

On Bias and Fairness in LMs

Fatma Elsafoury

Mentor: Bhatta, **Manager:**
Hans



Introduction

Motivation

- Language Models are biased.
- Intrinsic Bias
 - Measuring intrinsic bias using one specific metric [1].
 - Larger models are more biased than smaller models [2].
 - Removing Hate, offense, profanity (HAP) might make the models less biased.
- Extrinsic Bias
 - Threshold-based metrics (Equalized Odds) is the metric to use when measuring extrinsic bias [1].
 - There is/is not correlation between intrinsic and extrinsic bias [1,3].

Contribution: In this work, I tested these claims to see which one hold.

[1] Upstream Mitigation: Is Not All You Need.

[2] TruthfulQA: Measuring How Models Mimic Human Falsehoods

[3] Intrinsic Bias in Language Models with English and Chinese

Intrinsic vs. extrinsic bias

- Intrinsic bias: is the bias where the model learns biased representations of different groups of people.
 - For example: “The **nurse** came to the room, ...**she**.... is nice.”
vs “The **doctor** came to the room, ...**he**.. Is nice.”
- Extrinsic bias (fairness): the unfair decision made by the model.
 - For example: For 2 CVs with the same skills and qualification but with different names (male vs. female), Amazon AI recruiting tool, would recommend the CV with the male name [1].

[1] <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

Intrinsic Bias Metrics

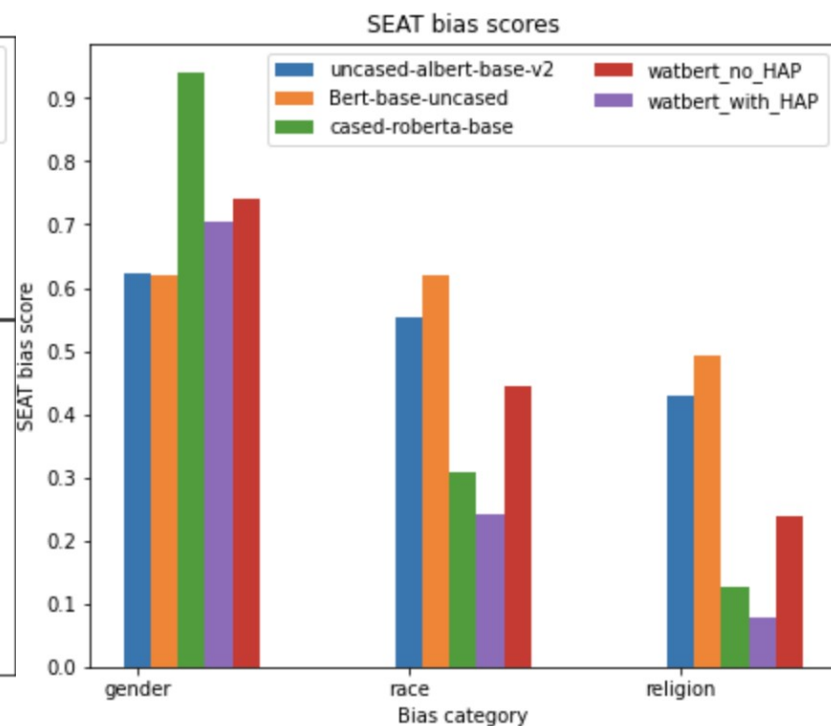
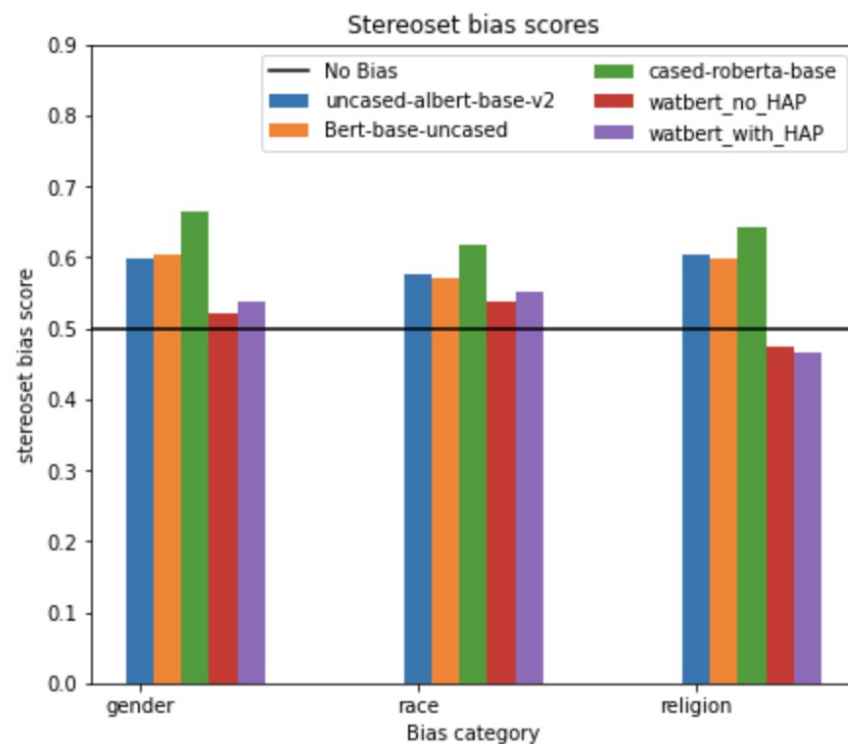
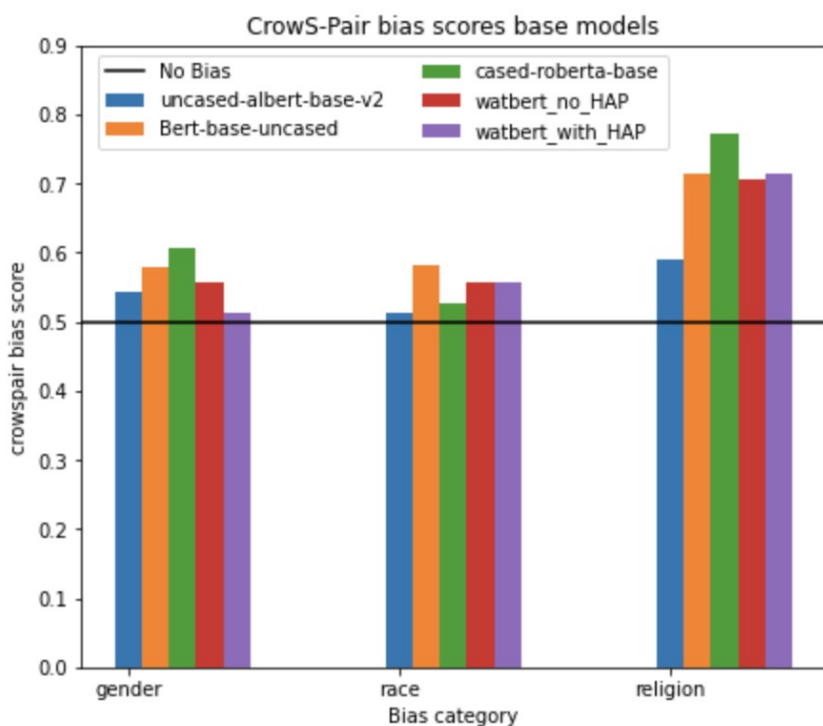
	CrowS-Pairs [2]	Stereoset [3]	SEAT [4]
Data	Human generated stereotyped vs non-stereotyped sentences	Human generated stereotyped vs non-stereotyped sentences	Bleached/Template This is [target], This is [attribute]
Task	MLM	MLM	encoding
e.g.	P(is a nurse she)	P(she is a nurse)	Cos('This is John', 'This is a doctor') - Cos('This is John', 'This is a nurse') Cos('This is Jane', 'This is a doctor') - Cos('This is Jane', 'This is a nurse')
Bias type	9 types	4 types	3 types

- [2] Crows-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models
 [3] StereoSet: Measuring stereotypical bias in pre-trained language models
 [4] On Measuring Social Biases in Sentence Encoders

Intrinsic Bias Metrics

Models	Pre-training data
Bert-base-uncased	Books Corpus and English Wikipedia
Roberta-base	Books Corpus, CC-NEWS, OPEN-WEB-TEXT, Stories
Albert-base	Books Corpus and English Wikipedia
Watson Roberta (WatBERT + HAP)	Book Corpus, English Wikipedia, OPEN-WEB-TEXT, CC-NEWS, WebHose, NMT
WatBERT	Same data as Watson Roberta but with Hateful/Abusive/Profane (HAP) removed (1.9% of the original data was removed).

Do bias metrics agree on the most biased models?



1. Do bias metrics agree on the most biased models?

- Bias scores don't agree
- Bias scores changed from the original papers with Transformers 4.
- Best practice to use more than one metric and go with the majority.

Answer: No, the bias metrics don't agree on the bias in the different models. They tell us about the existing of bias but they can't tell us about the exact amount of bias and may be not suitable for comparing models.

Bias and model size

[5] demonstrates that larger models are more biased **but** they only used auto-regressive models (GPT-3). **Does this claim hold for MLMs?**

	Layers	Hidden	Attention heads	No. Parameters
Bert-base-uncased	12	768	12	110M
Bert-large-uncased	24	1024	16	340M
Roberta-base	12	768	12	123M
Roberta-large	24	1024	16	355M
Albert-base-v2	12	768	12	11M
Albert-xlarge-v2	12	4096	64	233M

[5] TruthfulQA: Measuring How Models Mimic Human Falsehoods

Bias and model size

CrowS-Pair

	BERT		RoBERTa		ALBERT	
	Base	Large	Base	Large	Base	xx-Large
Gender	0.580	0.553	0.606	0.572	0.541	0.649**
Race	0.581	0.600	0.527	0.620**	0.513	0.643**
Religion	0.714	0.685	0.771	0.714	0.590	0.752**

The bias score (% of stereotyped sentences that received higher probability by the model)
** is statistically significant between individual sentences

Bias and model size

Stereoset

	BERT		RoBERTa		AIBERT	
	Base	Large	Base	Large	Base	xx-Large
Gender	0.602	0.632	0.663**	0.535	0.599	0.664**
Race	0.570	0.571	0.616**	0.546	0.575	0.611**
Religion	0.597	0.599	0.642**	0.508	0.603	0.696**

The bias score (% of stereotyped sentences that received higher probability by the model)
** is statistically significant between individual sentences

Bias and model size

SEAT

	BERT		RoBERTa		AIBERT	
	Base	Large	Base	Large	Base	xx-Large
Gender	0.620	0.331	0.939	0.627	0.622	0.387
Race	0.620	0.516	0.307	0.432	0.551	0.309
Religion	0.491	0.185	0.126	0.386	0.430	0.458

The bias score (The mean of the absolute effect size)
** is statistically significant between individual sentences

2. Do Larger models contain more social bias than smaller models?

- Bert-Base vs Bert-Large:
 - No difference according to the 3 metrics.
- Roberta-base vs Roberta-large:
 - No difference according to CrowS-Pair and SEAT.
 - Roberta-base is more biased according to stereoset.
- Albert-base vs Albert-xx-large:
 - Albert-xx_large is more biased according to CrowS-Pair and Stereoset.

Answer: For MLM, large models are not more socially biased than based models. May be if the model size continues to increase that might lead to a more biased model. More investigation need.

HAP removal

CrowS-Pair

	WatBERT + HAP	WatBERT
Gender	0.511	0.557
Race	0.556	0.558
Religion	0.714	0.704

The bias score (% of stereotyped sentences that received higher probability by the model)

** is statistically significant between individual sentences

HAP removal

Stereoset

	WatBERT + HAP	WatBERT
Gender	0.538	0.520
Race	0.552	0.538
Religion	0.467	0.474

The bias score (% of stereotyped sentences that received higher probability by the model)

** is statistically significant between individual sentences

HAP removal

SEAT

	WatBERT + HAP	WatBERT
Gender	0.705	0.739
Race	0.242	0.443
Religion	0.076	0.237

The bias score (The mean of the absolute effect size)
** is statistically significant between individual sentences

3. Does removing HAP from training datasets makes the language models less socially biased?

- According all the bias metrics, there is no statistically significant difference between WatBERT + HAP and WatBERT.
- Removing Hateful, abusive and profane content does not mean removing social bias.

No, removing HAP from training datasets does not lead to a less biased language model

Extrinsic Bias

- Downstream task: Toxicity Detection.
- Fine-tune the base models on the Jigsaw dataset.
- Measure the extrinsic bias in these models' predictions.

Extrinsic Bias Toxicity detection

- Jigsaw dataset:
 - Kaggle challenge
 - Release by the Conversation AI team (Jigsaw & Google)
 - ~ 2M Wikipedia comments.
 - labeling:
 - labelled as toxic or not.
 - with information on the identity of the target of the comment:
 - Religion, Sexual orientation, Gender, and Disability.

Extrinsic Bias

Toxicity detection

- Jigsaw data:
 - We don't know who are the annotators and how the annotations were collected.
 - Demographics information provided by both crowdsourced and automatically generated.
 - Demographic information are not clean. E.g. the target of the same comment could be labeled as "male " and "female".

Extrinsic Bias

Toxicity detection

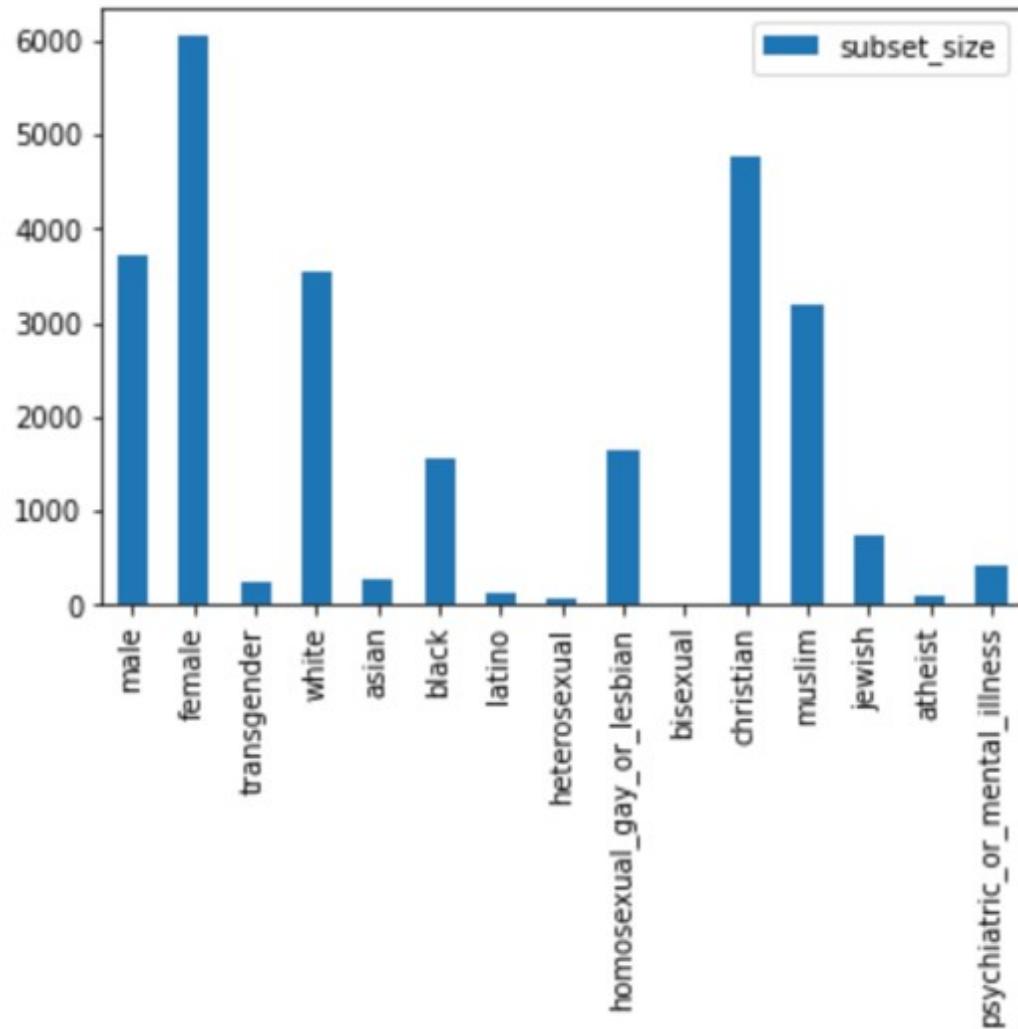
- Pre-processing Jigsaw data:
 - Use only data items with human annotations for demographic information (400K).
 - 70% training and 30% test.
 - For extrinsic bias: Filtered the test data to make sure that only one identity is present for 3 categories: gender, race, religion, and sexual-orientation (~21K).
 - However, the data is still not perfect.

Extrinsic Bias Toxicity detection

Model	F1
Bert-base	0.557
RoBERTa-base	0.570
Albert-base	0.525
WatBERT	0.567
WatBERT + HAP	0.561

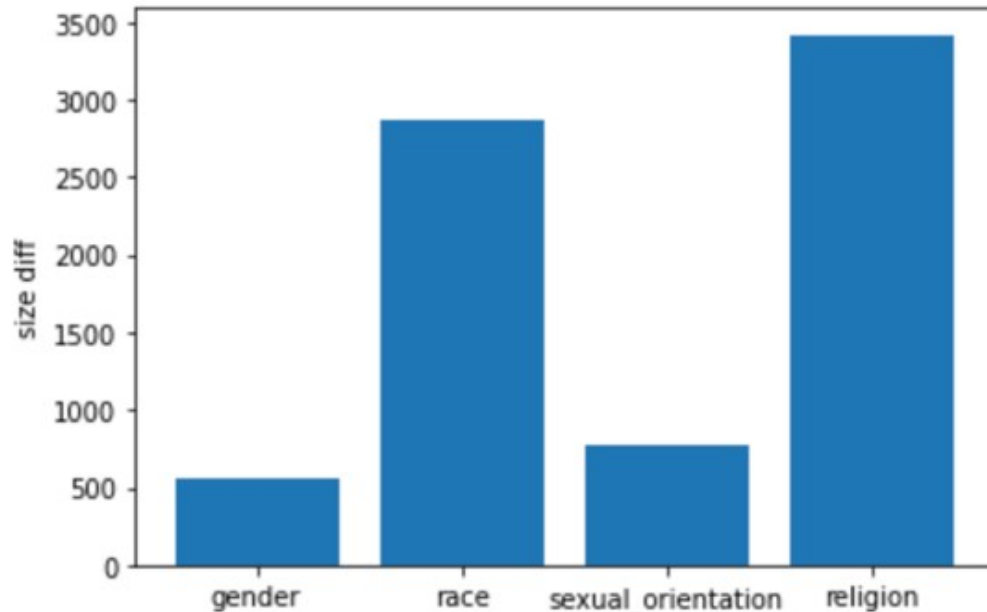
Performance of the different models on the Jigsaw dataset.

Extrinsic Bias Fairness



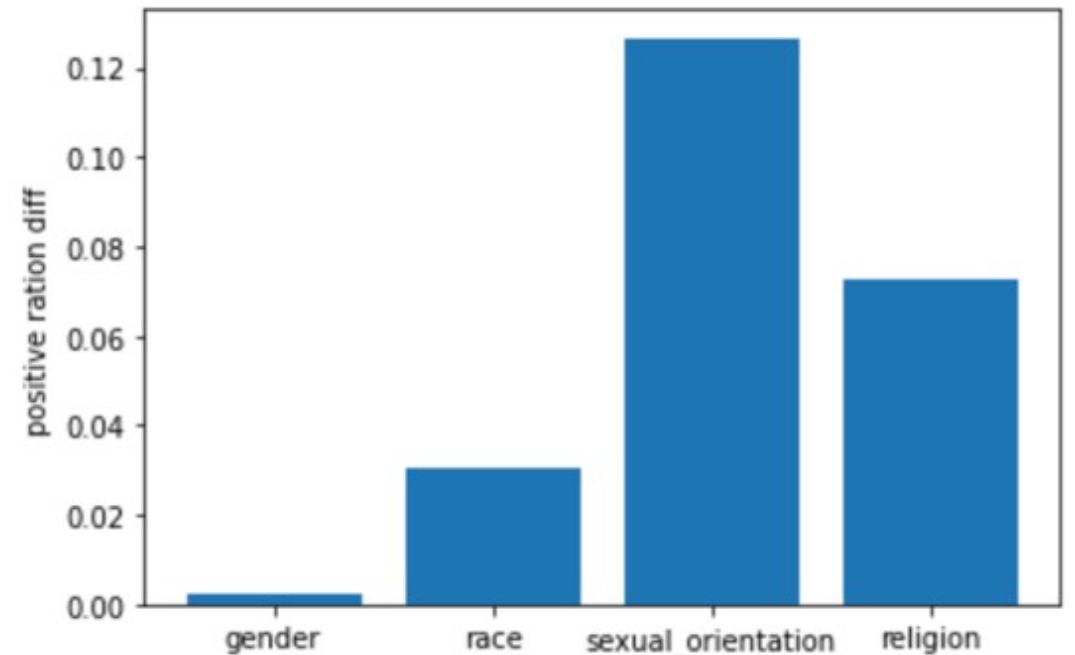
	Marginalize d	Non- marginalize d
Gender	Female and transgender	male
Sexual-orientation	Bi-sexual and gay	heterosexual
Ethnicity	Black, Asian and Latino	White
Religion	Jewish, Muslims, Atheists	Christians

Extrinsic Bias Fairness



The difference in No. comments between the marginalized and non-marginalized groups in each category

Bias in the Dataset



The difference in % of the positive (toxic) comment between the marginalized and non-marginalized groups in each category

Extrinsic Bias Fairness

- Extrinsic bias metrics in the literature:
 - Threshold-based metrics: Equalized odds:
 - $\text{Diff} (\text{FPR} (g), \text{FPR} (g'))$ [5].
 - $\text{Diff} (\text{TPR} (g), \text{TPR}(g'))$ [6].
 - $\text{Max} (\text{FPR_diff}, \text{TPR_diff})$ [7].
 - Threshold-agnostic metrics:
 - AUC (subgroups) [8].
- In this work, I use both threshold bases and agnostic metrics on Toxicity classification and compare their results.

[5] On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextual

[6] Bias in Representation: A Case Study of Semantic Representation bias in High-

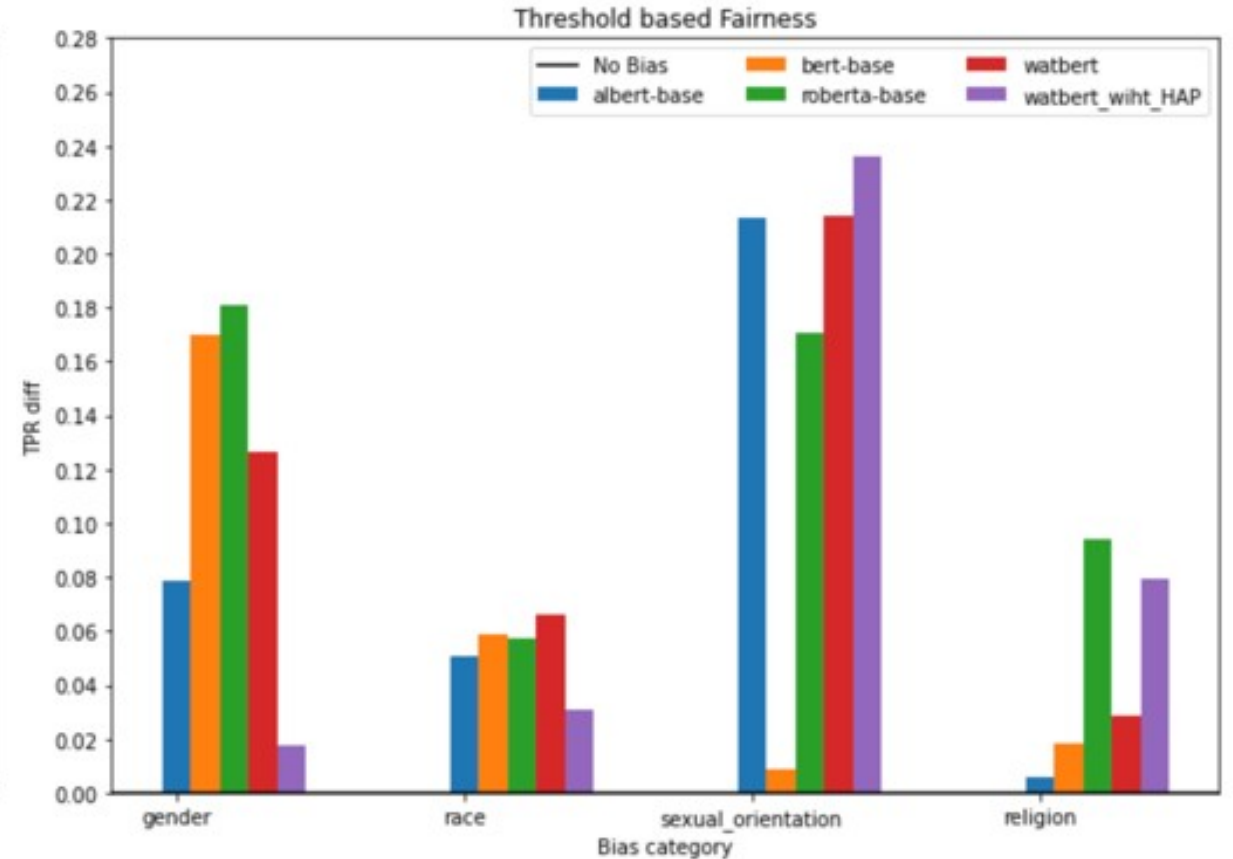
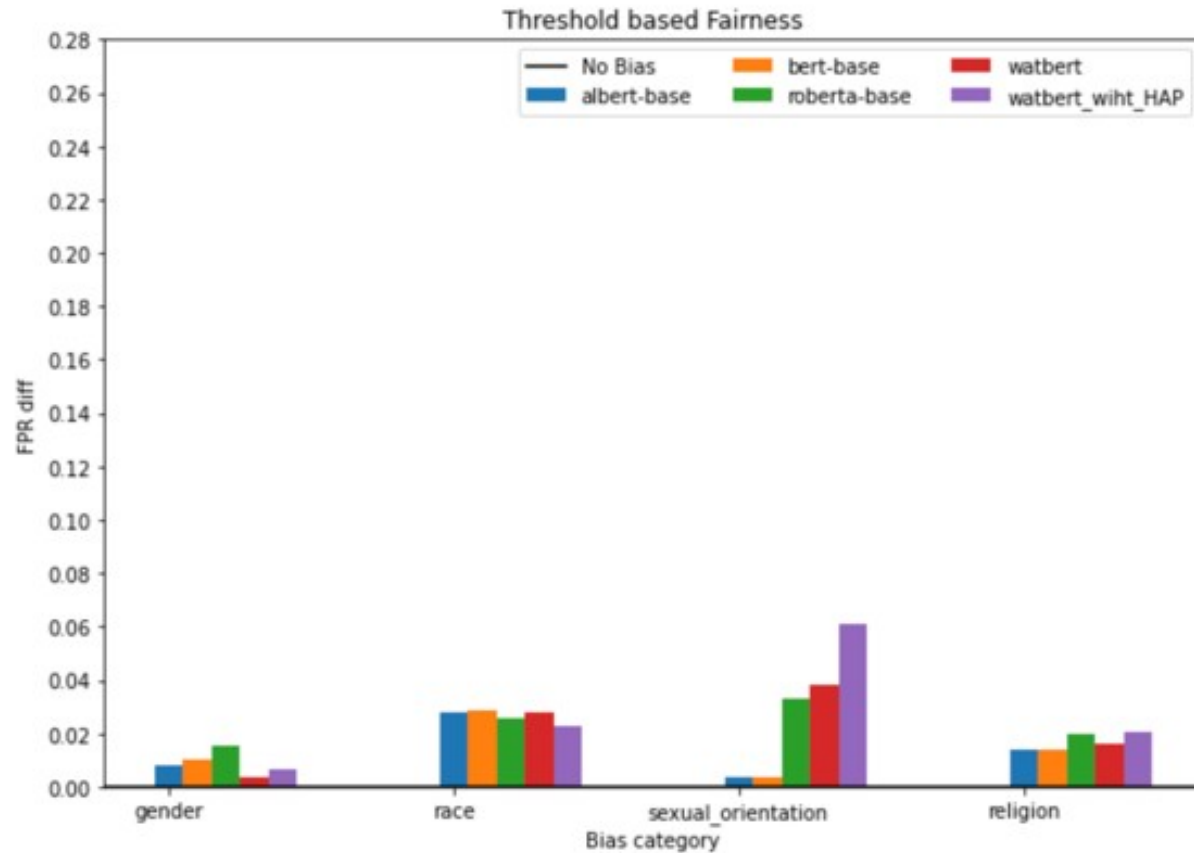
[7] Stake Settings May Vary: Pre-trained Language Model Fairness in Toxic

[8] Classification Metrics for Measuring Unintended Bias with Real Data For Text

Classification

Extrinsic Bias

Fairness (threshold-based metrics)



Extrinsic Bias

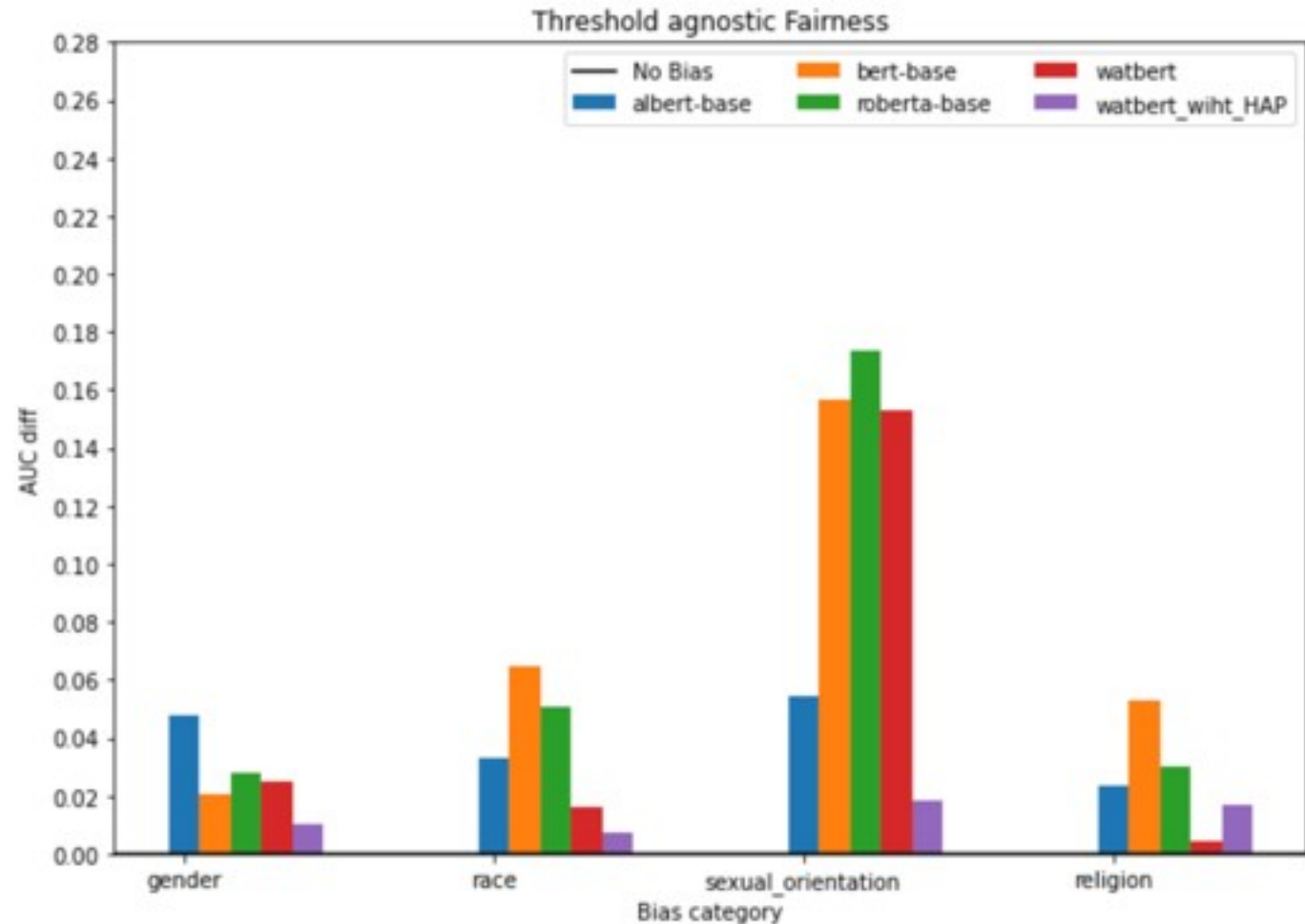
Fairness (threshold-based metrics)

Based on a threshold > 0.5

Albert-base	size	TN	FP	FN	TP	FPR	TPR
Bi-sexual	10	8	0	1	1	0	0.5
homosexual	1644	1179	89	253	123	0.07	0.327
heterosexual	57	50	2	4	1	0.038	0.2
$EO_{tp_albert} = ((0.5 + 0.327) / 2) - 0.2 = 0.213$							
BERT-base	size	TN	FP	FN	TP	FPR	TPR
Bi-sexual	10	8	0	1	1	0	0.5
homosexual	1644	1180	88	270	106	0.06	0.28
heterosexual	57	50	2	3	2	0.038	0.4
$EO_{tp_bert} = ((0.5 + 0.28) / 2) - 0.1 = -0.01$							

Extrinsic Bias Fairness (threshold-agnostic metrics)

$AUC_gap = \text{Diff} (AUC (g), AUC (g'))$



Extrinsic Bias Fairness (threshold-agnostic metrics)

Albert-base	size	AUC
Bi-sexual	10	0.937
homosexual	1644	0.725
heterosexual	57	0.776
BERT-base	size	AUC
Bi-sexual	10	0.625
homosexual	1644	0.746
heterosexual	57	0.842

$$\text{AUC_gap} = ((0.937 + 0.725) / 2) - 0.776 = 0.055$$

$$\text{AUC_gap} = ((0.625 + 0.746) / 2) - 0.842 = -0.156$$

Extrinsic Bias Fairness

- Correlate between Bias in the dataset and the different fairness metrics.
- **Speculation:** Threshold-based metrics (EO) show the bias that the model learned during fine-tuning from the dataset.
- While Threshold-agnostic metrics (AUC-gap) shows the bias in the underlying model.

	EO_FP R	EO_TP R	AUC_g ap
Bias in the dataset	0.611**	0.35	-0.538**

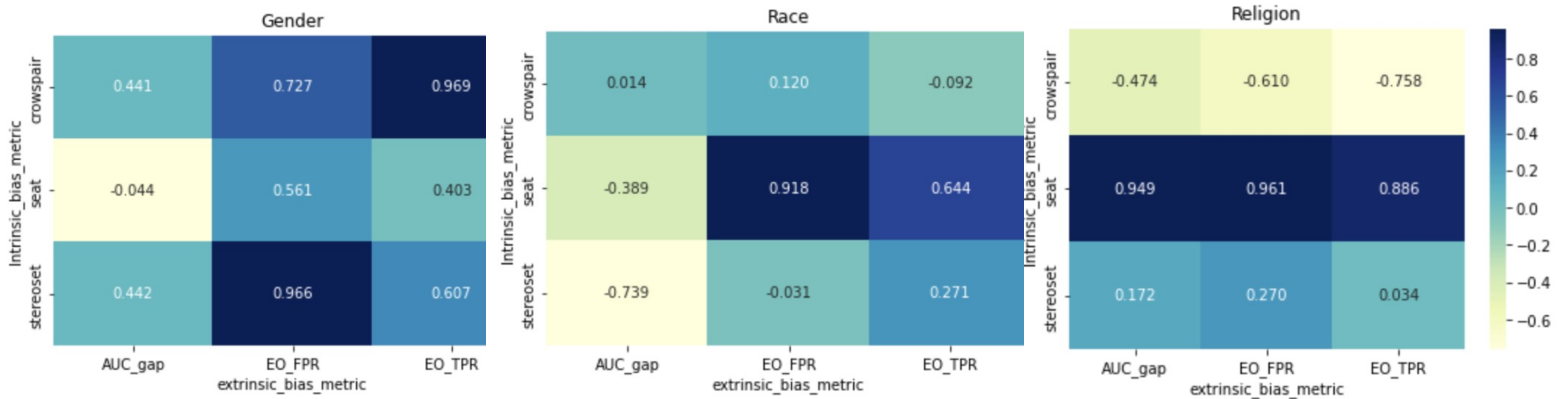
Pearson correlation between the difference in bias in the dataset and the fairness measures.
** is statistically significant.

Should we use threshold-based to measure fairness or threshold-agnostic metrics?

- Threshold-agnostic metrics mitigate the imbalance in the data items that belong to a subgroup.
- Threshold-based metrics correlate positively with the bias in the dataset while threshold-agnostic does not.

It is not straightforward to answer this question but may be it is not always right to use threshold-base metrics and depending on the task and the dataset, we might consider using threshold-agnostic metrics.

Intrinsic vs. Extrinsic Bias



Most of the positive correlations are actually statistically insignificant.

Is there a correlation between intrinsic and extrinsic bias?

- Not consistent positive correlation across all bias types.
- Not statistically significant.
- It is only for this dataset and those metrics.

The results are inconclusive and we'd need to do more experiments with more datasets and other types of downstream tasks to be able draw any conclusion.

Conclusion

What have we learned?

- Published results in the literature on bias and fairness :
 - Not necessarily true, replicate-able or generalizable.
- Different bias metrics don't agree:
 - Best practice would be to use more than one and spot a pattern or go with majority.
- HAP does not indicate social bias:
 - So removing it does not lead to a less biased model.
- Large language models are not more socially biased than base models.
 - Bert and Roberta.
 - xxlarge models could be more biased. More investigation is needed.
- On the correlation between intrinsic and extrinsic bias:
 - Not conclusive.
 - For now, focus on extrinsic bias as it is easier to interpret.

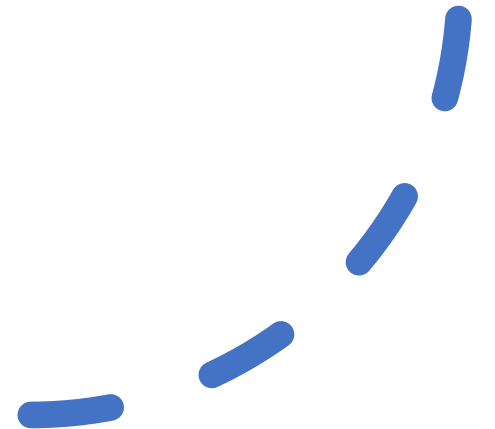
Conclusion


What is next?

- Investigate the effect of debiasing the model on the extrinsic bias:
 - Debiasing the pre-trained models to remove biased representations.
 - Debiasing the fine-tuning dataset.
 - Which is more effective?
- Fairness in Sentiment analysis task:
 - Create a dataset to measure fairness in sentiment analysis tasks.
- Correlation is not causation:
 - Bias in NLP from causal perspective.

Acknowledgement

- Special thanks to :
 - Interns: Katy, Kofy,
 - IBMers: Aashka, Salim, Ioana, and Kush.
 - Bhatta and Hans.





Thanks!
Questions?

Fatma Elsafoury



@fatmaElsafoury

Fatma.elsafoury@IBM.com

Fatma.elsafoury@uws.ac.uk