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

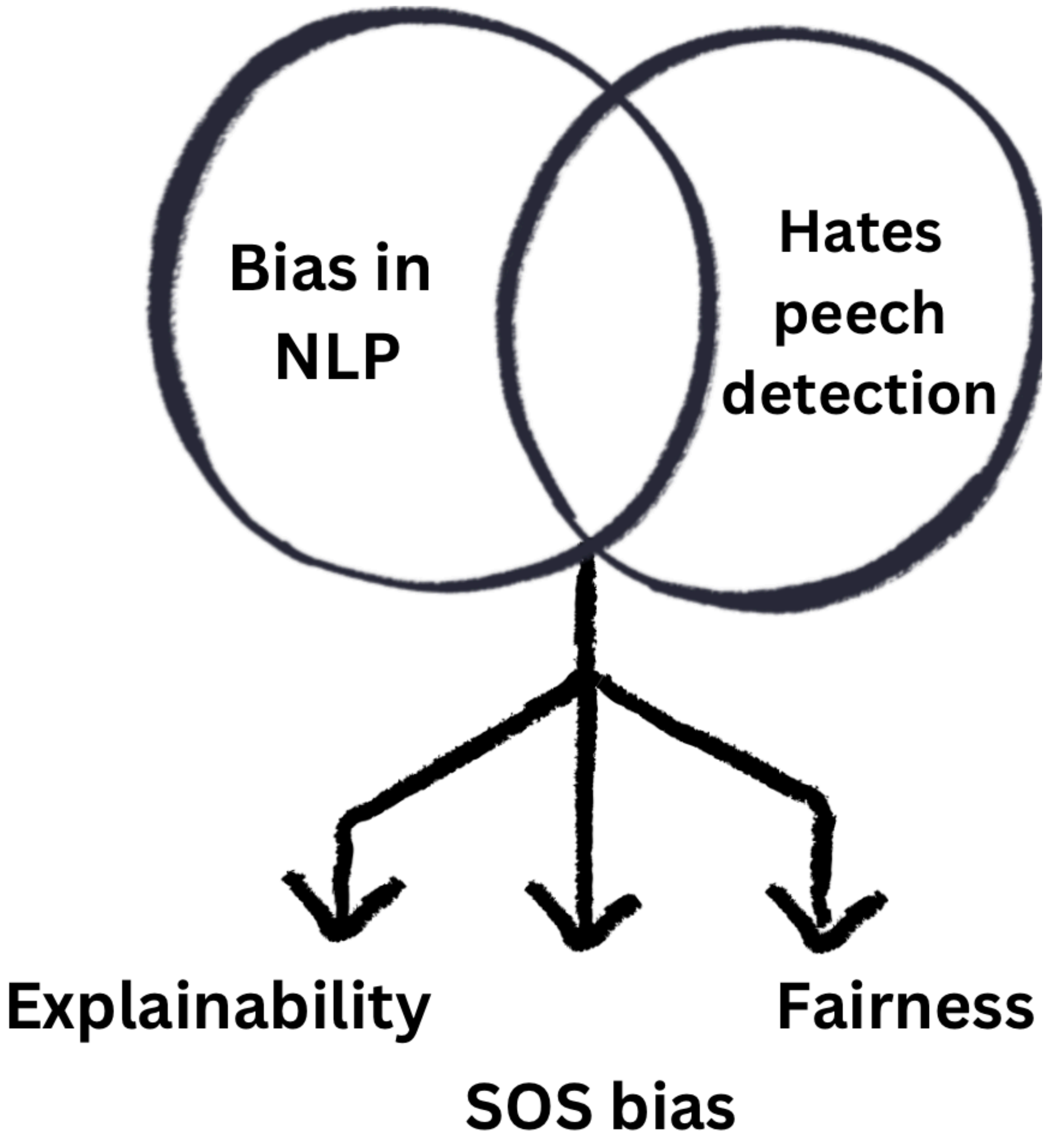
Fatma Elsafoury

fatma.elsafoury@fokus.fraunhofer.de
@FatmaElsafoury

Research Problem

- **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.



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

Findings

❓From the **Explainability perspective**, it is inconclusive that the social bias in NLP models explains the performance of hate speech detection models **due to limitations in the proposed metrics to measure social bias.**

✅From the **Offensive stereotyping bias perspective**, the results demonstrate that word embeddings, static and contextual, are systematic offensive stereotyping (SOS) biased.

✅From the **Fairness perspective**, the results show that the inspected types of bias have an impact on the fairness of the models on the task of hate speech detection, especially the downstream sources of bias.

What Have We Learned?

💡 These findings assert the notion that the bias in NLP models negatively impacts hate speech detection models.

💡 We need to mitigate those biases so that we can ensure the reliability of hate speech detection models.

💡 **However**, 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.

What Have We Learned?

As a short-term solution to improve the fairness of hate speech detection and text classification tasks, **I provide these guidelines:**

1. Measure and remove downstream bias.

2. To reliably measure fairness, use counterfactual fairness metrics.

3. Trade-off between performance and fairness.

What Have We Learned?

For a long-term solution, I provide the following **recommendations**:

1. Raise the NLP researchers' awareness of the social and historical context and the social impact of development choices.

2. Encourage workshops on re-imagining NLP models with an emphasis on fairness and impact on society.

3. Encourage shared NLP models audit asks.

4. Reproducibility and ethical badges.

5. Encourage interdisciplinary workshops between NLP and social sciences.

Future Research Directions

- Widening the study of bias in NLP.
- Investigate the impact of social bias causes on the bias in NLP.
- Studying the impact of bias on NLP tasks using causation instead of correlation.
- Studying the intersectionality of bias in NLP models.

Thank you!

fatma.elsafoury@fokus.fraunhofer.de

@FatmaElsafoury