

Assignment 2

NLP Course Project

Jacopo Francesco Amoretti, Roberto Frabetti, Ivo Rambaldi

Master’s Degree in Artificial Intelligence, University of Bologna

{ jacopo.amoretti, roberto.frabetti, ivo.rambaldi }@studio.unibo.it

Abstract

We evaluate Phi-3-mini-4k-instruct and Mistral-7B-Instruct-v0.3 on five-class sexism detection using zero- and few-shot prompting via a controlled template, with predictions mapped to labels.

On 300 test samples, Phi-3-mini better performs in zero-shot (Macro-F1=0.396), while Mistral-7B improves with few-shot (Macro-F1=0.439). Generally low fail-ratios show reliable instruction adherence. Confusion matrices reveal challenges distinguishing derogation, animosity, and prejudiced. Few-shot benefits only the larger model, indicating scale mediates demonstration utility.

1 Introduction

Note: as per instructions, this being an assignment report, the problem description and importance has been skipped

Sexism detection, key for online safety, demands distinguishing subtle harm. Traditional supervised methods need annotated data, often scarce, thus motivating zero/few-shot LLM prompting. Their reliability on multiclass taxonomies is unclear.

We test 300 samples across four configurations (2 models \times zero/few-shot). Zero-shot performs well for both; few-shot boosts Mistral-7B but hurts Phi-3-mini, showing capacity dependence.

2 Background

Unlike binary toxicity, sexism detection categorizes harm into subclasses (threats, derogation, animosity, prejudiced). Instruction-tuned LLMs generalize without fine-tuning, but reliability depends on prompts, size, and complexity.

Few-shot improves via examples, but smaller LLMs struggle while larger benefit from richer representations—motivating our cross-scale comparison.

3 System description

We evaluate large language models for sexism classification using a prompting-based approach. No task-specific fine-tuning is performed.

Two instruction-tuned causal language models are used:

- **Phi-3-mini-4k-instruct** (3.8B parameters),
- **Mistral-7B-Instruct-v0.3.**

Both models are loaded with 4-bit quantization (NF4) using `bitsandbytes` to reduce memory usage. Inference is performed via greedy decoding with low temperature ($T = 0.2$).

The task is framed as *instruction-following classification*. Each input text is embedded in a structured prompt defining five mutually exclusive labels (not-sexist, threats, derogation, animosity, prejudiced), with explicit rules enforcing a single-label textual output.

Two prompting regimes are evaluated:

- **Zero-shot:** the model receives only label definitions.
- **Few-shot:** two labeled examples per class are injected into the prompt.

All prompting logic, response parsing, and metric computation are implemented by us. Model architectures and weights are reused without modification.

4 Data

Experiments are conducted on the official Assignment 2 test set, consisting of 300 English social media posts. Each instance is annotated with one of five fine-grained sexism categories.

The label distribution is moderately imbalanced, with `not-sexist` as the majority class. Ground-truth labels are mapped to integer IDs for evaluation only; models operate purely on text.

For few-shot prompting, an auxiliary dataset of labeled demonstrations is provided. From this dataset, we uniformly sample two examples per class to construct in-context demonstrations.

No text preprocessing or filtering is applied beyond prompt formatting, to preserve the original linguistic signal. Dataset links are provided separately as required by the assignment.

5 Experimental setup and results

Each model is evaluated under zero-shot and few-shot prompting. Predicted textual outputs are post-processed and mapped to label IDs. Unparseable outputs are treated as errors.

Model	Macro F1	Fail Ratio
Phi-3-mini (zero-shot)	0.397	0.023
Mistral-7B (zero-shot)	0.367	0.013
Phi-3-mini (few-shot)	0.364	0.063
Mistral-7B (few-shot)	0.439	0.017

Table 1: Performance comparison across models and prompting strategies.

Zero-shot performance is comparable across models, with Phi-3-mini slightly outperforming Mistral-7B. Introducing demonstrations substantially improves Mistral-7B, while degrading Phi-3-mini, suggesting that smaller models may struggle to exploit in-context examples effectively.

6 Discussion

Quantitatively, the results highlight a strong dependence on model scale. While both LLMs follow the classification template consistently, only Mistral-7B benefits from demonstrations. Its few-shot Macro-F1 (0.439) surpasses all other configurations. Phi-3-mini, conversely, appears overloaded by additional context, leading to worse performance.

Error analysis (confusion matrices in the notebook) shows persistent confusion among derogation, animosity, and prejudiced, likely due to overlapping linguistic cues and subtle distinctions not easily separable without deeper semantic grounding. Mistral-7B handles prejudiced substantially better in the few-shot case, suggesting demonstrations help clarify abstract category definitions.

Typical errors include:

- Mislabeling derogatory insults as animosity,

- Treating ideological statements as non-sexist,
- Overgeneralizing examples seen in few-shot prompts.

7 Conclusion

This assignment shows that prompting alone can achieve reasonable performance on a fine-grained sexism classification task. Zero-shot performance is modest but consistent, while few-shot prompting significantly enhances the accuracy of larger models such as Mistral-7B. Smaller models like Phi-3-mini do not always benefit from additional context, underscoring the importance of model capacity.

Limitations include reliance on strict prompt formatting, confusion among semantically similar categories, and absence of controlled hyperparameter tuning. Future work could explore optimized demonstration selection, chain-of-thought guidance, or lightweight fine-tuning approaches such as LoRA to improve discrimination among difficult classes.

References