Assignment 2

Mattia Buzzoni, Mirko Mornelli and **Riccardo Romeo**
Master's Degree in Artificial Intelligence, University of Bologna
{ mattia.buzzoni, mirko.mornelli, riccardo.romeo }@studio.unibo.it

Abstract

Comparative analysis of instruction based popular models for sexist text classification: evaluating Phi3-mini and Mistral v3 models across two English datasets, with a focus on Zero-Shot and Few-Shot as prompting techniques.

1 Introduction

This work analyzes the performance of two instruction-based open-source models in classifying English texts as sexist or not. We implemented Zero-Shot and Few-Shot inference methods, testing Few-Shot with both two and four demonstrations to evaluate sensitivity to the number of examples.

Our findings indicate that both models performed worse in the Zero-Shot scenario, with Phi3-mini slightly outperforming Mistral-v3 in this context. However, in Few-Shot inference, Mistral-v3 surpassed Phi3-mini, although increasing examples significantly improved Phi3-mini's performance, while the effect on Mistral-v3 was less pronounced.

Surprisingly, the model with more parameters did not significantly outpace the one with fewer. Both models struggled with the Assignment 2 dataset but performed better on the Assignment 1 dataset.

2 System description

We downloaded the dataset and converted it into Pandas DataFrames. We created three datasets: one for demonstrations and two containing texts to be injected into the models. Notably, one of these datasets was derived from the texts used in Assignment 1.

Subsequently, we selected two models to download from Hugging Face: Phi3-mini (Marah Abdin et al., 2024) and Mistral v3 (Albert Q. Jiang et al., 2023). This choice was made to ensure we had a model with a smaller number of parameters alongside one with a larger parameter set.

3 Experimental setup and results

We evaluated both models using Zero-Shot and Few-Shot Inference, measuring metrics such as Accuracy, macro-averaged F1-Score, and Fail Ratio.

To accomplish this, we utilized the prompt templates presented in the Assignment, tokenizing them with the appropriate tokenizer for each model. The Few-Shot examples were randomly selected from the demonstration dataset.

Specifically, we tested the two models in Few-Shot Inference with varying numbers of examples to determine if an increased number of examples would enhance performance. Initially, we used two demonstrations, followed by four.

We also assessed the models on the dataset containing tweets from the first Assignment to compare their behaviors across different datasets. It's important to note that we employed the same demonstration dataset for Few-Shot Inference in both cases to ensure a fair comparison. Additionally, we constructed the dataset from Assignment 1 to match the number of samples in the dataset from Assignment 2 for the same reason.

Finally, we organized the results into tables, plots, and confusion matrices to facilitate a comparison between the models across both inference scenarios and datasets.

# demo	Phi3-mini		Mistral v3	
	Acc.	F1	Acc.	F1
2	0.64	0.61	0.69	0.70
4	0.69	0.70	0.69	0.70

Table 1: Models performances in terms of Accuracy and F1-macro, evaluated on A2 dataset during Few-Shot.

4 Discussion

In all our experiments, both models consistently achieved a null Fail Ratio, indicating they never failed to complete the tasks.

Dataset	Prompt tech.	Phi3-mini					Mistral v3				
		Acc.	F1	FP	FN	TOT	Acc.	F1	FP	FN	TOT
A1 dataset	Zero-Shot	0.73	0.72	0.31	0.19	82	0.75	0.75	0.31	0.18	74
	Few-Shot	0.77	0.77	0.22	0.28	70	0.76	0.76	0.21	0.29	71
A2 dataset	Zero-Shot	0.59	0.52	0.40	0.01	123	0.56	0.45	0.44	0	132
	Few-Shot	0.67	0.65	0.22	0.087	91	0.70	0.69	0.39	0.0067	119

Table 2: Models performances in terms of Accuracy, F1-Macro, normalized False Positives (FP), normalized False Negatives (FN) and the total number of missclassifications (TOT) evaluated on A1 and A2 datasets.

When testing the models with Few-Shot prompting on the Assignment 2 dataset, we observed that Mistral-v3 showed limited sensitivity to the increase in the number of demonstrations within the prompt. It performed similarly in both scenarios, while Phi3-mini demonstrated improved performance with more examples (see Table 1).

Our experiments revealed that Few-Shot inference consistently yielded better results compared to Zero-Shot inference in terms of Accuracy, F1-score, and total misclassifications (see Table 2).

Notably, with Few-Shot inference, the normalized number of False Positives decreased for both models, regardless of the dataset used. In contrast, the normalized number of False Negatives exhibited a complementary trend.

Overall, both models performed better on texts from the Assignment 1 dataset compared to those from Assignment 2, as evidenced by improvements in Accuracy, F1-score, and the total number of misclassifications.

Interestingly, Few-Shot inference led to significantly better results with the second dataset, whereas its effectiveness was less pronounced with the first dataset.

Upon reviewing several texts from the Assignment 2 dataset, we found them more challenging to classify without contextual information. This is evident in the false positives generated by the models. For instance, some texts that could be considered genuinely sexist were labeled by annotators as non-sexist. An example of this is: *"Import the third world and become the third world. You won't see feminists protesting about this. [URL]"*.

5 Conclusion

In this work, we conducted an evaluation of two models, Phi3-mini and Mistral-v3, using both Zero-Shot and Few-Shot inference on datasets derived from two Assignments. Our findings indicated that both models consistently achieved a null Fail Ratio, demonstrating their reliability in task completion. Notably, Few-Shot inference outperformed Zero-Shot across all metrics, including Accuracy and F1-score, with Phi3-mini showing significant improvement as the number of examples increased. However, Mistral-v3's performance remained stable regardless of the number of demonstrations, which was unexpected.

Despite these promising results, several limitations were identified. The models exhibited challenges in accurately classifying texts from the Assignment 2 dataset. Suggesting to use a more representative demonstration dataset during Few-Shot inference. Additionally, while Few-Shot inference proved beneficial, its effectiveness varied between datasets, highlighting the greater complexity of the second dataset.

Future work could explore employing the Chain of Thoughts prompting technique to enhance performance on both datasets. Moreover, trying to increase the number of examples in Few-Shot inference could be interesting. Additionally, experimenting with a more representative demonstration dataset may yield further improvements.

6 Links to external resources

- Phi-3-mini-4k-instruct
- Mistral-7B-Instruct-v0.3

References

Arthur Mensch Albert Q. Jiang, Alexandre Sablayrolles et al. 2023. Mistral 7b. *arXiv*.

Hany Awadalla Marah Abdin, Jyoti Aneja et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*.