

Title: Transfer Learning for Automatic Sexism Detection with Multilingual Transformer Models

1 Summary

1.1 Motivation/purpose/aims/hypothesis

This research delves into the pervasive issue of sexism on online platforms, motivated by the need to overcome the challenges posed by diverse forms and cultural/language barriers. The primary goal is to provide an effective solution to combat sexism in the digital landscape.

1.2 Contribution

A significant contribution of this study lies in the strategic use of pre-trained transformers, including mBERT, XLM-R, and T5, for identifying sexism. The incorporation of data augmentation techniques and the exploration of bilingual datasets contribute to the robustness and effectiveness of the models.

1.3 Methodology

The methodology encompasses a meticulous two-step process involving the pre-training and fine-tuning of transformer models. Models like mBERT, XLM-R, and T5 are harnessed, showcasing their ability to comprehend complex linguistic patterns. The paper also underlines the positive impact of label aggregation on refining classification results.

1.4 Conclusion

In conclusion, the study emphasizes the crucial importance of addressing sexism in online platforms. Through the strategic use of data augmentation and pre-trained transformers, the research showcases promising results in sexism identification, underscoring potential avenues for future enhancements.

2 Limitations

2.1 First Limitation/Critique

Acknowledging the potential bias in the pre-training data, the paper recognizes its impact on the model's performance and the reinforcement of existing biases. This acknowledgment emphasizes the ongoing need for refinement in addressing biases in machine learning models.

2.2 Second Limitation/Critique

The study openly admits the challenge of addressing culture and language-specific conditions. This limitation underscores the necessity for continuous efforts to improve the model's adaptability to diverse linguistic and cultural contexts.

3 Synthesis

The ideas presented in this paper hold significant implications for practical applications. The developed models exhibit promise in creating more inclusive and respectful online environments. The synthesis underscores the potential real-world applications of these approaches, offering a pathway to address sexism and enhance the safety of digital spaces.

In summary, this research unfolds valuable insights into leveraging transfer learning for automatic sexism detection. While acknowledging its limitations, the study lays a strong foundation for future endeavors, contributing to the ongoing pursuit of refining models and fostering equality in online communications.