Assignment 2: Sexism Detection via LLM Prompting

Hassan Mujtaba Ahmed, Xiyan Wang, and Muhammad Talha Sohail Chattha
Master's Degree in Artificial Intelligence, University of Bologna
{hassan.ahmed5, xiyan.wang, muhammadtalha.chatha}@studio.unibo.it

Abstract

This report addresses the EDOS Task B on sexism detection, a multi-class classification problem involving five categories: *threats, derogation, animosity, prejudiced discussion*, and *not-sexist*. Moving away from traditional supervised learning, we evaluate the capabilities of Large Language Models (LLMs) using Zero-shot and Few-shot prompting strategies. We experiment with two open-source models: **Mistral-7B-Instruct-v0.3** and **TinyLlama-1.1B-Chat-v1.0**. Our experiments reveal a significant performance gap based on model scale. While Mistral-7B achieves a competitive Macro F1-score of 0.45 with few-shot prompting, TinyLlama-1.1B struggles significantly, exhibiting a 95% failure rate in instruction following under zero-shot conditions. The results highlight the critical importance of model size and in-context learning for nuanced text classification tasks.

1 Introduction

Sexism detection in online discourse is a challenging Natural Language Processing (NLP) task due to the subtle and subjective nature of hate speech. This assignment focuses on EDOS Task B, which requires classifying text into one of four sexist categories (*threats, derogation, animosity, prejudiced discussion*) or labeling it as *not-sexist*.

Unlike Assignment 1, which utilized fine-tuned BERT-based architectures, this study explores the efficacy of generative Large Language Models (LLMs) via prompting. We aim to assess how well general-purpose models can adapt to specific safety classification guidelines without parameter updates (fine-tuning). We compare two distinct model sizes to evaluate the trade-off between computational resources and reasoning capability.

2 System Description

2.1 Models and Quantization

We selected two popular open-source models available on the HuggingFace Hub:

- **Mistral-7B-Instruct-v0.3**: A 7-billion parameter model known for strong instruction-following capabilities.
- **TinyLlama-1.1B-Chat-v1.0**: A smaller 1.1-billion parameter model designed for efficiency.

To accommodate hardware constraints (single GPU), both models were loaded using 4-bit quantization (NF4 format) via the `bitsandbytes` library. This reduced memory usage while maintaining inference capabilities.

2.2 Prompting Strategies

We employed a structured chat template for all interactions:

- **Zero-shot:** The model was provided with a system instruction defining the role ("annotator for sexism detection") and the definitions of the five categories. No examples were provided.
- **Few-shot:** We injected 2 demonstration examples per class (10 examples total) into the context window to guide the model's output format and reasoning. These examples were sampled from the provided `demonstrations.csv`.

2.3 Inference Pipeline

The inference loop generated a maximum of 10 new tokens per sample with a temperature setting optimized for deterministic output. A post-processing function parsed the generated text to

map textual answers (e.g., "Answer: threats") to numeric labels (0-4). Responses that did not contain a valid category were marked as failures (label -1).

3 Experimental Setup

- **Dataset:** The models were evaluated on a balanced test set of 300 samples (`a2_test.csv`), containing an equal distribution of the five categories.
- **Metrics:** We report Macro F1-score, Accuracy, and Fail-Ratio (the percentage of generated responses that could not be parsed into a valid label).
- **Environment:** Experiments were run on a GPU-accelerated environment using PyTorch and HuggingFace Transformers.

4 Results

Table 1 summarizes the performance of both models across zero-shot and few-shot settings.

Model	Prompting	Macro F1	Acc	Fail-Ratio
Mistral-7B	Zero-shot	0.37	0.38	0.00
Mistral-7B	Few-shot	**0.45**	**0.46**	0.00
TinyLlama-1.1B	Zero-shot	0.08	0.20	0.95
TinyLlama-1.1B	Few-shot	0.07	0.20	0.21

Table 1: Classification performance comparison. Fail-Ratio indicates the proportion of invalid responses.

5 Discussion and Error Analysis

5.1 The Impact of Model Scale

Our experiments demonstrate a massive capability gap between the 7B and 1.1B parameter models.

- **TinyLlama's Collapse:** In the zero-shot setting, TinyLlama failed to follow the output formatting instructions in 95% of cases, often hallucinating new categories or continuing the prompt indefinitely. While few-shot prompting reduced the fail-ratio to 21%, the model collapsed into a trivial solution, predicting "not-sexist" for almost all inputs (Random Guessing), resulting in an F1-score of 0.07.
- **Mistral's Competence:** Mistral-7B followed instructions perfectly (0.00 fail ratio) in both settings. It achieved a baseline F1 of 0.37 in zero-shot, which is respectable given the task's complexity.

5.2 Few-Shot Learning Effectiveness

Consistent with literature (Brown et al., 2020), Mistral benefited significantly from in-context learning. Providing 10 demonstration examples improved the Macro F1-score from 0.37 to 0.45. The confusion matrices indicate that few-shot examples helped the model better distinguish between *animosity* and *threats*, likely due to the examples clarifying the specific "intent to harm" required for the *threats* category.

5.3 Class Confusion and Semantic Overlap

Despite the improvements, error analysis reveals persistent confusion between semantically similar categories:

1. **Animosity vs. Derogation:** Mistral frequently misclassified *derogation* as *animosity*. This is understandable as both involve negative sentiment towards women; *derogation* implies a belittling description, while *animosity* involves slurs. The model often defaulted to *animosity* for general insults.
2. **Prejudiced Discussion:** Mistral performed relatively well on identifying *prejudiced discussion* (F1 ≈ 0.62 in few-shot), suggesting it can detect ideological statements supporting mistreatment better than it can distinguish between specific types of slurs.

6 Conclusion

This assignment highlighted the limitations of smaller LLMs for multi-class classification tasks requiring strict adherence to complex definitions. TinyLlama-1.1B proved insufficient for this task without fine-tuning. Conversely, Mistral-7B demonstrated that with careful prompt engineering and few-shot demonstrations, open-source LLMs can achieve competitive performance (F1 0.45) compared to traditional supervised baselines, offering a viable path for data-efficient sexism detection.

References

[1] Kirk, H. R., et al. (2023). SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS).

[2] Jiang, A. Q., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

[3] Zhang, P., et al. (2024). TinyLlama: An Open-Source Small Language Model. arXiv preprint arXiv:2401.02385.

[4] Brown, T., et al. (2020). Language models are few-shot learners. NeurIPS.