# NLP Assignment 2 report

**Tara Sabooni, Yasaman Samadzadeh, Bahar Hamzehei, Hamidreza Bahmanyar**
Master's Degree in Artificial Intelligence, University of Bologna
{ tara.sabooni, yasaman.samadzadeh, bahar.hamzehei, hamidreza.bahmanyar }@studio.unibo.it

## Abstract

This report explores sexism detection in user-generated text using two large language models, Phi3-mini and Llama v3.1, with zero-shot and few-shot prompting. Phi3-mini outperforms Llama v3.1 in few-shot mode, achieving 65.67% accuracy compared to Llama's 58%. An error analysis highlights challenges with subtle or context-dependent sexism, suggesting that refined prompts and domain-specific data can enhance model performance.

## 1 Introduction

Detecting sexism is crucial for fostering inclusive environments and combating discrimination. Our project leverages prompting with large language models(LLMs) to identify sexist sentences effectively.

## 2 Background

This project aims to develop a straightforward yet impactful solution: a system that labels text as sexist or not sexist. We utilized two Large Language Models (LLMs). By using prompting—providing specific instructions to the models, we streamline the process, saving time and adapting the models effectively to the task.

## 3 System description

**Architecture**: The system is designed to run experiments with both zero-shot and few-shot approaches. The architecture of our system is based on a prompt-based inference pipeline. The core components include:

- **Preprocessing Module:** This component prepares the input text by tokenizing it and formatting it as a prompt for the LLM.
- **Prompting Module:** The assignment provided prompt templates to guide LLMs in generating binary responses ("YES" for sexist, "NO" for non-sexist) based on input text.
- **Post-processing Module:** The post-processing module maps model responses to binary labels (1 for sexist, 0 for non-sexist) and uses -1 for ambiguous outputs.
- **Evaluation Module:** This module computes performance metrics such as accuracy and fail ratio, and generates confusion matrices for error analysis.

Figure 1 shows the system's pipeline.

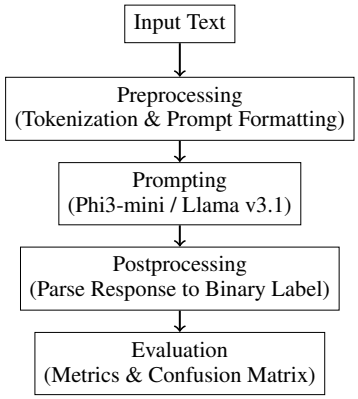


Figure 1: Model architecture

## 4 Data

### 4.1 Preprocessing

We performed the following preprocessing steps to ensure compatibility with the LLMs used in the project:

1. **Loading and Encoding**: The datasets were loaded into `Pandas DataFrames` and encoded using the Hugging Face `transformers` library for tokenization.
2. **Prompt Formatting**: Text samples were formatted into structured prompts to elicit consistent binary responses.

See Table 1 for a summary of the datasets.

| Dataset | Samples | Label Distribution |
|---|---|---|
| a2_test.csv | 300 | 50% / 50% |
| demonstrations.csv | 1000 | 50% / 50% |

Table 1: Summary statistics of the datasets.

## 5 Experimental setup and results

We conducted experiments with two Large Language Models(LLMs): Phi3-mini and Llama v3.1, evaluated in both zero-shot and few-shot configurations. For the few-shot setting, we employed a *demonstration* set of 1,000 samples and included $N \in \{2, 3, 4\}$ demonstration examples per class, ultimately choosing the best $N$ based on accuracy. We applied 4-bit NF4 quantization to fit each model in single-GPU memory.

Our primary metrics were accuracy owing to the balanced dataset and a fail-ratio metric that captures the irrelevant or malformed answers. We summarize these numerical results in Table 2.

| Model | Accuracy | Fail Ratio |
|---|---|---|
| Phi3-mini Zero | 64.33% | 0.00% |
| Phi3-mini Few | 65.67% | 0.00% |
| Llama v3.1 Zero | 64.67% | 1.67% |
| Llama v3.1 Few | 58.00% | 0.00% |

Table 2: Model performance metrics in zero-shot and few-shot setups.

## 6 Discussion

### 6.1 Observations on Misclassifications

Upon analysis, several patterns emerged from the misclassified examples:

1. Ambiguity in Non-Sexist Texts: Some texts labeled as non-sexist contained controversial language that the models flagged as sexist. For instance:

   - *Text:* "I was just looking at another post where a female had sex with men to get grades. Why support that?"
     *True Label:* Non-sexist
     *Predicted Label:* Sexist

   This was likely flagged due to the gendered language and stereotype.

2. Handling of Offensive Language: There was a struggle with explicitly offensive content, failing to generate binary labels for them. For example:

   - *Text:* "The ultimate stupid n*****, his cockroach mother is on benefits too... Useless vermin."
     *True Label:* Sexist
     *Predicted Label:* Fail (-1)

   This example shows that internal filters may have blocked the models from processing such harmful outputs.

3. **Contextual Understanding**: Both models showed limited ability to understand context. Sarcastic or hyperbolic statements were often misclassified. For instance:

   - *Text:* "Old man mogs everyone in this sub"
     *True Label:* Non-sexist
     *Predicted Label:* Sexist

   This statement was flagged as sexist due to surface-level features like the word "mog".

### 6.2 Model Comparison

Llama v3.1 performs better in zero-shot setups due to its strong generalization capabilities and extensive prior knowledge from pre-training on large datasets. However, in few-shot setups, while it shows effective adaption with no failures, its accuracy is lower than Phi3-mini's, likely due to overtraining. This suggests that Llama v3.1 relies too heavily on patterns from its pre-training, making it less adaptable to new, diverse inputs introduced in few-shot prompts.

## 7 Conclusion

This project explored sexism detection using Phi3-mini and Llama v3.1 under zero-shot and few-shot prompting. Phi3-mini showed better performance in few-shot setups, while Llama v3.1 struggled in this mode. Misclassifications revealed challenges with ambiguous and offensive language in Llama zero-shot model, highlighting limitations in nuanced understanding.

## References


[1] Microsoft. Phi-3-mini-4k-instruct. Available at: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct.

[2] Meta. Llama-3.1-8B-Instruct. Available at: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.