# NLP: Assignment 2

**Alberto Genovese, Simone Reale,** and **Alessandro Tutone**
Master's Degree in Artificial Intelligence, University of Bologna
{ alberto.genovese5, simone.reale3, alessandro.tutone }@studio.unibo.it

## Abstract

This assignment aims to address the **EDOS Task A** (*Explainable Detection of Online Sexism*) on sexism detection. Given an input text sentence, the task is to label the sentence as *sexist* or *not sexist*, namely we have to perform a binary classification task. Unlike the previous assignment, this one requires a different approach. The binary classification is tackled using two different LLMs: ***Mistral v3*** and ***Phi3-mini***. In order to categorize a sentence as *sexiest* or *not sexiest* we used two prompting approaches: *zero-shot prompting* and *few-show prompting*. The assignment requires to use the two prompting approaches with both the two LLMs elected; for each model, an evaluation and an error analysis is carried out to assess the performance of each model with each prompting approach.

## 1 Introduction

Sexism is a widespread issue in online social networks. While human supervision has traditionally been the primary approach to deal with this problem, thanks to the rapid advancement of artificial intelligence, the detection of such harmful contents can now be automatized. Several models such as *SVMs*, *RNNs*, *LSTMs*, and much more, are able to perform classification in various contexts. In the previous assignment, we tackled this problem by using *bidirectional LSTMs*; now the task is accomplished by using prompt engineering. *Zero-shot* and *few-shot* are the two prompting techniques employed. Since we are addressing a binary classification task, the two LLMs are required to tell if a given sentence is *sexist* or *not sexist* without adding any explanation to their predictions. The presence or absence of demonstrations, inside the prompt template, leads models to behave differently and accordingly to have different performance. The labels are not always correct; indeed, the nuanced nature of sexism, often expressed subtly or implicitly, arises further challenges for automated detection. To address these limitations, a combination of artificial intelligence and human expertise is essential. Human moderators can provide valuable feedback to refine models and ensure accurate detection of sexist content.

For each model evaluated (***Phi3-mini*** and ***Mistral v3***), metrics were used to assess the performance achieved for both the *zero-shot* and *few-shot prompting* approaches. A final evaluation table was used to compare and analyze the results, to determine which model and which prompting approach better fit the task.

## 2 System description

The two LLMs used were taken from *Hugging Face* (HuggingFace, 2024), a machine learning and data science platform and community. Both the two models follow a similar pipeline: given the two *model_cards*, the pre-trained version is downloaded as well as the tokenizer. Considering the huge dimensions of the LLMs, quantization is performed to allow both models to fit into a single GPU. Doing so, memory and computational costs are reduced by representing weights and activations with lower-precision data types. Once the models are loaded, two prompt templates are used to address the two different prompting techniques: *zero-shot* and *few-shot prompting*. To address task A, all tweets had to be pre-processed in order to make them fit into the given prompts: for the *zero-shot prompting* each tweet was just inserted in the given space after the task description, on the other hand, for the *few-shot prompting* an additional step was required to insert some demonstrations following the task description. Once the prompts were ready, the tokenizer for each model was used first to encode the sentence and then to decode the answer given by the LLM. Raw responses (namely, text responses), had to be converted into numbers allowing the evaluations to be carried out.

## 3 Experimental setup and results

The two LLMs taken from *Hugging Face* are ***Mistral v3***, a pre-trained generative text model with 7 billion parameters, and ***Phi3-mini***, a 3.8 billion parameters, lightweight, state-of-the-art open model trained with the Phi-3 datasets. In particular, both models were an instruct fine-tuned version of the their baseline model.

The *zero-shot* technique was carried out using the assignment 2 template, whereas for the *few-shot* technique, since a small degree of freedom was given, it has been decided to use 4 demonstrations. The following results were achieved with the two models and the two prompting techniques:

| LLM | zero-shot | few-shot |
|---|---|---|
| **Mistral v3** | *accuracy*<br>0.59<br>*fail-ratio*<br>0.01<br>*f1-score*<br>0.52 | *accuracy*<br>0.73<br>*fail-ratio*<br>0.02<br>*f1-score*<br>0.72 |
| **Phi3-mini** | *accuracy*<br>0.64<br>*fail-ratio*<br>0<br>*f1-score*<br>0.64 | *accuracy*<br>0.67<br>*fail-ratio*<br>0<br>*f1-score*<br>0.65 |

Table 1: Models results

## 4 Discussion

The results obtained are not completely satisfactory compared to those reached with previous models: *Transformer-based* and *Bidirectional LSTM*. LLMs have demonstrated remarkable success in various NLP tasks, largely due to their ability to address problems without explicit training data.

Given their strong performance in many NLP contexts, before trying to do the assignment, we hypothesized that LLMs might also prove effective in sentence classification. In the previous assignment, sexism detection was successfully tackled using models that produced interesting results; this success raised our expectations for applying LLMs to the same task. However, it is important to note that these models were trained for general purpose use, so when implied for a specific task, such as detecting sexism, they tend to perform worse, especially without a fine-tuning phase. Although the use of prompting techniques optimizes the performance of the models, their results do not come close to those of the models used in assignment 1. With zero-shot prompting, the results obtained were barely satisfactory. Slight improvements were observed with *few-shot prompting*, achieved by including four examples in the prompt. Despite these improvements, several errors were observed in the LLMs analyzed.

The *Mistral* model tends to classify non-sexist tweets correctly, but struggles significantly with sexist ones. This is particularly concerning, as accurately identifying sexist content is the primary objective. While the accuracy is better than random guessing, the model's performance is inadequate for this purpose.

The *Phi3* model shows a more balanced distribution of correct classifications, although non-sexist tweets are still classified more accurately. A notable difference between zero-shot and few-shot prompting lies in the distribution of correct predictions: when demonstrations are added, the model tends to classify more non-sexist tweets correctly. However, this improvement comes at the expense of performance in identifying sexist tweets, making the model less effective for our specific task.

Lastly, our analysis explored the most common words in misclassified texts. For both types of errors (False Positive/Negative), the frequent words were quite similar. This suggests that the model does not exhibit a strong bias toward particular words in the text when making misclassifications.

As a final extension of this work, sexism detection was applied to *Assignment 1 tweets* using the same models. A comparison was conducted to examine the differences between the two datasets and to evaluate how these variations affected the model's performance.

## 5 Conclusion

Based on the performance achieved, LLMs cannot yet be considered a reliable alternative to *transformer-based* or *bidirectional LSTM* models for addressing sentence classification tasks, such as sexism detection. While in-context learning can enhance the performance of LLMs, their results still are not comparable to those of fine-tuned models. Although future advancements in LLMs will undoubtedly improve their capabilities in sentence classification, for now, more established methods remain the better choice for such tasks.

## References

HuggingFace. 2024. https://huggingface.co.