Assignment 2

Matteo Fasulo, Maksim Omelchenko, Luca Babboni and **Luca Tedeschini**
Master's Degree in Artificial Intelligence, University of Bologna
{ matteo.fasulo, maksim.omelchenko, luca.babboni2, luca.tedeschini3 }@studio.unibo.it

Abstract

This work investigates the effectiveness of in-context learning through prompt-based evaluation on two advanced large language models, *Mistral-7B-Instruct-v0.3* and *Llama-3.1-8B-Instruct*.
We explore both zero-shot and few-shot settings, employing *BM25* scoring to select the most relevant K demonstrations for learning with few shots. Our approach aims to enhance model performance by providing contextually relevant examples. The evaluation focuses on the models' ability to detect sexist language, comparing their accuracy and error patterns across different prompting strategies.
Key findings reveal that while few-shot prompting generally improves performance, BM25 scoring can lead to increased false positives, highlighting the need for balanced and diverse demonstration selection.

1 Introduction

In-context learning (ICL) has revolutionized the way large language models (LLMs) tackle new tasks, allowing them to adapt without the need for further fine-tuning (Chowdhery et al., 2022). By incorporating task-specific instructions and demonstrations directly into the input prompt, LLMs can effectively handle a variety of tasks. This approach is particularly beneficial in situations where obtaining labeled data is either too costly or impractical. Traditional ICL methods typically use zero-shot prompts, which provide only task instructions, or few-shot prompts, which include both instructions and a set of labeled examples called demonstrations (Brown et al., 2020). Although zero-shot prompting is straightforward and user-friendly, it often falls short for tasks that require deeper contextual understanding. Few-shot prompting can significantly boost model performance by including relevant examples, but the selection of these examples is crucial. In this work, we evaluate two pre-trained LLMs, *Mistral-7B-Instruct-v0.3* and *Llama-3.1-8B-Instruct*, in detecting sexist text using zero-shot and few-shot settings that address EDOS Task A on sexism detection (Kirk et al., 2023).
Our experimental design involves a detailed comparison of zero-shot and few-shot performance using two different demonstration selection procedures: through stratified random sampling and BM25 scoring. Preliminary results suggest that, contrary to our initial hypothesis, the BM25 scoring negatively affected the model performance, leading to higher false positives. The models often misclassified non-sexist texts as sexist when demonstrations were selected based on lexical similarity, indicating an overemphasis on shared terms rather than contextual meaning.

2 System description

For all experiments, the pipeline follows a consistent structure: we begin by importing the LLM, then prepare the prompt using a chat template approach, tokenize the prompt, and finally use the model to predict whether the comments are sexist (see Figure 1). The predictions are stored in a pandas DataFrame for subsequent analysis. Both models, *Mistral-7B-Instruct-v0.3* and *Llama-3.1-8B-Instruct*, are quantized to 4-bit using 16-bit brain floating point precision (bfloat16) to fit within the constraints of a single RTX 3080 GPU with 10 GB of VRAM. Predictions are extracted from the model output using regular expressions to identify the strings "YES" and "NO".
The examples provided to the model are selected using two methods: stratified random sampling and top-k selection based on BM25 scoring. Although randomly chosen demonstrations may not always capture the necessary task-specific patterns, retrieval-based methods like BM25 scoring offer a more systematic, and data-driven approach. BM25 scoring ranks demonstrations based

LLM Model	Shots	Fail Ratio	Accuracy
Mistral-7B	0	0.0	0.616
Mistral-7B	2	0.0	0.766
Mistral-7B	4	0.0	0.770
Mistral-7B	8	0.0	0.746
Llama-3.1-8B	0	0.0	0.610
Llama-3.1-8B	2	0.0	0.633
Llama-3.1-8B	4	0.0	0.660
Llama-3.1-8B	8	0.0	0.646

Table 1: Results of the LLMs

on lexical overlap and term frequency-inverse document frequency (TF-IDF) heuristics (Robertson and Zaragoza, 2009), allowing us to select examples most similar to the input text. After scoring all possible demonstrations, only the top k are included in the prompt fed to the model.

3 Experimental setup and results

All models were evaluated by performing inference on a test set consisting of 300 observations. After obtaining predictions from each model, we computed two key metrics: accuracy and fail-ratio. Accuracy was chosen as a primary metric due to the balanced nature of the dataset across classification classes. The fail-ratio, defined as $1 - \text{accuracy}$, indicates how often the LLM fails to follow instructions and provides incorrect responses that do not address the classification task.

In few-shot inference, we ensured a balanced number of demonstrations per class to avoid introducing bias into the prompt. To prevent prompts from having examples with the same label placed consecutively, we shuffled the order of the examples before incorporating them into the prompt. This shuffling introduces stochasticity, helping to prevent the LLM from detecting patterns and correlations within the labels of the demonstrations, thereby enhancing the robustness and reliability of the model's predictions.

4 Discussion

From our results in Table 1, we observe that few-shot prompting significantly boosts model accuracy compared to zero-shot baselines. Providing examples in the prompt enhances the models' ability to classify sexist content. However, despite these improvements, both *Mistral-7B-Instruct-v0.3* and *Llama-3.1-8B-Instruct* exhibit a high rate of false positives, indicating a tendency to over-predict the "sexist" label.

Our intuition behind the high false positive rate is that models are overly sensitive to certain keywords, misclassifying non-sexist content. This is particularly evident in the *Mistral* model, likely due to its training dataset, aimed at avoiding unethical outputs[1]. Moreover, including too many demonstrations does not lead to better performance since the 8-shot version performs worse. The fail ratio (which is the number of times the model does not respond in a compliant manner with respect to the classification task) is 0 for all the evaluations. This means that both models are effective in following the given instructions. A possible improvement might involve modifying the prompt to include a precise definition of sexism in the context of our classification. Such a refinement could better condition the model to align with the intended challenge of identifying sexist text from Gab and Reddit.

5 Conclusion

In this study, we evaluated the use of ICL with prompt-based evaluation on foundation LLMs for detecting sexist language. Few-shot prompting improved accuracy over zero-shot baselines, but both models exhibited a high rate of false positives, likely due to sensitivity to certain keywords. While the accuracy improvements were anticipated, the high rate of false positives was unexpected, revealing limitations in the models' nuanced language understanding. The use of BM25 for selecting similar demonstrations based on input text actually increased the error rate, suggesting that lexical overlap alone is insufficient for effective demonstration selection. Our study highlights the need for more sophisticated demonstration selection methods that balance relevance and diversity, potentially incorporating semantic similarity measures rather than relying solely on lexical overlap. This approach could enhance the models' ability to accurately classify nuanced language and reduce false positives.

[1] Is mistral uncensored?

6 Links to external resources

- Github repository

Images

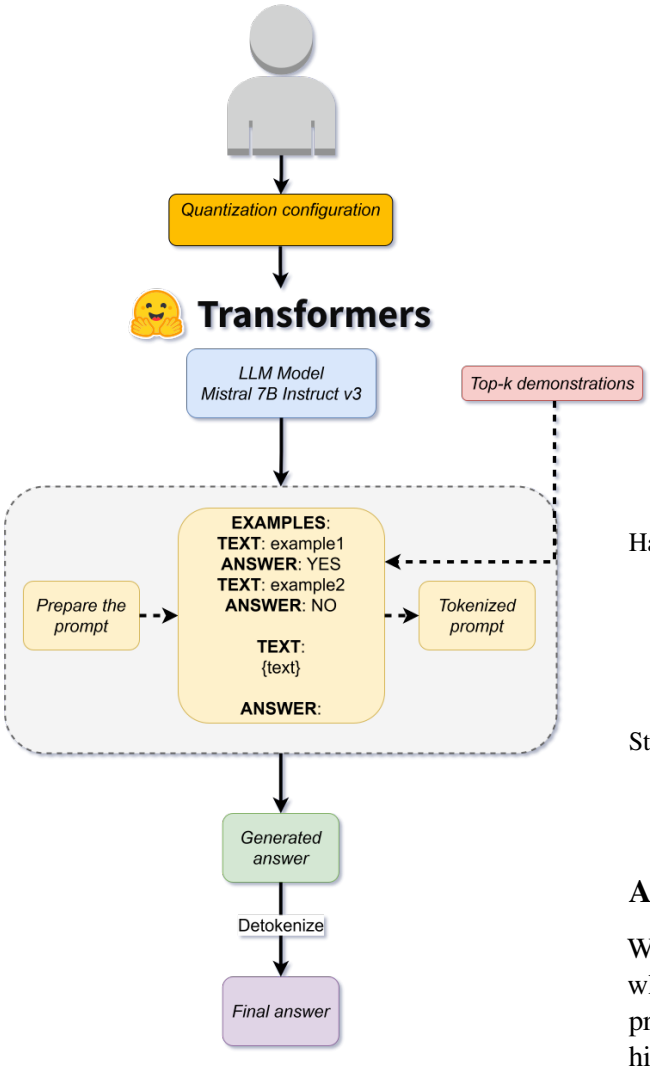

Figure 1: System pipeline illustrating the process from quantization configuration, through the selection of top-k demonstrations, to the final answer generation by the LLM model.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Appendix

We also tested with the *Dolphin Mistral* model [2], which uses a less biased dataset, we noticed an improved accuracy and balance in zero-shot settings, highlighting the impact of training data bias.
In few-shot settings, the issue persisted, likely due to the nature of the prompt: the *Dolphin Mistral* is primarily trained to address coding tasks. For this reason, we believe that a more specific prompt, providing detailed instructions on how to tackle the task (e.g., defining what constitutes sexism, explaining how to identify it, and clarifying what can be considered sexist), would improve the model's performance in the few-shot task.

[2] cognitivecomputations/dolphin-2.9.3-mistral-7B-32k