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

Fatma Elsafoury (fatma.elsafoury@uws.ac.uk)

UNIVERSITY OF THE WEST of SCOTLAND
UWS

Thesis Statement

Hate Speech detection models aim at providing a protective environment for people from different backgrounds to express themselves. However, the **bias in** hate speech detection models could **lead to associating hate with people from marginalized** backgrounds (Women, LGBTQ, non-white ethnicity) and hence falsely flag their content as inappropriate. **In this thesis,** I aim to understand and investigate the performance and the biases of hate speech and abuse detection models.

Research Objectives(ROs)

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.

RO1: Understand the performance of SOTA

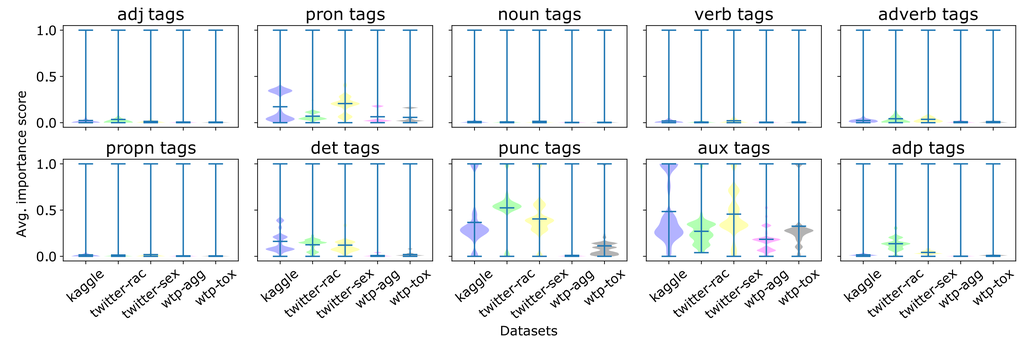
Performance on Hate speech detection models

Dataset	LSTM	Bi-LSTM	BERT
Kaggle-insults	0.6420	0.653	**0.768**
Twitter-sexism	0.6569	0.649	**0.760**
Twitter-racism	0.6400	0.678	**0.757**
WTP-aggression	0.7110	0.679	**0.753**
WTP-toxicity	0.7230	0.737	**0.786**

Table 1: F1-scores

In the The Figure to the right, I analysed BERT's importance scores for the part-of-speech (POS) tags in the datasets. I hypothesised that BERT assigns the highest importance scores to informative POS tags for the task of cyberbullying detection, e.g. **Nouns**, and **Adjectives**. Results show that **the most important POS** according to BERT are **Auxiliaries** and **Punctuation**. This suggests that BERT **relies on syntactic biases rather than linguistic features related to hate speech**.

What are the features that BERT relies on for its performance?

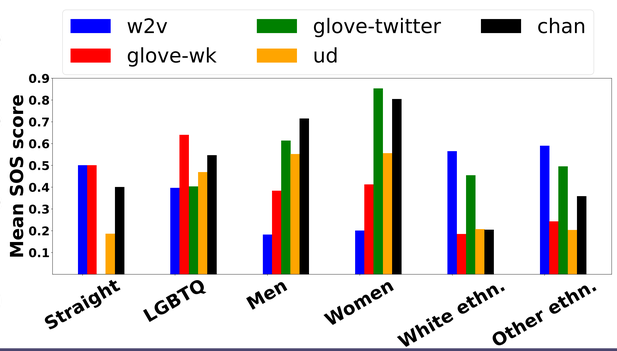


RO2: Inspect other biases than social biases in word embeddings

Systematic Offensive Stereotyping (SOS) Bias

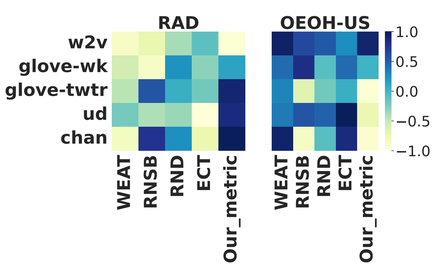
- $W_{NOI} = \{w_1, w_2, w_3, ...w_n\}$ be the list of identity words
- $W_{sw} = \{o_1, o_2, o_3, ...o_m\}$ be the list of swear words
- $\overrightarrow{\mathbf{W^{we}_{sw}}}$ is the average vector of the swear words.

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



Validating SOS Bias

1. The measured SOS bias correlates positively with published statistics on online harassment (RAD) and online hate and extremism in the US (OEOH-US).
2. I also compare the proposed metric to measure SOS bias to state of the art bias metrics (WEAT, RNSB, RND, ECT).
3. Our bias metric **reveal the difference between the word embeddings** in contrast to other metrics.



RO3: Investigate Intersectonal bias and causal inference of the bias

This research goal can be achieved by answering the following research question:

1. How to **measure the intersectional bias** in pre-trained contextual word embeddings?
2. 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**?

Take Away Messages

1. Language models like **BERT rely on syntactical biases** for the good performance.
2. All inspected word embeddings **contain SOS bias towards marginalized groups**.
3. It is **not conclusive** how the different **biases influence** downstream tasks.