Assignment 2

NLP Course Project | 3-cfu Project Work | NLP Course Project & Project Work

Roberto Giordano, Alessio Pittiglio, Federico Marchi and **Nicolas Ivan Cridlig**
Master's Degree in Artificial Intelligence, University of Bologna
{ roberto.giordano2, alessio.pittiglio, federico.marchi9, nicolasivan.cridlig }@studio.unibo.it

Abstract

This report evaluates the performance of two large language models, Llama 3.1 and Mistral 7B, in the context of a binary classification task aimed at detecting sexism in text. The analysis was conducted under zero-shot and few-shot learning paradigms to assess the models' capabilities in generalizing to this complex and sensitive task with no task-specific training data. Comparative results highlight the strengths and limitations of each model in terms of accuracy, fail-ratio, and overall robustness.

1 Introduction

This project evaluates the performance of Llama 3.1 and Mistral 7B in detecting sexism under zero-shot and few-shot settings. Traditional approaches, such as rule-based systems or supervised models, often fail to generalize effectively to diverse or unseen data. Transformer models offer better generalization but typically require substantial labeled data and computational resources.

Our approach leverages prompt-based learning to minimize fine-tuning while maintaining task adaptability (Brown et al., 2020). To further enhance efficiency, we adopt 4-bit quantization techniques, reducing computational overhead with minimal performance loss.

Experiments were conducted on a balanced dataset of 300 sentences labeled as "sexist" or "not sexist," using zero-shot and few-shot configurations with fixed and random prompts. Tests were performed on an NVIDIA GTX 1080 GPU (8GB) locally and a T4 GPU via Google Colab. Results highlight the computational advantages of 4-bit quantization and the critical role of carefully curated few-shot examples in improving consistency in resource-constrained tasks.

2 System description

The system integrates pre-existing transformer-based architectures with customized pipelines. We utilized the pre-trained Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3 models from Hugging Face as the backbone. Our primary contribution lies in adapting these architectures for 4-bit quantization and evaluating their performance in sexism detection under resource constraints.

The codebase combines external libraries with significant original contributions. Hugging Face's Transformers library was employed for model loading and tokenization, while custom scripts automated 4-bit quantization and optimized dataset loading for small-memory GPUs. Prompt templates have been furnished.

The system operates in four stages:

1. **Model Setup:** Pre-trained Llama-3.1 and Mistral-7B models were loaded in their quantized (4-bit) variants to optimize memory usage while maintaining performance.

2. **Prompt Configuration:** Both zero-shot and few-shot settings were explored using templates, with examples drawn from our dataset.

3. **Inference:** For each experimental setup, the models were evaluated on a balanced dataset of 300 sentences, generating predictions for "sexist" vs. "not sexist" labels.

4. **Metrics and Error Analysis:** Predictions were assessed using accuracy and Fail-ratio scores.

3 Experimental setup and results

Our experiments evaluated Llama 3.1 and Mistral-7B in 4-bit quantized versions to optimize computational efficiency. Models were tested in zero-shot and few-shot settings using 2 to 4 input prompts. Few-shot prompts were loaded either in a *fixed* manner (same examples across runs) or *randomly*

selected. To ensure reproducibility, random seeds were fixed at values 42 and 17.

The hyperparameters used in the experiments included default maximum sequence lengths for Llama and 100 tokens for Mistral, with default temperature values. Since no fine-tuning was performed, models relied entirely on pre-trained weights and prompt-based inference. For evaluation, two metrics were employed: *accuracy*, which represents the proportion of correct predictions over the total samples, and *fail ratio*, calculated as the percentage of incorrect answers where the model responded differently from the request.

Results

Table 1 summarizes the experimental results across different configurations. Results are reported for both models in zero-shot and few-shot settings.

Table 1: Results for sexism detection

Model & Setting	Accuracy (%)	Fail Ratio (%)
Llama 3.1 (Zero-shot)	64.3	1.66
Llama 3.1 (2-shot)	72.0	0.0
Llama 3.1 (3-shot)	70.3	0.0
Llama 3.1 (4-shot)	71.0	0.0
Mistral-7B (Zero-shot)	59.0	1.33
Mistral-7B (2-shot)	73.0	0.0
Mistral-7B (3-shot)	71.7	0.0
Mistral-7B (4-shot)	69.3	0.0

4 Discussion

Both models demonstrated significant improvements in accuracy when transitioning from zero-shot to few-shot settings. Few-shot prompts were constructed by incorporating example snippets from another dataset. These curated examples generally produced slightly more consistent results compared to using randomly selected ones. The results were consistent between seeds (42 and 17), indicating robustness to randomness in prompt sampling and model behavior.

In terms of quantitative metrics, Llama 3.1 achieved 64.3% accuracy in the zero-shot scenario, peaking at 72% with 2-shot prompting, and maintaining strong performance around 70% with additional shots. Mistral-7B, starting lower at 59% in the zero-shot setup, achieved the best result – and the highest overall – of 73% with 2-shot prompting, before slightly decreasing in the 3-shot (71.67%) and 4-shot (69.33%) setups. Few-shot setups also produced consistent results, with a 0% fail ratio across all configurations for both models. This indicates their ability to reliably adhere to binary "yes" or "no" responses.

However, qualitative error analysis revealed several limitations. For instance, Llama 3.1 (zero-shot) exhibited a small fail ratio (1.66%), primarily due to non-binary responses, such as ethical disclaimers. These responses, while demonstrating the model's caution, skew the results by contributing to its fail ratio. Additionally, Llama 3.1 (zero-shot) struggled with high false negatives (91), suggesting a tendency to underclassify sentences as sexist. False negatives were significantly reduced in few-shot setups, with the 2-shot model achieving a favorable balance (70 false negatives) and further stability in the 3-shot and 4-shot scenarios.

Mistral-7B displayed a similar trend: in the zero-shot setup, it recorded a high number of false negatives (120) and a very low number of false positives (3), indicating an overly conservative bias toward non-sexist classifications. Few-shot prompting helped balance these errors, with the 2-shot model achieving the most effective trade-off between false positives and false negatives. Finally, a peculiar behavior was observed: in four instances, Mistral-7B included an ethical disclaimer in its responses despite correctly classifying the content.

5 Conclusion

This work demonstrated that both Llama 3.1 and Mistral-7B show significant improvements in accuracy when transitioning from zero-shot to few-shot settings, with Mistral-7B (2-shot) achieving the highest accuracy (73%) and no fail ratio. Few-shot setups were shown to be robust across different seeds and provided consistent results, further emphasizing their adaptability. Despite these promising results, certain limitations emerged. Llama 3.1 (zero-shot) occasionally introduced ethical disclaimers instead of providing binary answers, highlighting a trade-off between sensitivity and reliability. Both models struggled with high false negatives in zero-shot settings, reflecting challenges in detecting sexist content when context or patterns are underrepresented in the data. Future work should focus on fine-tuning with more diverse and balanced datasets, experimenting with alternative architectures and models.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.