Paper Title: Cyberbullying Detection With Fairness Constraints.

Paper Link: https://ieeexplore.ieee.org/abstract/document/9076550

Summary:

1. **Motivation:** This study tackles cyberbullying detection with machine learning, but acknowledges bias in these algorithms. They propose a new training method to reduce bias while still effectively detecting cyberbullying.

2. **Contribution:** This research introduces a method for training cyberbullying detection models with fairness constraints. This approach effectively reduces bias across datasets and contexts (gender, language) without sacrificing performance. It's data-agnostic and doesn't require group labels, making it generalizable for building fairer AI in cyber-social health. This work paves the way for ethical and transparent machine learning solutions for cyberbullying detection.

3. **Methodology:** This study incorporates fairness into cyberbullying detection models. FNED and FPED metrics measure bias in error rates across groups. To reduce bias, the authors train the model with penalties for unequal error rates, but address complexity with simpler proxy constraints. Training involves two data streams: regular training and fairness monitoring. This achieves fairness without needing group data later. Model performance and fairness are evaluated using metrics like F1 score, while bias is quantified with FNED/FPED. Overall, this work offers a promising approach for fairer models, considering the fairness-performance trade-off.

4. **Conclusion:** Study demonstrates fair training constraints for cyberbullying detection models effectively reduce bias while maintaining accuracy across datasets. This method offers generalizability but requires further research on real-world application.

Limitations:

This study's limitations are:

- **Data Scarcity:** Relies on public datasets, which might be limited and not reflect real-world situations.
- **Generalizability concerns:** More work is needed to see how well the method performs with different groups and real-world data.
- **Broad Cyberbullying Definition:** The definition used here might be too broad and miss the nuances of real-world cyberbullying.

Synthesis: The article addresses the challenge of cyberbullying detection using machine learning algorithms and the presence of unintended social biases in these models. It proposes a model training scheme that incorporates fairness constraints to mitigate biases without compromising model quality. The study validates this approach with various datasets, demonstrating that biases related to gender, language, and recency can be reduced. The approach is agnostic to data modality and does not require group labels during inference, contributing to the development of ethical machine learning solutions for cyber-social health. The article emphasizes the importance of equitable algorithms in the context of cyberbullying detection and the ongoing need to develop methods to reduce existing biases.