

ALC - MIARFID

Laboratory Lesson 2

Sexism Identification in Memes

6th March 2025

Objectives

In this laboratory session, we aim to design and implement a computational system capable of automatically detecting sexism in Memes. Unlike previous work, we address the current laboratory session from a different perspective:

- **Exploring the potential of multimodal approaches** by integrating additional data from annotator participants using electroencephalography, eye-tracking, and hear-rate devices.
- **Exploiting the use of other datasets** to boost our model's overall performance, even when they do not contain the same modality inputs, via in-domain fine-tuning or direct inclusion during training.

As this laboratory aligns with the scope and objectives of the NLP-oriented subject, **the use of image-based approaches** —i.e., directly using the **memes** — **is optional**. Of course, you are welcome to explore them if you are interested; however, please note that this will not impact the evaluation of your laboratory work.

The EXIST2026 Challenge

This lab session is designed to simulate the participation in the **Task 2.1: Sexism Identification in Memes** of the EXIST 2026 Challenge¹. This binary classification task, as detailed below, includes memes in both Spanish and English, encouraging the development of multilingual models. Figure 1 shows examples of the memes we can find in the training dataset of EXIST2026, illustrating the multi-lingual nature of the challenge task.

¹<https://nlp.uned.es/exist2026/>



Figure 1: Examples of memes in the EXIST 2025 Meme Dataset, showcasing the inherent challenges of the proposed multimodal and multilingual task.

Notably, **in this challenge edition, multimodal data was collected**, comprising features that capture neural, behavioral, and physiological signals:

- **Electroencephalography (EEG):** Power in Delta, Theta, Alpha, Beta, and Gamma bands across 16 channels using an OpenBCI headset.
- **Eye-Tracking (ET):** Pupil diameter statistics, fixation, blink and saccade durations, and reaction times, collected with a Pupil Labs Neon.
- **Heart-Rate (HR):** Overall statistics including mean, standard deviation, min/max heart rate values, measured using a Garmin Venu 3 watch.

These multimodal recordings provide a rich set of features for interesting analyses. An overview of the data collection setup is shown in Figure 2.



Figure 2: Overview of the multimodal data collection setup.

The MAMI Dataset

As introduced earlier, this laboratory session also focuses on leveraging similar datasets to enhance our model’s overall performance. While the common belief is that “*the more data, the better*,” this is not always the case. You are encouraged to explore various approaches, such as directly incorporating all MAMI

samples during training or performing preliminary in-domain fine-tuning of the model before adapting it to your target task.

The MAMI dataset was designed to detect misogynous memes on the web by leveraging both text and images. For further information, you can refer to the original paper [1] and its corresponding GitHub repository. **A key distinction between this corpus and the EXIST challenge is that the MAMI dataset includes only English memes.** Moreover, the fact that the corpus also lacks multimodal data further complicates their integration and highlights the need for careful handling when leveraging it.

Across different tasks, a common practice is to leverage English-based pre-trained models to address less-represented languages. Do you think this approach could work in the context of misogyny detection in memes?

Lab Materials

All necessary data and metadata required for this laboratory session can be accessed through the subject's PoliformaT platform, where we can find the:

- `dataset_task2_exist2026/` directory, which contains two JSON files specifying our training and test samples of the EXIST2026 Task 2.1 challenge dataset, along with their associated metadata. Among this metadata, apart from the sample ID and multiple labels, we can find the *sensorial* field, which **stores the multimodal features described above** for each subject. Another relevant field is *text*, which contains the OCR transcription of the text displayed in the meme image. **Besides, this directory also includes** two CSV files containing the BLIP²-based image captions for both data sets. *Please note that while some memes are in Spanish, all the BLIP-generated image captions are always in English.*
- `golds_task2_exist2026/` directory, which contains a JSON file specifying the gold standards for the training set. They are the reference ground-truth labels to evaluate the performance of your models.
- `mami_dataset.zip` compressed file containing both the OCR³-based transcription of the text displayed in the meme and the BLIP-based image captions for the official training and test samples of the MAMI dataset.

An important point to note is that, as in a real challenge, **labels for the test set are not provided**. *Please refer to the implementation details outlined in the previous laboratory session* to address strategies for handling model evaluation, multiple annotation labels, and the use of PyEvALL.

²<https://huggingface.co/Salesforce/blip-image-captioning-large>

³<https://cloud.google.com/vision/docs/ocr>

If you're interested, the meme images for EXIST2026 are available in PoliFormaT. Specifically, the "[OPTIONAL] Memes of EXIST2026/" directory contains ZIP files with the images from both our training and test sets. Remember, the use of memes is optional and will not affect the evaluation of your laboratory work.

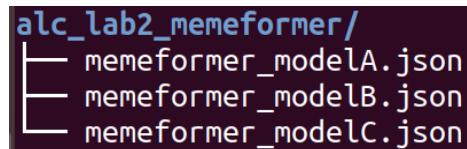
What Should You Do?

In this lab session, you should: (i) study the effectivity of integrating multimodal signals for sexism detection in memes; (ii) explore the use of external yet similar datasets, as well as distinct fine-tuning approaches, to boost your model's overall performance; (iii) and familiarize with model checkpoint savings and loading when working with machine/deep-learning models. This may involve reading scientific papers related to the task in order to gather insights on which fine-tuning methods might be useful.

For reference, you can consult the official EXIST2025 challenge overview paper [2]. In particular, focus on Subsection 6.4.2 (Table 13), where you will find the leaderboard with the final ranking of the task under hard-label settings. A brief description of the models involved is found in Section 5.

As you can observe, the overall F_1 -score of the top-performing models is around 77%. The aim of this lab session is not to surpass the state of the art, but to learn. Once you finish your experiments, you must submit via PoliFormaT:

- **A ZIP file with your top-3 models.** This compressed file should thereby contain three JSON files, one for each of your top models. Each JSON file must include the model's predictions for the evaluation dataset, formatted according to the guidelines provided in the first lab bulletin. *This will be part of our in-home challenge :) Who will be the winner?* Please ensure that your JSON files are well-organized, as exemplified by the image below, from which we can infer that our files should follow the naming template: <GROUP-ID>_<MODEL-ID>.json⁴.



```
alc_lab2_memeformer/  
├── memeformer_modelA.json  
├── memeformer_modelB.json  
└── memeformer_modelC.json
```

- **A single-page PDF report briefly describing your approach.** This should include details on the ablations studies conducted, fine-tuning methods explored, the rationale behind your choices, and hypotheses regarding why your approach is or is not working. Perhaps, analyzing the

⁴**Warning:** any model name that can be considered offensive or inappropriate may result in a penalty or, in severe cases, disqualification of the laboratory work.

data distribution of some feature values as a function of the labels might provide some valuable clues. Not everything is training models :)

Deadline: 11th March 2026, 23h55

You can work individually or in groups of up to two people.

In the case of group submissions, only one member should upload the final work, clearly indicating the names of all participants in the corresponding task available and prepared in PoliFormaT.

References

- [1] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022. [Link to PDF](#).
- [2] Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos. In *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, pages 266–289. Springer, 2025. [Link to PDF](#).