UNIVERSITY OF THE WEST of SCOTLAND
UWS

🆘Systematic Offensive Stereotyping (SOS) bias in Word Embeddings

Fatma Elsafoury

Agenda

1. Offensive stereotyping bias.
2. Static word embeddings: Measure, Validation, Explanation.
3. Contextual word embeddings.
4. Findings.
5. Limitations.
6. Future work

Bias in NLP

Offensive stereotyping

- Using swear words to describe groups of people aiming at stressing on the inferiority of the identity of that groups [1].
- Since the internet is rife with swear words and slurs, it is important to study how ML models encode this offensive stereotyping.
- **In this work,** we study this offensive stereotyping in static and contextual word embeddings.

Fatma Elsafoury, Steven R. Wilson, Stamos Katsigiannis, and Naeem Ramzan**.**
SOS: Systematic Offensive Stereotyping Bias in Word Embeddings.
COLING ’22**.**

[1] Kukla, Rebecca. "Slurs, interpellation, and ideology." *The Southern Journal of Philosophy* 56 (2018): 7-32.

SOS Bias
Definition

Systematic Offensive Stereotyping (SOS) bias:

“*A systematic* ***association*** *in the word embeddings between* ***profanity*** *and* ***marginalized*** *groups of people”*

1.SOS Bias in Static Word Embeddings

SOS Bias
Measurement

- **Profanity:**
  - A list of 403 swear words.
- **Marginalized groups:**
  - Women, LGBTQ, Non-white-ethnicity.
  - Non-offensive identity words (NOI).
- **Association:**
  - cosine similarity.

Group	Words
LGBTQ*	lesbian, gay, queer, homosexual, lgbt, lqbtq, bisexual, transgender, tran, non-binary
Women*	woman, female, girl, wife, sister, mother, daughter
Non-white ethnicities*	african, african american, black, asian, hispanic, latin, mexican, indian, arab, middle eastern
Straight	heterosexual, cisgender
Men	man, male, boy, son, father, husband, brother
White ethnicities	white, caucasian, european american, european, norwegian, canadian, german, australian, english, french, american, swedish, dutch

*Marginalised group

Table 1: NOI words

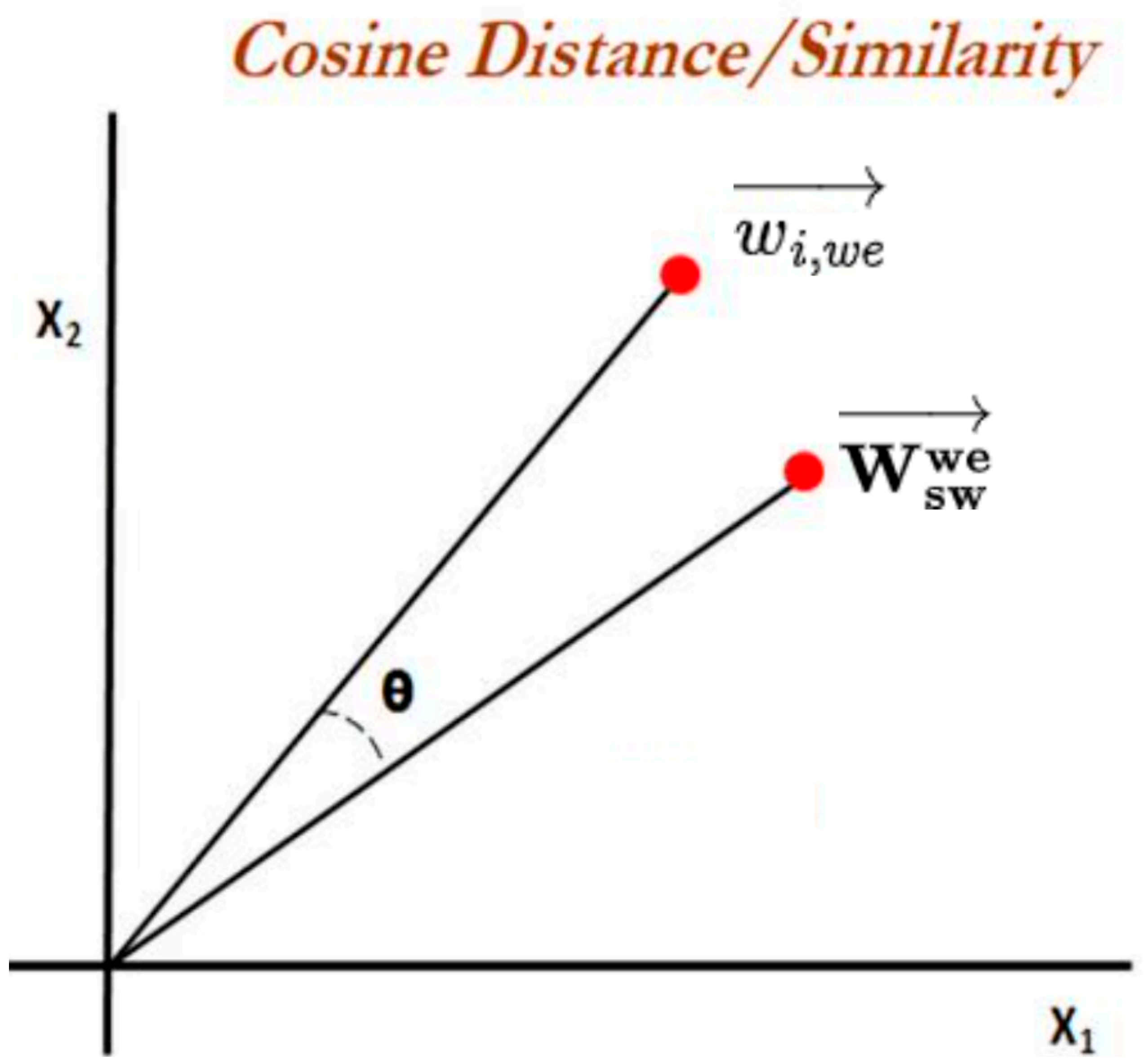
SOS Bias

Measurement

Normalised Cosine Similarity to Profanity (NCSP)

- we is a word embeddings model, e.g. W2V.
- $\overrightarrow{\mathbf{W}_{\mathbf{sw}}^{\mathbf{we}}}$ is the average of swear words for a word embedding (we).
- $\overrightarrow{w_{i,we}}$ is the word vector of the NOI word i for the word embeddings (we).
- Min-max normalization for a SOS scores in [0, 1].

$$SOS_{i,we} = \frac{\overrightarrow{\mathbf{W}_{\mathbf{sw}}^{\mathbf{we}}} \cdot \overrightarrow{w_{i,we}}}{||\overrightarrow{\mathbf{W}_{\mathbf{sw}}^{\mathbf{we}}}|| \cdot ||\overrightarrow{w_{i,we}}||}$$



SOS Bias
Word Embeddings

- 15 word embeddings.
- **Models:** Skip-gram, Glove, FastText.
- **Data:** Social media data, Wikipedia, google news, and common crawls.
- 3 de-biased word embeddings (gender bias removed).

Model	Dimensions	Trained on
W2V	300	100B words from Google News
Glove-WK	200	6B tokens from Wikipedia 2014 and Gigaword
Glove-Twitter	200	27B tokens collected from two billion Tweets
UD	300	200M tokens collected from the Urban Dictionary website
Chan	150	30M messages from the 4chan and 8chan websites
Glove-CC	300	42B tokens from Wikipedia 2014 and Gigaword
Glove-CC-large	300	840B tokens from Wikipedia 2014 and Gigaword
FastText-CC	300	600B common crawl tokens
FT-CC-sws	300	600B common crawl tokens with subwords information
FT-Wiki	300	16B tokens collected from Wikipedia 2017, UMBC, and statmt.org news dataset
FT-wiki-sws	300	16 billion tokens with subwords information collected from the Wikipedia 2017, UMBC, and statmt.org
SSWE	50	10M comments collected from Twitter
Debias-W2V	300	W2V model after the gender bias has been removed using the hard debiasing method
P-DeSIP	300	Debiased Glove-WK with the potential proxy gender bias removed.
U-DeSIP	300	Debiased Glove-WK word embeddings with the unresolved gender bias removed.

Table 2: examined word embeddings in our work

SOS Bias

Bias in word embeddings

In 14 out of the 15 word embeddings, there is higher SOS bias against marginalised groups

Gender Bias removed

Word embeddings	Mean SOS							
	Gender		Sexual orientation		Ethnicity		Marginalised vs. Non-marginalised	
	Women	Men	LGBTQ	Straight	Non-white	White	Marginalised	Non-marginalised
W2V	**0.293**	0.209	0.475	**0.5**	**0.456**	0.390	**0.418**	0.340
Glove-WK	**0.435**	0.347	**0.669**	0.5	**0.234**	0.169	**0.464**	0.260
Glove-Twitter	**0.679**	0.447	**0.454**	0*	**0.464**	0.398	**0.520**	0.376
UD	**0.509**	0.436	**0.582**	0.361	**0.282**	0.244	**0.466**	0.319
Chan	**0.880**	0.699	**0.616**	0.414	**0.326**	0.176	**0.597**	0.373
Glove-CC	**0.567**	0.462	**0.480**	0.195	**0.446**	0.291	**0.493**	0.339
Glove-CC-large	**0.318**	0.192	**0.472**	0.302	**0.548**	0.278	**0.453**	0.252
FT-CC	**0.284**	0.215	0.503	**0.542**	**0.494**	0.311	**0.439**	0.301
FT-CC-sws	**0.473**	0.422	**0.445**	0.277	**0.531**	0.379	**0.480**	0.384
FT-Wiki	**0.528**	0.483	0.555	**0.762**	**0.393**	0.265	**0.496**	0.385
FT-Wiki-sws	0.684	0.684	0.656	**0.798**	0.555	**0.579**	0.632	**0.635**
SSWE	0.619	**0.651**	**0.438**	0*	**0.688**	0.560	**0.569**	0.537
Debias-W2V	**0.205**	0.204	0.446	**0.5**	**0.471**	0.420	**0.386**	0.356
P-DeSIP	**0.266**	0.220	**0.615**	0.491	**0.354**	0.314	**0.434**	0.299
U-DeSIP	**0.266**	0.220	**0.616**	0.492	**0.343**	0.299	**0.431**	0.283

*Glove-Twitter and SSWE did not include the NOI words that describe the "Straight" group.

Table 3: Mean SOS scores of the different groups for all the word embeddings.

SOS Bias

Bias in word embeddings

Some word embeddings are more SOS biased against certain groups

Word embeddings	Mean SOS		
	Women	LGBTQ	Non-white
W2V	0.293	**0.475**	0.456
Glove-WK	0.435	**0.669**	0.234
glove-twitter	**0.679**	0.454	0.464
UD	0.509	**0.582**	0.282
Chan	**0.880**	0.616	0.326
Glove-CC	**0.567**	0.480	0.446
Glove-CC-large	0.318	0.472	**0.548**
FT-CC	0.284	**0.503**	0.494
FT-CC-sws	0.473	0.445	**0.531**
FT-WK	0.528	**0.555**	0.393
FT-WK-sws	**0.684**	0.656	0.555
SSWE	0.619	0.438	**0.688**
Debias-W2V	0.205	0.446	**0.471**
P-DeSIP	0.266	**0.615**	0.354
U-DeSIP	0.266	**0.616**	0.343

Most biased against LGBTQ

Most biased against women

Most biased against Non-white ethnicity

Table 4: Mean SOS scores of the different groups for all the word embeddings.

SOS Bias

Bias in word embeddings

- **Social bias:** Gender and Racial bias
- **Metrics:** WEAT[1], RND[2], RNSB[3], and ECT[4]

SOS bias reveals different information from the ones revealed by social bias

SOS bias vs. Social bias

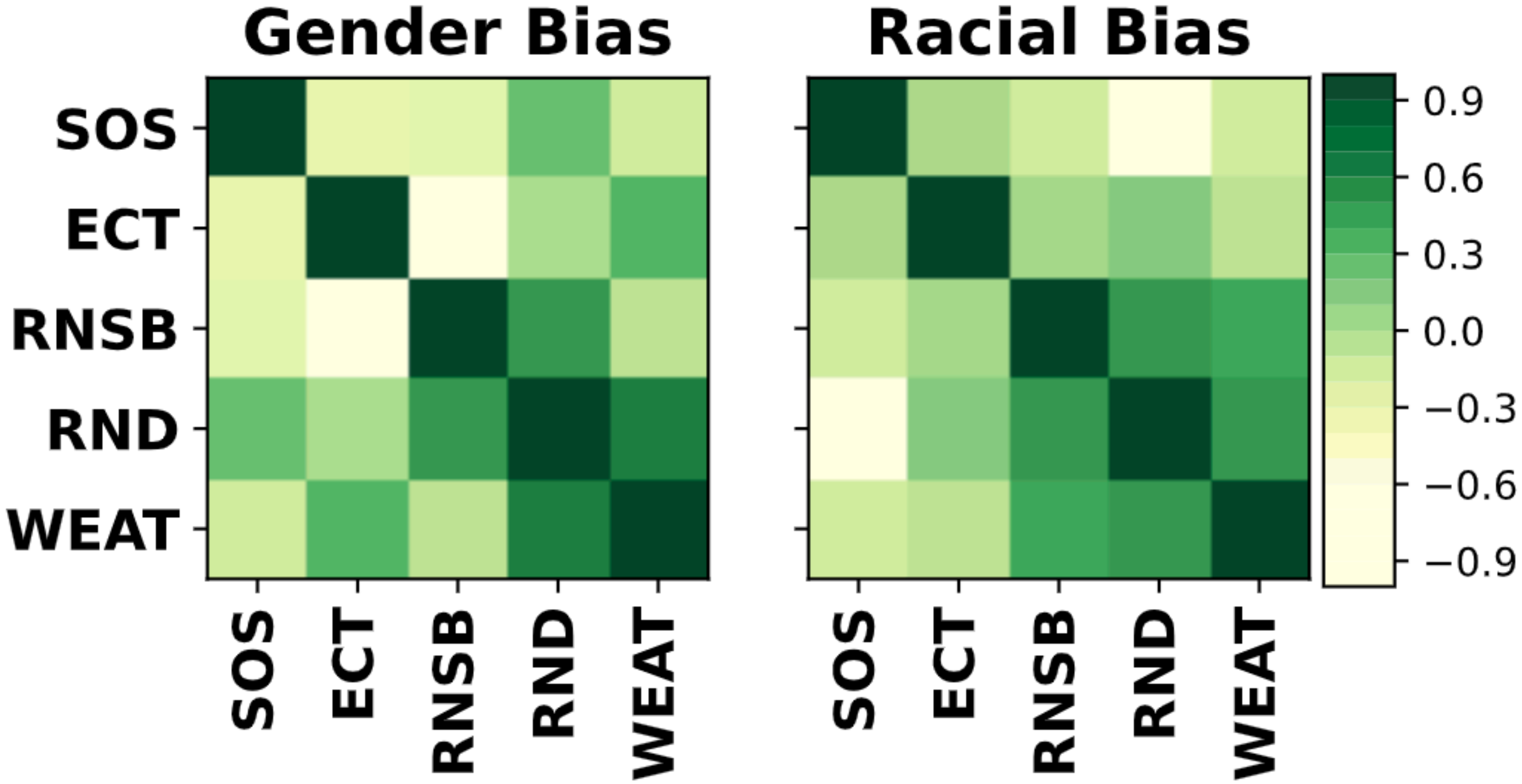


Figure 1: Spearman's correlation

[1] Caliskan, Aylin and Bryson, Joanna J. and Narayanan, Arvind "Semantics derived automatically from language corpora contain human-like biases".
[2] Garg, Nikhil and Schiebinger, Londa and Jurafsky, Dan and Zou, James "Word embeddings quantify 100 years of gender and ethnic stereotypes".
[3] Sweeney, Chris and Najafian, Maryam "A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings".
[4] Kamalika Chaudhuri and Masashi Sugiyama "Attenuating Bias in Word vectors"

SOS Bias
Validation

1. SOS bias and online hate [1].

2. Our proposed method **(NCSP)** versus other bias metrics **(WEAT, RND,RNSB, ECT)** to measure the SOS bias.

[1] Hawdon, James, Atte Oksanen, and Pekka Räsänen. "Online Extremism and Online Hate." *NORDICOM* (2015): 29.

SOS Bias

SOS bias vs. Online hate statistics

According to the online hate stats:

- LGBTQ (61%).
- Non-White ethnicity (60%).
- Women (44%).

Country	Sample size	Ethnicity	LGBTQ	Women
Finland	555	0.67	0.63	0.25
US	1033	0.6	0.61	0.44
Germany	978	0.48	0.5	0.2
UK	999	0.57	0.55	0.44

Table 5: The percentage of examined groups that experience online hate in different countries [1].

The expected pattern of positive correlation is:

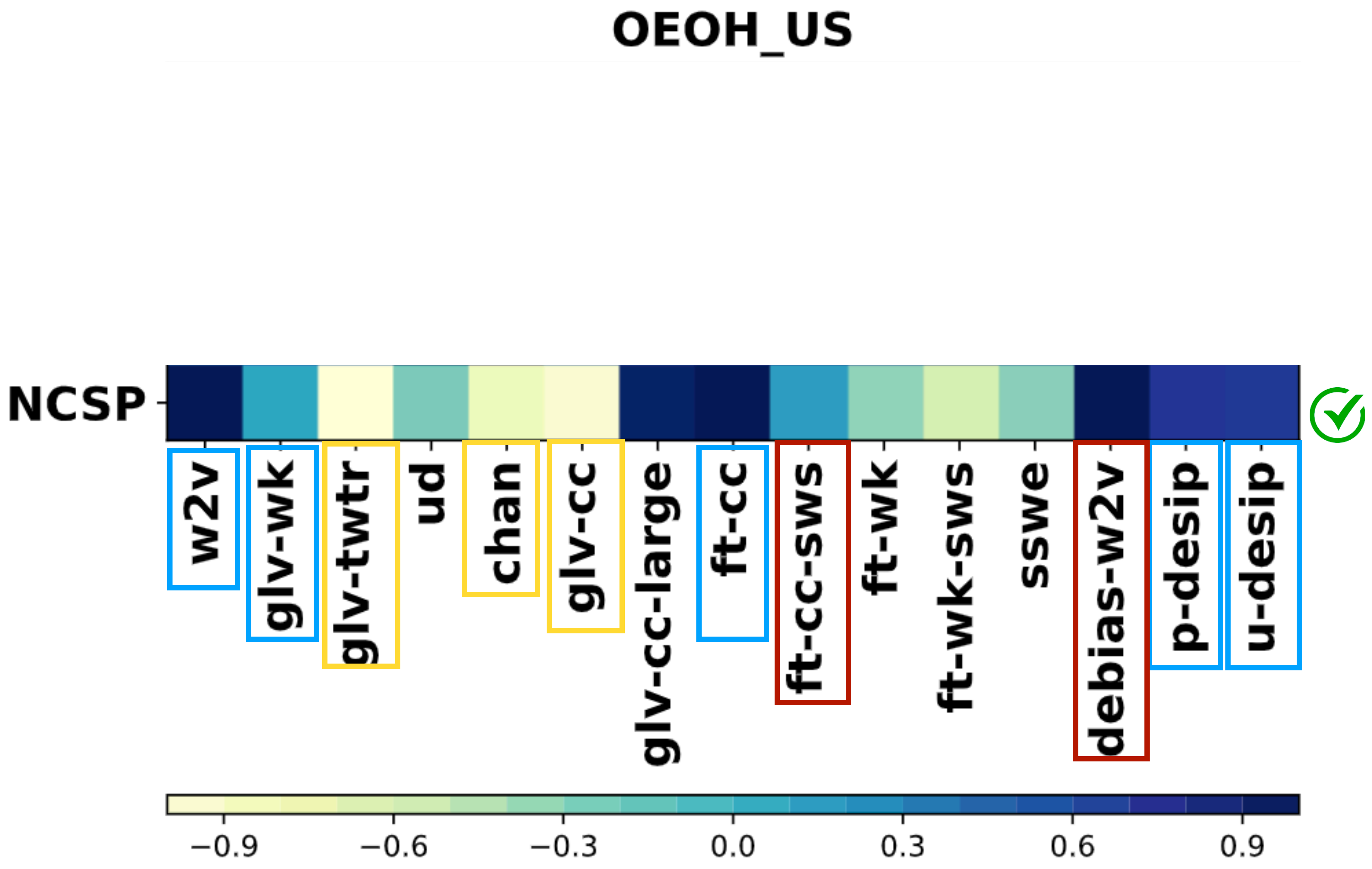
- The word embeddings most biased against LGBTQ and Non-White ethnicities correlate positively.
- The word embeddings most biased against women correlates negatively.

[1] Hawdon, James, Atte Oksanen, and Pekka Räsänen. "Online Extremism and Online Hate." *NORDICOM* (2015): 29.

SOS Bias

SOS bias vs. Online hate statistics

SOS bias scores are representative of the online hate experienced by marginalised groups.



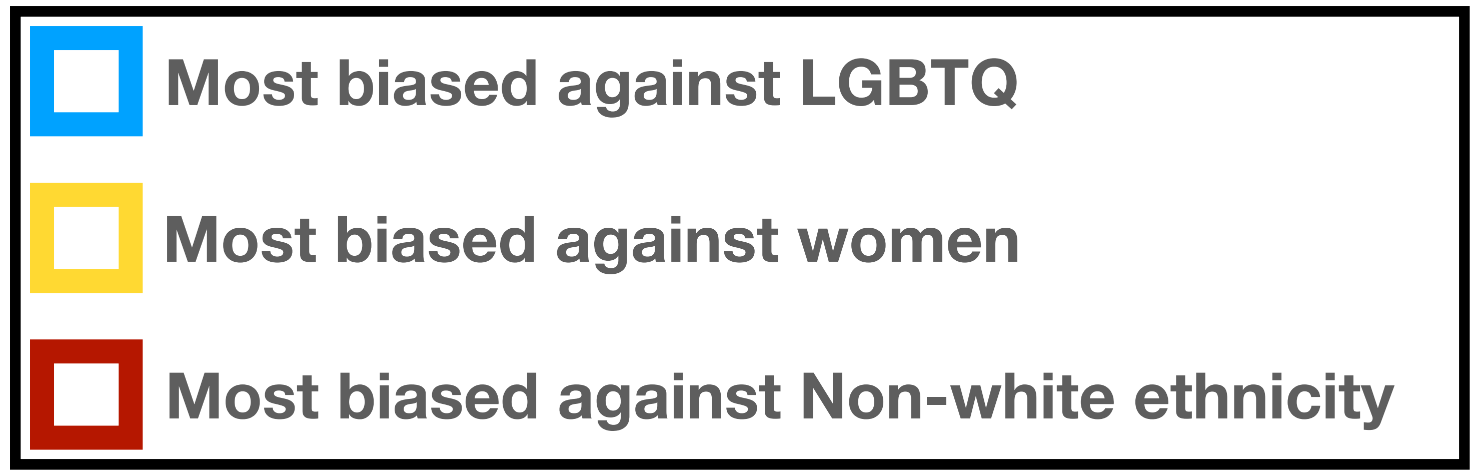


Figure 2: Pearson's correlation between SOS bias scores and published stats on online hate.

SOS Bias

SOS bias vs. Online hate statistics

Our SOS bias metric (NCSP) is the most reflective of the SOS bias in the different word embeddings

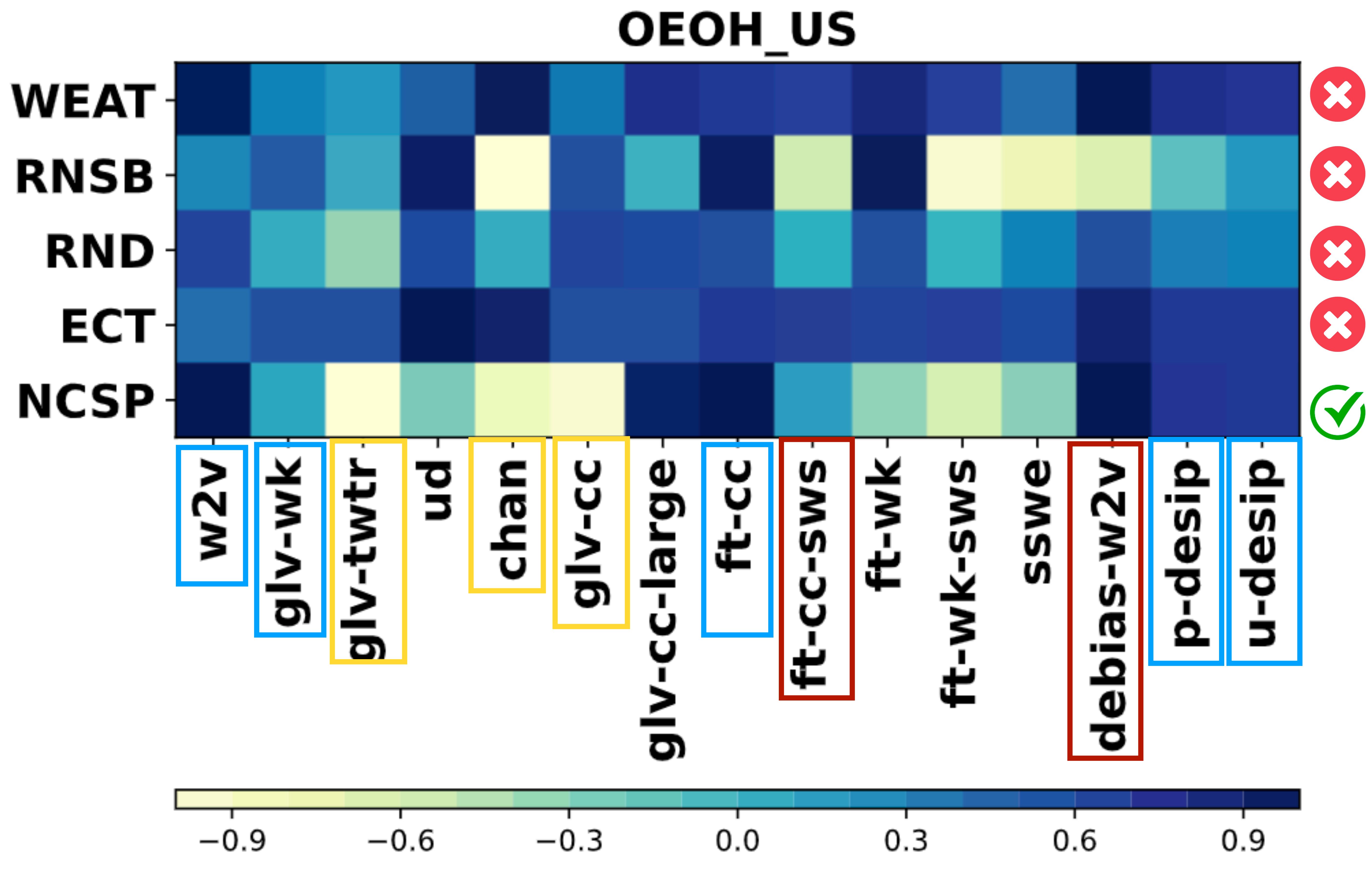


Figure 2: Pearson's correlation between SOS bias scores and published stats on online hate.

- Most biased against LGBTQ
- Most biased against women
- Most biased against Non-white ethnicity

SOS Bias

Does it explain the performance of hate speech detection models?

- MLP and Bi-LSTM models + frozen embeddings layer.
- 4 Hate speech datasets.
- Correlate SOS bias scores with F1 scores

No! SOS bias does not explain the performance of Hate speech detection models.

Dataset	Model	WEAT	RNSB	RND	ECT	NCSP
HateEval	MLP	**0.277**	0.223	-0.100	0.019	0.230
	BiLSTM	0.377	**0.540***	0.094	-0.030	0.100
Twitter Sexism	MLP	**0.157**	0.030	-0.216	-0.039	0.121
	BiLSTM	0.109	**0.266**	0.093	-0.361	0.246
Twitter Racism	MLP	0.042	0.017	-0.336	-0.223	**0.241**
	BiLSTM	-0.264	**0.135**	-0.210	-0.103	0.110
Twitter Hate	MLP	0.107	0.218	-0.164	-0.148	**0.223**
	BiLSTM	**0.507**	0.475	0.289	-0.217	0.396

*Statistically significant at $p < 0.05$.

Table 6: Pearson's correlation coefficient of the SOS bias scores measured using different metrics and the F1 scores of the model

2.SOS Bias in Contextual Word Embeddings

Bias in LM

Measurement

	CrowS-Pairs [1]
Data	Human generated stereotyped vs non-stereotyped sentences
Task	MLM
e.g.	Score (S) = P(is | she) + P(a| she) + P(nurse| she) Score (S') =P(is | he) + P(a| he) + P(nurse| he)
Bias type	9 types

Table 7: Used intrinsic bias metrics

[1] Crows-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

Bias in LM

Measurement

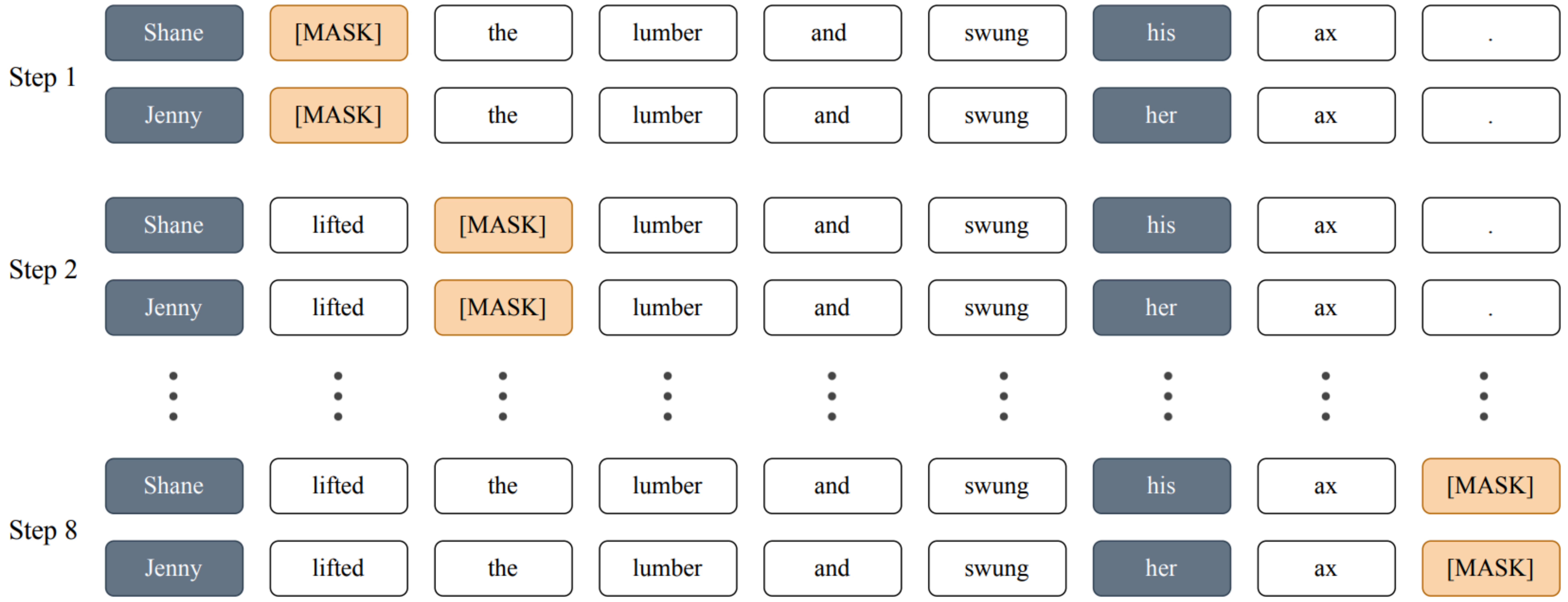


Figure 3: Crows-Pairs Example [1]

[1] Crows-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

Bias in LM

Measurement

$$Score(S) = \sum_{i}^{C} logP(u_i | M, \theta)$$

u_i is an unmodified token, non-identity words, i where $u_i \in U$

M is the modified tokens which are words that describe an identity group.

S is the sentence where $S = U \cup M$. It could be a sterotypical or non-stereotypical sentence.

[1] Crows-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

Bias in LM

Measurement

$$\text{Bias score} = \frac{Count(Score(S) > Score(S'))}{N}$$

S is the sterotpical sentence
S' is the non-sterotpical sentence
N is the Number of stereotpical sentences

Score	Meaning
0.5	No bias.
> 0.5	The model is biased towards the stereotype
< 0.5	The model is biased against the stereotype

SOS Bias

Definition

Systematic Offensive Stereotyping (SOS) bias:

“*A systematic* ***association*** *in the word embeddings*

between ***profanity*** *and* ***marginalized*** *groups of people”*

SOS Bias in LM
Measurement

- **Profanity vs niceness:**
  - A list of 21 swear words
  - A list 21 nice words.
- **Marginalized groups:**
  - Women, LGBTQ, Non-white-ethnicity, disability, religion, occupation.
  - Non-offensive identity words (NOI).
- **Association:**
  - Log-liklihood MLM.

SOS Bias in LM

Measurement

Create synthesised data:

Template1 : "You are ……"

Template2 : "I……"

Template3 : "……"

$\Bigg\}$ + **Nice Word, Verb** **Or** **Profane Word, hateful Verb** + **Identity group**

SOS Bias in LM

Measurement

Create synthesised data: Example:

Template1

Profane sentence: You are a dumb Arab

Nice sentence: You are a nice Arab

Template2

Profane sentence: I hate Arab

Nice sentence: I love Arab

Template3

Profane sentence: Kill Arab

Nice sentence: Hug Arab

SOS Bias in LM

Measurement

Create synthesised data:

- **21 Template** with nice and profane words.
- **84 marginalised and non-marginalised identities** gender, racial, religion, sexual-orientation, ablism, social class(occupations).
- **1764 nice and profane** sentences to measure the SOS bias.

SOS Bias in LM

Measurement

	Crows-Pairs to measure SOS bias
Data	Templated sentences: profane vs nice sentences
Task	MLM
e.g.	Score (S) = P(you | dumb) + P(are| dumb) + P(a| dumb) + P(arab| dumb) Score (S') = P(you | nice) + P(are| nice) + P(a| nice) + P(arab| nice)
Bias type	SOS bias for 6 sensitive attributes.

Table 7: SOS intrinsic bias metrics in LM

SOS Bias in LM

Measurement

$$Score(S) = \sum_{i}^{C} logP(u_i | M, \theta)$$

$$\text{SOS Bias score} = \frac{Count(Score(S) > Score(S'))}{N}$$

S is the profane sentence
S' is the nice sentence
N is the Number of profane sentences

SOS Bias in LM

Measurement

SOS Bias Score	Meaning
0.5	No bias.
> 0.5	The model associates **profanity** with the **identity group** present in the sentence
< 0.5	The model associates **niceness** with the **identity group** present in the sentence.

SOS Bias in LM

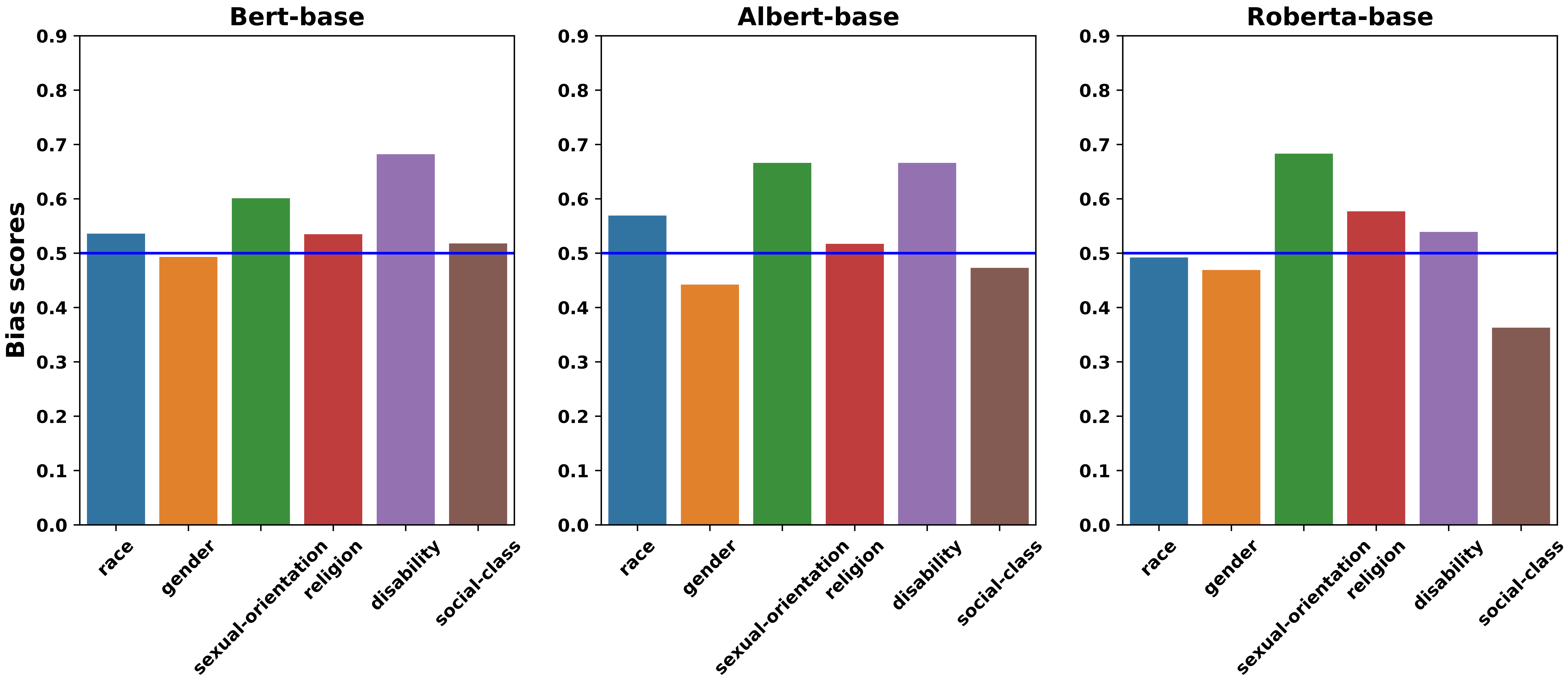
LM Models

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

Table 8: Used Language models

SOS Bias in LM

Scores

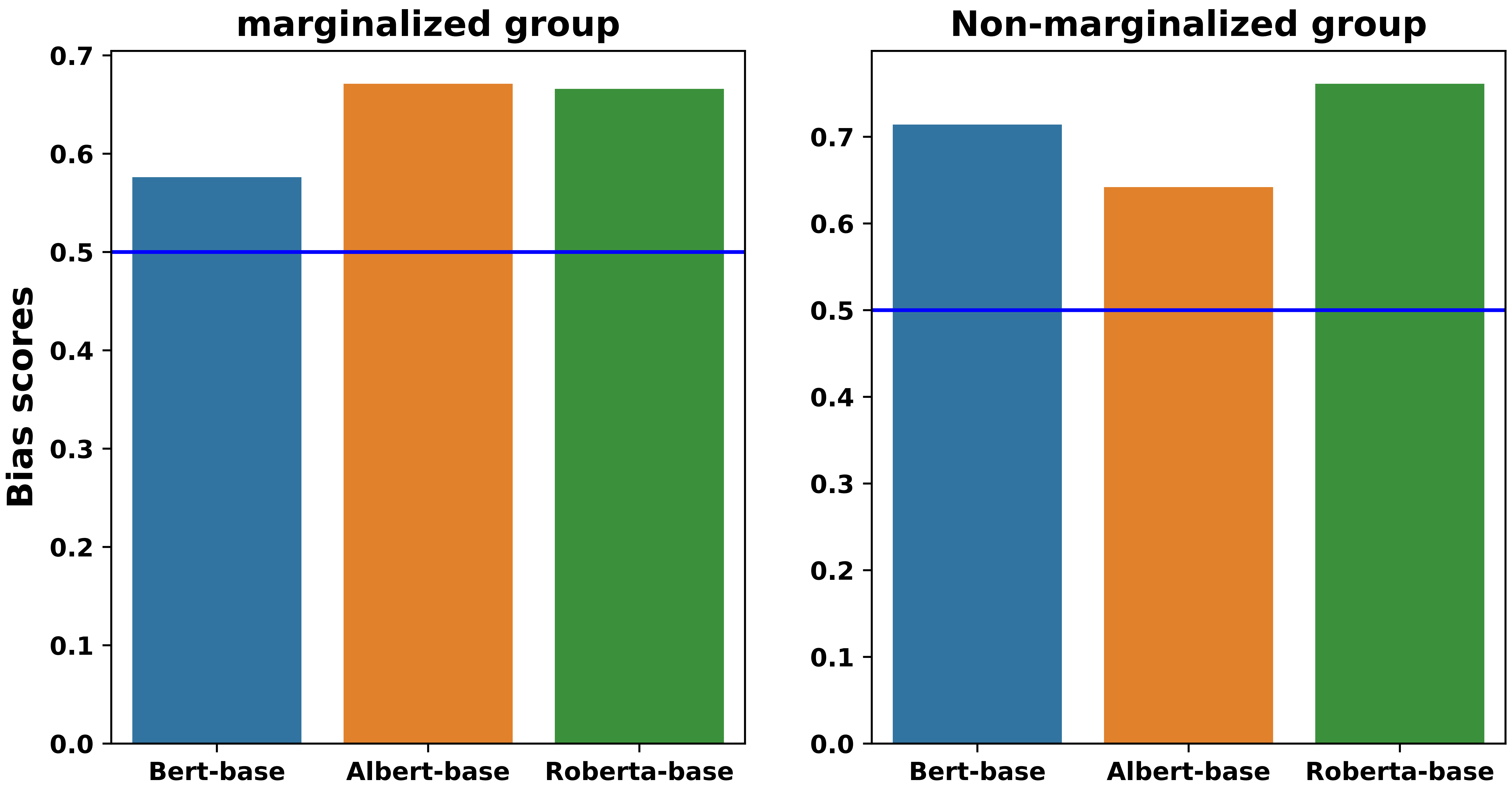


The attributes that received the most bias are: Sexual orientation, Disability, Race, and Religion

SOS Bias in LM

Scores

Homosexuals



Heterosexuals

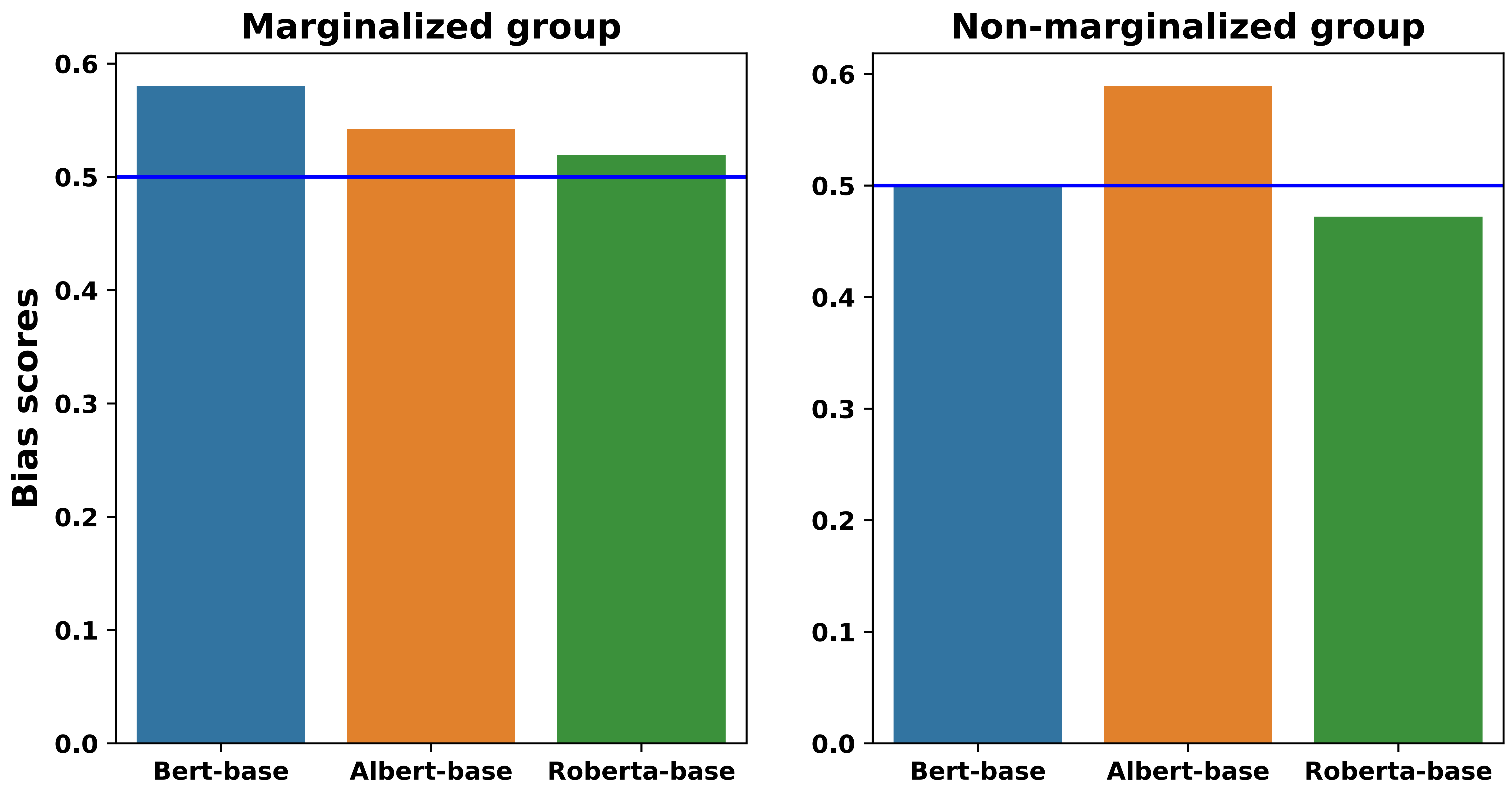
Sensitive attribute: Sexual orientation

High SOS bias scores towards both marginalised and non-marginalised

SOS Bias in LM

Scores

Non-white



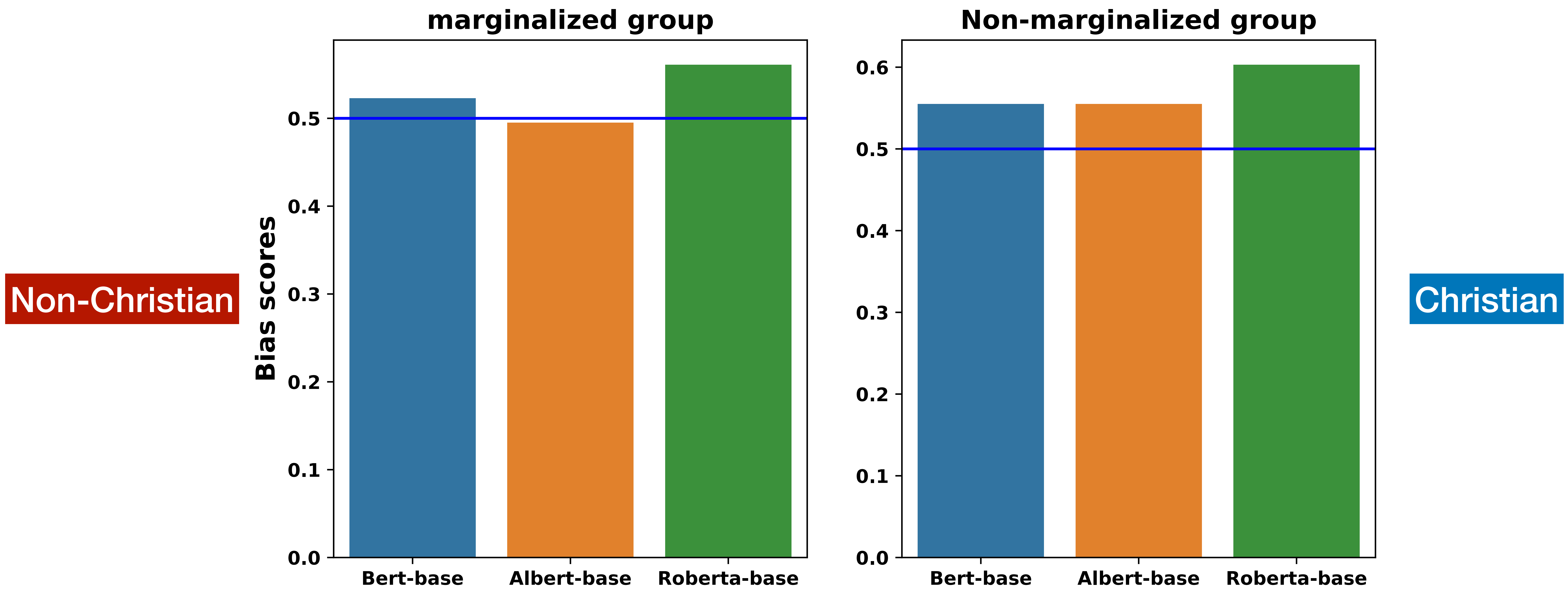
White

Sensitive attribute: Race

Higher SOS bias scores towards marginalised

SOS Bias in LM

Scores



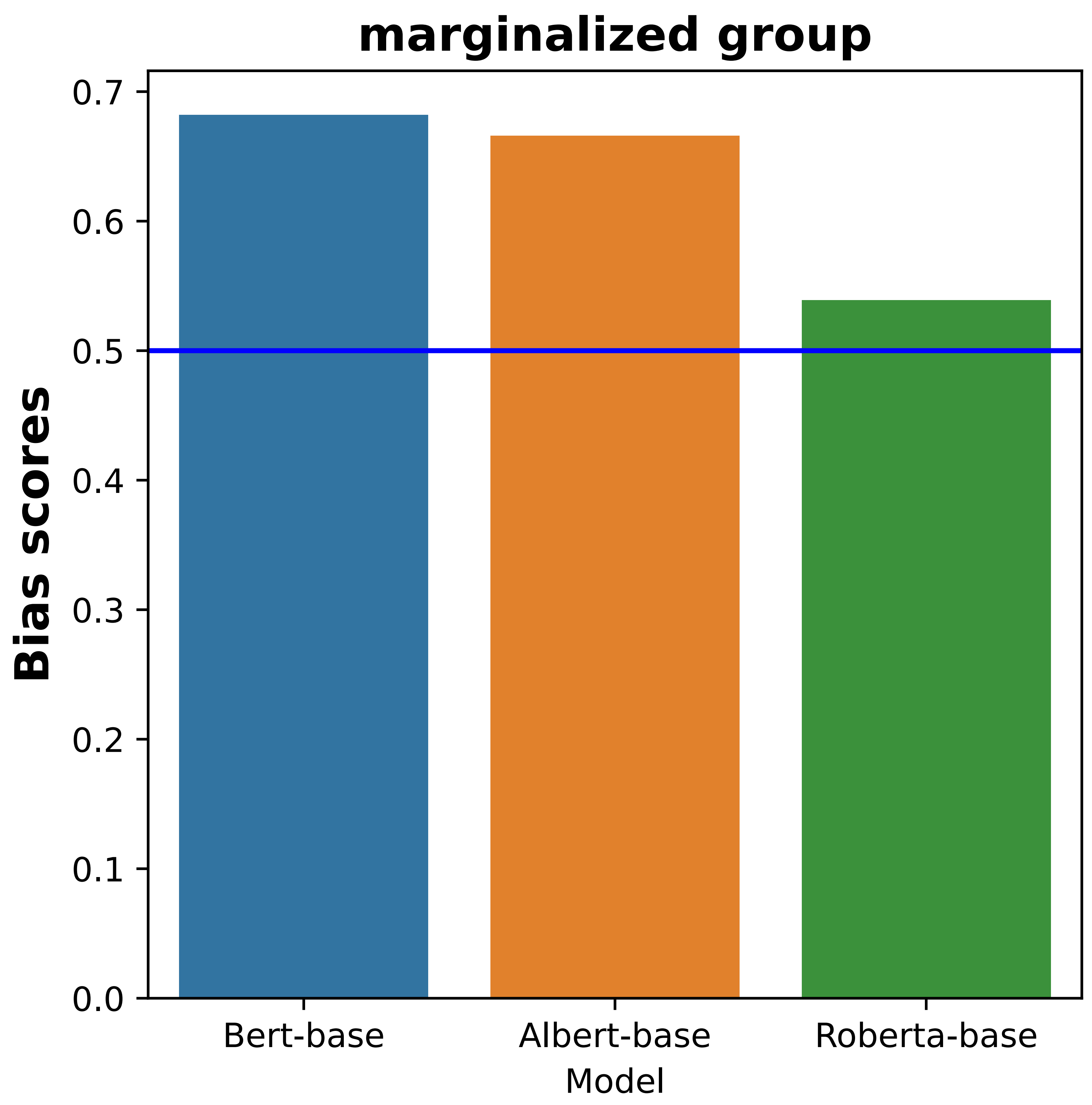
Sensitive attribute: Religion

Higher SOS bias scores towards non-marginalised

SOS Bias in LM

Scores

Deaf, blind, disabled



Sensitive attribute: Disability

How to describe abled people?

SOS Bias in LM

Does it explain the performance of hate speech detection models?

- Fine-tune Albert-base, Bert-base, Roberta-base
- Jigsaw-toxicity dataset: 400K, 40% training, 30% validation and 30% test.
- Correlate mean SOS bias for all sensitive attributes and scores with F1 scores

Models	F1-scores
Bert-base-uncased	0.582
Roberta-base	0.589
Albert-base	0.558

Table 9: Performance on hate speech detection.

No! SOS bias does not explain the performance of hate speech detection models.

SOS Bias
Findings

1. There is **SOS bias** in **Static** and **contextual word embeddings.**

2. **SOS bias** is **higher** towards **marginalised groups** (Women, LGBTQ, and Non-white-ethnicity) in most of the examined **static word embeddings** but **not Contextual word embedding**.

3. The SOS bias **is reflective** of the **online hate** that marginalised groups of people experience in **static word embeddings**.

4. **SOS bias does not explains** the performance of the different word embeddings Static or contextual on **hate speech detection**. However, That could be because of **other biases** in the hate speech **datasets**.

SOS Bias

Limitations

1. Our proposed metrics are limited to the English language and the bias from a Western perspective.

2. The proposed SOS bias metrics measures the existence of bias not its absence. Low scores don not mean the model is unbiased.

3. The use of template sentences do not provide real context.

4. Using the log-likelihood with MLM task to measure bias gives different scores between Transformers 3 and 4.

5. Measuring intrinsic bias is important but at the moment our tools to measure it are not reliable.

SOS Bias
What is Next

1. Measure Fairness in downstream tasks.
2. Investigate the impact of different sources of bias on the downstream fairness.
3. Investigate the impact of different debasing methods on the downstream fairness.

SOS Bias
Future Work

- Studying Bias and fairness from a non-Western perspective:
    1. Language.
    2. Culture.

Thanks!

Questions?

Fatma Elsafoury

@FatmaElsafoury
fatma.elsafoury@uws.ac.uk
UNIVERSITY OF THE WEST of SCOTLAND UWS