UWS UNIVERSITY OF THE WEST of SCOTLAND

Darkness can't drive out darkness: Investigating Bias in Hate Speech Detection Models

Fatma Elsafoury

Research Problem

- Hate Speech is a problem.
- Bias in NLP.
- Hate speech detection models could lead to associating hate with people from marginalized backgrounds (Women, LGBTQ, non-white ethnicity).

Research Objectives

1. Understand the performance of state-of-the-art hate speech and abuse detection models.
2. Inspect other biases than social stereotypical bias in commonly used static word embeddings.
3. Investigate intersectional bias in contextual word embeddings and the causal effect of social and intersectional bias on the task of hate speech detection.

Research Objective 1:
Understand the performance of SOTA

Dataset	LSTM	Bi-LSTM	BERT
Kaggle-insults	0.642	0.653	**0.768**
Twitter-sexism	0.656	0.649	**0.760**
Twitter-racism	0.640	0.678	**0.757**
WTP-aggression	0.711	0.679	**0.753**
WTP-toxicity	0.723	0.737	**0.786**

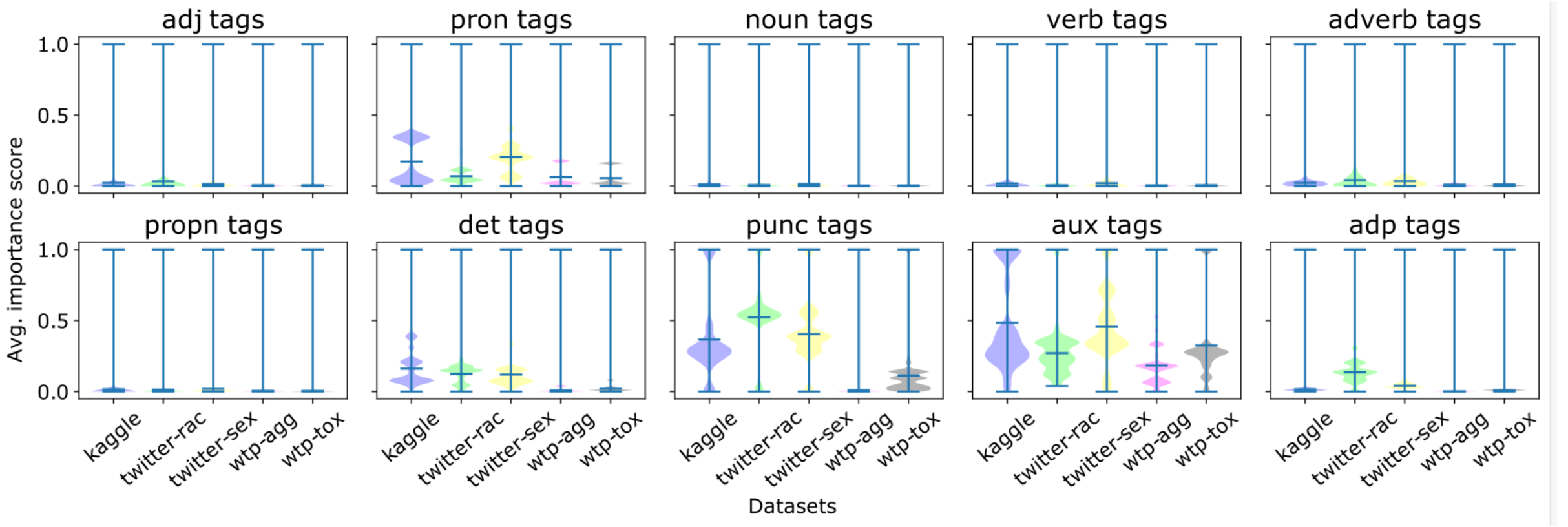
F1 scores of the different models on each dataset

Research Objective 1:
Understand the performance of SOTA

- To understand BERT's performance, I look at BERT's feature importance scores using Gradient-based methods.
- BERT's attention weight.
- Hypothesis: BERT assigns high importance scores to POS tags that are informative to the task of hate speech detection like Nouns and Adjectives.

Research Objective 1:

Understand the performance of SOTA



Research Objective 2:

Inspect other biases than social biases in word embeddings

- Most of the literature focus on social bias like racial or gender bias.
- Using racial slurs and third person profanity to describe groups of people aiming at stressing on the inferiority of the identity of the marginalized group [1].
- The internet is rife with slurs and profanity, it is important to study how machine learning models encode this offensive stereotyping.

[1] Slurs, interpellation, and ideology. The Southern Journal of Philosophy, 56:7–32

Research Objective 2:

Inspect other biases than social biases in word embeddings

- Systematic Offensive Stereotyping (SOS) bias:
  - A systematic association in the word embeddings between profanity and marginalised groups of people e.g. women, LGBTQ, and non-white-ethnicities.
- We look the SOS bias in 5 word embeddings:
  - Word2vec, glove-WK, glove-twitter, UD, and Chan.

Research Objective 2:

Inspect other biases than social biases in word embeddings

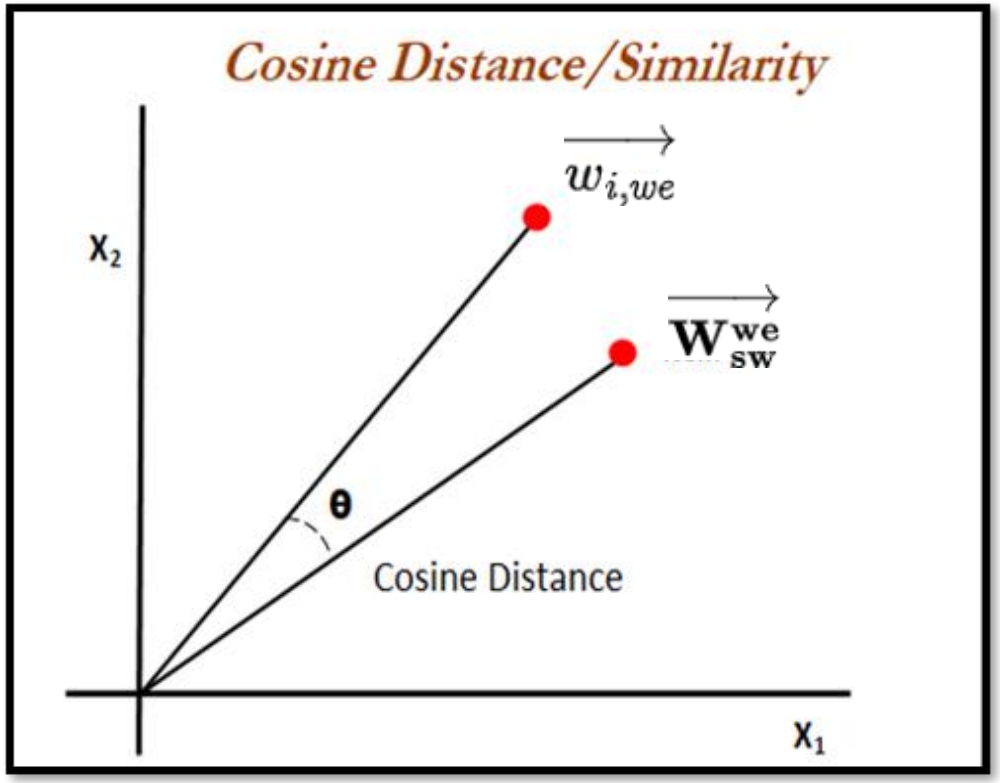
- Measure SOS Bias:

we Is a word embeddings model e.g. word2vc, glove-wk, glove-twitter, ud, and chan.

$\overrightarrow{\mathbf{W}_{\mathbf{sw}}^{\mathbf{we}}}$ Profanity vector is the average vector of the 427 swear words for a word embeddings

$\overrightarrow{w_{i,we}}$ Word vector of identity word for the word embeddings

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



Research Objective 2:

Inspect other biases than social biases in word embeddings

Word embedding	Mean SOS	
	Marginalised	Non-marginalised
Word2Vec	0.403	**0.430**
Glove-WK	**0.448**	0.281
Glove-Twitter	**0.558**	0.461
UD	**0.407**	0.320
Chan	**0.558**	0.393

Table 2: Mean SOS score of the different groups.

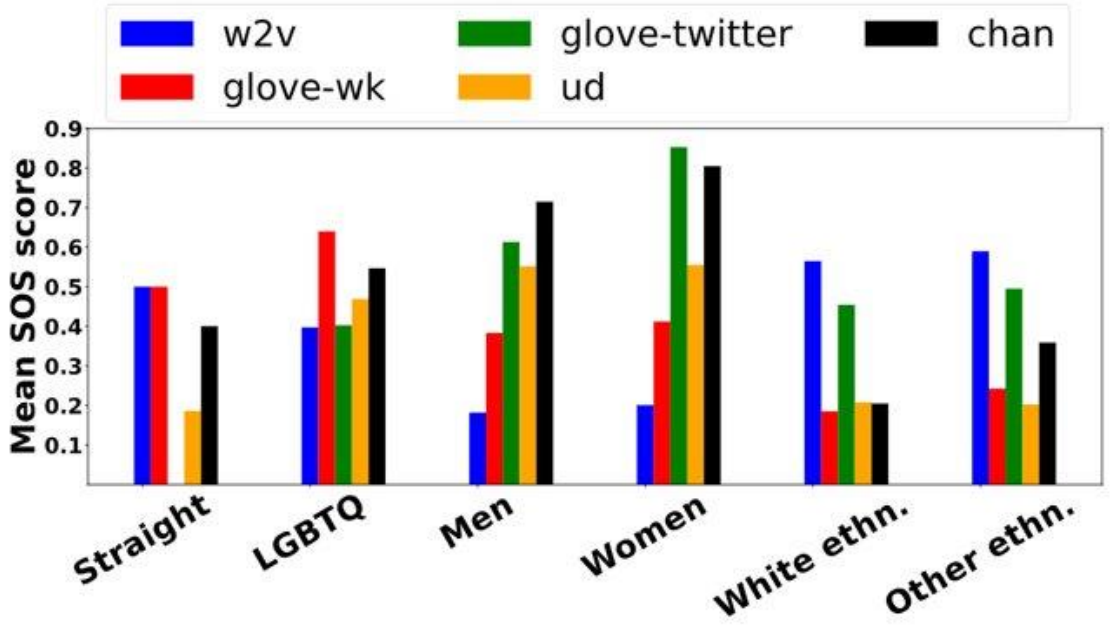
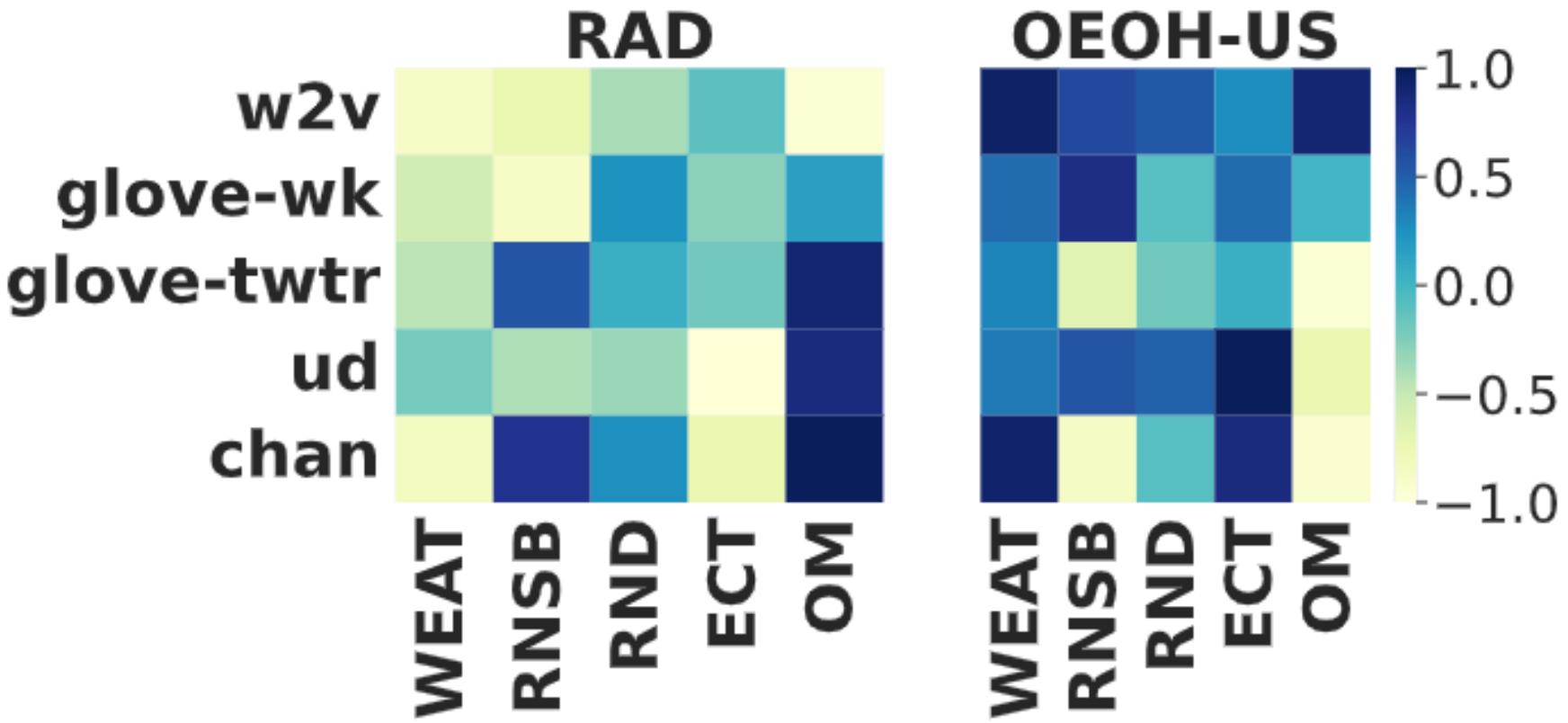


Figure 1: Mean SOS scores for the examined word embeddings and groups.

Research Objective 2:

Validating SOS Bias

Online Abuse **Online extremism**



Research Objective 3:

Investigate Intersectonal bias and causal inference of the bias

- This research goal can be achieved by answering the following research question:
  - How to measure the intersectional bias in pre-trained contextual word embeddings?
  - What is the causal influence of bias, in the pre-trained contextual word embeddings on the task of hate speech detection? and how harmful that bias is it on the models' fairness?

Research Objective 3:

Investigate Intersectonal bias and causal inference of the bias

- A tool to measure intersectional bias in the contextual word embeddings.
- Understand how the bias causally influence the performance and the unfairness of the hate speech detection models.
- Developing more effective and targeted debias techniques that address the unfairness caused by the bias.

Conclusion

1. Language Models like BERT rely on syntactical biases in the dataset for its good performance rather than learning linguistic features related to hate speech.
2. All the inspected word embeddings contain SOS bias towards marginalized groups.

Thank you

Fatma Elsafoury
@FatmaElsafoury
Fatma.elsafoury@uws.ac.uk