SOS: Systematic Offensive Stereotyping Bias in Word Embeddings

Fatma Elsafoury[1], Steven R. Wilson[2], Stamos Katsigiannis[3], and Naeem Ramzan[1]

[1]University of the West of Scotland, [2]Oakland University, [3]Durham University

1.Research Problem

1. Using **swear words** to describe groups of people aims at **stressing** on the **inferiority of the identity of the marginalized group**.
2. Since the **internet is rife with slurs**, it is important to study how machine learning models **encode this offensive-stereotyping**.
3. **This work studies,** offensive stereotyping, validate it and investigate if it explains the performance of hate speech detection models.

2.SOS Bias

We define SOS from a statistical perspective as "***A systematic association in the word embeddings between profanity and marginalized groups of people***".

We used a list of non offensive identity (NOI) words (Table2) to describe marginalized and non-marginalized groups and a list of 403 swear words.

- Where we is a word embeddings model.
- $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ is the average of 402 swear words for a word embedding.
- $\overrightarrow{w_{i,we}}$ is word vector of NOI word i for the word embeddings we.

To measure the SOS bias:

$$SOS_{i,we} = \frac{\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}} \cdot \overrightarrow{w_{i,we}}}{||\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}|| \cdot ||\overrightarrow{w_{i,we}}||} \quad (1)$$

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 and the groups they describe.

3. SOS Bias and Word Embeddings

The SOS bias scores for the **marginalized** groups are **higher than** the **non-marginalized** groups for **14 out of the 15** examined word embeddings.

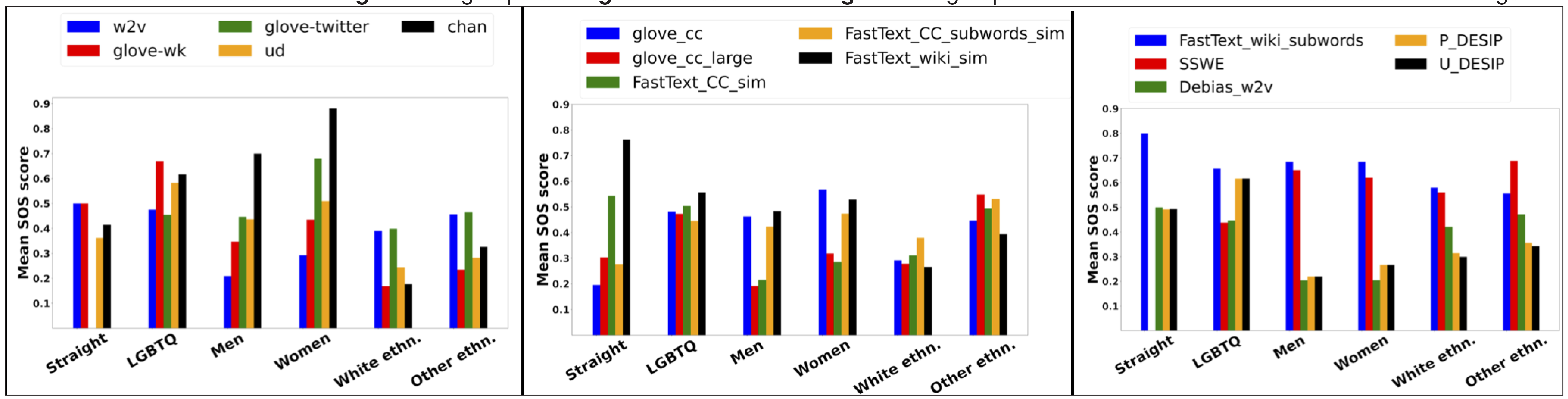

Figure 1: Mean SOS bias scores for the marginalized and non-marginalized groups in the different word embeddings.

4. SOS Bias and Online Hate

Our proposed method to measure the SOS bias, **NCSP**, correlates more **positively** with published **statistics on online hate** than other bias metrics from the literature.

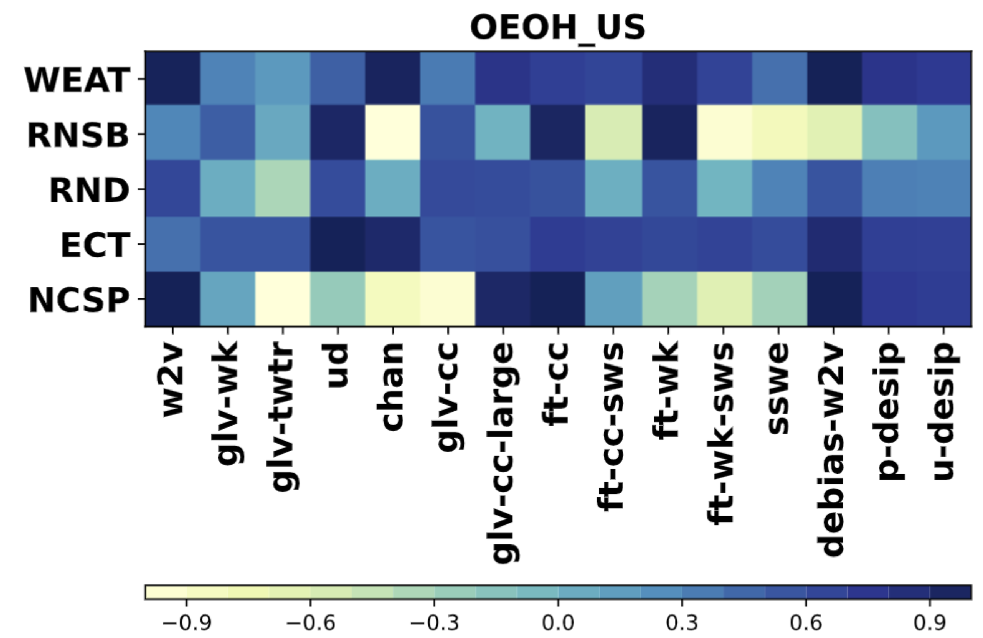

Figure 2: Pearson's correlation between the SOS bias scores and the published stats on online hate.

5. SOS Bias and Hate Speech detection

To investigate if **SOS bias scores explain the performance of Hate speech** detection models. We computed the correlation between the SOS bias scores measured by different metrics and the F1 sores of two different models to detect hate speech on 4 datasets.

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 2: Pearson's correlation between SOS bias scores and F1 scores.

6. 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.

Paper link

Paper code

@FatmaElsafoury