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

Text classification for toxic content detection is a critical task in Natural Language Processing (NLP). Traditional approaches 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 approximately 300 texts [ADD INFORMATION ABOUT DATASET/REFERENCE?] using two different strategies:

1. **Zero-Shot:** The model is given the task description and input text only.
2. **Few-Shot:** 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 strictly follow formatting constraints out-of-the-box.
- 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.
   - Zero-Shot: Instruction + target text.
   - Few-Shot: Instruction + labeled examples + target text.
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 performance metrics obtained during the prompting experiments, for Zero-Shot and Few-Shot Inference respectively.

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 in standard inference settings, TinyLlama's metrics are artifacts of the evaluation pipeline: the model consistently failed to produce valid labels, triggering a fallback to class 0 (*not-sexist*) which artificially matched the ground truth about 20% of the time. We utilized Prompt Tuning to address this specific syntactic failure. While the tuned model successfully learned to respect the output format (generating only valid integers), it exhibited a severe semantic collapse, explicitly predicting class 0 for nearly all inputs. Consequently, the final metrics remained effectively identical to the baseline; the model merely shifted from a fallback-driven default to a learned prediction bias, confirming its inability to discriminate between sexist categories regardless of formatting compliance.

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

In our tests, TinyLlama failed to follow the prompt instructions in the default, pre-tuning configuration, consistently producing syntactically invalid outputs, in both Zero-Shot and Few-Shot Inference.

SHOW EXAMPLE OF WRONG CLASSIFICATION

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 the inclusion of in-context demonstrations (Few-Shot prompting) 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, as all 300 examples in the test dataset were classified with a single, valid label; nevertheless, the model still failed to learn the semantics, leading to unsatisfying performance once again. This dual failure—first in formatting, then in discrimination—strongly suggests that the 1.1B parameter size acts as a fundamental bottleneck for the 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 1

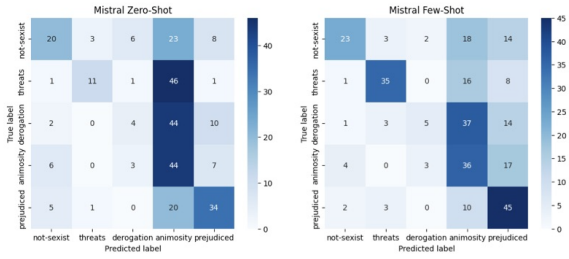

Figure 1: Confusion matrices for Minstral in Zero-Shot and Few-Shot Inference.

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 lacked the representational depth to discriminate 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 a strong bias toward generic "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 architecture.

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.