Detecting and Moderating Sexist Content in Social Media

1 Introduction

In this project, we developed a system for detecting sexist content on Facebook, focusing on both text-based and image-based posts. The approach involves training, fine-tuning, and implementing machine learning models to identify and moderate inappropriate content.

2 Text-Based Classification

For text classification, we fine-tuned a pre-trained model, BERTweet-large-sexism-detector, on the EDOS dataset. BERTweet is a variant of the BERT model, specifically optimized for social media text, utilizing a transformer architecture with a bidirectional attention mechanism. This architecture enables effective contextual understanding of social media language, making it well-suited for detecting sexist language in Facebook comments.

2.1 Model Fine-Tuning

Fine-tuning was conducted to adapt BERTweet to the specific nuances of sexist language as presented in the EDOS dataset. This process involved adjusting model weights to improve classification performance on this specific task.

2.1.1 Evaluation Results Before and After Fine-Tuning

The table below presents key evaluation metrics before and after fine-tuning the BERTweet model on the EDOS dataset:

Metric	Before Fine-Tuning	After Fine-Tuning
Accuracy	0.8903	0.9412
Loss	0.5346	0.4374
F1 Score	0.4202	0.6723
Precision	0.4262	0.7215
Recall	0.4143	0.6295

Table 1: Comparison of Initial and Final Evaluation Metrics for BERTweet Model

3 Image-Based Classification

For image-based content, we developed a more complex workflow to identify sexist memes and blur them if needed.

3.1 Model Workflow

The workflow consists of the following steps:

1. **Image Captioning and OCR:** To extract textual content from images, we used the BLIP image captioning model from Hugging Face, along with Optical Character Recognition (OCR).

2. **Multimodal Classification:** The extracted text from the image, combined with any additional context text, is passed to a multimodal classifier. This classifier was fine-tuned on the MAMI dataset, which contains annotated samples of sexist memes.

4 Conclusion

The final system successfully classifies and moderates both text and image-based sexist content on social media. By leveraging BERTweet for text classification and BLIP along with OCR for image captioning, our multimodal approach provides a comprehensive solution for moderating sexist content.