0.014	0.127	0.028
		Balanced	↓ 0.003	↓ 0.011	↓ 0.021
Religion	ALBERT	Original	0.019	0.060	0.042
		Balanced	↓ 0.009	↑ 0.108	↓ 0.020
	BERT	Original	0.016	0.027	0.051
		Balanced	↓ 0.008	↑ 0.062	↓ 0.012
	RoBERTa	Original	0.027	0.030	0.0369
		Balanced	↓ 0.021	↑ 0.160	↓ 0.027

Table 3: The fairness scores of the examined models on the original and the balanced c
community fairness datasets. (↑) denotes that the fairness score increased, and the fairn
worsened. (↓) denotes that the fairness score decreased, and the fairness improved.

What is the impact of different sources of bias on the Fairness of toxicity detection?

Representation bias

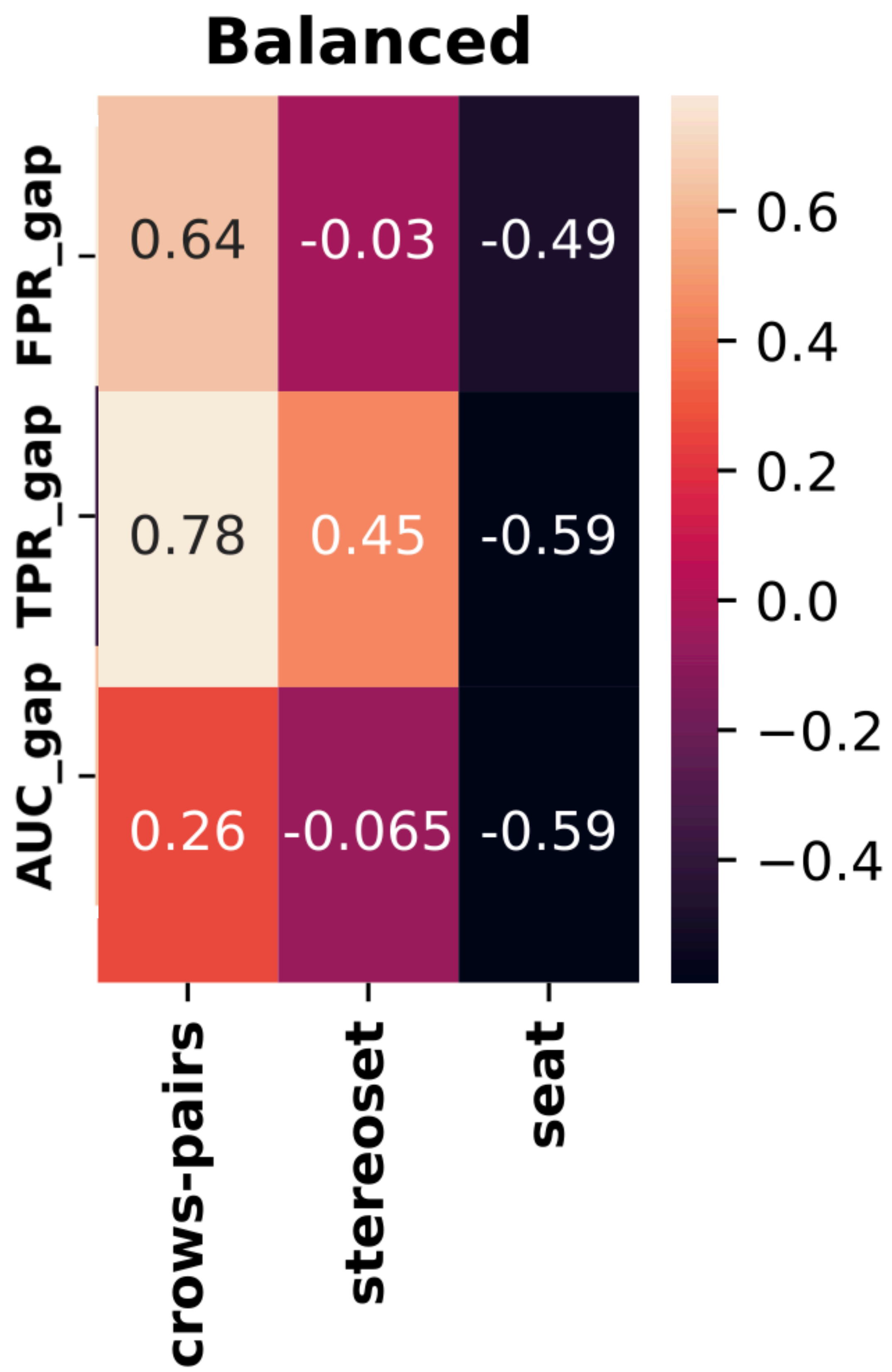
Measurement & Impact

Different bias metrics give different results.

	CrowS-Pairs		
	Gender	**Race**	**Religion**
AlBERT	0.541	0.513	0.590
BERT	0.580	**0.581**	0.714
RoBERTa	**0.606**	0.527	**0.771**
	StereoSet		
	Gender	**Race**	**Religion**
AlBERT	0.599	0.575	0.603
BERT	0.607	0.570	0.597
RoBERTa	**0.663**	**0.616**	**0.642**
	SEAT		
	Gender	**Race**	**Religion**
AlBERT	0.622	0.551	0.430
BERT	0.620	**0.620**	**0.491**
RoBERTa	**0.939**	0.307	0.126

Bias scores

There is positive correlation between fairness metrics and Crows-Pairs scores.



Correlation scores between bias scores and fairness scores

Selection bias

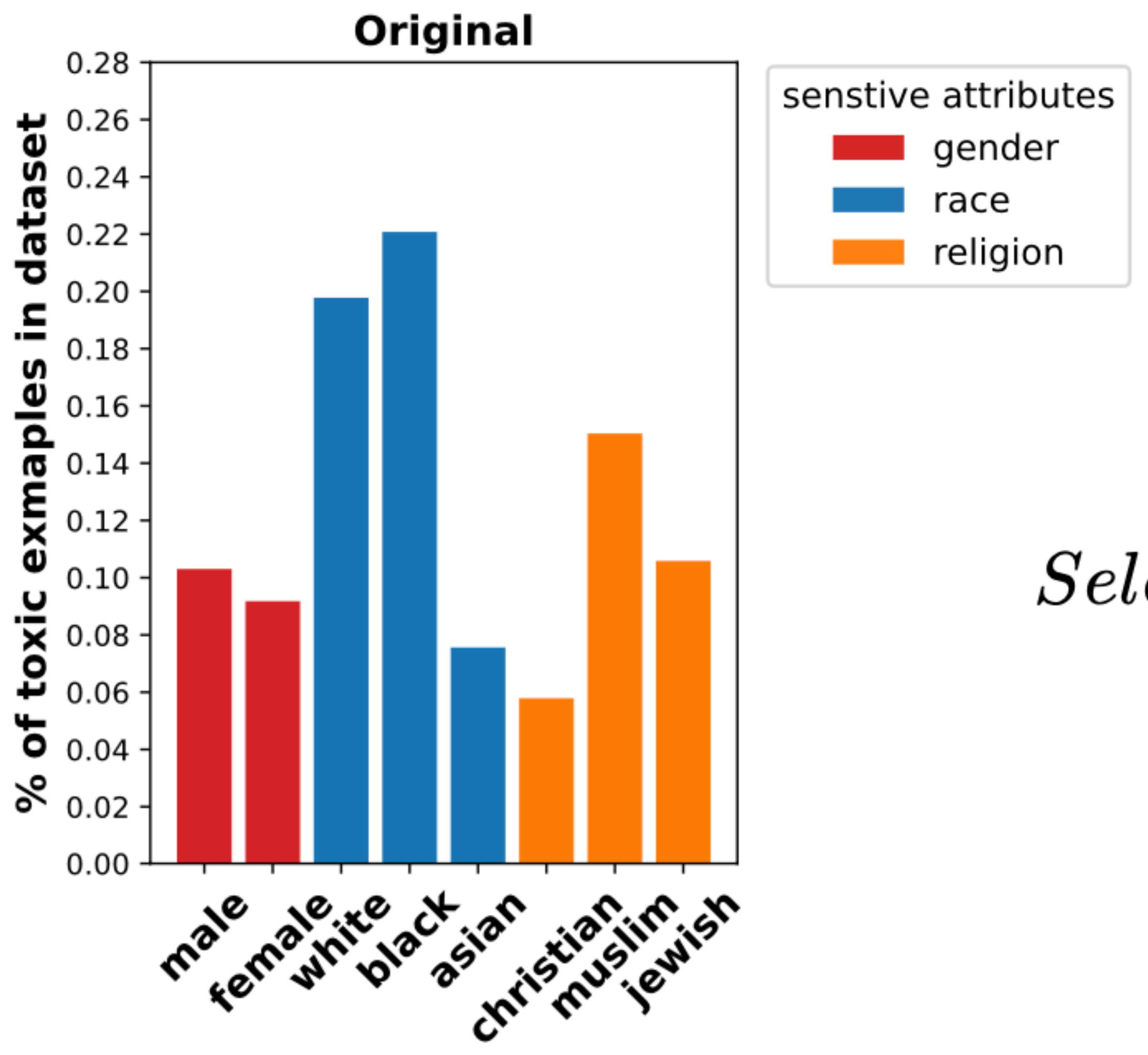
Measurement & Impact

There is positive correlation between fairness metrics and Selection bias.

But not for all the models

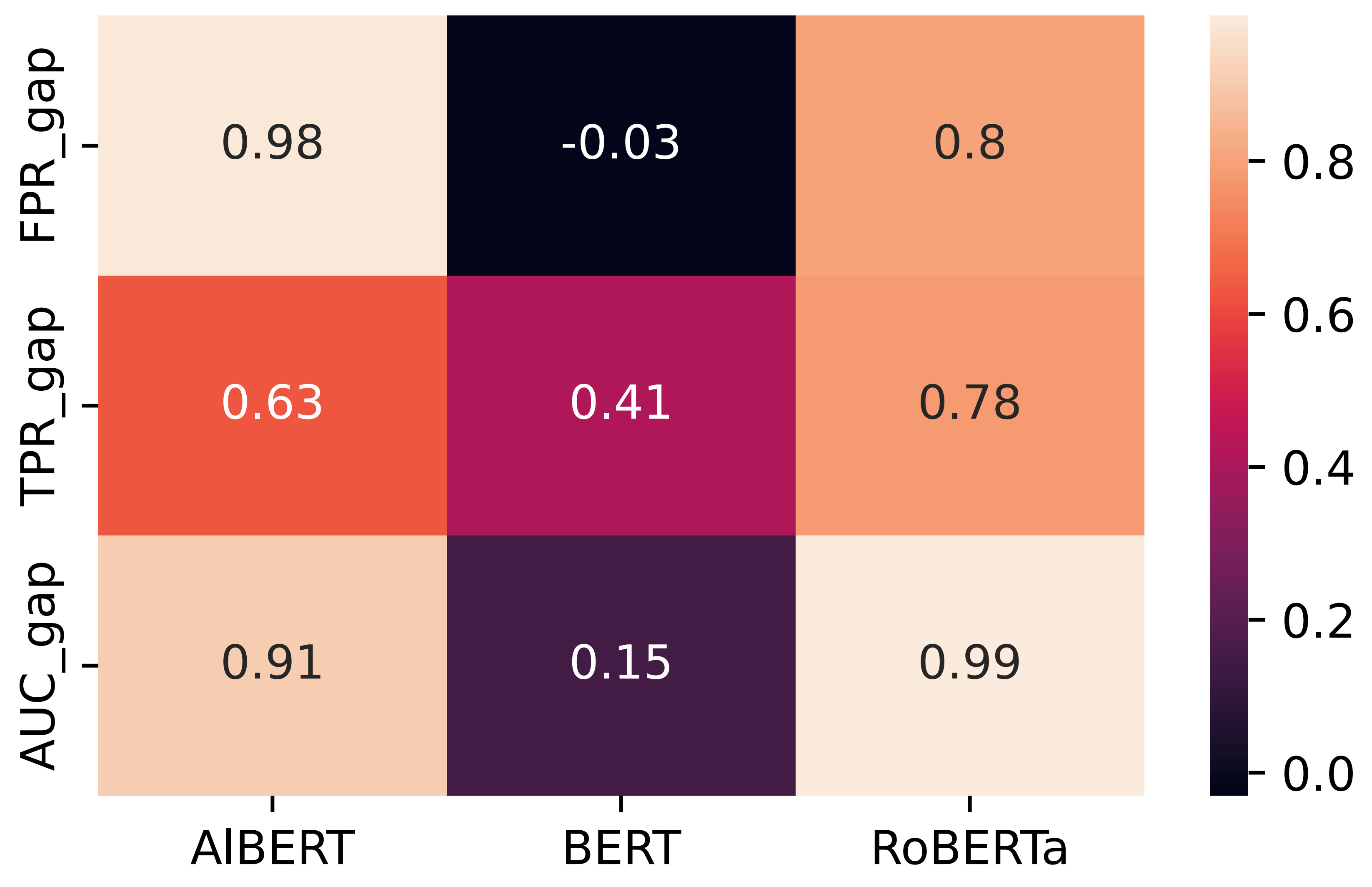
<u>**Selection/Sample bias[1]**</u> *: is a result of non-representative observations in the training datasets used in downstream tasks.*

<u>***For toxicity detection***</u> : *The over-representation of a certain group with the toxic label.*



Jigsaw Training Dataset

$$Selection_{g,\hat{g}} = \left| \left(\frac{N_{g,toxicity=1}}{N_g} \right) - \left(\frac{N_{\hat{g},toxicity=1}}{N_{\hat{g}}} \right) \right|$$



Correlation scores between bias scores and fairness scores

[1] Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

Overamplification bias

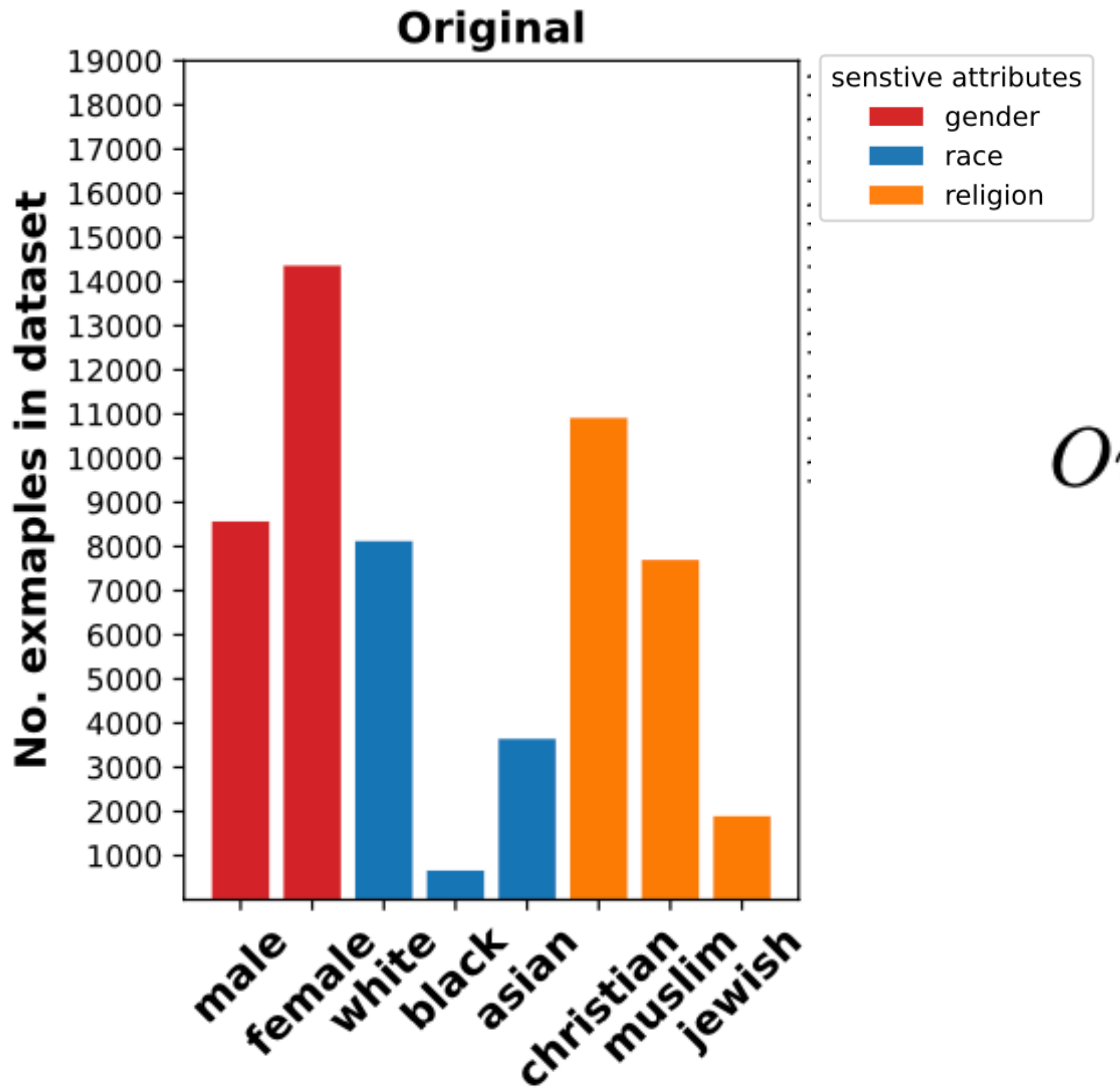
Measurement & Impact

There is positive correlation between fairness metrics and Overamplification bias.

But not for all the models

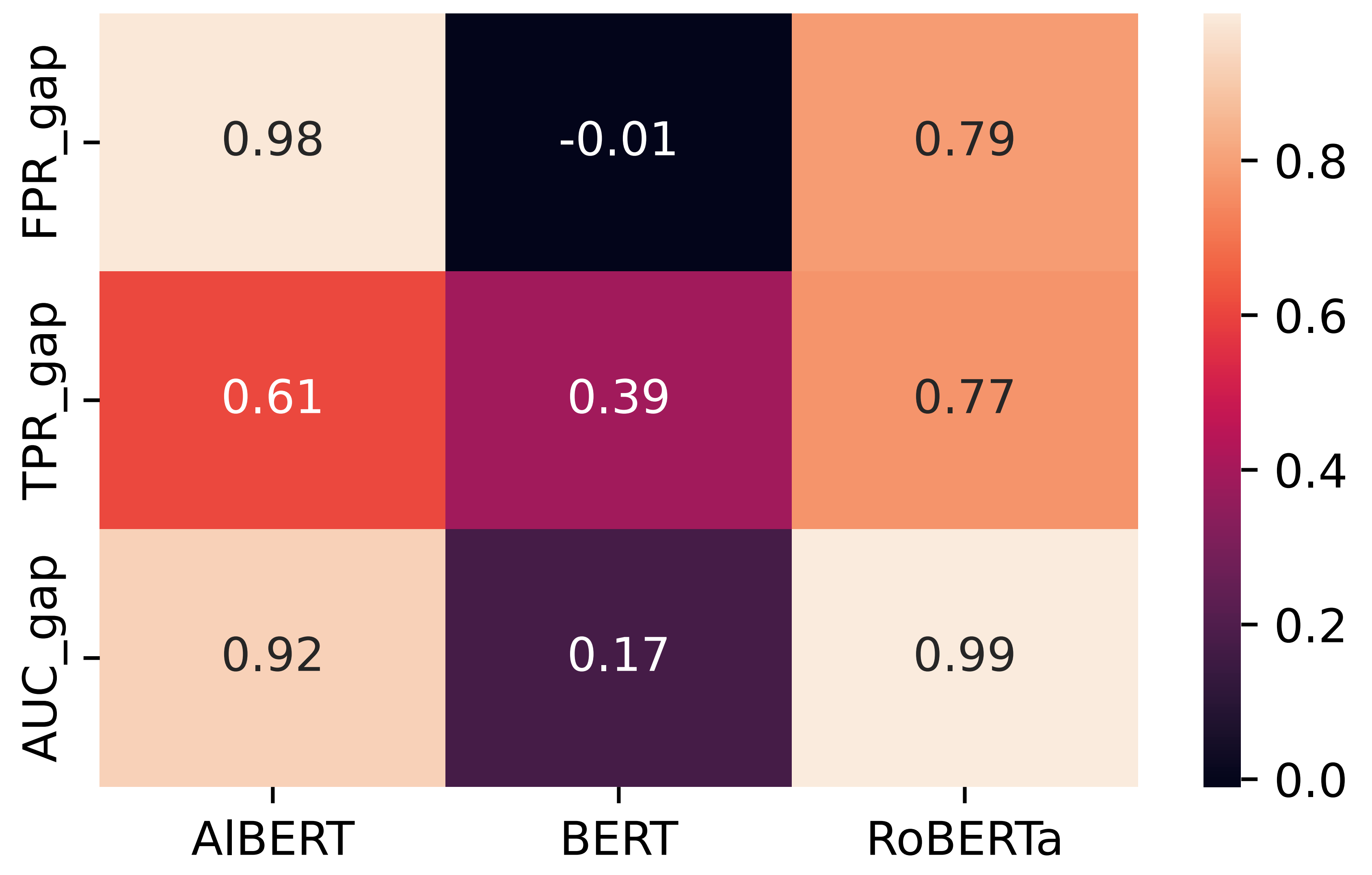
<u>**Overamplification bias**</u>[1] *:During training, LMs amplify small differences between different groups.*

<u>For toxicity detection</u> : *The over representation of certain identity group with a certain context*



Jigsaw Training Dataset

$$Overamplification_{g,\hat{g}} = |N_g - N_{\hat{g}}|$$



Correlation scores between bias scores and fairness scores

[1] Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

Sources of bias

What is the impact of different sources of bias on the Fairness of toxicity detection?

All sources of bias have an impact of the fairness of toxicity detection.

Downstream sources (selection & oversimplification) of bias are more impactful than representation bias.

The results are not consistent across all models or metrics.

AlBERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	0.466	**0.999**	0.233
Selection	0.984	0.633	0.911
Overamplification	**0.988**	0.613	**0.921**
BERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	-0.536	**0.819**	-0.369
Selection	-0.037	0.418	0.150
Overamplification	**-0.011**	0.395	**0.175**
RoBERTa			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	**0.972**	**0.980**	0.555
Selection	0.809	0.785	0.992
Overamplification	0.794	0.770	**0.995**

Pearson Correlation Coefficient between different bias scores and fairness of toxicity detection

What is the impact of **removing** different sources of bias on the Fairness of toxicity detection?

Sources of bias

Bias removal methods

1. Remove Representation Bias

Use SentDebias[1] to remove gender, racial, and religious bias (Upstream-SentDebias)

2. Remove Selection Bias

Stratification[2]: Data augmentation used to create more positive examples.

3. Remove Overamplification Bias

- Data Perturbation[3]: Creating counterfactuals
- SentDebias after fine-tuning (Downstream-SentDebias)

4. Remove Downstream Sources of Bias

5. Remove all Sources of Bias

[1] Liang, Paul Pu, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Association for Computational Linguistics, Online.

[2] Zmigrod, Ran, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Association for Computational Linguistics, Florence, Italy.

[3] Webster, Kellie, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report, Google Research.

Sources of bias

Bias removal impact on fairness

Remove Representation Bias
Remove Overamplification Bias
Remove Selection Bias
Remove All Downstream Bias
Remove all Sources of Bias

	AlBERT-base			BERT-base			RoBERTa-base		
Debias approach	gender	race	religion	gender	race	religion	gender	race	religion
Upstream-SentDebias	✗	✗	✗	✗	✗	✗	✗	✗	✓
Downstream-SentDebias	✗	✗	✓	✓	✓	✗	✗	✓	✓
Downstream-perturbed-data	✗	✓	✓	✓	✗	✓	✓	✗	✓
Downstream-stratified-data	✗	✗	✗	✓	✗	✗	✗	✗	✓
Downstream-perturbed-stratified-data	✗	✗	✓	✓	✗	✓	✓	✗	✓
Upstream-sentDebias-Downstream-all-data-debias	✗	✗	✓	✓	✗	✓	✓	✗	✓

Summary of the most effective debiasing method according to all the fairness metrics for all the models and all the sensitive attributes.

Removing Representation bias did not have an impact on improving fairness.

Sources of bias

Bias removal impact on fairness

Remove Representation Bias

Remove Overamplification Bias

Remove Selection Bias

Remove All Downstream Bias

Remove all Sources of Bias

	AlBERT-base			BERT-base			RoBERTa-base		
Debias approach	gender	race	religion	gender	race	religion	gender	race	religion
Upstream-SentDebias	✗	✗	✗	✗	✗	✗	✗	✗	✓
Downstream-SentDebias	✗	✗	✓	✓	✓	✗	✗	✓	✓
Downstream-perturbed-data	✗	✓	✓	✓	✗	✓	✓	✗	✓
Downstream-stratified-data	✗	✗	✗	✓	✗	✗	✗	✗	✓
Downstream-perturbed-stratified-data	✗	✗	✓	✓	✗	✓	✓	✗	✓
Upstream-sentDebias-Downstream-all-data-debias	✗	✗	✓	✓	✗	✓	✓	✗	✓

Summary of the most effective debiasing method according to all the fairness metrics for all the models and all the sensitive attributes.

Removing Overamplification bias using SentDebias after fine-tuning led to the worst performance.

Sources of bias

Bias removal impact on fairness

Remove Representation Bias

Remove Overamplification Bias

Remove Selection Bias

Remove All Downstream Bias

Remove all Sources of Bias

	AlBERT-base			BERT-base			RoBERTa-base		
Debias approach	gender	race	religion	gender	race	religion	gender	race	religion
Upstream-SentDebias	✗	✗	✗	✗	✗	✗	✗	✗	✓
Downstream-SentDebias	✗	✗	✓	✓	✓	✗	✗	✓	✓
Downstream-perturbed-data	✗	✓	✓	✓	✗	✓	✓	✗	✓
Downstream-stratified-data	✗	✗	✗	✓	✗	✗	✗	✗	✓
Downstream-perturbed-stratified-data	✗	✗	✓	✓	✗	✓	✓	✗	✓
Upstream-sentDebias-Downstream-all-data-debias	✗	✗	✓	✓	✗	✓	✓	✗	✓

Summary of the most effective debiasing method according to all the fairness metrics for all the models and all the sensitive attributes.

Removing Overamplification bias using data perturbation was the most effective. More than removing all downstream bias or even all sources of bias

Sources of bias

Bias removal impact on fairness

- To confirm these results, we use counterfactual fairness metric (SenseScore).

Sentence
% of men lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the men are charged you sir you are one of the men that is the problem I can tell !
% of women lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the women are charged you mademoiselle you are one of the women that is the problem I can tell !

Example of a sentence where the original target is a Male (top) and when the gender is swapped to Female (bottom).

$$SenseScore = |Mean_{x \in X}(f(\hat{x}) - f(x))|$$

Using counterfactual fairness confirmed our same results

What have we learned?

To improve the task of toxicity detection

- *Know the data*: measure bias in fine-tuning datasets.
- *Remove downstream sources of bias:* Using data perturbations.
- Balance the fairness datasets.
- Use more than one metric.

What have we learned?

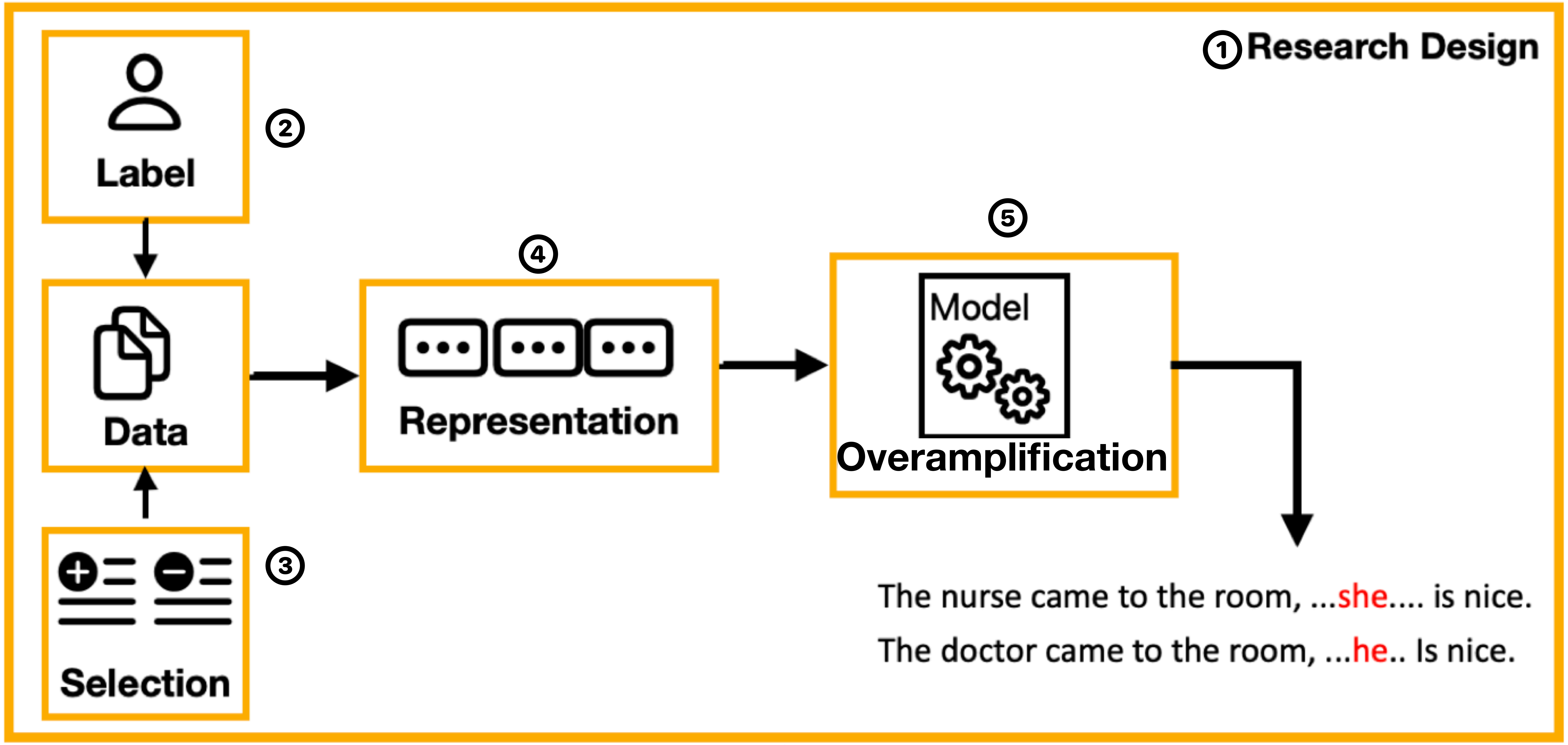
Challenges

1. Failing to define what the metrics actually measure[2].
2. Different results from different bias metrics.
3. Published results are not replicatable.
4. Bias metrics measure the existence of bias not it's absence[1].
5. Ineffective representation bias removal methods.

[1] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.
[2] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*
[3] Hedden, B. (2021), On statistical criteria of algorithmic fairness. Philos Public Aff, 49: 209-231. https://doi.org/10.1111/papa.12189

What are the *Origins* of Bias?

Sources of Bias in NLP



Conceptual framework of five sources bias in NLP models [1,2]

[1] Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

[2] Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

Origins of Bias

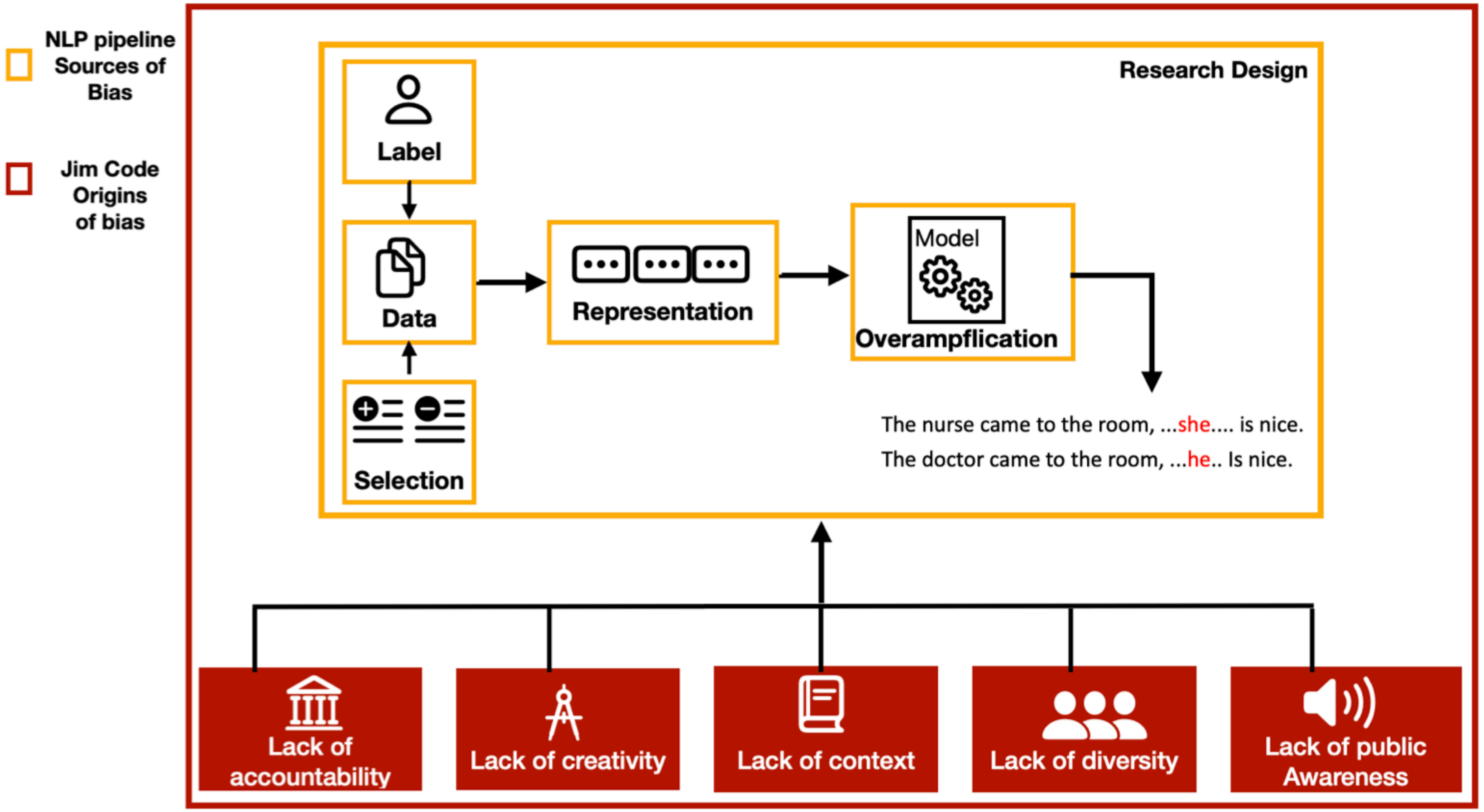
We build this list to origins of bias from studies in

- digital humanities,
- critical race theory,
- gender studies,
- and sociology.

Fatma Elsafoury, Gavin Abercrombie. ”On the Origins of Bias in NLP through the Lens of the Jim Code”. A long paper **arXiv preprint arXiv:2305.09281, 2023.**

Origins of Bias

1. Lack of context.
2. Lack of creativity.
3. Lack of accountability.
4. Lack of diversity.
5. Lack of awareness.



The origins of bias in supervised NLP models

Origins of Bias

1. Lack of context is when **social and historical contexts** are not **considered** during data collection or the research design .

<u>For example:</u>

- Using data collected in the **50s, 60s** without regard to the **discriminatory laws and racial and gender divid** in societies back then.
- Or even **now** using machine **generated text** to train new NLP models without regard the **biases** those generated texts reproduce.
- Using NLP models to make decisions on **eligibility jobs** on criteria that might end up increasing **the wealth gap**.

Origins of Bias

1. Lack of context is when **social and historical contexts** are not **considered** during data collection or the research design .

NLP Sources of Bias

For example:

- Using data collected in the **50s, 60s** without regard to the **discriminatory laws and racial and gender divid** in societies back then.
- Or even **now** using machine **generated text** to train new NLP models without regard the **biases** those generated texts reproduce.
- Using NLP models to make decisions on **eligibility jobs** on criteria that might end up increasing **the wealth gap**.

Research Design Bias

Selection Bias

Overamplification Bias

Representation Bias

Origins of Bias

2. Lack of creativity is when we building **NLP** systems on top of **discriminatory** systems.

For example

- Recommendation systems use "*Culture segregation*" to infer information about a person's **ethnicity** to **personalise** the recommendations using **ethnicity as a proxy for individuality**.

Origins of Bias

2. Lack of creativity is when we building **NLP** systems on top of **discriminatory** systems.

NLP Sources of Bias

- Research Design Bias
- Selection Bias
- Overamplification Bias
- Representation Bias

For example

- Recommendation systems use "*Culture segregation*" to infer information about a person's **ethnicity** to **personalise** the recommendations using **ethnicity as a proxy for individuality**.

Origins of Bias

3. Lack of accountability leads to big tech priotrise profit maximisation over societal impact.

<u>For example</u>

- When the **Justice League** launched **the Safe Face pledge** to ensure that computer vision is not used to **discriminate** between people**, no major tech company** was willing to sign it.

- The **Exploitation** of **Data/Platform workers**.

Origins of Bias

3. Lack of accountability leads to big tech priotrise profit maximisation over societal impact.

NLP Sources of Bias

- Research Design Bias
- Label Bias

For example

- When the **Justice League** launched **the Safe Face pledge** to ensure that computer vision is not used to **discriminate** between people, **no major tech company** was willing to sign it.
- The **Exploitation** of **Data/Platform workers**.

Origins of Bias

3. Lack of accountability leads to big tech priotrise profit maximisation over societal impact.

NLP Sources of Bias

- Research Design Bias
- Label Bias

<u>**For example**</u>

- When the **Justice League** launched **the Safe Face pledge** to ensure that computer vision is not used to **discriminate** between people, **no major tech company** was willing to sign it.

- The **Exploitation** of **Data/Platform workers**.

Origins of Bias

4. Lack of diversity as the major companies and research institutes are in Western countries.

For example:

- Lack of NLP and recommendation systems for **indigenous languages or dialects**.
- Translation tools and **content moderation** tools **failing** to work with **indigenous languages**.

Origins of Bias

4. Lack of diversity as the major companies and research institutes are in Western countries.

NLP Sources of Bias

Research Design Bias

Label Bias

For example:

- Lack of NLP and recommendation systems for **indigenous languages or dialects**.
- Translation tools and **content moderation** tools **failing** to work with **indigenous languages**.

Origins of Bias

Jim Code perspective

5. Lack of awareness leads to technochauvinism or believing that computational solutions are considered superior to all other solutions.

For example

- Developing tools to remove bias in LMs instead of spending time to collect more representative data.

Origins of Bias

Jim Code perspective

5. Lack of awareness leads to technochauvinism or believing that computational solutions are considered superior to all other solutions.

NLP Sources of Bias

For example

- Developing tools to remove bias in LMs instead of spending time to collect more representative data.

Research Design Bias

Selection Bias

Overamplification Bias

Representation Bias

Label Bias

How do we **mitigate some** of the **origins** of **bias** and in turn the **sources** of **bias** in NLP?

What have we learned?

Long-term Recommendations

- Interdisciplinary research
- Raising awareness of social and historic contexts.
- Raising awareness of thinking about the social impact of development decisions.
- State level regulations.

Thank You!

Questions?

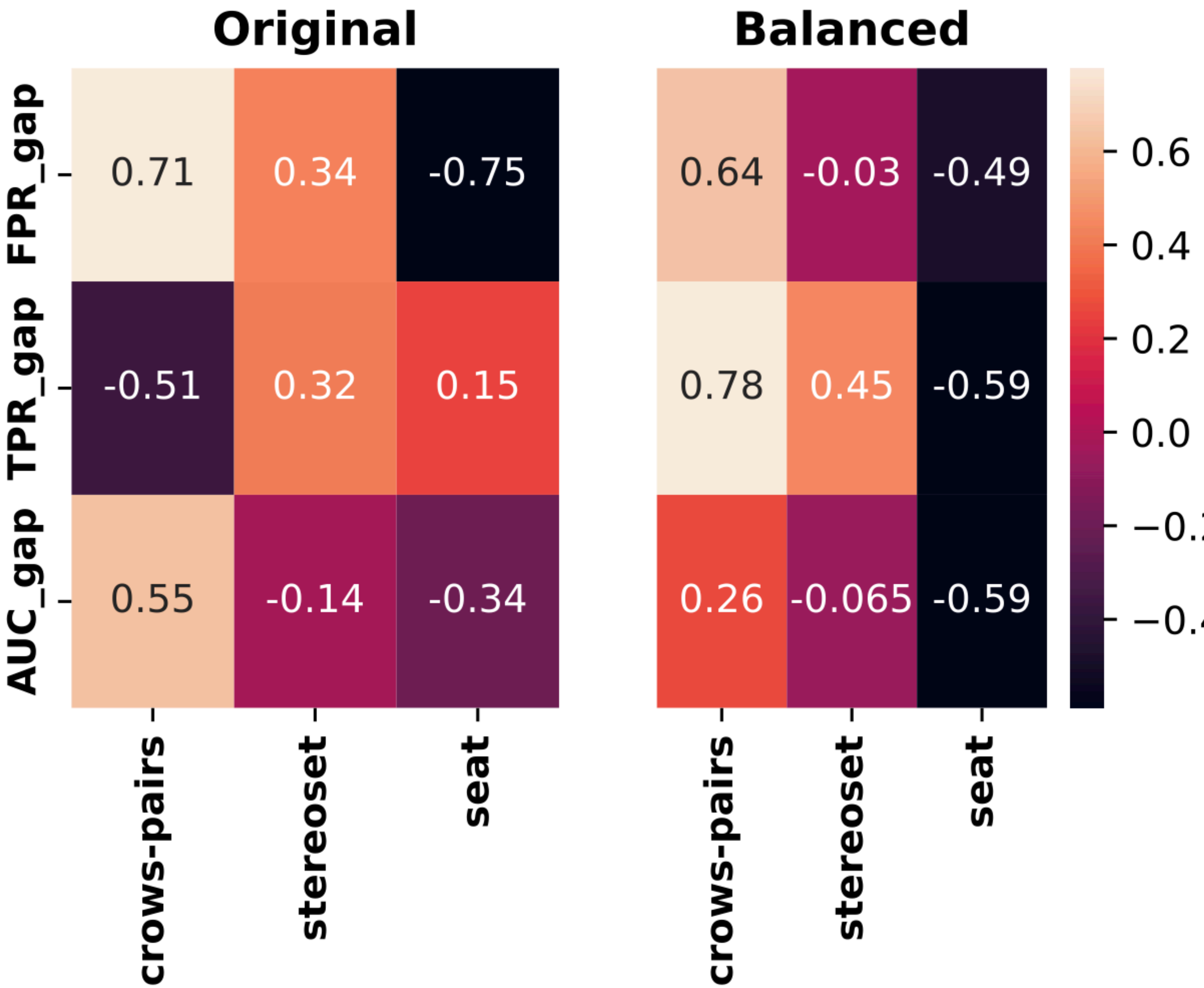
Fatma Elsafoury

Representation bias

Measurement & Impact

1. Positive correlation between fairness and Bias scores measured using Crows-Pairs

2. More consistent correlation results for the balanced fairness datasets.



Representation bias

Debias & Impact

1. Lack of consistency across different metrics.
2. According to CrowS-Pairs, SentDebias worsened in some cases.
3. Unlike the published results in [1], The scores have not changed for SEAT.

Model	CrowsPairs			StereoSet			SEAT		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
AlBERT-base	0.541	0.513	0.590	0.599	0.575	0.603	0.622	0.551	0.430
+ SentDebias-gender	↓0.461	↓0.436	↓0.466	↓0.517	↓0.552	↓0.586	0.622	0.551	0.430
+ SentDebias-race	↑0.564	↓0.440	↑0.666	↓0.542	↓0.521	↓0.555	0.622	0.551	0.430
+ SentDebias-religion	↑0.549	↑0.660	↓0.581	↓0.501	↓0.529	↓0.510	0.622	0.551	0.430
BERT-base-uncased	0.580	0.581	0.714	0.607	0.5702	0.597	0.620	0.620	0.491
+ SentDebias-gender	↓0.427	↓0.555	↓0.647	↓0.475	↓0.476	↓0.504	0.620	0.620	0.491
+ SentDebias-race	↓0.534	↓0.398	↓0.704	↓0.467	↓0.562	↓0.489	0.620	0.620	0.491
+ SentDebias-religion	↓0.534	↑0.675	↓0.561	↓0.469	↓0.511	↓0.399	0.620	0.620	0.491
RoBERTa-base	0.606	0.527	0.771	0.663	0.616	0.642	0.939	0.307	0.126
+ SentDebias-gender	↓0.467	↑0.691	↓0.561	↓0.518	↓0.497	↓0.477	0.939	0.307	0.126
+ SentDebias-race	↓0.429	↓0.467	↓0.419	↓0.485	↓0.488	↓0.486	0.939	0.307	0.126
+ SentDebias-religion	↓0.413	↓0.478	↓0.352	↓0.516	↓0.497	↓0.486	0.939	0.307	0.126

Table 4: Representation bias scores in the examined models using different bias metrics before and after removing bias using the SentDebias algorithm. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased, and the fairness improved.

[1] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Representation bias

Debias & Impact

1. Performance did not change much.
2. Debias led to more positive predictions in general (FP & TP).
3. Fairness did not necessarily improve across all metrics except for removing religion bias from RoBERTA.
4. No statistically significant difference.

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ upstream-sentDebias-gender	0.840	0.006	↓ 0.032	0.004
	BERT	0.830	0.090	0.036	0.010
	+ upstream-sentDebias-gender	**0.841**	↓ 0.011	↑ 0.049	↓ 0.006
	RoBERTa	0.851	0.005	0.032	0.011
	+ upstream-sentDebias-gender	**0.856**	↑ 0.006	↓ 0.022	↓ 0.003
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ upstream-sentDebias-race	0.838	↓ 0.003	↑ 0.003	↓ 0.013
	BERT	**0.830**	0.016	0.002	0.026
	+ upstream-sentDebias-race	0.829	↑ 0.021	↑ 0.005	↓ 0.024
	RoBERTa	0.851	0.003	0.011	0.021
	+ upstream-sentDebias-race	**0.854**	↑ 0.017	↓ 0.009	0.021
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ upstream-sentDebias-religion	0.837	↑ 0.019	↓ 0.094	↓ 0.016
	BERT	0.830	0.008	0.063	0.012
	+ upstream-sentDebias-religion	**0.833**	↑ .015	↑ 0.084	↑ 0.017
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ upstream-sentDebias-religion	0.843	↓ 0.021	↓ 0.100	↓ 0.003

Table 5: Fairness scores of the models on Toxicity detection, after removing representation bias

Selection bias

Measurement & Impact

Selection Bias in the training dataset is:

- Religion (0.08)
- Race (0.05)
- Gender (0.03)

	Fairness metrics		
Model	**FPR_gap**	**TPR_gap**	**AUC_gap**
AlBERT	0.98	0.63	0.91
BERT	-0.03	0.41	0.15
RoBERTa	0.80	0.78	0.99

Pearson Correlation coefficient between Selection bias scores and fairness scores

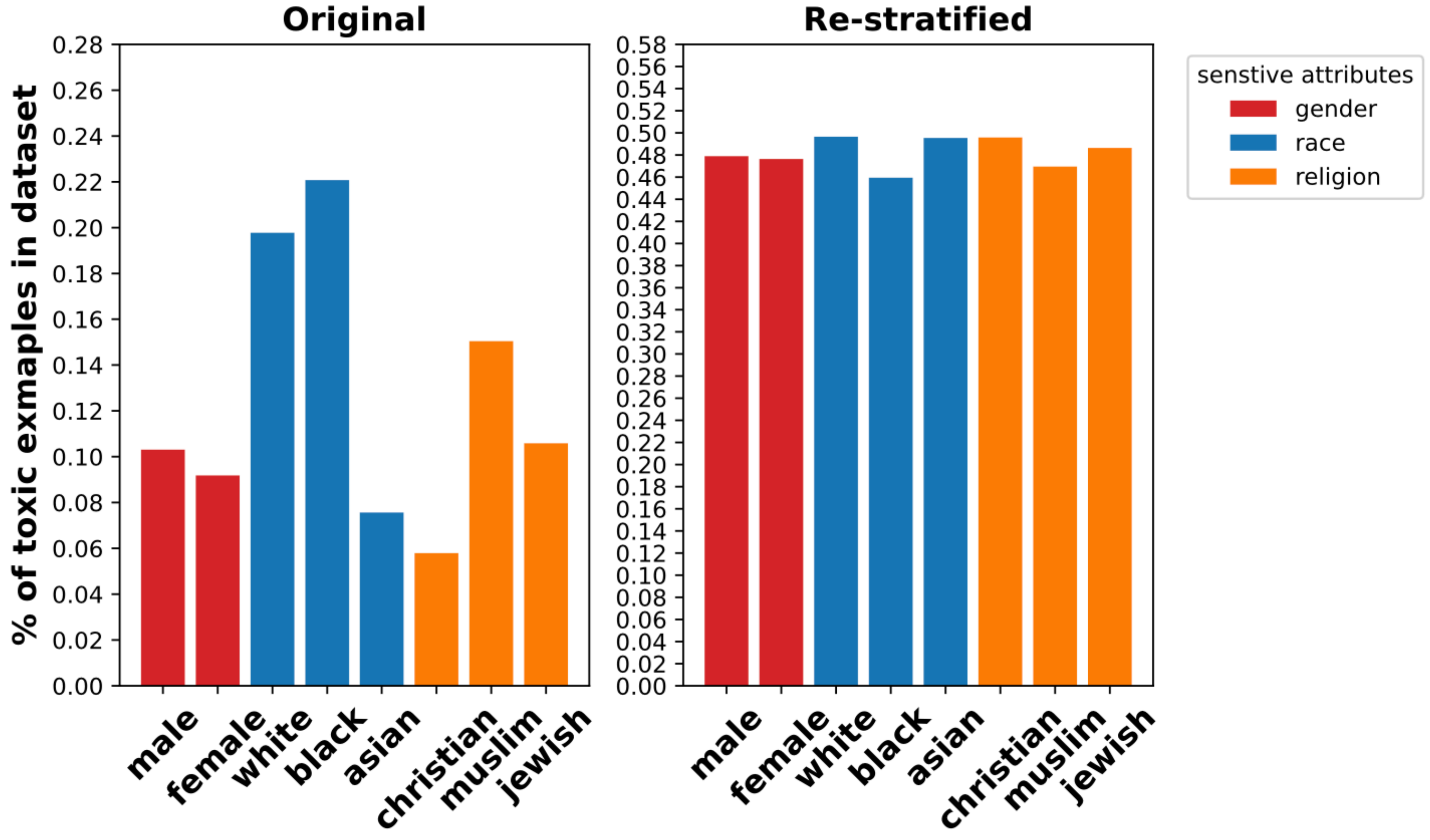
For AlBERT and RoBERTa, there is a **strong positive correlation** between Selection bias scores and fairness scores measured using different metrics. But not BERT.

Selection bias

Debias & Impact

To remove selection bias, minimise the mismatch in class representation between different identities.

- Data augmentation used to create more positive examples.
- NLPAUG[1] tool used to create word substitutions to augment the positive examples.
- Create dataset with balanced positive to negative examples for all groups.
- Size of training dataset 443K.



Jigsaw Training Dataset

[1] [1] https://github.com/makcedward/nlpaug

Selection bias

Debias & Impact

1. Performance got worse.
2. Debias led to more positive predictions in general (FP & TP) and less TN.
3. Inconsistent results except for the AUC_gap metric.

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ downstream-stratified-data	0.816	↓ 0.005	↓ 0.003	↑ 0.005
	BERT	**0.830**	0.090	0.036	0.010
	+ downstream-stratified-data	0.817	↓ 0.007	↓ 0.006	↓ 0.006
	RoBERTa	**0.851**	0.005	0.032	0.011
	+ downstream-stratified-data	0.842	↑ 0.006	↓ 0.005	↓ 0.002
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ downstream-stratified-data	0.816	↑ 0.022	↑ 0.026	↓ 0.008
	BERT	**0.830**	0.016	0.002	0.026
	+ downstream-stratified-data	0.817	↓ 0.010	↑ 0.018	↓ 0.008
	RoBERTa	**0.851**	0.003	0.011	0.021
	+ downstream-stratified-data	0.842	↑ .014	0.011	↓ 0.014
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ downstream-stratified-data	0.816	↑ 0.030	↓ 0.058	↓ 0
	BERT	**0.830**	0.008	0.063	0.012
	+ downstream-stratified-data	0.817	↑ 0.020	↓ 0.049	↓ 0.006
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ downstream-stratified-data	0.842	↓ 0.019	↓ 0.071	↓ 0.001

Table 6: Toxicity detection performance and fairness scores for all models before and after removing selection bias. **Bold** values refer to higher AUC scores and better performance. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.

Overamplification bias

Measurement & Impact

Selection Bias in the training dataset is:

- Religion (1)
- Race (0.97)
- Gender (0.94)

	Fairness metrics		
Model	**FPR_gap**	**TPR_gap**	**AUC_gap**
AlBERT	0.98	0.613	0.92
BERT	-0.01	0.39	0.175
RoBERTa	0.79	0.77	0.99

Pearson Correlation coefficient between Selection bias scores and fairness scores

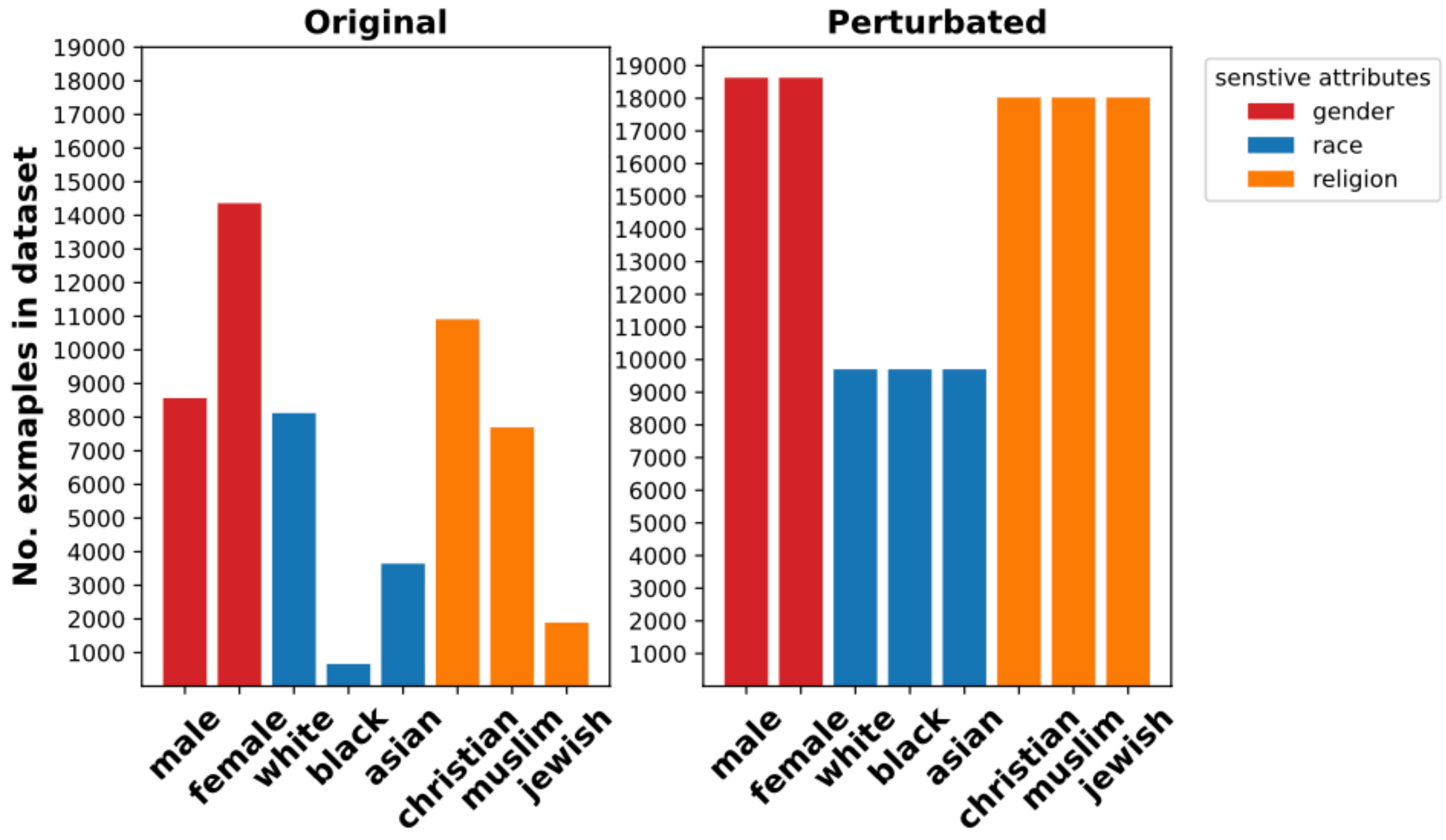
For AlBERT and RoBERTa, there is a **strong positive correlation** between Overamplification bias scores and fairness scores measured using different metrics. But not BERT.

Overamplification bias

Debias & Impact

To remove oversimplification bias, train the model on a dataset with balanced semantic representations.

- Data perturbations
- Train a text-to-text model on PANDA dataset to automatically generate perturbations. ROUGE-2 = 0.9 But results were not good.
- Lexical word replacement.
- Size of training dataset 382K.



Jigsaw Training Dataset

Overamplification bias

Debias & Impact

1. Downstream debias performance was random.

2. Using perturbed data improved the performance and the fairness

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	0.847	0.006	0.039	0.004
	+ downstream-sentDebias-gender	0.524	↓ 0	↓ 0.008	↑ 0.011
	+ downstream-perturbed-data	**0.848**	↓ 0.001	↓ 0.010	0.004
	+ downstream-perturbed-stratified-data	0.803	↓ 0.005	↓ 0.006	↑ 0.008
	BERT	0.830	0.09	0.036	0.01
	+ downstream-sentDebias-gender	0.478	↓ 0	↓ 0.001	↓ 0.004
	+ downstream-perturbed-data	**0.837**	↓ 0.003	↓ 0.005	↓ 0.003
	+ downstream-perturbed-stratified-data	0.810	↓ 0.003	↓ 0.003	↓ 0.005
	RoBERTa	0.851	0.005	0.032	0.011
	+ downstream-sentDebias-gender	0.520	↑ 0.015	↓ 0.019	↓ 0.004
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.009	↓ 0.002
	+ downstream-perturbed-stratified-data	0.825	↓ 0	↓ 0.005	↓ 0.007
Race	ALBERT	0.847	0.008	0.002	0.019
	+ downstream-sentDebias-race	0.421	↓ 0	↑ 0.004	↓ 0.001
	+ downstream-perturbed-data	**0.848**	↓ 0.003	↓ 0.001	↓ 0.003
	+ downstream-perturbed-stratified-data	0.803	↑ 0.004	0.002	↓ 0.002
	BERT	0.830	0.016	0.002	0.026
	+ downstream-sentDebias-race	0.504	↓ 0	↓ 0	↓ 0.002
	+ downstream-perturbed-data	**0.837**	↓ 0.009	↑ 0.019	↓ 0.003
	+ downstream-perturbed-stratified-data	0.810	↓ 0.002	0.002	↓ 0.002
	RoBERTa	0.851	0.003	0.011	0.021
	+ downstream-sentDebias-race	0.561	↓ 0	↓ 0	↓ 0.005
	+ downstream-perturbed-data	**0.873**	↑ 0.018	↑ 0.038	↓ 0.003
	+ downstream-perturbed-stratified-data	0.825	0.003	↓ 0.006	↓ 0.001
Religion	ALBERT	0.847	0.010	0.109	0.020
	+ downstream-sentDebias-religion	0.507	↓ 0.004	↓ 0	↓ 0.002
	+ downstream-perturbed-data	**0.848**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-perturbed-stratified-data	0.803	↓ 0	↓ 0.002	↓ 0.002
	BERT	0.830	0.008	0.063	0.012
	+ downstream-sentDebias-religion	0.447	↓ 0	↓ 0	↑ 0.030
	+ downstream-perturbed-data	**0.837**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-perturbed-stratified-data	0.810	↓ 0	↓ 0.001	↓ 0.003
	RoBERTa	0.851	0.022	0.160	0.027
	+ downstream-sentDebias-religion	0.523	↓ 0	↓ 0	↓ 0
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.003	↓ 0.002
	+ downstream-perturbed-stratified-data	0.825	↓ 0.001	↓ 0.004	↓ 0.001

Sources of bias

Bias removal impact on fairness

- Using perturbed data to balance the representation of different groups is the most effective in improving fairness.
- Using perturbed data improved the fairness without harming the performance unlike stratification.

	SenseScore		
Model	Gender	Race	Religion
AlBERT-base	$6.9e^{-05}$	0.032	0.006
+ downstream-perturbed-data	↓ $4.2e^{-05}$	↓ 0.002	↓ 0.001
+ downstream-stratified-data	↑ 0.042	0.032	↑ 0.009
+ downstream-perturbed-stratified-data	↑ 0.013	↓ 0.003	↓ 0.0007
BERT-base	0.001	0.03	0.001
+ downstream-perturbed-data	↓ 0.0007	↓ 0.003	0.001
+ downstream-stratified-data	↑ 0.025	↓ 0.022	↑ 0.004
+ downstream-perturbed-stratified-data	↑ 0.002	↓ 0.002	↓ 0.0008
RoBERTa-base	0.001	0.024	0.003
+ downstream-perturbed-data	↓ 0.0008	↓ 0.006	↓ 0.001
+ downstream-stratified-data	↑ 0.038	↑ 0.036	0.003
+ downstream-perturbed-stratified-data	↑ 0.003	↓ 0.002	↓ 0.0003

SenseScores of the difference models before and after the different debiasing methods.