Assignment 2

Tommaso Bernardini , Samuele Gasbarro, Carmine Santella
Master's Degree in Artificial Intelligence, University of Bologna
{ tommaso.bernardini6, samuele.gasbarro, carmine.santella }@studio.unibo.it

Abstract

This report addresses the task of fine-grained sexism detection (EDOS Task B), which involves classifying text as non-sexist or assigning it to one of four specific sexist categories. We evaluate the zero-shot performance of three open-source LLMs, i.e. Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct and Qwen3-1.7B. Our methodology involves designing specific instruction prompts that define each category and processing the generated text to map model outputs to class labels. We measure performance using Macro F1-score and a Fail-Ratio metric to assess instruction-following capabilities. Our results indicate that while models like Mistral demonstrate a capability to classify sexism with a Macro F1 of approximately 0.34, other models such as Llama exhibit significant safety-filter triggers, resulting in high refusal rates that hinder performance on this specific downstream task.

1 Introduction

This report addresses **EDOS Task B (Explainable Detection of Online Sexism)**, a multi-class classification challenge. The goal is to categorize user-generated sentences into one of five classes: not-sexist, or if sexist, into four specific categories. Standard approaches to this problem typically involve **fine-tuning pre-trained language models like BERT (Devlin et al., 2019).** In this assignment, we focus on Large Language Models using **prompting techniques**, specifically zero-shot and few-shot approaches. This methodology allows us to evaluate the models' inherent reasoning capabilities without the need for a traditional training set. Our experimental setup involves testing **three open-source models** available on Hugging Face: **Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024) , and Qwen3-1.7B (Yang et al., 2025)**. To accommodate hardware limitations, all models were loaded using **4-bit quantization (Dettmers et al., 2023)**. We conducted experiments on a **test set of 300 samples**. For the few-shot prompting phase, we utilized a **separate batch of 1,000 demonstrations**, testing the models by providing an increasing number of examples per class (from 1 to 4) to measure the impact on performance. Our results highlight that **few-shot prompting significantly enhances performance**, with Llama-3.1 leading in zero-shot capabilities despite occasional generation failures. However, distinguishing between specific sexist sub-categories remains a challenge, as models frequently misclassify complex threats or prejudices as generic animosity.

2 System description

The system is built on the Hugging Face *Transformers* (Wolf et al., 2020) library and optimized for consumer hardware. We evaluate three distinct architectures: **Mistral-7B-Instruct-v0.3**, **Llama-3.1-8B-Instruct**, and the smaller **Qwen3-1.7B**. To adhere to memory constraints (Single T4 GPU), we employ **4-bit quantization** via the *BitsAndBytes* library. The configuration uses the NormalFloat4 (NF4) data type, double quantization, and `float16` compute dtype to balance efficiency and precision.

We utilize the EDOS Task B dataset, mapping inputs to five labels: *non-sexist*, *threats*, *derogation*, *animosity*, and *prejudiced discussion*.

To ensure reproducibility, inference is performed using **greedy decoding** (`do_sample=False`) with a strict limit of 20 new tokens. We handle model-specific requirements, such as disabling the "thinking" logic in Qwen's chat template and enforcing left-padding for Llama to maintain batch integrity. A post-processing module maps the generated text to class IDs via string matching, assigning a "fail" status (-1) to refusals or unparsable outputs.

3 Experimental setup and results

We implement multiple prompting strategies to evaluate in-context learning and reasoning capabilities.

Few-Shot Prompting: For the few-shot experiments, we retrieve demonstration examples from a separate dataset. We construct prompts by prepending k randomly selected context-label pairs ($k \in \{1, 4\}$) to the target query. Moreover, we also tried to use the Assignment 1 dataset to make zero-shot and few-shot inference. The results were remarkable (F1 ~0.7); we thought this could be due to the binarization of the labels.

Chain-of-Thought (CoT): To enhance reasoning for ambiguous cases, we modify the system prompt to require a logical deduction step (Wei et al., 2023). The model is instructed to generate a "Step-by-step reasoning:" block before predicting the final category.

Prompt Tuning (PEFT): As an alternative to static prompting, we employ Parameter-Efficient Fine-Tuning (PEFT) (Belanec et al., 2025). We freeze the base model weights and optimize a small set of virtual tokens (soft prompts) using the AdamW optimizer with a learning rate of $3e-4$ over 5 epochs. To evaluate performance, we employ two primary metrics, which are **F1-score** and **Fail-Ratio**.

Table 1 summarizes the performance of Mistral, Llama-3.1, and Qwen3 across all experimental settings.

Model	Methodology	Macro F1	Fail-Ratio
Mistral-7B	Zero-Shot	0.34	0.0%
	Few-Shot ($k = 1$)	0.43	0.3%
	Few-Shot ($k = 2$)	0.51	4.0%
	Few-Shot ($k = 3$)	0.53	4.7%
	Few-Shot ($k = 4$)	**0.54**	5.0%
	Bonus: CoT ($k = 2$)	0.52	0.0%
	Bonus: Prompt Tuning	**0.58**	2.7%
Llama-3.1	Zero-Shot	0.49	5.7%
	Few-Shot ($k = 1$)	0.40	0.0%
	Few-Shot ($k = 2$)	0.42	4.0%
	Few-Shot ($k = 3$)	0.43	5.3%
	Few-Shot ($k = 4$)	0.44	5.3%
Qwen3	Zero-Shot	0.23	0.0%
	Few-Shot ($k = 1$)	0.32	0.3%
	Few-Shot ($k = 2$)	0.32	1.3%
	Few-Shot ($k = 3$)	0.34	1.3%
	Few-Shot ($k = 4$)	0.34	1.3%

Table 1: Performance comparison across models and strategies. Mistral-7B benefits most from few-shot examples and parameter-efficient tuning.

4 Discussion

Our results demonstrate that **Mistral-7B** benefits most from in-context learning, with F1 scores scaling linearly from 0.34 (Zero-Shot) to 0.54 ($k = 4$). This confirms its high adaptability to the task schema. Conversely, **Llama-3.1** exhibited an inverse trend ($0.49 \rightarrow 0.44$), suggesting a conflict between the provided examples and its rigid internal safety alignment.

Regarding advanced strategies, **Prompt Tuning (PEFT)** emerged as the superior method (F1: 0.58), outperforming both static few-shot prompting and Chain-of-Thought (CoT). While CoT ($k = 2$) improved reasoning for complex cases (F1: 0.52), the increased token generation led to parsing errors that offset the gains.

Error Analysis. We identified three primary failure modes:

- **Safety Refusals:** Llama-3.1 frequently failed to distinguish between *analyzing* and *generating* hate speech, outputting refusals (e.g., "I cannot annotate...") that severely impacted its Fail-Ratio.
- **Specificity and Imbalance:** **Qwen3** defaulted to the majority class (*non-sexist*) due to limited capacity. **Mistral (Zero-Shot)** often confused *animosity* (slurs) with *derogation* (descriptive attacks), though few-shot examples successfully corrected this distinction.
- **Implicit Bias:** All quantized models struggled with sarcasm (e.g., "Women are like tea bags..."). Only PEFT showed consistent improvement in capturing these implicit patterns.

5 Conclusion

Experiments confirm that while Llama-3.1 offers strong zero-shot priors, its strict safety alignment is detrimental for toxicity detection. Mistral-7B proves to be the most robust candidate, especially when adapted via few-shot learning or parameter-efficient tuning. The superior performance of **PEFT** (F1: 0.58) highlights that optimizing soft prompts is more effective than static context for aligning quantized models to fine-grained schemas. Future work should focus on mitigating safety triggers and exploring advanced reasoning pipelines like Self-Consistency or DSPy to address sarcasm detection.

References

Robert Belanec, Ivan Srba, and Maria Bielikova. 2025. Peft-factory: Unified parameter-efficient fine-tuning of autoregressive large language models.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.