SOS:Systematic Offensive Stereotyping Bias in Word Embeddings

Fatma Elsafoury, Steven R. Wilson, Stamos Katsigiannis, and Naeem Ramzan

Introduction

Bias in NLP

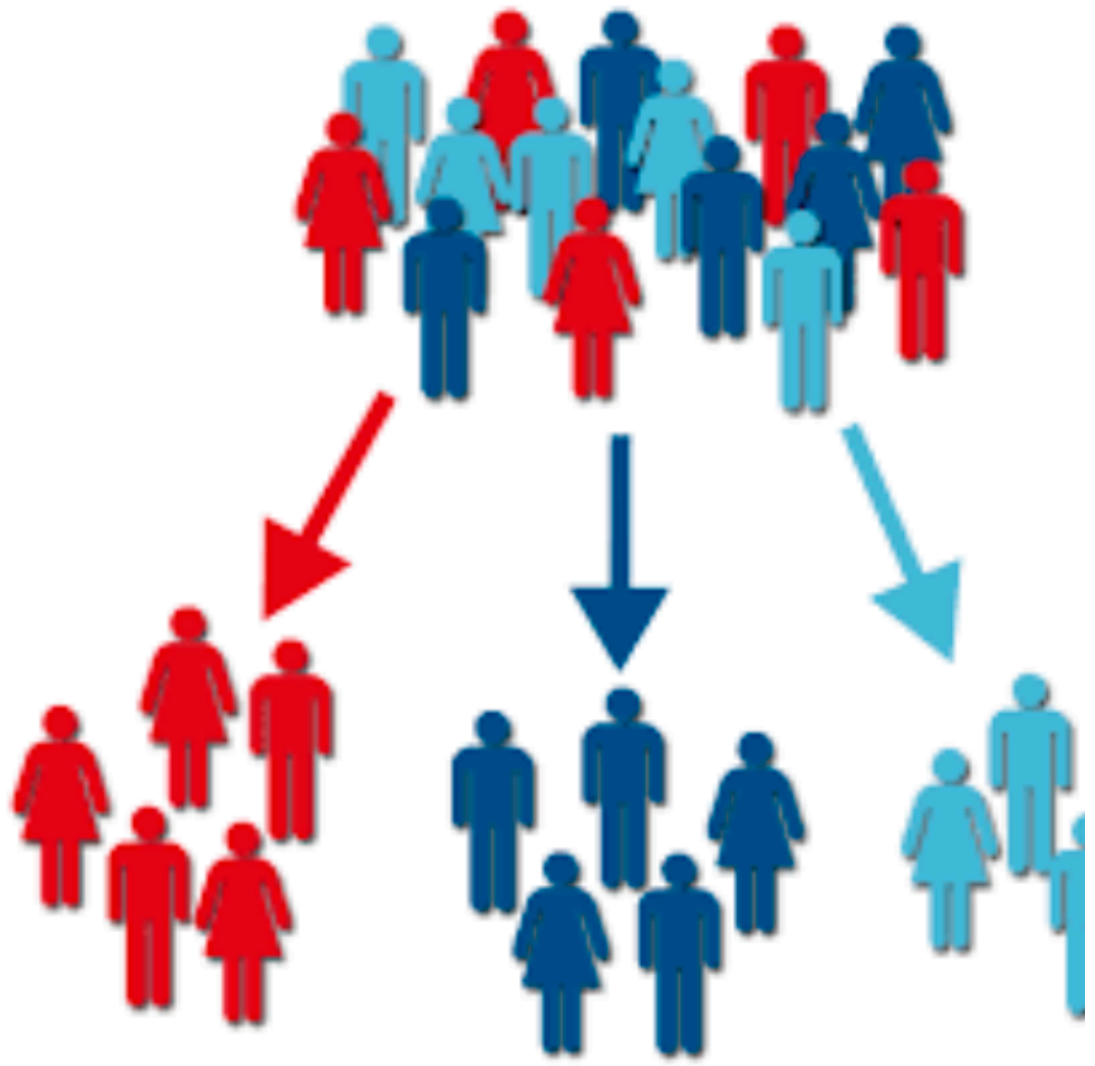
- In 2021, Claudia Wagner et al., define the term **Algorithmically infused societies** as "*The societies that are shaped by algorithmic and human behaviour*", such as social media platforms [1].
- The data collected from those societies, is biased [2].
- Unsupervised word embedding models encode these biases during training [3].

[1] Wagner, C., Strohmaier, M., Olteanu, A. *et al.* Measuring algorithmically infused societies. *Nature* **595**, 197–204 (2021). https://doi.org/10.1038/s41586-021-03666-1
[2] Olteanu A, Castillo C, Diaz F, Kıcıman E. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front Big Data. 2019 Jul 11;2:13. doi: 10.3389/fdata.2019.00013. PMID: 33693336; PMCID: PMC7931947.
[3] Brunet, Marc-Etienne, et al. "Understanding the origins of bias in word embeddings." *International conference on machine learning*. PMLR, 2019.

Introduction

Social Bias in NLP

- Most studied in the literature of bias in NLP.
- To group people in pre-defined groups based certain characteristics e.g., gender bias and racial bias [4].
- Metrics used to measure social bias static word embeddings include:
  - WEAT[5], RNSB[6], RND[7], and ECT[8].



[4] The End of Bias, Nordell 2021.
[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan.Semantics derived automatically from language corpora contain human-like biases.Science, 356(6334):183–186, 2017
[6]Chris Sweeney and Maryam Najafian. A transparent framework for evaluatingunintended demographic bias in word embeddings.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1662–1667, 2019.
[7]Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115.16 (2018): E3635-E3644.
[8]Dev, Sunipa, and Jeff Phillips. "Attenuating bias in word vectors." The 22nd international conference on artificial intelligence and statistics. PMLR, 2019.

Introduction

Research Problem

- Using swear words to describe groups of people aiming at stressing on the inferiority of the identity of that groups [9].
- 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 word embeddings.

[9] 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”*

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 ethnic-ities	white, caucasian, european american, european, nor-wegian, canadian, german, australian, english, french, american, swedish, dutch

*Marginalised group

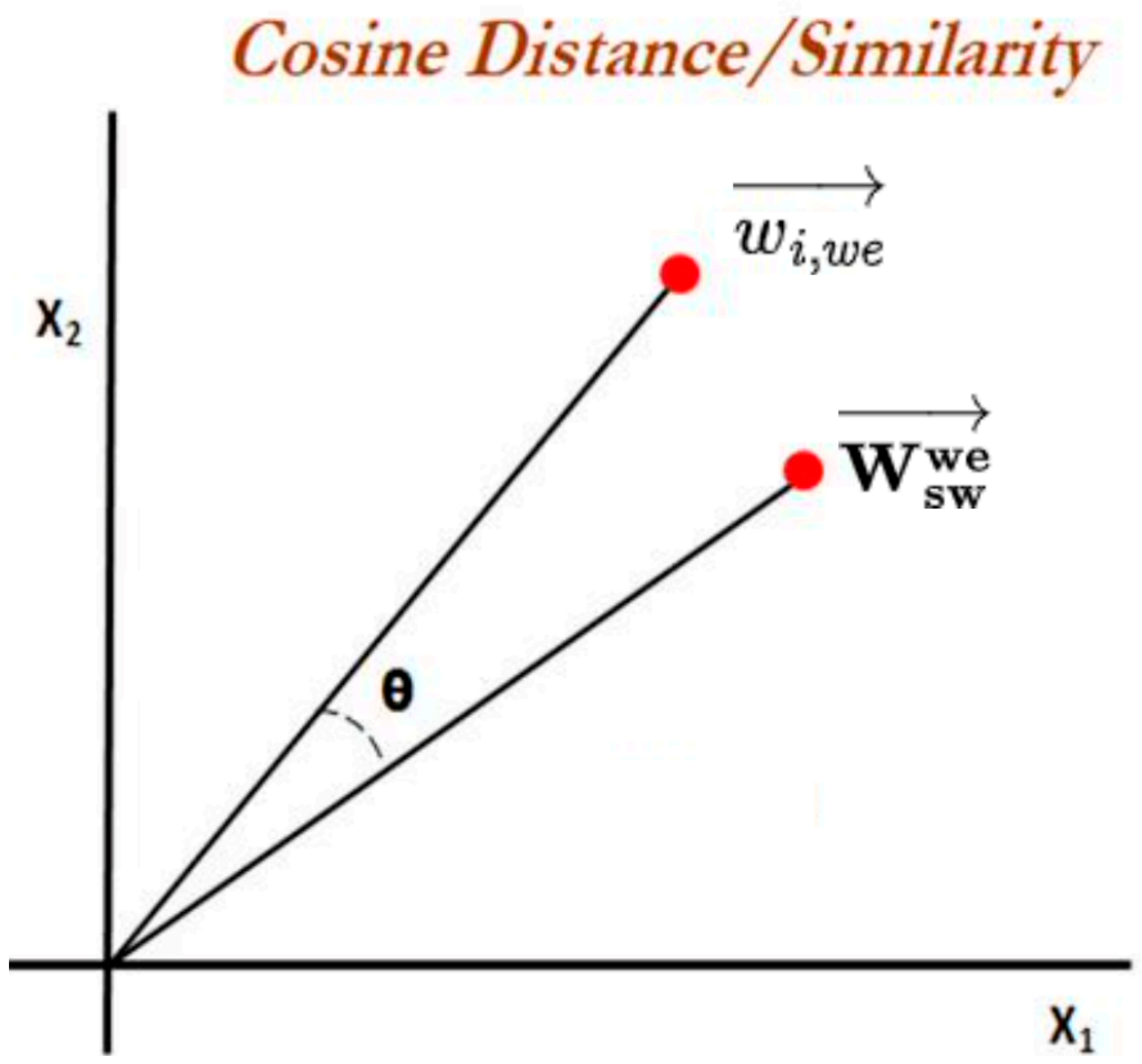
Table1: NOI words

SOS Bias

Measurement

- 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).

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



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 1: examined word embeddings in our work

SOS Bias

Bias in word embeddings

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

SOS Bias

Bias in word embeddings

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

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

SOS Bias

Bias in word embeddings

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

SOS Bias

Bias in word embeddings

SOS bias vs. Social bias

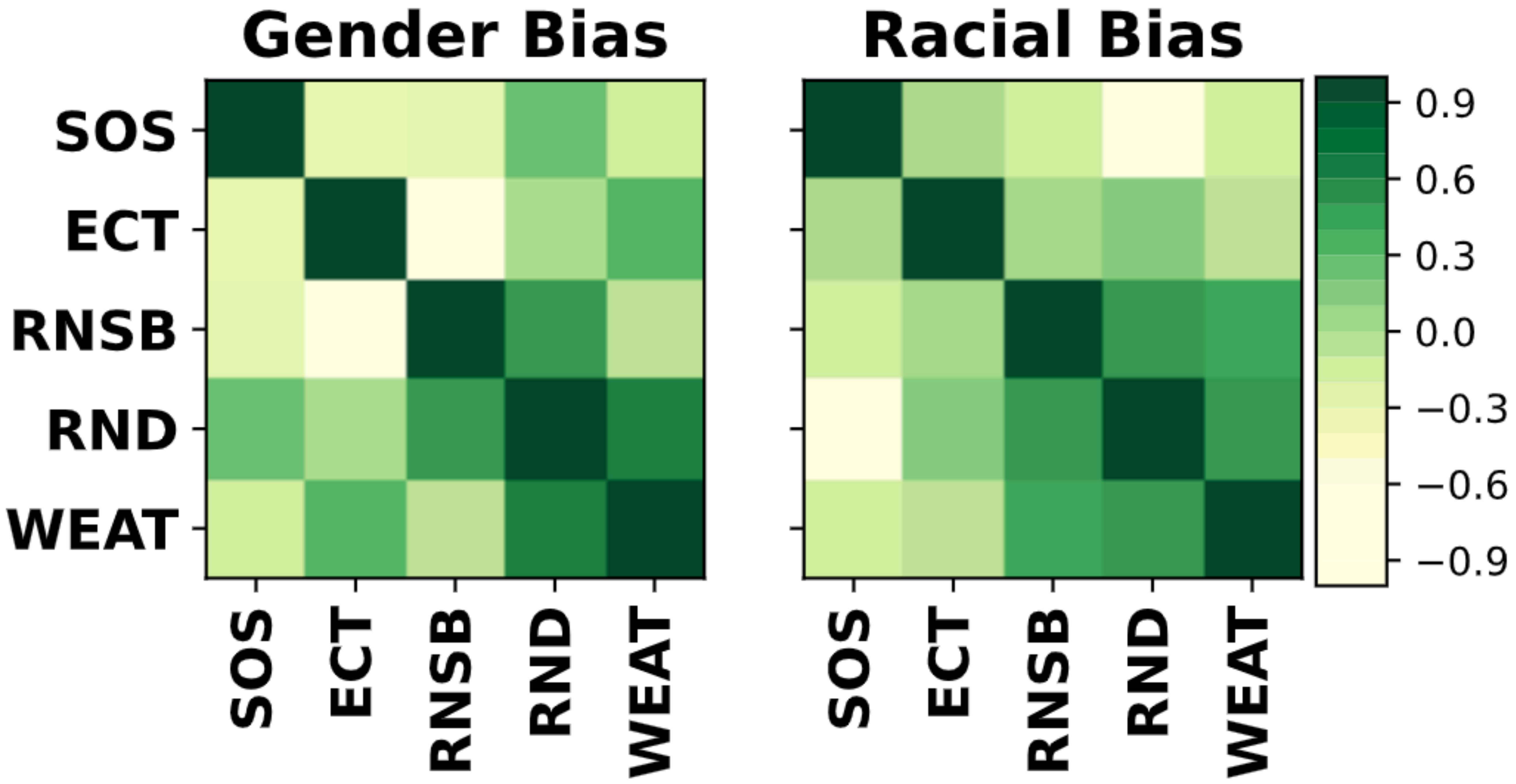


Figure 1: Spearman's correlation

SOS Bias
Validation

1. SOS bias and online hate.

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

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 4: The percentage of examined groups that experience online hate in different countries [10].

[10] 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, we find that the community that experience online hate the most in order are:
    - LGBTQ (61%).
    - Non-White ethnicity (60%).
    - Women (44%).
- 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.

SOS Bias

SOS bias vs. Online hate statistics

- According to the online hate stats, we find that the community that experience online hate the most in order are:
  - LGBTQ (61%).
  - Non-White ethnicity (60%).
  - Women (44%).

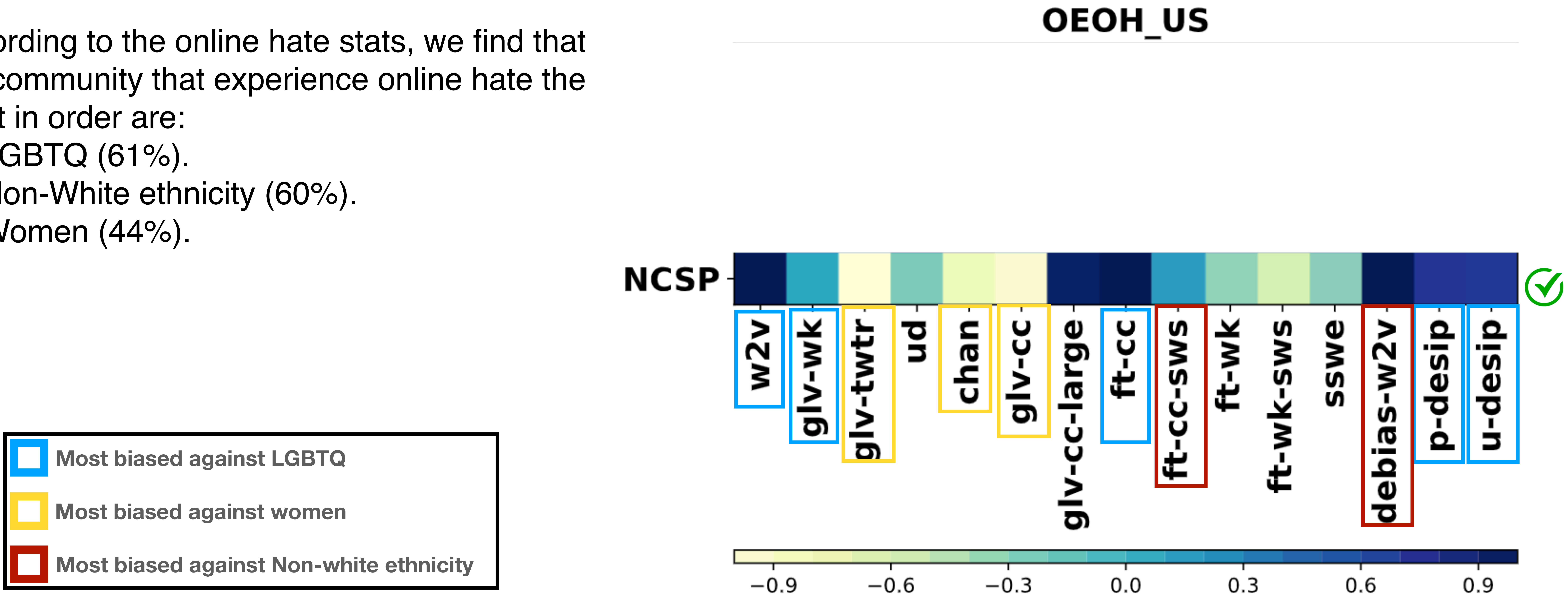


Figure 2: Pearson's correlation between SOS bias scores and published stats on online hate.

SOS Bias

SOS bias vs. Online hate statistics

- According to the online hate stats., we find that the community that experience online hate the most in order are:
    - LGBTQ (61%).
    - Non-White ethnicity (60%).
    - Women (44%).

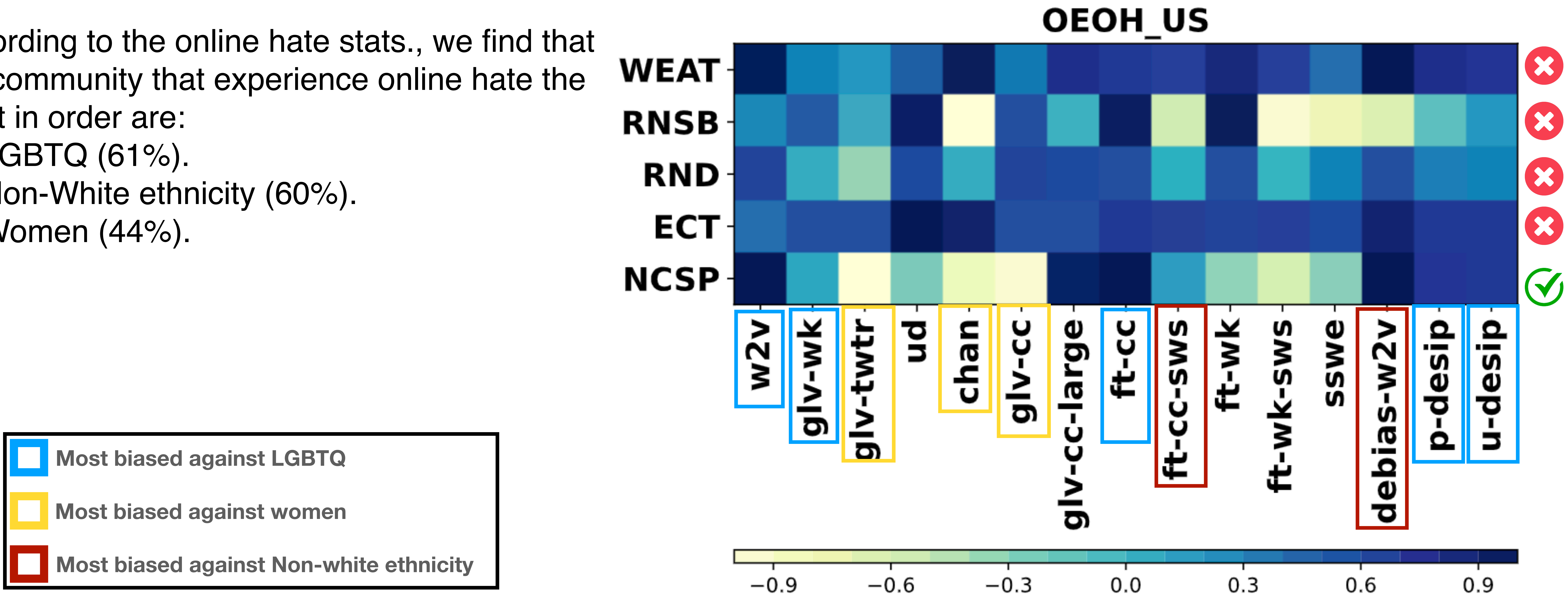


Figure 2: Pearson's correlation between SOS bias scores and published stats on online hate.

SOS Bias

Does it explain hate speech detection models?

Word embeddings	HateEval		Twitter-Hate		Twitter-racism		Twitter-sexism	
	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM
W2V	0.593	0.663	0.681	0.772	0.683	0.717	0.587	0.628
Glove-WK	0.583	0.651	0.713	0.821	0.681	0.727	0.587	0.641
Glove-Twitter	0.623	0.671	0.775	0.851	0.680	0.699	0.589	**0.668**
UD	0.597	0.652	0.780	0.837	0.679	0.698	0.578	0.632
Chan	**0.627**	0.661	0.692	0.840	0.650	0.712	0.563	0.647
Glove-CC	0.625	**0.675**	0.778	0.839	0.695	0.740	0.577	0.648
Glove-CC-large	0.626	0.674	0.775	**0.860**	**0.709**	0.724	0.593	**0.668**
FT-CC	**0.627**	**0.675**	**0.792**	0.843	0.701	0.741	**0.607**	0.654
FT-CC-sws	0.605	0.660	0.746	0.830	0.701	**0.746**	0.588	0.657
FT-WK	0.606	0.650	0.784	0.827	0.699	0.706	0.601	0.653
FT-WK-sws	0.606	0.650	0.723	0.820	0.689	0.736	0.561	0.633
SSWE	0.558	0.628	0.502	0.715	0.324	0.666	0.171	0.548
Debiased-W2V	0.626	0.652	0.678	0.741	0.674	0.715	0.564	0.638
P-DeSIP	0.575	0.657	0.697	0.817	0.673	0.731	0.538	0.650
U-DeSIP	0.598	0.649	0.702	0.815	0.673	0.726	0.548	0.638

Table 5: F1 scores of the hate speech detection models using the inspected word embeddings.

SOS Bias

Does it explain 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

SOS Bias

Take Away Messages

1. There is SOS bias towards marginalized groups (Women, LGBTQ, and Non-white-ethnicity) in most of the examined word embeddings.

2. The proposed SOS bias metric reveals different information than the types of bias measured by existing social bias metrics.

3. The SOS bias scores correlates positively with published statistics on online hate experienced by the marginalized groups.

4. No evidence that the SOS bias explains the performance of the different word embeddings on hate speech detection.

Thank You!

Questions?

fatma.elsafoury@uws.ac.uk

@FatmaElsafoury