Assignment 2

Mohammad Pourtaheri, Fatemehzahra Ghafari Ghomi, Shafagh Rastegari and **Mehregan Nazarmohseni Fakori**
Master's Degree in Artificial Intelligence, University of Bologna
{ Mohammad.pourtaheri, fatemezahra.ghafari, shafagh.rastegari, mehrega.nazarmohseni }@studio.unibo.it

Abstract

The objective of this project was to classify input text sentences as either sexist or not sexist, addressing a binary classification task. To tackle this problem, zero-shot and few-shot learning approaches were explored using three open-source large language models (LLMs): Mistral v3, Llama v3.1, and Phi3-Mini.

1 Introduction

This project addresses the binary classification of text sentences as sexist or not sexist using zero-shot and few-shot learning with three LLMs: Mistral v3, Llama v3.1, and Phi3-Mini. Challenges such as sensitivity to input structure, prompt complexity, and overfitting were observed, particularly in Llama 3.1 and Phi3-Mini. To improve accuracy and efficiency, techniques such as intelligent example injections, modified example distributions, ensemble models, and reduced tokenization overhead were applied, highlighting the potential and limitations of LLMs for this task.

2 System description

- Zero-shot Learning: Two prompts were utilized: the original provided in the assignment and a modified version. Llama 3.1 struggled with the original prompt but performed well with the modified version. Conversely, Phi3-Mini only performed well with the original prompt, failing to work with the modified one. Mistral v3 effectively handled both prompts.

- Few-shot Learning: The original prompt was employed for few-shot learning and successfully generated responses across all three models.

3 Experimental setup and results

Two mentioned approaches applied to three models and we observed in few-shot inference adding more examples did not improve accuracy for Llama 3.1 and Phi3-Mini, reflecting their sensitivity to input structure and prompt length. In contrast, Mistral v3 showed promising accuracy gains with additional examples. However, comparing the models' best few-shot inference results with their zero-shot results shows that, at best, their performance remains the same. In this way different methods are applied for trying to improve the accuracy and time efficiency.

- we ran few-shot inference on 300 data points, gradually increasing the number of balanced examples from 2 to 14. We also used indices that were misclassified in the initial runs. The result was a gradual increase in the number of incorrect predictions, due to a more complex prompt.

- we tried to experiment with different example distributions.

- more intelligent example injections were applied in few-shot inference. To implement this, three methods were employed to identify similar samples: (1) TF-IDF, (2) Word2Vec, and (3) a BERT-based model. Overall, by employing better example injections, where the examples are more closely related to the text being examined, both accuracy and time efficiency can be optimistically improved.

- the normal voting ensemble method, the weighted ensemble method, and the meta-ensemble method (logistic regression). Among these, the meta-ensemble model demonstrated the best results in our tests. However, it did not yield any significant improvement in accuracy.

Model	Best zero shot	Best few shot
LLaMA	67	56
Phi-3	61	61
mistral	71	71

Table 1: Comparison performance between zero-shot prompting and few-shot prompting.

- it was hypothesized that reducing tokenization overhead could decrease time consumption. This hypothesis was based on the observation that certain constant sections of each prompt were repeatedly tokenized for every injection. To address this, the constant sections were tokenized once, and in subsequent tokenization processes, only the variable parts of the prompt were tokenized.

4 Discussion

In this study, we examined the performance of three large language models—LLaMA-3.1, Phi-3, and Mistral—in classifying texts as sexist or not-sexist using both zero-shot and few-shot prompting strategies. It was observed that increasing the number of examples in few-shot inference did not enhance accuracy for LLaMA-3.1 and Phi-3. Additionally, a comparison of the models' best few-shot results with their zero-shot performance revealed that their accuracy, at best, remained unchanged. The results are summarized in 1.

Additionally, intelligent example selection methods (BERT-based, Word2Vec, TF-IDF) were tested, revealing that BERT embeddings marginally improved accuracy for LLaMA-3.1 and Phi-3 but did not benefit Mistral. However, this approach incurred a substantial time cost, as shown in 2. Notably, certain prompt formats prevented Phi-3 and LLaMA from generating outputs, highlighting the models' sensitivity to prompt style and design.

The models were also evaluated on the A1 dataset (from Assignment 1), where a best accuracy of 77% was achieved, compared to the 83% accuracy obtained in Assignment 1. This result suggests that, for simpler datasets, better performance may be achieved with transformer-based models rather than prompt-based approaches.

5 Conclusion

This project demonstrated the potential and limitations of optimizing LLMs for binary classification. While methods like BERT-based retrieval and TF-IDF offered insights, none improved both accuracy and time efficiency simultaneously. Challenges such as model bias and sensitivity to prompt length further impacted results. Future work should explore varying parameters like temperature, increasing num_return_sequences, employing beam search, and testing continuous prompts to enhance accuracy and contextual alignment.

Moreover, The analysis revealed a significant bias in the model's predictions, favoring the "sexist" class. False positives were observed to be approximately 3 to 10 times higher than false negatives, with misclassifications for "not-sexist" texts ranging from 250–400, compared to 10–60 for "sexist" texts. This bias became more pronounced when class imbalance was introduced during few-shot prompting, such as with a 6:1 ratio of "not-sexist" to "sexist" examples, leading to an increase in incorrect "not-sexist" predictions. These findings highlight the importance of addressing class imbalance and improving model robustness in future work.

Model	TF-IDF	Word2vec	Bert Based	Previous result
mistral(7:7)	68% (9min)	67% (11min)	69% (41min)	70% (9min)
LLaMA(1:1)	56% (4min)	56% (6min)	58% (36min)	56% (4min)
Phi-3(1:1)	64% (3min)	62% (5min)	66% (33min)	64% (7min)
Average	62% (5min)	61% (7min)	64% (36min)	63% (6min)

Table 2