NLP Assignment 2

Francesco Baiocchi, Christian Di Buò and **Leonardo Petrilli**
Master's Degree in Artificial Intelligence, University of Bologna
{francesco.baiocchi2, christian.dibuo, leonardo.petrilli}@studio.unibo.it

Abstract

This work addresses the SemEval 2023 Task-10 on binary sexism detection, focusing on classifying tweets as sexist or not sexist using instruct LLMs and prompt engineering techniques. Our analysis focuses on comparing the performance of two LLMs—Phi3 (developed by Microsoft) and Mistral v0.3 (developed by Mistralai)—using different prompting techniques. The results show how the models' performance increases when examples are provided and illustrate how these examples affect the quality of predictions.

1 Introduction

Online sexism can inflict harm on people, especially women, who are often targeted. Automatically detecting whether a sentence contains sexist references is an increasingly important task with applications in multiple fields, such as law and social media. SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis. To address online sexism issue, (Kirk et al., 2023) introduce SemEval Task 10 on the Explainable Detection of Online Sexism. Numerous approaches to this problem have been proposed, many of these rely on Transformer-based fine-tuned architectures such as RoBERTa, which has achieved impressive results (Vallecillo-Rodríguez et al., 2023).

We have tested the performance of Phi-3-mini-4k-instruct and Mistral-7B-Instruct-v0.3, two decoder-only transformer-based architectures, in recognizing whether a sentence contains sexist references. Our work explores different prompting techniques, specifically Zero-Shot, Few-Shot, and Chain-of-Thought.

2 System description

Our tests were run on two models taken from Hugging Face, namely Phi-3-mini-4k-instruct and Mistral-7B-v0.3. Both models are transformer decoder-only architecture, in particular:

- Phi3 mini (Abdin et al., 2024), developed by Microsoft, is the smallest version of the Phi3 family, with 3.8 Billion parameters. It has a vocabulary size of 32064 and uses a hidden dimension of 3072, with 32 heads and 32 layers. It was trained on a total of 3.3T tokens.

- Mistral v0.3 7B (Jiang et al., 2023), developed by MistralAI, has instead 7 Billion parameters and, with respect to the previous versions, it has an extended vocabulary of 32768.

Both models are already chat fine-tuned and the chat template is: "<|user|> \n Question <|end|>\n <|assistant|>".

3 Experimental setup and results

The pipeline of our experiments is as follows. First, we load two datasets: a2test and demonstrations. The former contains the sentences to be classified, while the latter provides example sentences for the few-shot prompting technique. Next, we instantiate a 4-bit quantized version of each model along with its respective tokenizer. Finally, we configure the models so that they respond with a single sentence containing at most 100 tokens.

The prompt template used was designed to simulate an annotator tasked with detecting sexism in text. Specifically, the prompt structured the model's role as that of an annotator who must classify input text as either containing sexism or not. The prompt explicitly constrained the model's response to only two options: YES or NO. Prompting requires an input pre-processing phase where we convert each input sentence into a specific instruction prompt. For the few-shot prompting technique, we randomly sampled two sentences from

	accuracy_score	fail_ratio	f1-score
Phi3_zero_shot	0.643	0	0.642
Phi3_few_shot	0.67	0	0.668
Phi3_CoT	0.73	0	0.723
Mistral_zero_shot	0.59	0	0.516
Mistral_few_shot	0.697	0	0.679
Mistral_CoT	0.647	0	0.608

Table 1: Prediction metrics on the test set

the demonstrations dataset and used those two sentences to format all prompts. The model then generated an answer to the provided prompt. Then we let the model generate an answer to the provided prompt. The models' performances were evaluated using fail-ratio and accuracy as metrics. The results are shown in table 1.

4 Discussion

We observed that the models performed very well in terms of the fail ratio; indeed, all the answers were either YES or NO, as required by the prompt. Furthermore, both models showed improvements in the few-shot prompting setup. This was an expected result, as providing examples should help the model achieve a better understanding of the context. When comparing the models' performances, we found that the Phi3 model performs better than Mistral v0.3 in the zero-shot setup, despite having about half the number of parameters compared to its counterpart (3.8B vs. 7B). However, in the few-shot prompting setup, Mistral shows a significant improvement, surpassing Phi3. This substantial improvement could be attributed to Mistral's higher number of parameters, which likely helps the model better understand the examples provided.

Furthermore, we observed that the errors made by Phi3 are balanced between sexist and non-sexist classifications, whereas Mistral shows a stronger bias toward categorizing sentences as sexist. This behaviour is more pronounced in the zero-shot prompting setup, with slight improvement in the few-shot setup.

To better understand the errors made by the models, we looked for incorrect answers that differed from a simple 'Yes' or 'No'. While Phi3 does not produce such answers, Mistral sometimes provides responses followed by explanations. Although these explanations are not entirely incorrect, they suggest that the model struggles to grasp the implicit undertone of the sentences. An example of such answer is the following one:

- This statement objectifies women and reduces them to their physical appearance, which is a form of sexism.

Next, we analysed the correlation between the model performances and the examples injected in the prompts during the few-shot technique. As shown in Figure 1, varying the examples leads to different results. To address this issue, we decide to use random examples for each prompt in every run. This approach helps reduce fluctuations in the models' performance. Furthermore, we explored the connection between model performance and the number of sentences provided as examples, aiming to determine whether there is a correlation between performance and the number of given examples. Our findings indicate that the performance is not directly related to the number of samples.

We attempt to improve the models' performance using the Chain of Thought technique, which elicits reasoning in Large Language Models (Wei et al., 2022). We provide the models with some general rules to guide their reasoning on the key points for deciding how to classify a sentence. As shown in Table 1, both models show improvements.

5 Conclusion

The few-shot prompting technique proves to be the most effective for this type of task. The models show a strong correlation between performance and the type of examples provided. This highlights that, to significantly improve model performance, it may be necessary to study which examples should be included in the prompt. Alternatively, using different examples for each prompt could help achieve more consistent and less biased performance. The Chain of Thought technique also has a positive impact on model performance.

Furthermore, experiments on a smaller dataset (CLEF EXISTS Task 1, 2023) lead us to conclude that fine-tuning is essential for these tasks. Fine-tuning allows the use of "lightweight" models, which outperform larger LLMs, a theory supported by recent research (Bucher and Martini, 2024) and confirmed by our tests.

A possible future case of study could be understanding which examples characteristics increase the model performances and combine them with the chain of thought technique, giving an explanation of the examples instead of only the correct answer.

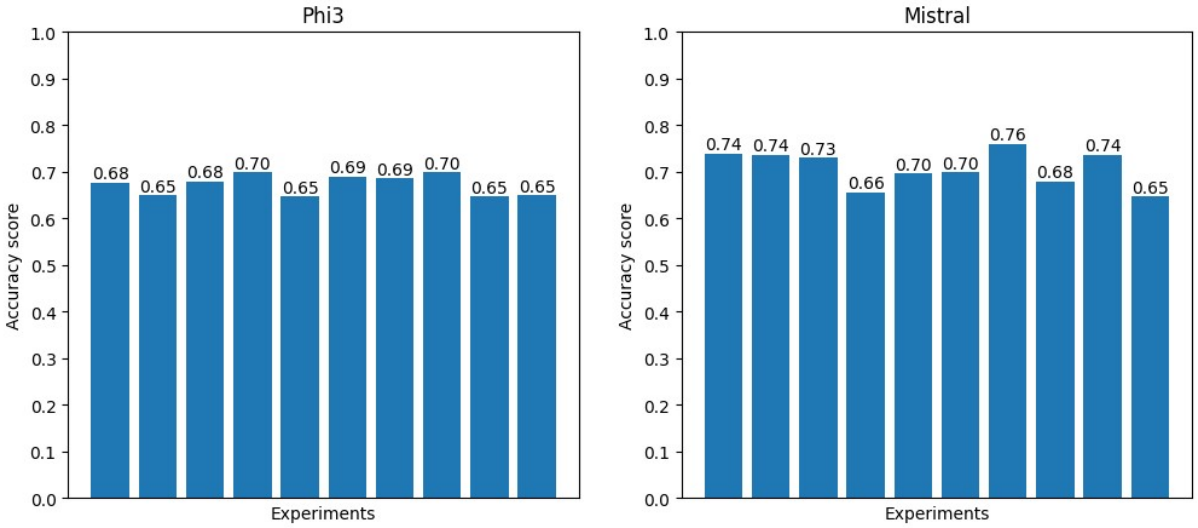

Figure 1: Models performances variability changing the examples per class

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

CLEF EXISTS Task 1. 2023. EXIST 2023 - Explainable Sexual Misogyny Identification Task.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.

María Estrella Vallecillo-Rodríguez, Flor Miriam Plaza del Arco, Luis Alfonso Ureña-López, María Teresa Martín-Valdivia, and Arturo Montejo-Ráez. 2023. Integrating Annotator Information in Transformer Fine-tuning for Sexism Detection. [Online; accessed 8-Jan-2025].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.