Assignment 2
NLP Course Project

Habib Kazemi, Hesam Sheikh Hassani, Ehsan Ramezani
Master's Degree in Artificial Intelligence, University of Bologna
{ habib.kazemi2, hesam.sheikhhassani, ehsan.ramezani }@studio.unibo.it

Abstract

This report covers our approach to tweet sexism detection using Large Language Models (LLMs). Two models, **Mistral-7B-Instruct-v0.3** and **Phi-3.5-mini-instruct** were tested under *zero-shot*, *two-shot*, and *four-shot* configurations. The findings indicate that **Mistral-7B-Instruct-v0.3** achieved the highest accuracy of **72%** in the two-shot configuration.

This study demonstrates the capability of LLMs to address sensitive tasks like sexism detection, where classification isn't black-and-white but involves subtle judgment, even with a limited number of examples.

1 Introduction

Traditional text classification methods, such as SVMs and RNNs, depend heavily on large, annotated datasets. LLMs have transformed natural language processing by capturing rich context and linguistic nuances, enabling effective performance even with limited labeled data. These models can be adapted through fine-tuning or prompting; in this work, we focus on prompting.

This study evaluates the performance of two open-source LLMs, Mistral-7B-Instruct-v0.3 and Phi-3.5-mini-instruct, for sexism detection (MistralAI) (Microsoft). The models were tested using a fixed prompt template in zero-shot, two-shot, and four-shot. Few-shot learning was used to enhance performance with minimal labeled examples.

Experiments were conducted on a smaller subset of the EDOS Task A dataset and across six configurations, the models' classification accuracy and fail ratio were assessed.

The results indicate that two-shot prompting provided the best accuracy for both models. Interestingly, accuracy decreased in the four-shot configuration for both models and was even lower than zero-shot for Phi-3.5-mini-instruct. Zero-shot prompting showed a bias toward classifying inputs as sexist, which was significantly reduced with few-shot setups.

2 System description

The system consists of four main components:

1. **Model Initialization:** Models were downloaded from HuggingFace and quantized for 4-bit precision to optimize inference in hardware-limited environments. (HuggingFace) They were used without further modifications.

2. **Prompt Setup:** A fixed prompt template was used for all experiments. In the few-shot setups, balanced examples of "sexist" and "not sexist" labels were randomly selected from the demonstration dataset and included in the prompt. Specifically, the two-shot setup included two examples per label (four examples in total), while the four-shot setup included four examples per label (eight examples in total).

3. **Inference Pipeline:** A custom pipelines processed input data, tokenized, generated responses, and parsed responses into sexist or not sexist labels.

4. **Evaluation:** Model outputs were evaluated against ground truth labels using accuracy, F1-score, and fail ratio.

3 Experimental setup and results

Quantization was applied to optimize inference performance. The models were loaded in 4-bit precision, using double quantization to reduce memory usage while maintaining numerical accuracy. Batch sizes of 8 or 16 were used to enhance inference efficiency.

The evaluation was performed on a subset of the EDOS Task A dataset, comprising 300 balanced

samples. Model performance was assessed using Accuracy (percentage of correct predictions), Fail Ratio (proportion of responses that did not follow the prompt or failed to respond), and F1-score.

The performance of the models is summarized in the table below:

Shots	Accuracy		Fail Ratio	
	Mistral	Phi	Mistral	Phi
Zero Shot	0.59	0.590	0.003	0.0
Two Shot	0.72	0.650	0.000	0.0
Four Shot	0.70	0.586	0.000	0.0

Table 1: Performance of Mistral-7B-Instruct-v0.3 and Phi-3.5-mini-instruct on the EDOS Task A dataset.

4 Discussion

Accuracy Analysis: Two-shot inference delivered the best performance for both models. Mistral achieved the highest accuracy of 0.72 in two-shot inference, while four-shot led to a slight decrease to 0.70, remaining higher than zero-shot (0.59). Phi3 performed best in two-shot inference with 0.65 accuracy, but four-shot dropped to 0.586, even lower than zero-shot (0.59).

Detection Analysis: Zero-shot inference detected the highest true positives (sexist), but two-shot and four-shot achieved better balance with higher true negatives (not sexist). Both models had the highest false positives in zero-shot, indicating a bias toward "sexist," which was reduced in two- and four-shot setups. False negatives increased in few-shot setups, particularly for Phi3.

F1-Score Analysis: Mistral achieved its highest F1-score of 0.75 in two-shot inference, while Phi3 peaked in zero-shot (0.70) but declined in few-shot configurations (two-shot: 0.65, four-shot: 0.59).

Best Configuration: Mistral's two-shot inference proved to be the most effective, achieving the highest accuracy (0.72) and F1-score (0.75).

Error Analysis

Formatting Issues:
Text: "Typical praising of women and bashing of men. They're so predictable. And dumb."
True Label: Sexist
Zero-shot (Mistral): Correctly classified but output as "Yes" instead of "YES."
Few-shot: Correctly classified as "YES."
This example was the only instance where the Mistral model failed to follow the instructions, unlike Phi3, which adhered to the format. This highlights zero-shot limitations in maintaining strict output formatting, a challenge resolved in the few-shot setup through provided examples.

Misclassifications:
Text: "Don't think that using the word 'feminine' wasn't strategic, Mo... men are a gift from God."
True Label: Sexist
Zero-shot: Correctly classified as "Sexist."
Few-shot: Misclassified as "Not sexist.", likely due to the examples influencing the models toward "Not sexist."
Text: "Old man mogs everyone in this sub."
True Label: Not sexist
Zero-shot: Misclassified as "Sexist."
Few-shot: Correctly classified as "Not sexist."
This exemplifies the observed tendency in zero-shot inferences to classify inputs as "Sexist."

Future work should focus on carefully selecting few-shot examples and analyzing how the choice of different examples impacts classification performance and biases.

5 Conclusion

This study evaluated the performance of two pre-trained large language models, Mistral-7B-Instruct-v0.3 and Phi-3.5-mini-instruct, on sexism detection using zero-shot, two-shot, and four-shot prompting. Two-shot inference consistently provided the best results, with Mistral achieving the highest accuracy (0.72) and F1-score (0.75). However, by increasing the number of examples in four-shot configurations the performance reduced. In this work we chose few shot examples by random, future work should explore how different examples impact performance.

References

HuggingFace. Quantization concept guide. https://huggingface.co/docs/optimum/concept_guides/quantization. Accessed: January 2, 2025.

Microsoft. Phi-3.5-mini-instruct. https://huggingface.co/microsoft/Phi-3.5-mini-instruct. Accessed: January 2, 2025.

MistralAI. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed: January 2, 2025.