NLP Assignment 2

Sexism Detection via Few-Shot Prompting with LLMs

Nadia Farokhpay - 0001111417

Master's Degree in Artificial Intelligence, University of Bologna

nadia.farokhpay@studio.unibo.it

Abstract

This report investigates binary sexism detection using Large Language Models (LLMs) on the EDOS Task A dataset. We compare the performance of Phi-2 and Llama 3.1 under zero-shot, two-shot, and four-shot prompting. Results show that few-shot prompting generally improves accuracy for Llama 3.1, with the two-shot configuration yielding the best performance. The study highlights the LLMs' ability to follow instructions and the critical role of prompt design and example selection in sensitive text classification tasks.

1 Introduction

Online sexism is a pervasive issue that can cause significant harm. Automatically detecting sexist content in text, such as tweets, is an increasingly vital task with applications in social media content moderation and legal analysis. Traditional text classification methods often rely on large, extensively annotated datasets. However, Large Language Models (LLMs) have revolutionized natural language processing by capturing rich contextual and linguistic nuances, enabling effective performance even with limited labeled data through techniques like prompting.

This study evaluates the performance of two prominent open-source LLMs, Phi-2 and Llama 3.1, for binary sexism detection (EDOS Task A). My objective is to classify an input text sentence as either "sexist" or "not sexist." I explore the effectiveness of different prompting strategies, specifically zero-shot and few-shot (two-shot and four-shot) configurations, to assess the LLMs' capabilities in this sensitive classification problem without extensive fine-tuning. The findings will demonstrate the potential of LLMs in addressing such tasks and highlight the impact of prompting techniques on their performance.

2 Methodology

The system comprises four main components: model initialization, prompt setup, inference pipeline, and evaluation.

Model Initialization: Phi-2 (microsoft/phi-2) and Llama 3.1 (meta-llama/Llama-3.1-8B) were downloaded from Huggingface. To optimize inference in hardware-limited environments, both models were loaded with 4-bit quantization using BitsAndBytesConfig and torch.bfloat16. Text generation pipelines were configured with `max_new_tokens=30`, `do_sample=False`, and `temperature=0.0` to ensure deterministic and concise responses. The loaded and quantized models are then saved locally for future reuse.

Data Loading: Evaluation was performed on a balanced subset of the EDOS Task A dataset, comprising 300 samples for testing (`a2_test.csv`) and 1000 samples for demonstrations (`demonstrations.csv`).

Prompt Setup: A `zeroshot_prompt` template was defined to instruct the LLMs as "sexism detection annotators," requiring a "YES" or "NO" response. For few-shot experiments, a `fewshot_template` was used, into which a specified number of balanced examples from `demonstrations.csv` were injected. This process involved dynamically formatting input texts into the structured instruction prompts.

Inference and Evaluation: The inference pipeline processed tokenized input data, generated responses, and parsed them into "sexist" (1) or "not sexist" (0) labels. Model performance was evaluated using Accuracy (percentage of correct predictions) and Fail-ratio (proportion of responses that failed to follow the prompt format).

3 Experimental Results

The performance of Phi-2 and Llama 3.1 across zero-shot, 2-shot, and 4-shot configurations is summarized below:

Model	Prompt Type	Accuracy	Fail Ratio
Phi-2	Zero-shot	0.55	0.00
Phi-2	2-shot	0.57	0.00
Phi-2	4-shot	0.47	0.00
Llama 3.1	Zero-shot	0.54	0.00
Llama 3.1	2-shot	0.61	0.00
Llama 3.1	4-shot	0.59	0.00

4 Discussion

Accuracy Analysis: Both Phi-2 and Llama 3.1 demonstrated modest performance on the sexism detection task, with accuracies ranging from approximately 47% to 61%. Llama 3.1 consistently benefited from few-shot prompting, with the 2-shot configuration yielding its highest accuracy (0.61). This suggests that providing a few examples helps Llama 3.1 better understand the task context. In contrast, Phi-2 showed a slight improvement with 2-shot prompting but a notable decrease in accuracy with 4-shot (0.47), even performing worse than its zero-shot baseline. This indicates that for Phi-2, an increased number of examples might introduce noise or lead to overfitting on the provided demonstrations.

Fail Ratio: A consistent 0.0 fail ratio across all experiments for both models highlights the effectiveness of the prompt engineering in guiding the LLMs to produce responses in the desired "YES" or "NO" format. This is critical for automated systems requiring structured outputs.

Error Analysis: Classification reports revealed a consistent trend: both models exhibited higher recall for the "Sexist" class and lower precision for the "Not Sexist" class. This implies that the models are more prone to identifying non-sexist content as sexist (false positives) rather than missing actual sexist content (false negatives). This bias towards "sexist" classification was more pronounced in zero-shot setups and was generally reduced in few-shot configurations, suggesting that examples help in balancing the classification. The lower precision for "Not Sexist" suggests that the models might be overly sensitive or generalize too broadly from the concept of sexism.

5 Conclusion

The assignment successfully demonstrated the application of LLMs for binary text classification using zero-shot and few-shot prompting. While the models consistently adhered to the response format (zero fail-ratio), their classification accuracy for sexism detection remains moderate. Llama 3.1 benefited from few-shot examples, while Phi-2′s performance can degrade with more examples, highlighting the sensitivity of LLMs to prompt design and the selection of demonstration examples. Further work could involve more sophisticated prompt engineering, exploring different LLM architectures, or fine-tuning the models on a larger, task-specific dataset to improve classification performance.

6 References

- Microsoft Phi-2: https://huggingface.co/microsoft/phi-2
- Meta Llama-3.1: https://huggingface.co/meta-llama/Llama-3.1-8B
- Hugging Face Transformers: https://huggingface.co/docs/transformers/
- BitsAndBytes Quantization: https://github.com/TimDettmers/bitsandbytes