Leveraging DistilBERT and BERT for the Detection of Online Sexism: A Comparative Analysis [Paper URL]

1 Summary

1.1 **Motivation:** Online sexism leads to hostile user experiences of women and can cause various mental challenges in day-to-day life. Detecting such acts fast can make the online space a bit safer for women and everyone.

1.2 **Contribution:** The paper presents a comparative analysis of using the BERT and DistilBERT transformer models for the detection of online sexism, with BERT outperforming DistilBERT in terms of F1-score while DistilBERT being more efficient in terms of training time.

1.3 **Methodology:** The work used a dataset of 14,000 comments divided into training, validation, and test sets and then preprocessing the text data through tokenization, padding, truncating, and segment embeddings. They used the PyTorch deep learning library to train BERT and DistilBERT models. The BERT model considers the entire context of a word by looking at it from both directions and the DistilBERT model is a compressed version of BERT with fewer parameters.

1.4 **Conclusion:** According to their finding, the BERT model had a higher F1 score than other works. On the other hand, DistilBERT had reached similar results with lower training time. Also, the electricity cost was shown at 8$, much higher than similar tasks. In comparison to other research using LSTM and GRU models, the BERT model performed better.

2 Limitations

2.1 This work performed only Task A (binary classification) on the EDOS Challenge. Task B (Multi-Category of Sexism), and Task C (Classifying Fine-grained Vector of Sexism) were left.

2.2 Exploring the implications of the tradeoff between computational speed and precision, as demonstrated by DistilBERT and BERT, for real-world scenarios where operational limits intersect with the demand for immediate digital surveillance.

2.3 The work failed to optimize the training process, resulting in a much larger training time of 12 hours for BERT. DistilBERT had a lower training time, 5 hours.

3 Synthesis

3.1 Incorporating a wider array of transformer models beyond BERT and DistilBERT to further evaluate their performance in detecting online sexism.

3.2 Integrating the explainability of the predicted outputs can greatly increase the model's acceptability. This can also aid in understanding if the model is just making random predictions instead of learning.