Management Summary: Online Sexism Detection

Draganic Vanja, Milicic Sofija, Forman Henry and Pfennigbauer Johannes

January 26, 2025

1. Overview of the Task

The objective of this project was to develop a machine learning solution for detecting and classifying online sexism. Using the **EDOS (Explainable Detection of Online Sexism) dataset**, which contains labeled examples of sexist content, our objective was to create models capable of identifying sexist remarks to serve in moderating online platforms.

2. Challenges Faced

Several challenges were encountered during the project:

- **Class imbalance:** The dataset exhibited a significant imbalance between classes, with the sexist category underrepresented.
- **Label aggregation:** Every sentence was labeled by three individual experts, as the true label is inherently ambiguous, and there are no objective criteria to determine whether a sentence is sexist or not.
- **Contextual understanding:** Detecting sarcastic or hidden forms of sexism requires models capable of understanding complex linguistic contexts.
- **Computational constraints:** Training large language models (LLMs) like BERT and Llama requires considerable computational resources.

3. Implemented Solution

To implement, train and compare the models, several external resources were needed. The key external resources utilized in this project include:

- **EDOS Dataset:** A labeled dataset for online sexism detection.
- **pretrained transformer models:** BERT variants (from Hugging Face model repository) and Llama 3.2

After a careful exploration and analysis of the data, the text was cleaned and tokenized for further usage. Then, we implemented and compared the following approaches:

- **Baseline Models:** Majority class, Rule-Based, **Naive Bayes**, Logistic regression, XGBoost and **LSTM**.
- **BERT-based Models:** RoBERTa, DeBERTa, **HateBERT**, DistilBERT
- **Large Language Models:** Llama 3.2 3B, Llama 3.1 8B, Phi-4 14B

The models were evaluated on the same validation set using the balanced accuracy, precision and recall as key metrics. The best performance was achieved using the **fine-tuned Llama 3.1 8B model**, which outperformed both baseline models and BERT based models in terms of balanced accuracy and precision for the sexist class.

	LSTM	HateBERT	Llama 3.1 8B
Recall (sexist)	0.593	0.761	0.875
Balanced Accuracy	0.713	0.777	0.870

Table 1: Evaluation on the test set with provided aggregation.

4. Limitations

- **Dataset Bias:** The dataset reflects annotator bias, as the true label of whether a sentence is sexist or not is inherently ambiguous. There are no objective criteria for this classification, which makes the labels subjective and influenced by individual perceptions.
- **Scalability:** The baseline models are all rather small when it comes to memory and computational requirements, making them a cost effective solution. When using BERT-based models or LLMs we get a significant improvement across all metrics, at the cost of requiring a GPU for inference. Depending on if there are GPUs already available or the increased performance is worth the investment, those models are a better choice
- **Explainability:** LLMs provide unique ways of making them explainable: We can compute the probability for the next token, an indication of the models confidence in its prediction or let it generate an explanation of why it "thinks" a sentence is sexist. Another possibility is to analyze the attention scores and see which words pay particular attention to others. Additionally, we used SHAP on our BERT-based models to see which words had the most impact on the model's decisions. This made the model more transparent and helped us spot possible biases in how it classifies sexist content.

5. Potential Next Steps

- **Model Combination**: To achieve comparable performance to the LLMs, but with fewer computational resources, a combination of HateBERT and LSTM could be used. HateBERT is effective at identifying real sexism, though it sometimes overclassifies sentences as sexist, while LSTM is better at recognizing non-sexist sentences. To combine their strengths and address their weaknesses, we propose merging the hidden representations of both models and then training a simple neural network to make the final decision.
- **Model Compression**: An easy way to compress the models thereby reducing the memory footprint, computational cost and increasing the speed of inference, is to quantize them, i.e. reduce the precision of the internal parameters. However, this will reduce the model's performance. Further testing would need to be done to see by how much.
- **Model Explainability**: As a next step, SHAP could be used on our Llama model and other LLMs to further analyze their decision making and compare interpretability across models.

In conclusion, in this project we successfully developed and evaluated machine learning models for detecting online sexism. Despite challenges such as annotator bias and computational constraints, our fine-tuned Llama 3.1 8B model demonstrated a solid performance of almost 90% across key metrics, providing a robust solution for this complex task.
While the results are promising, they highlight the trade-offs between performance and computational efficiency. To address these challenges, future work could focus on combining model strengths, applying advanced model compression techniques, and improving explainability to ensure transparency in real-world applications. Ultimately, our work contributes to the improvement of automated moderation tools fostering safer digital environments.