Assignment 2

Erfan Samieyan Sahneh
Master's Degree in Artificial Intelligence, University of Bologna
{ erfan.samieyansahneh}@studio.unibo.it

Abstract

This project investigates sexism detection in English tweets using large language models (LLMs) through prompt engineering. Two methods were employed: (1) Zero-shot prompting, where the model was directly prompted to classify text as sexist or non-sexist without any prior examples, and (2) Few-shot prompting, where labeled examples with expected outputs were provided alongside the prompt to guide the model's classification. The findings indicate that both methods enable LLMs to identify sexist content, with the few-shot approach offering improved accuracy by leveraging contextual examples. This project highlights the potential of prompt engineering to utilize pre-trained LLMs for societal applications without extensive fine-tuning.

1 Introduction

Sexism detection in social media is vital for promoting respectful online interactions. Traditional methods, such as machine learning classifiers and deep learning models, have been employed to identify sexist content. While effective, these approaches often require extensive labeled data and significant computational resources for training and fine-tuning.

In this project, we explore prompt engineering techniques with large language models (LLMs) to detect sexism in English tweets. We implemented two methods: zero-shot prompting, where the model classifies text without prior examples, and few-shot prompting, where the model is provided with a few labeled examples to guide its predictions.

Our experiments were conducted on a dataset of English tweets, comprising 300 samples, with 150 labeled as sexist and 150 as non-sexist. We evaluated the performance of both prompting methods using standard metrics such as accuracy, precision, recall, and F1-score.

The key findings from our study are:

- **Effectiveness of Prompt Engineering**: Both zero-shot and few-shot prompting with LLMs demonstrated the capability to detect sexist content without the need for extensive training data.
- **Performance Comparison**: Few-shot prompting yielded higher accuracy and F1-scores compared to zero-shot prompting, indicating that providing examples enhances the model's performance.
- **Resource Efficiency**: Utilizing LLMs with prompt engineering is a resource-efficient alternative to traditional methods, reducing the need for large labeled datasets and computationally intensive training processes.

These results suggest that prompt engineering with LLMs is a promising approach for automated sexism detection in social media, offering a balance between performance and resource utilization.

Recent studies have explored similar approaches, highlighting the effectiveness of prompt engineering in tasks like sexism detection. For instance, research utilizing GPT and Llama models has demonstrated the potential of prompt-based methods in identifying sexist language (Kumar et al., 2024).

2 System description

In this work, we utilized two large language models (LLMs), Mistral v3 and Phi3-mini, for sexism detection in English tweets. Both models were employed through prompt engineering, enabling us to classify text without extensive model fine-tuning.

Two methods were implemented for this task:

- **Zero-Shot Prompting:** The model was provided with a simple prompt to classify text as sexist or non-sexist without any prior examples.

Model	Class	Precision	Recall	F1-Score
Phi3-mini	0 (Non-Sexist)	0.60	0.83	0.70
	1 (Sexist)	0.72	0.45	0.56
Mistral v3	0 (Non-Sexist)	0.80	0.67	0.73
	1 (Sexist)	0.71	0.83	0.77
Overall Accuracy		Phi3-mini: 0.64, Mistral v3: 0.75		

Table 1: Performance metrics for Phi3-mini and Mistral v3 on sexism detection with few-shot prompting.

- **Few-Shot Prompting:** The model was given a few labeled examples from a dataset different from the testing dataset to guide its classification.

This approach leveraged the pre-trained capabilities of LLMs, allowing for efficient and effective sexism detection while avoiding overlap between training and testing data.

3 Experimental setup and results

3.1 Experimental Setup

Our experiments were conducted to evaluate the performance of large language models (LLMs), specifically Mistral v3 and Phi3-mini, for sexism detection in English tweets. The models were optimized using quantization techniques to enhance computational efficiency and reduce memory usage. Quantization was configured to utilize 4-bit precision, employing advanced techniques like double quantization and mixed precision arithmetic.

Results

As shown in Table 1, the Mistral v3 model outperforms Phi3-mini in terms of F1-score and accuracy for few-shot prompting.

These findings demonstrate that Mistral v3 outperforms Phi3-mini in terms of both precision and recall, particularly in balancing the classification of sexist and non-sexist tweets. Few-shot prompting further enhances model performance by leveraging contextual examples.

4 Discussion

Our experiments demonstrate that few-shot prompting significantly outperforms zero-shot prompting in the task of sexism detection. By providing contextual examples during few-shot prompting, the models were better equipped to understand nuanced patterns in the dataset, leading to improved precision, recall, and overall accuracy. The results clearly highlight the importance of including contextual examples to guide the models effectively, especially in complex tasks like sexism detection where the boundary between classes can be subtle.

4.1 Error Analysis

While the models achieved satisfactory performance, an analysis of errors reveals several limitations that need to be addressed for further improvement.

Missed Sexist Texts with Positive Language

One common failure case observed was the inability of the models to detect sexist texts that employ positive or motivational language. For instance:

> *Thanks for the response brotha, I'll work towards it. It's just another nasty remnant of my beta days. But like you said, I'm a man. Men conquer.*

The model classified this as non-sexist, although it contains implicit sexism through the reinforcement of gender stereotypes.

Another example:

> *Giving women all your money is good personality.*

This statement, while seemingly innocuous, perpetuates harmful stereotypes about gender roles. However, the model failed to detect its sexist implications.

False Positives Due to Offensive Words

Conversely, the models sometimes misclassified non-sexist texts as sexist due to the presence of offensive language, even when the context or intent did not convey sexism. For example:

> *I would love to suck her tits.*

While the text contains offensive language, its purpose is not inherently sexist and does not perpetuate

gender-based discrimination. Such misclassifications highlight the need for models to better discern between offensive language and genuine sexism.

5 Conclusion

These findings suggest that while few-shot prompting improves overall performance, there is still significant room for improvement in handling nuanced cases of sexism detection. Future work could explore incorporating external knowledge or fine-tuning models on curated datasets that include diverse examples of implicit and explicit sexism. Additionally, advanced techniques such as contextual embeddings or multimodal approaches could help mitigate errors arising from surface-level features like offensive words.

6 Links to external resources

GitHub page.

References

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and llm judges for gender bias detection in language models.