Thesis Distillation: Investigating the Impact of Bias in NLP Models on Hate Speech Detection

Fatma Elsafoury (fatma.elsafoury@fokus.fraunhofer.de)

Fraunhofer FOKUS weizenbaum institut

1.Thesis Statement

Hate Speech detection models aim at providing a protective environment for people from different backgrounds to express themselves. However, the impact of **bias in NLP** models on hate speech detection is still understudied. **In this thesis,** I aim to understand that impact from three perspectives **Explainability, Systematic Offensive stereotyping (SOS) bias, and Fairness**.

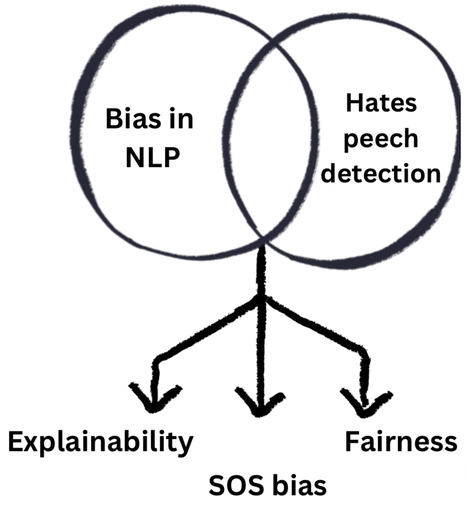
2.Findings

Explainability: Results are inconclusive due to limitations in bias metrics.

SOS bias: All Inspected static and contextual word embeddings are SOS biases .

Fairness: Downstream sources of bias are the most impactful on the fairness of text classification

2.Contributions



Explainability

Investigate whether the bias in NLP models explain the performance of Hate speech detection Models.

SOS bias

Invstiagte the systematic offensive stereotyping bias nstatic and contextual word embeddings.

Fairness

Investigate the impact of different sources of bias and tehir removal on the fairness of hate speech detection task

3.1 What have we learned?

These findings assert the notion that **bias in NLP models negatively impacts hate speech detection** models. I argue that the limitations of the currently used methods to measure and mitigate bias in NLP models are due to **fail to incorporate findings from the social sciences.**

3.2 What have we learned?

Shot-term solution guidelines

1. Measure and remove downstream bias.
2. Use Counterfatual fairness metrics.
3. Trade-off performance and fairness.

3.3 What have we learned?

Long-term recommendations

1. Raise NLP researchers awareness on social and historical contexts
2. Encourage interdiscplinlary conferences betweeen NLP and social sciences.
3. Promote diversity in NLP research teams.
4. Raise NLP researchers awareness on social and historical contexts

4. Future research direction

- Widening the study of bias in NLP.
- Incorporate social sciences in studying bias
- Use causalityto understand hte imact of bias .
- Studying the inter - sectionality of bias in NLP.

5. References

Paper Link

My Website

@FatmaElsafoury