Assignment 2

Shariq Mohd Ansari, Leonardo Chiarioni, Federico Faccioli
Master's Degree in Artificial Intelligence, University of Bologna
{ mohdshariq.ansari, leonardo.chiarioni, federico.faccioli2 }@studio.unibo.it

Abstract

The advent of social media has provided us with a platform to share our thoughts and ideas, but it also creates a space for harassment. With our project, we wish to utilize the capabilities of LLM in order to tackle sexism. In this report, we explore the application of state-of-the-art large language models (LLMs), to detect sexism in tweets using zero-shot and few-shot learning approaches. We observed the importance of model selection, prompt engineering, and the trade-off between computational efficiency and classification accuracy.

1 Introduction

Detecting sexism in online content, such as tweets, is an important yet challenging task that has gained significant attention in recent years. Social media platforms like Twitter often serve as a breeding ground for gender-based hate and bias. The ability to automatically detect and flag such content can have profound implications for creating safer and more inclusive digital spaces.

Traditional approaches to sexism detection in text have relied on rule-based systems or manually curated datasets. More recently, machine learning (ML) and deep learning (DL) models have been applied to this task, utilizing SVM, Random Forest to classify text based on large annotated corpora (Sreekumar et al., 2021). But these models do not generalize well on unseen data and face difficulties in handling the contextual nuance.

LLM understand and generate language in more flexible and context aware manner. For this reason, using a pre-trained LLM is a computationally efficient and scalable solution for real time content moderation and sexism detection on social media.

In our project, we explore the application of large language models, specifically **Mistral-7B-Instruct-v0.3** and **Phi-3-mini-4k-instruct**, to detect sexism in tweets using zero-shot and few-shot learning approaches. The primary objective of the project is to assess the effectiveness of these LLMs in classifying tweets as sexist or non-sexist based on carefully designed prompts.

To evaluate the performance of the LLM, we conducted a series of experiments using the balance annotated dataset of 300 tweet samples. We divided the experiments into 2 zero-shot and 6 few-shot experiments. Additionally, we ran an experiment using Langchain framework using **Mistral-7B-Instruct-v0.3**. To compare the performance we primarily used accuracy and fail-ratio. In addition to these metrics, we also used f1-score, precision, recall and confusion matrices as additional metrics and tools. We have utilized a single T4 GPU, along with 4-bit quantization to reduce the memory overhead.

2 System description

We implemented sexism detection by using the LLM models that we downloaded from HuggingFace API. More specifically, we used the following two models:

- **Mistral-7B-Instruct-v0.3** : It is a large language model developed by Mistral AI. It features 7 billion parameters and is designed to excel in instruction-following tasks.
- **Phi-3-mini-4k-Instruct** : It is a lightweight instruction-tuned model developed by Microsoft. It contains approximately 3 billion parameters, making it a faster alternative to larger LLMs. The model is designed to handle instruction-following tasks while maintaining a good balance between performance and resource consumption.

We also made use of bitsandbytes library for 4-bit quantization of the models. For tokenization, we implemented the tokenization available for the models on Huggingface without any alteration. The

work-flow takes its inspiration from the Tutorial 3 of the NLP course. In the end, we implemented **LangChain** framework to utilise the PromptTemplate, LLMChain module for concise implementation of prompting.

3 Experimental setup and results

The experimentation is divided into following major blocks:

- **Zero shot Learning** : We begin with preparing the data, then we parse the tweets into the prompt. Once the formatted prompts are ready, we supply them directly to the models. Responses obtained are sent to the post-processing step to enhance the fail-ratio metric. Apart from converting the response to lower case, we also search for "yes" or "no" string in the response because some of the response contain correct classification inside the string like (no?). The experiment is performed using the Mistral 7B and Phi-3-mini. After several experimentation, we tuned hyper-parameters to temperature = 0.2 to preserve deterministic nature, top-p=0.90 to obtain the response with 90 percent probability, top-k =50 to restrict the token selection. Also we set the do_shuffle as True to see the effect of top-k , top-p and temperature configuration.

- **Few shot Learning** : For Few-shot learning, the experiments are designed in the same manner as zero shot with additional step of adding examples in the prompt from the demonstration dataset. We run multiple experiments with $num\_per\_class \in [2, 4]$ to access the effect of examples on the performance of model.

4 Discussion

From the the Table 1, we can see that the Phi-3-mini is outperforming Minstral in zero shot learning, beside being trained on 3k tokens. On the other hand, we can observe that the few shot learning, enhances the performance of Mistral model and we were able to reach accuracy of 0.72 only with 3 examples of each class. During the experimentation, we can also observe that increasing the number of examples also enhances the performance but very slightly.

As shown from the confusion matrices in the error analysis Section of the notebook, we can be observe that in few shot learning, **Phi-3-mini-4k-instruct** model is biased towards not identifying correctly the tweets as sexist; this stems from the fact that we are working with the pre-trained model without any fine-tuning, and being a light-weight model, it struggles to generalize quickly in a few-shot setting, as it lacks the capacity to store and process diverse knowledge.

Overall, since we are quantizing the weights to 4-bits, that also affects the performance, as a lot of information has been lost during quantization.

Model Setting	Accuracy	Fail Ratio	Precision	Recall	F1 Score
Mistral zero_shot	0.59	0.0	0.55	0.98	0.70
Phi 3 zero_shot	0.64	0.0	0.63	0.69	0.66
Mistral 2shot	0.70	0.0	0.67	0.80	0.73
Phi 3 2shot	0.59	0.0	0.75	0.28	0.41
Mistral 3shot	0.71	0.0	0.68	0.76	0.72
Phi 3 3shot	0.64	0.003	0.71	0.47	0.56
Mistral 4shot	0.72	0.0	0.71	0.73	0.72
Phi 3 4shot	0.60	0.003	0.74	0.32	0.45

Table 1: Performance metrics for Mistral and Phi models in zero-shot and few-shot settings.

5 Conclusion

With our work we showed how we can use prompting techniques for classification task like sexism detection. We also highlighted how the light weight models like Phi3-mini can outperform bigger models in zero shot learning. We also demonstrated how the post processing of the response is a crucial steps for creating a fair evaluation of the models. We also showed how few shot learning can help improve performance of the prompting compared to zero shot learning. We experimented with Langchain frameworks specially the PromptTemplate, LLMChain modules for concise pipelines for prompting.

In order to further improve the result, We can combine the few-shot learning with chain-of-thought (CoT) (Wei et al., 2023) prompting. And further enhancement can be obtained by fine-tuning of the pre-trained model on the sexism specific corpora like 'Call me sexist but' Dataset (CMSB)(Samory, 2021).

References

Mattia Samory. 2021. The 'call me sexist but' dataset (cmsb). GESIS, Cologne. Data File Version 1.0.0, https://doi.org/10.7802/2251.

Murari Sreekumar, Shreyas Karthik, Durairaj Thenmozhi, Shriram Gopalakrishnan, and Krithika

Swaminathan. 2021. Sexism identification in tweets using machine learning approaches. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

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.