Sexism Classification with Large Language Models
A Comparative Study of Zero-Shot and Few-Shot Prompting

Ivo Rambaldi
University of Bologna
ivo.rambaldi@studio.unibo.it

Abstract

This report presents a study on sexism classification in social media using instruction-tuned large language models through prompting strategies. We compare two models—Phi-3-mini-4k-instruct and Mistral-7B-Instruct-v0.3—under both zero-shot and few-shot settings. The task involves classifying text into five categories: not-sexist, threats, derogation, animosity, and prejudiced. We evaluate the models using macro-F1 and fail ratio metrics, and provide error analysis through confusion matrices. Results show that while Phi-3-mini performs better in zero-shot (macro-F1 0.374), Mistral-7B benefits more from few-shot examples, achieving the best overall performance (macro-F1 0.471).

1 Introduction

Large language models have demonstrated strong capabilities in understanding and generating natural language, making them promising candidates for classification tasks through prompting rather than fine-tuning. This approach is particularly attractive for tasks with limited labeled data or where rapid prototyping is needed, as it requires no gradient updates to model parameters.

In this work, we investigate the effectiveness of instruction-tuned LLMs for fine-grained sexism classification. Unlike binary sexism detection, this task requires distinguishing between multiple types of sexist content: threats of harm, derogatory descriptions, use of slurs (animosity), and expressions of prejudice. We compare two models of different scales and assess whether providing in-context examples improves classification performance.

2 Dataset and Task Definition

The dataset consists of 300 test samples from the EXIST 2023 shared task, containing social media posts labeled with one of five categories:

- **not-sexist**: Content without sexist elements
- **threats**: Text expressing intent or desire to harm women
- **derogation**: Text describing women in derogatory or demeaning ways
- **animosity**: Text containing slurs or insults toward women
- **prejudiced**: Text expressing support for mistreatment or inequality of women

For few-shot learning, we use a separate demonstrations dataset containing labeled examples across all five categories. These examples are sampled to construct in-context demonstrations that precede the test instance in the prompt.

3 Models

3.1 Model Selection

We evaluate two instruction-tuned language models representing different points on the scale-capability spectrum:

- **Phi-3-mini-4k-instruct** (Microsoft): A compact 3.8B parameter model optimized for efficiency while maintaining strong instruction-following capabilities. The 4k context window is sufficient for our prompting strategy.
- **Mistral-7B-Instruct-v0.3** (Mistral AI): A 7B parameter model with demonstrated strong performance on instruction-following benchmarks, providing a comparison point at a larger scale.

Both models are loaded with 4-bit quantization using the NF4 (Normal Float 4) format to reduce memory requirements, enabling deployment on consumer hardware. This quantization uses

BitsAndBytes configuration with float16 compute dtype, maintaining reasonable inference quality while significantly reducing memory footprint.

3.2 Prompting Strategy

We design a structured prompt template with explicit instructions following best practices for LLM prompting:

- **System message**: Establishes the model's role as an expert annotator
- **Task definition**: Clear enumeration of the five categories
- **Category definitions**: Explicit descriptions of what each label means
- **Output constraints**: Rules requiring single-word answers without explanation

For few-shot prompting, we inject demonstration examples before the target text. We use K=2 examples per class (10 total examples), each formatted as TEXT/ANSWER pairs showing the input and expected output format. The demonstrations are randomly sampled from the training split with a fixed seed for reproducibility.

4 Experimental Setup

4.1 Inference Configuration

Generation uses the following parameters:

- Maximum new tokens: 30 (sufficient for single-word answers)
- Temperature: 0.2 (low temperature for more deterministic outputs)
- Sampling: Disabled (do_sample=False for greedy decoding)

Each model's tokenizer applies its specific chat template to format the prompt according to its training conventions. We extract only the content after "ANSWER:" from generated responses and apply post-processing to map outputs to category labels.

4.2 Response Processing

The `process_response` function handles various model output formats:

- Direct category names (e.g., "threats", "derogation")
- Variations with spaces or hyphens ("not sexist", "non-sexist")
- Numeric outputs (0-4) mapping to categories
- Substring matching for partial responses

Responses that cannot be mapped to any valid category are marked as parsing failures.

4.3 Evaluation Metrics

We report two complementary metrics:

- **Macro-F1**: The unweighted average of per-class F1 scores, treating all categories equally regardless of their frequency. This metric is particularly important for imbalanced datasets.
- **Fail Ratio**: The proportion of sexist samples (threats, derogation, animosity, prejudiced) that are misclassified as not-sexist. This captures the critical error of missing harmful content.

5 Results

Table 1 summarizes performance across both models and prompting strategies.

Model	Macro-F1	Fail Ratio
Phi-3-mini (zero-shot)	0.374	0.010
Mistral-7B (zero-shot)	0.332	0.017
Phi-3-mini (few-shot)	0.315	0.127
Mistral-7B (few-shot)	**0.471**	0.053

Table 1: Performance comparison across models and prompting strategies. Bold indicates best performance.

5.1 Zero-Shot Performance

In the zero-shot setting, Phi-3-mini slightly outperforms Mistral-7B in macro-F1 (0.374 vs 0.332), despite being roughly half the size. Both models achieve very low fail ratios (1.0% and 1.67% respectively), indicating strong instruction-following capability and adherence to the specified output format.

This result is somewhat surprising given that larger models typically perform better. The relatively small performance gap suggests that for this task under zero-shot conditions, the compact Phi-3 model's instruction-tuning is sufficiently effective, and additional scale does not provide substantial benefits.

5.2 Few-Shot Performance

The introduction of in-context examples produces dramatically different effects on the two models:

- **Mistral-7B** shows substantial improvement, increasing macro-F1 from 0.332 to 0.471 (+42% relative improvement). The fail ratio increases moderately to 5.33% but remains acceptable.
- **Phi-3-mini** experiences performance degradation, with macro-F1 dropping from 0.374 to 0.315 (-16% relative decline). More concerning, the fail ratio jumps to 12.67%, indicating the model now frequently misses sexist content.

This divergence suggests that the smaller Phi-3 model struggles to effectively leverage in-context examples. Possible explanations include:

- Limited capacity to maintain both the task definition and multiple examples in context
- Interference between the zero-shot instruction-following behavior learned during fine-tuning and the pattern-matching behavior encouraged by examples
- Insufficient context processing capability in the smaller architecture

The larger Mistral-7B model, conversely, effectively utilizes the demonstrations to improve its understanding of category boundaries, achieving the best overall performance.

6 Error Analysis

Figure **??** shows confusion matrices for all four experimental conditions.

6.1 Zero-Shot Error Patterns

Phi-3-mini shows relatively balanced confusion across sexism categories but struggles particularly with derogation, correctly identifying very few instances. Many derogation cases are misclassified as animosity or prejudiced, suggesting difficulty distinguishing between types of negative content about women.

Mistral-7B demonstrates better recognition of threats (correctly identifying most instances) but shows significant confusion between derogation and animosity, with 46 derogation samples misclassified as animosity. This indicates the model may focus on the presence of negative language rather than the specific semantic distinction between demeaning descriptions and direct insults.

6.2 Few-Shot Error Patterns

Mistral-7B (few-shot) achieves the strongest performance, particularly excelling at the prejudiced category with 55 correct classifications. The examples appear to help the model learn the subtle distinction between expressing prejudiced views and other forms of sexism. However, persistent confusion remains between derogation, animosity, and prejudiced categories, with some systematic misclassification of derogation as animosity.

Phi-3-mini (few-shot) shows increased confusion across all categories, with particular difficulty in the not-sexist class. The elevated fail ratio reflects a tendency to over-predict sexist categories, possibly because the model over-fits to the example patterns and loses the precision of its zero-shot instruction-following.

6.3 Category-Specific Challenges

Several patterns emerge across all conditions:

- **Derogation vs Animosity**: Frequent confusion between these categories across all models. Both involve negative language about women, but derogation focuses on demeaning descriptions while animosity involves direct insults or slurs. The semantic overlap makes this distinction challenging.
- **Prejudiced**: Difficult to detect in zero-shot but improves substantially with examples (particularly for Mistral). This category requires understanding systemic inequality concepts that benefit from demonstration.
- **Threats**: Generally the easiest category to identify, likely due to explicit violent or harmful language that provides clear lexical cues.
- **Not-sexist**: Most models handle this well in zero-shot, but few-shot Phi-3-mini shows degradation, suggesting the examples bias it toward over-detecting sexism.

7 Conclusion

This study demonstrates that instruction-tuned LLMs can perform fine-grained sexism classification through prompting, with performance varying significantly based on model scale and prompting strategy.

Key findings include:

- In zero-shot settings, the smaller Phi-3-mini performs comparably to the larger Mistral-7B (macro-F1 0.374 vs 0.332), suggesting that for simple instruction-following, scale advantages are limited.

- Few-shot prompting provides substantial benefits for the larger model (Mistral-7B improves to 0.471 macro-F1) but degrades performance for the smaller model (Phi-3-mini drops to 0.315 macro-F1).

- Both models struggle with distinguishing between derogation and animosity, indicating this semantic boundary is difficult to capture through prompting alone.

- The fail ratio metric reveals that while overall accuracy varies, most models maintain relatively low rates of missing sexist content entirely, except few-shot Phi-3-mini.

The divergent response to few-shot examples suggests that in-context learning effectiveness is not universal across model scales. Smaller models may require different prompting strategies or benefit more from fine-tuning rather than in-context learning.

Future work could explore:

- Chain-of-thought prompting to provide explicit reasoning steps

- Hierarchical classification (first binary sexist/not-sexist, then fine-grained type)

- Optimal example selection strategies rather than random sampling

- Alternative few-shot configurations (different K values) to find the sweet spot for smaller models

- Integration with retrieval systems to dynamically select relevant examples

For practitioners, our results suggest that model selection should consider the availability of demonstrations: zero-shot applications favor smaller efficient models, while few-shot scenarios benefit from larger models despite increased computational cost.

Figure 1: Confusion matrices for zero-shot and few-shot settings across both models. Rows represent true labels, columns represent predictions.