Sexist Intention Detection in Social Media: Assignment 2

A2: LLM prompting

Omid Nejati, Alireza Shahidiani, Matheus de Almeida
Master's Degree in Artificial Intelligence, University of Bologna
omid.nejati, alireza.shahidiani, matheus.dealmeida @studio.unibo.it

Abstract

This report describes our solution to Assignment 2 of the NLP course, which addresses the EDOS Task B "Category of Sexism" using large language model (LLM) prompting. We use two instruction-tuned LLMs, Phi-3-mini and Mistral-7B-Instruct, and treat the task as in-context learning instead of supervised fine-tuning. We design zero-shot and few-shot prompts that explain the task, define the four categories (not-sexist, threats, derogation, animosity, prejudiced discussion), and optionally include balanced demonstrations. On the official balanced test set of 300 posts (60 per class), we compare zero-shot and few-shot prompting for both models, reporting macro F1 and accuracy. Our best configuration, Mistral-7B with few-shot prompting, reaches a macro F1 of 0.4698 and an accuracy of 0.4800, outperforming its zero-shot version and the Phi-3 variants. The analysis highlights typical confusions (especially between derogation and animosity), the benefit of few-shot prompting for Mistral-7B, and the limited gains for Phi-3.

1 Introduction

The EDOS shared task focuses on detecting and categorising online sexism. In Task B, each English post must be assigned to one of four categories: threats, derogation, animosity, or prejudiced discussion, plus a not-sexist class for non-sexist content. Accurate automatic labelling can support moderation and analysis of online platforms.

In Assignment 1 we approached sexist intention detection with supervised BiLSTM and transformer models trained on EXIST. In this assignment we study whether LLM prompting alone, without any gradient-based training, can produce competitive performance on EDOS Task B. Prompting is attractive because it is simple to deploy and can be adapted to new tasks by editing the prompt instead of retraining a model.

We evaluate two open LLMs accessed via HuggingFace Transformers: microsoft/Phi-3-mini-4k-instruct and mistralai/Mistral-7B-Instruct-v0.3, both loaded with 8-bit quantization on a GPU. We compare zero-shot prompting, where the model only sees task instructions and category definitions, with few-shot prompting, where we additionally provide a small balanced set of labelled demonstrations.

2 System description

Data. We use the EDOS Assignment 2 files provided in the course material. The test set `a2_test.csv` contains 300 English posts, evenly balanced over the five labels (60 not-sexist, 60 threats, 60 derogation, 60 animosity, 60 prejudiced). The `demonstrations.csv` file provides a larger pool of labelled examples from which we sample few-shot demonstrations.

Labels are mapped to integers with a robust string-based function that detects the presence of label keywords (e.g., "threat", "derogat", "animosit", "prejudic") and maps them to 0 (not-sexist), 1 (threats), 2 (derogation), 3 (animosity), 4 (prejudiced).

Prompt templates. We follow the chat format expected by each model. All prompts share a system message declaring the model to be "an annotator for sexism detection", followed by a user message.

The zero-shot user message explains the task, lists the five labels, and includes short natural-language definitions of the four sexist categories. The model is instructed to respond only with one of the label names: *not-sexist, threats, derogation, animosity, prejudiced*.

For few-shot prompting we extend the user message with an `EXAMPLES` block. We build this block by sampling, for each class, two examples from `demonstrations.csv` (10 demonstrations in total, balanced across labels). Each example is formatted as two lines: `TEXT: <tweet>` and `ANSWER: <label>`. These are followed by the test instance and the final `ANSWER:` slot to be filled by the model.

Generation and parsing. We generate responses with the HuggingFace `generate` API, using a maximum of 100 new tokens, nucleus sampling (`top_p=0.95`) and temperature 0.7. For efficiency we process multiple prompts per batch.

Generated text is mapped back to label indices with a robust parser that lowercases the response and checks for several lexical variants (e.g., "non-sexist", "not sexist"). If no keyword is recognised, the prediction defaults to not-sexist (label 0), which is a conservative choice.

3 Experimental setup and results

Models. Both models are loaded with 8-bit quantization (BitsAndBytes) and executed on a single GPU:

- **Phi-3-mini**: microsoft/Phi-3-mini-4k-instruct
- **Mistral-7B**: mistralai/Mistral-7B-Instruct-v0.3

We keep the same generation parameters for all runs.

Evaluation protocol. For each model we run:

1. zero-shot prompting on all 300 test examples;
2. few-shot prompting with the 10-example balanced demonstration block.

We compute macro-averaged F1 and accuracy over the five labels, and also inspect per-class precision/recall and confusion matrices.

Results. Table 1 reports macro F1 and accuracy for all four configurations.

Model / Setting	Macro F1	Accuracy
Phi-3-mini (zero-shot)	0.4365	0.4467
Phi-3-mini (few-shot)	0.4147	0.4400
Mistral-7B (zero-shot)	0.3468	0.3767
Mistral-7B (few-shot)	**0.4698**	**0.4800**

Table 1: Classification performance on the EDOS test set (300 examples, balanced over 5 labels).

Phi-3-mini achieves similar accuracy in zero-shot and few-shot conditions, with a slight drop in macro F1 when demonstrations are added. Mistral-7B benefits more from few-shot prompting, improving macro F1 from 0.3468 to 0.4698 and accuracy from 0.3767 to 0.4800.

4 Discussion

Quantitative analysis. The results show that prompting alone can reach moderate performance on EDOS Task B, with the best configuration reaching macro F1 0.47. Few-shot prompting is clearly helpful for Mistral-7B but not for Phi-3-mini. This suggests that the larger Mistral model can better exploit the additional demonstrations, whereas Phi-3-mini may already be close to its capacity with the task description alone.

Per-class scores (from the classification reports) reveal strong variation across labels. For Phi-3 zero-shot, the model performs best on threats (F1 0.61) and prejudiced discussion (0.53) but struggles with derogation and animosity (F1 around 0.22–0.32). Mistral zero-shot heavily over-predicts animosity, achieving high recall but low precision for that class, and very low recall for threats and derogation.

Few-shot prompting helps Mistral-7B balance its predictions. Threats and prejudiced discussion both reach F1 above 0.56, and animosity improves to 0.39. However, all configurations still find derogation difficult, with F1 below 0.20 even in the best setting.

Error patterns. The confusion matrices show systematic confusions between derogation and animosity: derogatory statements about women that include insults are often predicted as animosity rather than derogation. There are also mistakes between not-sexist and prejudiced discussion when the text uses abstract or indirect language about gender roles. Some borderline examples are simply ambiguous, and the correct label depends on fine-grained interpretation of intent.

Another observation is that both models tend to classify many sexist posts as prejudiced discussion, probably because the definition mentions "support for mistreatment", which overlaps conceptually with the other categories. More precise label descriptions or additional examples focusing on the boundaries between categories might mitigate this.

5 Conclusion

We implemented an LLM-based system for EDOS Task B using zero-shot and few-shot prompting with Phi-3-mini and Mistral-7B. Our best configuration, Mistral-7B with few-shot prompting, achieves a macro F1 of 0.4698 and accuracy of 0.4800 on a balanced 300-example test set, without any task-specific training. While this is likely below the performance of fine-tuned transformers, it shows that prompting with open LLMs can provide a simple baseline for sexism categorisation.

The main limitations are the persistent difficulty in recognising derogation, the sensitivity to prompt wording, and the computational cost of running large models over many prompts. Future work could explore prompt optimisation, class-specific demonstrations aimed at difficult distinctions, and hybrid systems where LLM predictions are used to

train smaller supervised models.

Links to external resources

- EDOS task resources and dataset: `https://github.com/rewire-online/edos`
- HuggingFace model cards: `microsoft/Phi-3-mini-4k-instruct`, `mistralai/Mistral-7B-Instruct-v0.3`

References

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.