Assignment 2

Armando Renzullo, Cosimo Russo and Tommaso Perniola
Master's Degree in Artificial Intelligence, University of Bologna
{ armando.renzullo, cosimo.russo, tommaso.perniola}@studio.unibo.it

Abstract

This study evaluates in-context learning with two models, Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct, in zero-shot and few-shot scenarios, focusing on detecting sexist language. Results show that using multiple examples generally improves performance, but Mistral and Llama exhibit model-specific differences in accuracy and false positives. The findings highlight the need for tailored prompting strategies and diverse examples for optimal model performance.

1 Introduction

The rapid advancement of large language models (LLMs) has revolutionized natural language processing (NLP), offering unprecedented capabilities for tasks such as text generation, translation, and classification (Brown et al., 2020; Devlin et al., 2019). Among these innovations, in-context learning (ICL) has emerged as a powerful paradigm, enabling models to perform specific tasks by providing examples directly within the input prompt (Dong et al., 2022).

Traditional approaches to sexist language detection often rely on supervised learning, requiring large amounts of labeled data and fine-tuning, which can be resource-intensive. Moreover, biases in training data can perpetuate stereotypes or result in inaccuracies (OpenAI, 2020). In contrast, ICL offers a more efficient method, where models adapt to specific tasks by leveraging examples within the input prompt without the need for extensive retraining (Min et al., 2022; Zhao et al., 2021). However, the effectiveness of ICL in detecting biased language remains underexplored, especially across different model architectures and prompt configurations.

This study investigates the effectiveness of ICL in detecting sexist language, focusing on two state-of-the-art LLMs: Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct. These models are selected due to their cutting-edge architecture and demonstrated success in a variety of NLP tasks. We evaluate their performance in both zero-shot and few-shot settings, comparing their ability to detect sexist language in a dataset of social media comments. Specifically, we analyze (1) the accuracy of detection, (2) the prevalence of false positives and false negatives, and (3) the impact of different prompting strategies.

2 Background

The detection of biased or harmful language, such as sexism, has become a significant challenge in natural language processing (NLP), particularly in domains like social media, customer support, and online forums, where large volumes of unstructured text need to be processed in real-time. Traditional methods for this task often rely on supervised learning, where models are trained on large, manually labeled datasets that include explicit examples of sexist language. However, acquiring such datasets can be costly and time-consuming, and the models themselves may inherit biases from the data or fail to generalize to new forms of biased language. In recent years, the advent of large pre-trained models, particularly those based on the Transformer architecture, has opened up new possibilities for addressing these challenges with less reliance on extensive labeled datasets.

3 System description

The proposed system utilizes pre-trained language models such as Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct for sexism detection in text, with a flexible pipeline that operates in both zero-shot and few-shot modes, depending on the availability of labeled data. The pipeline begins by formatting input texts into structured prompts that guide the model in the classification task, distinguishing between sexist and non-sexist language.

In the zero-shot setting, the prompt contains only the task description, while in the few-shot setting, a small number of labeled examples are included to improve accuracy. Once the model generates a response, it is processed into a binary classification ("YES" or "NO"), which is then compared to the ground truth labels and evaluated using metrics such as accuracy and fail-ratio.

Mistral-7B-Instruct and Llama-3.1-8B-Instruct are both instruction-following models, but Mistral is typically smaller and more optimized for specific tasks, often providing faster inference for targeted applications. Llama, on the other hand, offers a more generalized performance across multiple languages and domains, making it more versatile for a variety of use cases.

4 Experimental setup and results

In our experiments, we compared the performance of Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct in detecting sexism in text across different few-shot prompting configurations (0-shot, 2-shot, 3-shot, 4-shot). These configurations allowed us to evaluate model performance as the amount of labeled data increased.

We assessed performance using several metrics: precision, recall, F1-score, false positive rate (FPR), false negative rate (FNR), and fail-ratio. Precision and recall provided insight into the models' ability to identify sexist and non-sexist language, while the F1-score balanced these metrics. The FPR and FNR indicated misclassification rates, and the fail-ratio gave an overall view of robustness.

Additionally, the quantization of the models plays a critical role in performance, influencing both the accuracy of results and the computational efficiency. Was used a 4-bit quantization to reduce memory usage and speed up inference, while maintaining model accuracy. Double quantization and NF4 format enhance precision, and using bfloat16 for computation ensures efficient processing with minimal loss in performance, making the model faster and more resource-efficient.

5 Discussion

Both Mistral and Llama initially struggle to accurately classify non-sexist texts, often mislabeling them as "sexist." For instance, at 0-shot, Mistral's recall is 0.98, while Llama's remains consistently above 0.84. As the number of examples increases, both models show improvement, with Mistral experiencing a more significant gain in precision—from 0.550 at 0-shot to 0.710 at 4-shot. Llama's precision increases more gradually, from 0.585 to 0.654 over the same range. However, despite these gains, Mistral starts to face challenges in correctly identifying "sexist" texts as the number of examples increases, with its recall dropping from 0.98 at 0-shot to 0.72 at 4-shot. In contrast, Llama's performance in identifying "sexist" texts remains stable. This trend suggests that Llama is the more stable and robust model overall, particularly evident in its steady improvement in the 4-shot configuration, where its F1-score increases from 0.692 at 0-shot to 0.738 at 4-shot. To further enhance performance, one potential improvement could be to adjust the prompt to provide a clearer definition of sexism within the context of the classification task.

6 Conclusion

The comparative analysis between Mistral and LLama highlights key differences in performance, stability, and sensitivity across various configurations. While LLama, as the larger model, demonstrates superior stability and adaptability due to its capacity for better generalization, Mistral shows a more pronounced improvement in the initial stages but struggles with recall and sensitivity as the number of examples increases.

Both models initially face challenges in accurately classifying non-sexist texts as non-sexist, often misclassifying them as sexist. This suggests inherent biases potentially caused by imbalances in the training data or annotation processes.

Future efforts to optimize these systems should focus on addressing these limitations through:

- **Data Augmentation:** Enhancing the diversity of training data to improve the models' ability to generalize across different contexts.

- **Bias Mitigation:** Balancing training datasets to reduce the tendency to favor one class over the other.

- **Meta-Learning Techniques:** Incorporating advanced learning strategies to improve adaptability, particularly in detecting subtle patterns in complex datasets.

References

Tom B. 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, Karthik Narasimhan, Richard L. Hennigan, Clement Delangue, Evan Schneider, Leo Gao, Xiang Lisa Li, Cheng Li, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Xu Dong, Wenxuan Zhou, Liqiang Nie, and Xuemin Lin. 2022. A survey on in-context learning. *arXiv preprint arXiv:2209.09566*.

Sungdong Min, Jaewoo Kang, Haoda Xu, Dheeraj Rajagopal, Zhou Yu, and Xiang Lisa Li. 2022. Rethinking the role of examples in few-shot learning: A practical guide. *arXiv preprint arXiv:2204.04564*.

OpenAI. 2020. Language models are few-shot learners.

Xing Zhao, Yichao Lu, Chia-Hsiu Chen, Hongyu Wu, and Zhiwei Liu. 2021. Calibrate your few-shot learner with contrastive prompts. *arXiv preprint arXiv:2101.01652*.