Assignment 2

Daniele Algeri, Martina Kenna, Joana Pimenta and **Matilde Simonini**
Master's Degree in Artificial Intelligence, University of Bologna
{ daniele.algeri, martina.kenna, joana.monteiro, matilde.simonini }@studio.unibo.it

Abstract

This work compares two open-source LLMs, *TinyLlama-1.1B-Chat-v1.0* and *Mistral-7B-Instruct-v0.3*, on multiclass sexism detection using zero-shot and few-shot prompting. Results highlight a critical dependency on model size: TinyLlama consistently failed to follow instructions in both inference settings and, even after Prompt Tuning, semantic reasoning was still absent, leading to a single-class prediction bias, although formatting errors were corrected. In contrast, Mistral-7B demonstrated effective instruction adherence. Although initially biased towards one class, Mistral's performance improved significantly with few-shot examples, raising the F1-score from 0.35 to 0.46.

1 Introduction

Traditional approaches for text classification often rely on supervised learning with models like BERT (Jacob Devlin, Ming-Wei Chang and Kenton Lee, 2018), which require extensive fine-tuning. Recently, Large Language Models (LLMs) have introduced the paradigm of "prompting," allowing models to perform classification tasks without weight updates, relying instead on in-context learning. In this work, we investigate the efficacy of LLMs in categorizing a set of texts into five distinct categories: *not-sexist*, or one among four sexist classes, namely *threats*, *derogation*, *animosity*, and *prejudiced discussion*. We compare a lightweight model (TinyLlama-1.1B) against a larger model (Mistral-7B). Our experimental setup involves classifying a balanced, supervised test set of 300 texts, using two different strategies: **Zero-Shot** prompting, where the model is given the task description and input text only, and **Few-Shot** prompting, where the model is provided with the task description and a small set of labeled examples.

In addition, we perform **Prompt Tuning** to TinyLlama to address its instruction-following limitations. Our experiment highlights that:

- The smaller model (TinyLlama, 1.1B parameters) lacks the capacity to follow the syntactic constraints out-of-the-box, and outputs invalid labels for every instance in the dataset.
- Prompt Tuning successfully forced such model to output valid labels, but could not induce the reasoning required for correct classification.
- Mistral tends to default to the *animosity* class in the zero-shot setting.

2 System description

We implemented an inference pipeline based on the Hugging Face `transformers` library. The pipeline for each example in the dataset consists of the following stages:

1. **Prompt Construction:** We use a strict template defining the role of the model, listing the five valid categories, and instructing the model to output *only* the label. The prompt consists of the instructions, the target text and, for the Few-Shot setting, the labelled examples.
2. **Inference:** The prompt is passed to the LLM.
3. **Output Parsing:** The raw text response is mapped to numerical IDs (0–4). If the model generates invalid text, the system defaults to **0** (*not-sexist*).

To address TinyLlama limitations, we implemented an additional Prompt Tuning pipeline for such model. This technique involves freezing the pre-trained model backbone and optimizing only a small set of continuous vector embeddings that are prepended to the input (corresponding to approx. 0.006% of parameters), aiming to align the model with the specific prompt structure.

3 Experimental setup and results

The performance is evaluated on the Macro F1 Score and the Accuracy Score. The following tables summarize the obtained metrics.

Table 1: **Zero-Shot** Inference Metrics

Model	F1-Score	Accuracy
Mistral	0.35	0.38
TinyLlama	0.07	0.20

Table 2: **Few-Shot** Inference Metrics

Model	F1-Score	Accuracy
Mistral	0.46	0.48
TinyLlama	0.07	0.20

It is crucial to note that TinyLlama's metrics are artifacts of the pipeline defaulting invalid outputs to class **0**. While Prompt Tuning successfully corrected the syntax, it resulted in a semantic collapse where the model explicitly predicted *not-sexist* for nearly all inputs. Consequently, the metrics remained identical, effectively shifting from a fallback-driven default to a learned prediction bias without improving discrimination.

4 Discussion

Our results show that Mistral achieves a respectable Zero-Shot performance which is further boosted by Few-Shot examples (F1 increase from 0.35 to 0.46). TinyLlama, conversely, shows static metrics (Accuracy 0.20, F1 0.07) across all setups, which match the statistical probability of a random classifier for a balanced 5-class dataset.

4.1 Error Analysis

The invalid outputs provided by TinyLlama, in both Zero-Shot and Few-Shot Inference, consist in long generative text rather than discrete class labels. A possible reason for this behavior is that TinyLlama-Chat is trained primarily for conversational settings, where verbose and explanatory answers are generally preferred. Even in-context demonstrations (Few-Shot) proved insufficient to enforce syntactic adherence, likely because this technique relies on a model's ability to perform in-context learning, which is known to scale strongly with model size.

Ultimately, the Prompt Tuning experiment was successful in making the model learn the syntax,

```
Text: "damn, i bet she likes tall dudes"
True Label: "not-sexist"
Predicted: "Sexist Categories Definition:
Threats: The text expresses intent [...]"
```

Figure 1: Example of invalid TinyLlama output

as all 300 dataset examples were classified with a valid label; nevertheless, the model still failed to learn the semantics, yielding poor performance. This dual failure—first in formatting, then in discrimination—strongly suggests that the 1.1B parameter size acts as a fundamental bottleneck for classification performance.

Concerning Mistral, the model presents a strong bias towards the *animosity* class in the Zero-Shot setting, which, however, was significantly mitigated by few-shot prompting, as shown by the two confusion matrices in Figure 2.

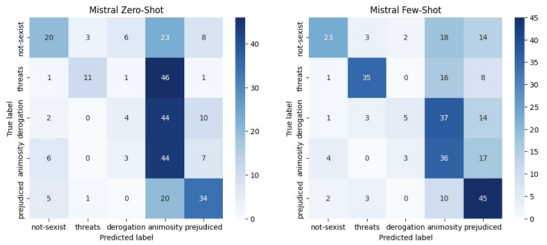

Figure 2: Confusion matrices for Mistral in Zero-Shot and Few-Shot Inference.

5 Conclusion

In this experiment, we evaluated the capabilities of LLMs for multiclass sexism detection. We observed that model size is a decisive factor not only for rigid instruction following (syntax) but also for the semantic understanding required to distinguish between complex classes. TinyLlama-1.1B proved unable to adhere to the output format in standard prompting; furthermore, even when Prompt Tuning successfully enforced syntactic compliance, the model failed in discriminating between categories, collapsing into a single-class prediction.

In contrast, Mistral-7B demonstrated that a sufficiently large model can effectively interpret constraints and leverage context. The transition from zero-shot to few-shot proved highly effective for Mistral, rectifying its bias toward *animosity* and allowing for better identification of specific categories like *threats*. This confirms that larger models possess the necessary capacity to refine decision boundaries via in-context learning—a capability that appears absent in the smaller 1.1B model.

References

Jacob Devlin, Ming-Wei Chang and Kenton Lee. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.