GrootWatch at EXIST 2025: Automatic Sexism Detection on Social Networks - Classification of Tweets and Memes

Notebook for the EXIST Lab at CLEF 2025

Nathan Nowakowski[1,*], Lorenzo Calogiuri[1], Előd Egyed-Zsigmond[1], Diana Nurbakova[1], Johan Erbani[1] and Sylvie Calabretto[1]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France

Abstract

This paper presents our participation in the EXIST (sEXism Identification in Social neTworks) challenge at CLEF 2025, focusing on the classification of tweets and memes. We participated in all the tasks for tweets and memes, including both hard and soft classifications for tweets and hard classification for memes. For tweet classification, we propose a multi-task headed BERT model enriched with relevant information surrounding the tweet, helping the model achieve a full understanding of the tweet and its context. For memes, the paper explores the use of a Vision-Language Model (VLM)-based application to detect and categorise sexism in different scenarios, leveraging the ability of such models to understand the relationship between images and text in situations where sexist ideas are often expressed subtly. Our solutions achieved excellent performance, ranking first in all soft-soft tweet classification tasks and second in all hard-hard meme classification tasks.

Content Warning: This paper includes examples of hateful, explicit and sexist language presented for illustrative purposes.

Keywords

Sexism Identification, Text Classification, Image Classification, Natural Language Processing, Transformers

1. Introduction

Sexism, in the form of pre-judges or hateful comments, is a prevalent form of digital violence that must be addressed in a context where social networks and digital platforms are ubiquitous. In 2024, 81% of French women reported experiencing sexist comments on these platforms [1]. This concerning situation presents a major societal challenge, creating a balance between the ethical expectations of moderation and the need to protect free expression. This work takes place against a backdrop in which platforms such as Meta are drastically relaxing their moderation policies, exacerbating the risks of polarisation and gendered hatred [2, 3]. At the same time, masculinist discourse is gaining in visibility, making it essential to develop tools capable of mapping and countering these dynamics in real time. Today's forms of sexism extend beyond verbal attacks, with diverse representations such as videos, comments, or images appearing on platforms like X (former Twitter), Instagram or TikTok [4].

Therefore, automatic identification of sexist content on social media becomes a crucial task. To foster such initiatives, the EXIST 2025 challenge [5, 6] comprises nine subtasks in two languages, English and Spanish, which are the same three tasks (sexism identification, source intention detection, and sexism categorisation) applied to three different types of data: text (tweets), image (memes), and video (TikToks).

- **Sexism Identification** (Subtasks 1.1, 2.1 and 3.1) : This task involves binary classification to determine whether a given tweet, meme, or TikTok video contains sexist expressions or behaviours. The categories are **YES** and **NO**.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

nathan.nowakowski@insa-lyon.fr (N. Nowakowski); lorenzo.calogiuri@insa-lyon.fr (L. Calogiuri); elod.egyed-zsigmond@insa-lyon.fr (E. Egyed-Zsigmond); diana.nurbakova@insa-lyon.fr (D. Nurbakova); johan.leydet@insa-lyon.fr (J. Erbani); sylvie.calabretto@insa-lyon.fr (S. Calabretto)

- **Source Intention Detection** (Subtasks 1.2, 2.2 and 3.2): Once a message has been classified as sexist, this task aims to categorise the message according to the intention of the author. For tweets and videos, the categories are **DIRECT**, **REPORTED**, and **JUDGEMENTAL**. For memes, due to their characteristics, the REPORTED label is virtually null, so systems should only classify memes with **DIRECT** or **JUDGEMENTAL** labels.
- **Sexism Categorisation** (Subtasks 1.3, 2.3, 3.3): This task involves classifying sexist content into one or more categories: **IDEOLOGICAL AND INEQUALITY**, **STEREOTYPING AND DOMINANCE**, **OBJECTIFICATION**, **SEXUAL VIOLENCE**, and **MISOGYNY AND NON-SEXUAL VIOLENCE**.

The categories of sexism used in this study are defined on the EXIST 2025 challenge website. Given the complexity and the need for comprehensive detection tools, we decided to tackle both tweet-based subtasks (1.1, 1.2, 1.3) and meme-based subtasks (2.1, 2.2, 2.3) in our work. To address this challenge, we evaluated and compared state-of-the-art techniques, incorporating our insights to propose two tailored solutions: one for textual classification and another for meme classification.

The remainder of the paper is organised as follows. In Section 2, we provide a brief overview of approaches used for automatic detection of sexist content. We then describe the dataset and the evaluation metrics in Section 3. We describe our proposed solutions for tweets and memes in Section 4. We report the results of our experiments in Section 5. Section 6 concludes the paper and outlines the directions for future work.

2. Related Work

In this section, we present the different approaches used to detect online sexism. These methods fall into four broad categories: traditional approaches, Deep Learning-based approaches, transform-based approaches (BERT and LLM) [7, 8], and multimodal approaches.

Before the emergence of deep architectures, a number of studies used classic machine learning methods - such as Logistic Regression, SVMs or Random Forests. These methods were generally combined with feature extraction techniques (N-Grams, TF–IDF, Static Word Embeddings) [9]. While these approaches provided reasonable performance, they were limited in their ability to handle contextual variations and language evolution.

Deep Learning models have made it possible to capture complex patterns using specialised architectures. CNN-BiLSTM architectures, combining convolutional neural networks (CNNs) to detect local patterns (e.g. offensive N-Grams) and BiLSTMs to model long-term contextual dependencies, marked a significant advance [10].

The advent of transformers has revolutionised the detection of sexism thanks to their ability to encode the overall context of text:

- **BERT and derivatives**: Models such as RoBERTa [11] or DeBERTa [12], pre-trained on massive corpora, capture semantic nuances and sexist undertones [13].
- **LLM and contextual reasoning**: LLMs (e.g., Llama-3 [14]) fine-tuned with methods like LoRA [15] incorporate advanced reasoning capabilities, essential for interpreting emerging cultural references or sarcasm [16].
- **Enrichment by sentiment analysis**: Sentiment analysis techniques are used to enrich transform models in order to detect emotional nuances and tonality. This approach proves effective in spotting sexist comments sometimes disguised under a veneer of positive or neutral sentiment [17].

Existing datasets have played a crucial role in advancing the field of online sexism detection. Notable examples include:

- **Sexist Stereotype Classification (SSC)** [18]: Collected from Instagram hashtags like #bloodymen and #metoo, this English dataset contains 5,544 comments annotated manually and through active learning.

- **Semeval 2023 Task 10** [19]: Focused on explainable detection of online sexism, this dataset includes 20,000 English comments from Gab and Reddit, annotated by 19 female annotators with expert review for disagreements.
- **EXIST 2021-2025** [20, 21, 22, 23, 5]: These datasets comprise tweets in English and Spanish, with the detailed annotator demographics included starting from 2023. Notably, the definition of sexism varies across sociocultural contexts and annotator biases. The adoption of paradigms like Learning with Disagreements (LeWiDi) enables consideration of multiple and sometimes contradictory annotations, thus improving model robustness [24].

These datasets have contributed significantly to our understanding of online sexism, enabling researchers to develop more accurate and robust detection methods. With the rise of visual social media platforms, sexism is increasingly conveyed through multimodal forms such as memes, which blend text and images to encode prejudice in subtle, culturally loaded ways. This shift has spurred research into models capable of understanding both modalities simultaneously. Several key datasets support this area:

- **MAMI (SemEval-2022)** [25]: A benchmark dataset with 10,000 memes annotated for sexism and fine-grained categories (e.g., shaming, objectification).
- **MIMIC** [26]: A Hindi-English code-mixed dataset tackling misogyny in multilingual, multimodal memes with classification tasks.
- **EXIST 2024–2025** [23, 5] : A shared task that extended sexism detection to memes and, more recently, TikTok videos, leveraging the LeWiDi paradigm for multilingual and multimodal challenges.

The correlation between textual and visual elements in memes make VLMs (Vision Language Models), architectures built by combining large language models and vision encoding, suitable for the task. Several studies have been done on applying VLMs to social media memes for semantic understanding [27] and hate speech detection both in a zero-shot [28] and fine-tuning paradigm [29], showing the effectiveness of this approach. Among the methodologies utilised, the principal models applied in this research fall into the following categories:

- **Transformer-based multimodal systems** combining textual encoders, such as BERT, with visual representations often extracted via CLIP [30].
- **LLMs** such as GPT-4 are integrated in the classification pipeline to enrich memes with inferred context and deeper semantic understanding [23].
- **Multi-task VLMs** such as Florence 2 [31] and Qwen 2.5 VL [32] shows strong generalisation for cross-modal inputs.
- **Lightweight and multilingual models** like Mistral 3.1 Small [33] and Aya Vision 8B [34], offer high performance with lower resource requirements, supporting deployment across varied linguistic and visual settings.

Despite notable advances, several challenges persist:

1. **Knowledge obsolescence:** Pre-trained models possess frozen knowledge that may not always capture recent language and usage developments in tweets, limiting their relevance in current contexts. Valavi et al. [35] emphasises the need to periodically refresh training data to maintain high performance.
2. **Contextual dependence:** Correct classification often relies on information not present in the text itself (e.g., current events, cultural references, emerging trends).
3. **Oversight of visual cues:** Many methods overlook the information present in images, relying mostly on accompanying text for meme analysis [36, 37].
4. **Costly integration:** Some approaches integrate image features using large, proprietary models like GPT-4 [38], but this comes at a significant computational cost and with limitations.

In this paper, we tackle the aforementioned limitations by proposing two novel approaches that enhance the performance and robustness of sexism detection models in social media, as well as in tweets and in memes. The details of our methodology are presented in Section 4, which outlines how we address these challenges and advance the state-of-the-art in sexism detection.

3. Dataset and Evaluation Overview

Our study is based on the EXIST 2025 dataset, which offers a rich collection of tweets and memes annotated for online sexism detection. We draw upon these subsets to train and evaluate our approach. Tables 1 and 2 resume the datasets for tweets and memes respectively.

Table 1
EXIST 2025 tweets dataset language distribution and splitting (Number of tweets)

Language	Training	Development	Test
English	3260	489	978
Spanish	3660	549	1098

Table 2
EXIST 2025 memes dataset language distribution and splitting (Number of memes)

Language	Training	Development	Test
English	2010	-	513
Spanish	2034	-	540

In the subsequent experimental phase, we will conduct model fine-tuning using the labelled training set, followed by evaluation on the development dataset. Since the meme dataset does not explicitly provide for a development dataset, the training set was divided into two 80/20 partitions (seed: 1234), respectively for fine-tuning and for evaluating results on never-before-seen data. Ultimately, our approach will be benchmarked against challengers using the held-out test set.
The official evaluation metric for this challenge is the Information Contrast Measure (ICM) [39]. Throughout this report, we will employ the normalised variant, ICM Norm, to assess the performance of our models. We opted for ICM Norm due to its enhanced readability, which results from its normalisation to a maximum value of 1. Due to the class imbalance in the dataset, as showed in Table 3, we also provided the F1 score for hard classification to better capture the trade-off between precision and recall. With respect to the given tables, the Unlabelled class corresponds to records where annotator consensus was not reached, thereby precluding a definitive ground truth assignment. Furthermore, the percentages for subtasks 1.3 and 2.3 do not add up to one hundred, due to the nature of the task as a multi-label task.

4. Methodology

This methodology is structured into two distinct components. The first part focuses on our approach to tweet analysis, while the second part details our method for meme analysis.

4.1. Tweets

4.1.1. Data processing

The previous comprehensive literature review on classification techniques revealed that BERT and LLM models are at the forefront of natural language processing tasks. Given their state-of-the-art performance, we focused our efforts on these models. Our initial step involved conducting multiple tests to determine the optimal formatting for tweets to be processed by BERT. This process ensured

Table 3
EXIST 2025 class distribution on the training dataset

Label	Class proportion Tweets (%)	Class proportion Memes (%)
Subtasks 1.1 and 2.1		
YES	38.97	50.39
NO	48.65	34.17
Unlabelled	12.37	15.44
Subtasks 1.2 and 2.2		
NO	48.65	34.17
DIRECT	18.70	33.50
JUDGEMENTAL	5.43	11.38
REPORTED	6.63	-
Unlabelled	20.58	20.95
Subtasks 1.3 and 2.3		
NO	48.65	34.17
IDEOLOGICAL-INEQUALITY	16.08	20.84
STEREOTYPING-DOMINANCE	20.56	25.29
OBJECTIFICATION	20.56	23.34
SEXUAL-VIOLENCE	22.12	11.94
MISOGYNY-NON-SEXUAL-VIOLENCE	12.37	9.81
Unlabelled	12.57	15.75

that the input data was structured to maximise the model's performance. For LLM, Quan and Thin [16] indicated that extensive formatting was unnecessary, simplifying our preprocessing pipeline. At the end, tweets for BERT were pre-processed using the steps described in Table 4.

Table 4
Data cleaning process for BERT model: removal of mentions (@), URLs, superfluous spaces, emojis and HTML characters, conversion of text to lower case, and word decensoring (identifying and correcting censored words, e.g., '*f**k*'→'*fuck*').

id	Before formatting	After formatting
200176	Feel #blessed that I have raised a caring & loving 13 yo who is our Next Gen Feminist & Ally. I was crying inside when I got this text. Not only we must #BreakTheBias for women, we need to do it for our children. 🗼🍣💎🧋 @GlobalFundWomen @UN_Women @womensday @WomeninID https://t.co/UJvloR0IP	feel blessed that i have raised a caring loving 13 yo who is our next gen feminist ally. i was crying inside when i got this text. not only we must breakthebias for women, we need to do it for our children.

4.1.2. Annotator Information Analysis

To investigate the impact of annotator characteristics on sexism detection, we conducted a comprehensive analysis of annotator information using Chi-Squared tests and Logistic Regression models with feature importance. To improve the model's understanding of the subjective nature of sexism, we identified study level, country of origin, and ethnicity as relevant annotator attributes through our analysis. By integrating these attributes, we aimed to enhance the model's ability to capture diverse perspectives and biases. To achieve this, we vectorised the selected annotators' information and embedded it into the CLS token of the BERT model, prior to passing it to the classification head.
Table 5 illustrates a simplified example of the vectorisation process of annotator information, featuring three annotators for clarity. Note that in our actual implementation, the final vector is 65 elements

long, encompassing a more extensive range of ethnicities (more than 3), study levels (more than 2), and countries (more than 2). This simplified representation is intended to facilitate understanding and presentation.

Table 5
Vectorisation of Annotator Information for a Representative Tweet Evaluated by Three Fictive Annotators. (1) Each evaluated tweet includes annotator information categorised accordingly. (2) Each category is one-hot coded. (3) For each category, the corresponding one-hot-coded vectors are summed to generate a category vector. (4) The created category vectors are concatenated into a single annotator information vector. (5) Finally, the resulting vector is normalised.

Annotator information	Encoding
"ethnicities_annotators": ["White or Caucasian", "Hispano or Latino", "Asian"],	**ethnicities_annotators**: [1,0,0] [0,1,0][0,0,1] = [1,1,1]
"study_levels_annotators": ["High school degree or equivalent", "Master's degree", "High school degree or equivalent"],	**study_levels_annotators**: [1,0] [0,1][1,0] = [2,1]
"countries_annotators": ["Spain", "Portugal", "Portugal"]	**countries_annotators**: [1,0] [0,1][0,1] = [1,2]
	⇒ **Concatenation**: [1,1,1,2,1,1,2]
	⇒ **Normalisation**: [0.1,0.1,0.1,0.2,0.1,0.1,0.2]

4.1.3. Fine-Tuning and Initial Results

Our fine-tuning efforts with both BERT and LLM models yielded promising results (cf. Table 6), closely approaching the top performances achieved in the previous year [23]. The specific experience configurations employed are detailed in Appendix A. Notably, we drew inspiration from last year's edition [16] and complemented this with empirical tests on our side to determine the optimal hyperparameters and prompts.

Table 6
Subtask 1.1 Initial results for hard binary classification

Model	Configuration	ICM-Hard Norm	Macro F1
XLM Roberta Large	Fine-tuned	0.70	0.80
XLM Roberta Large	Fine-tuned + Annotator information	0.77	0.85
Llama-3.2-3B-Instruct	Zero-shot	0.44	0.59
Llama-3.2-3B-Instruct	Fine-tuned	**0.78**	**0.86**

However, we sought to further enhance our approach. An analysis of misclassification revealed that certain tweets were incorrectly classified due to their ambiguity or references to recent topics not present in the training data. For instance, tweets referencing very recent events or slang not included in the model's vocabulary posed significant challenges.

4.1.4. Leveraging AI Agents for Contextual Information

To overcome the limitations of traditional models, we leveraged the capabilities of AI agents that can dynamically interact with their environment with tools, plan actions, and integrate external data in real-time [40]. Our approach is exemplified in Figure 1, which illustrates the workflow of our agent when faced with an ambiguous tweet referencing a meme about a pregnant woman in Oklahoma. Initially

misclassified as sexist by our base model, which we hypothesise was due to the presence of keywords like 'woman' and 'pregnant', as our analysis of the TF-IDF representation of misclassified samples revealed that these words tend to dominate the feature space, leading to incorrect sexist classifications. To address this limitation, we propose an innovative solution: our AI agent intervenes to identify the need for context (1) and dynamically queries a search engine (2) to gather relevant information. The agent then analyses the search results (3), and extracts crucial context (4), enabling the capture of sexist–or non-sexist–connotations or nuances linked to recent events that are invisible to static models. If no additional context is required, the agent indicates "No external context needed.". By harnessing the potential of AI agents, we aim to improve the relevance and robustness of sexism detection, adapting to the rapid evolution of language and diverse contexts on social media platforms.

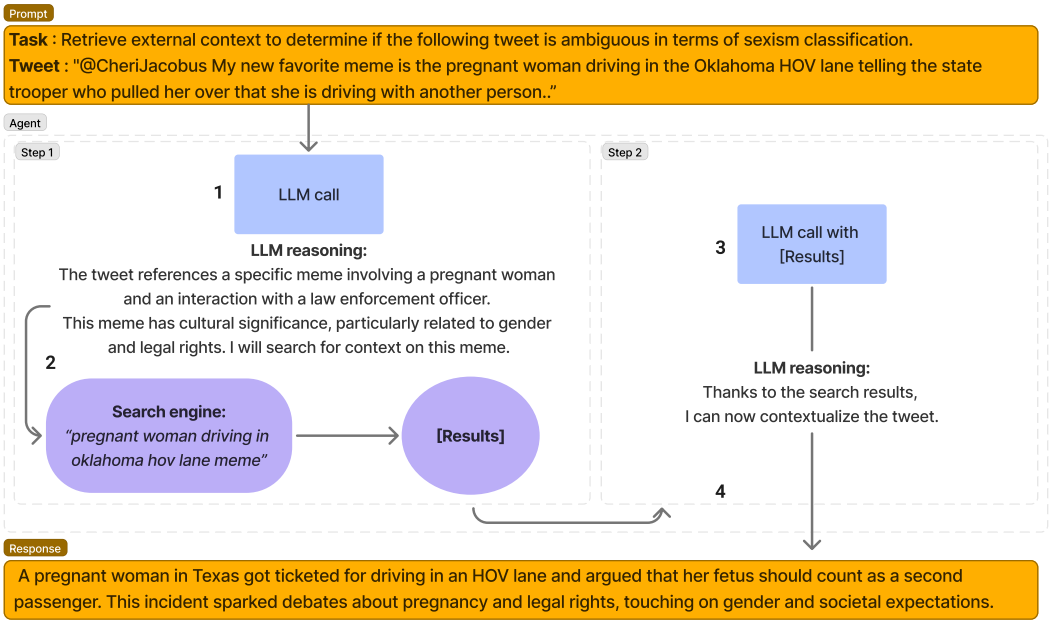

Figure 1: Detailed workflow of an agent analysing a tweet with an external call to a search engine.

We equipped our agent with the DuckDuckGo web search tool. We considered several options for utilising this AI agent:

- **Direct Classification by the Agent**: The agent classifies the tweet directly using relevant information gathered from the web search. Its user prompt is available in Appendix B
- **Context Retrieval by the Agent**: The agent retrieves contextual information around the tweet using the web search (the AI agent user prompt is available in Appendix C), which can then be fed into a BERT or LLM.
  - **BERT-based Architecture**: We fed the retrieved context into a BERT model, employing a Siamese Dual Encoder architecture (SDE) [41]. This design choice was motivated by our empirical findings, as alternative architectures yielded inferior results.
  - **LLM-based Approach**: We incorporated the retrieved context into the prompt, as detailed in Appendix E, and then fine-tuned an LLM to classify the tweet.

To optimise the LLM performance and output format for each experience configuration, various prompts have been empirically tested. Furthermore, a system prompt is appended to the LLM Agent by the library, facilitating correct parsing of its output for tool calls. For more information on the library details, see Section 5.1.

We employed two distinct LLMs in our approach: one for powering the autonomous AI agent and another for fine-tuning to classify tweets.

- **Autonomous AI Agent**: We chose the `Llama-3.3-70B-Instruct` model for its ability to handle complex tasks, which requires well-formatted responses to effectively leverage tools.
- **Fine-Tuned LLM for Classification**: Due to computational limitations, we used a smaller LLM, `Llama-3.2-3B-Instruct`, for fine-tuning. Despite using 4-bit quantisation, we lacked the necessary computational resources to fine-tune a 70B model.

Our experiments, presented in Table 7, reveal that BERT models augmented with contextual information outperform LLM with context, underscoring the efficacy of contextual enrichment on encoder-only architectures. In contrast, the incorporation of context into fine-tuned LLM appears to degrade performance, potentially due to the phenomenon of *context hijacking* [42], where the model overemphasises contextual cues. Nonetheless, the AI agent direct classification surpasses the zero-shot baseline in Table 6. Consequently, we will pursue a BERT-based architecture to fully leverage the potential of contextual research, as the LLM approach does not seem to yield comparable performance gains.

Table 7
Subtask 1.1 for hard binary classification, comparative experiments on development dataset of the context search and its impact on models

Model	Configuration	ICM-Hard Norm	Macro F1
Llama-3.3-70B-Instruct	AI agent alone	0.69	0.80
Llama-3.2-3B-Instruct	Fine-tuned + Context	0.60	0.73
XLM Roberta Large	Fine-tuned + Annotator Information + Context	**0.81**	**0.87**

Moving forward, our approach will leverage contextual information retrieved and formatted by an AI agent. As this external data is generated, evaluating its quality is a necessary consideration. An initial evaluation of the generated contexts is presented in Appendix D. While we do not delve further into this aspect in this paper, as it is not the primary focus, additional analysis may be merited for this case and future applications.

4.1.5. Soft Label Learning

One of the significant challenges we encountered was annotator disagreement, the Unlabelled data. When there was no clear majority—such as three "YES" and three "NO" or three "DIRECT" and three "JUDGEMENTAL"—we could not use these data points because we were training the model for hard label classification. This was not a trivial detail, as the amount of training data can impact model performance [43]. For instance, for the first task, we were losing around 10% of the data, and this loss increased with tasks 1.2 and 1.3.
A solution we identified was to train the model with probabilities rather than hard labels, aligning with the principles of soft label learning (SLL) as explored in [44]. This study demonstrated that incorporating information about the uncertainty of the outcome in classification models can significantly enhance performance compared to the standard approach of hard label learning (HLL). For example, when a tweet had annotations of five "YES" and one "NO," we previously provided "YES" as the training input. With probabilities, the input would be [0.83, 0.17]. This new formatting approach allowed us to achieve two key improvements: taking into account the whole training dataset and better capturing annotator discordance, aligning more closely with the LeWiDi paradigm. Our experiments demonstrated that this method improved the ICM-Hard Norm by 1 point and the ICM-Soft Norm by 2 points.

4.1.6. Model Runs and Performance

Now that we have selected the BERT architecture, in Figure 2 we conducted extensive runs with various of these models, including XLM-Roberta [11], Deberta V3 [12] and ModernBERT [45] variants. XLM-

Roberta emerged as the best-performing model with contextual injection and annotator information.

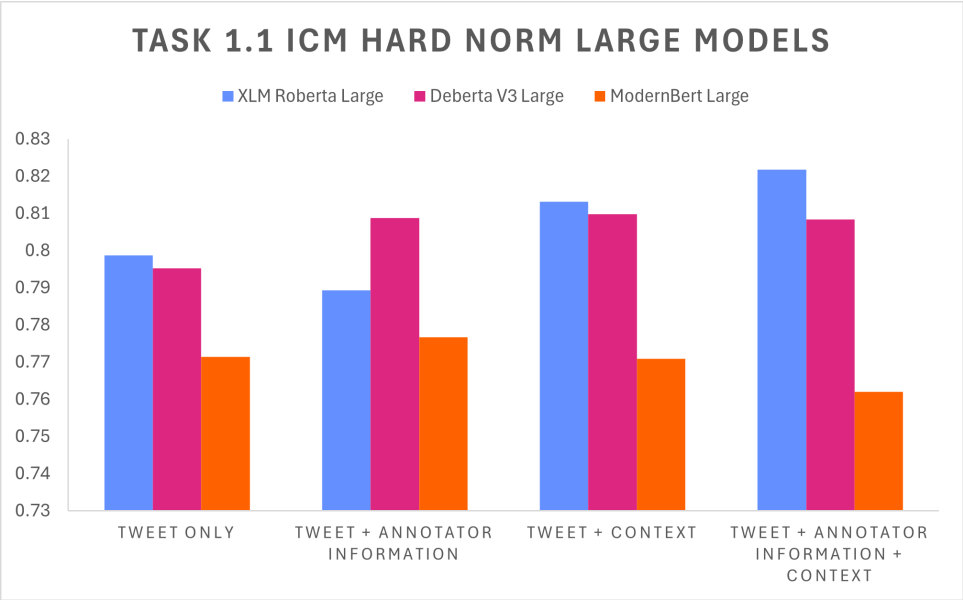

Figure 2: Comparison of three state-of-the-art BERT models with different input configurations on development dataset

4.1.7. Multi-Task BERT Architecture

One of the key advantages of selecting the BERT architecture is that, with minimal additional effort and computational resources, we can accommodate all three tasks and both hard and soft labels within a single multi-task BERT model [46, 47]. This design enables knowledge sharing across tasks by leveraging the base layers of the BERT model, while task-specific output heads capture the unique characteristics of each task.

Building upon the best existing approach, which employed a multi-task BERT [13], we sought to further improve it. Notably, our analysis revealed that the probability of a "NO" label remains consistent across all three tasks. This observation led us to propose a novel 2+1 architecture (cf. Figure 3), wherein one classification head is dedicated to softmax labels (subtasks 1.1 and 1.2) and another to sigmoid labels (subtask 1.3). Specifically, this design allocates the first two tasks to the first classification head and the third task to the second classification head.

A crucial aspect of our proposed architecture is that we leverage the consistency of "NO" probability across all three tasks. By recognising this consistency, we adapted our training approach to compute the loss of the Classifier B (subtask 1.3) only when the tweet is classified as sexist by the Classifier A. This hierarchical design enables us to filter out non-sexist examples and focus on the relevant samples for subtask 1.3, thereby improving performance and establishing coherence between the two classification heads despite their distinctness.

In contrast, our experiments with a single classification head for all categories did not perform well, likely due to the large number of categories. Similarly, attempting to predict only "YES" labels and computing "NO" labels by doing 1-"YES" also yielded subpar results.

Notably, this 2+1 architecture significantly impacted the performance of our results for subtasks 1.2 and 1.3. While subtask 1.1 results remained relatively consistent, our proposed architecture demonstrated substantial improvements for the latter two tasks. In particular, it led to a substantial improvement in soft classification, with an increase of two to three ICM Soft Norm points. The final results of our model are presented in the Section 5.2.

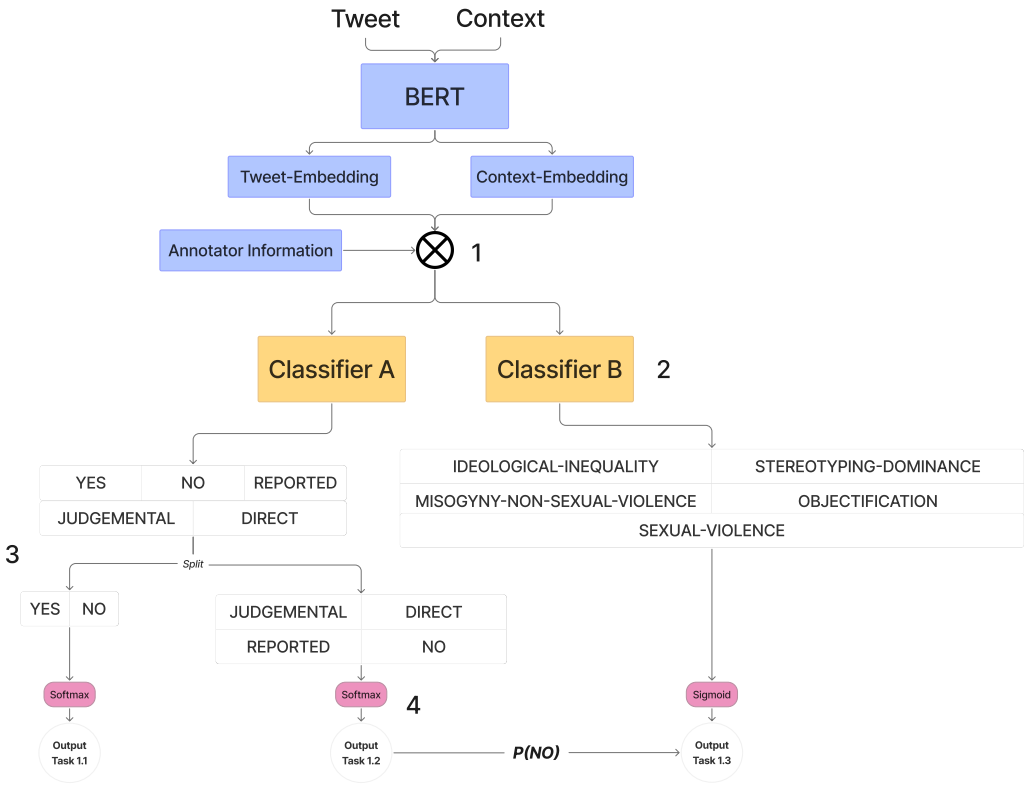

Figure 3: This figure illustrates our proposed model architecture, which consists of several key components. (1) The concatenation of tweet and context embeddings from the SDE architecture with the computed vector of annotator information (cf. 4.1.2). (2) Two classification heads applied to this concatenated representation. (3) The output of Classification Head A is further split into two components, corresponding to the challenge exception output. (4) Finally, the probability of "NO" from Classification Head A is passed to the labels of subtask 1.3 to complete it.

4.1.8. Result Formatting

To format the results, we rounded the probabilities to the nearest 1/6–as there are six annotators–and ensured that the sum of probabilities was 1 for subtasks 1.1 and 1.2. For hard classification, we adopted the following strategies: for subtasks 1.1 and 1.2, we selected the feature with the maximum probability; for subtask 1.3, a multi-label classification task, we chose all features with probabilities exceeding 0.25. This threshold was determined through testing on the training and development datasets during soft-to-hard label conversion.

In summary, our methodology involved a thorough literature review, extensive testing, and innovative use of AI agents to enhance contextual understanding. It also incorporates annotator information to address subjectivity and employs a multi-task headed approach, sharing base layers across tasks while capturing unique characteristics through specific output heads.

4.2. Memes

4.2.1. Data preprocessing

Regarding the Meme Dataset, we first wanted to verify the accuracy of the text and image pairs provided together. For each meme, we extracted the superimposed text using Florence-2 Large and we then compared it with the provided one. Average Jaccard similarity in terms of unigrams and bigrams showed respective values of 0.9518 and 0.9495, marking a minor difference that could be explained as follows:

- for unigrams, since diacritics matters, two semantically equal words could be treated as different

(e.g. "tenia" vs "tenía")
- for bigrams, being defined as sequences of two adjacent words, the sequence of words has an effect on the computed Jaccard similarity

Through this comparative analysis of the extracted and given texts, we observed that the superimposed texts provided with the data exhibit superior transcription quality compared to those extracted using Florence-2. Notably, these texts feature proper accentuation and sequentiality, resulting in a readability closer to human standards. An exemplary illustration of these findings is presented in Figure 4. Consequently, we opt to utilise the provided text instead of relying on a specific extraction technique.

Given text

"2+2=5 MITO Albert Einstein tenía bajo rendimiento en la escuela. VERDAD 2+2=4 CAN is El feminismo de hoy en día defiende la estupidez humana y no los derechos de las mujeres quemo ellas afirman "

Extracted text

"MITO VERDAD 2+2=5 2 +2=4 El feminismo de hoy en día Albert Einstein defiende la estupidez tenia bajo rendimiento humana y no los derechos de en la escuela. las mujeres quemo ellas afirman"

Figure 4: A comparison between the extracted text and the given text for a structurally complex meme. While Florence-2 focuses on a static top-to-bottom extraction, the text provided in the training set follows the human readability by considering also the left-to-right structure of the content, while providing a more precise transcription of accents.

4.2.2. Approach overview

To tackle meme classification, informed by our literature review which highlighted the necessity of exploring multiple approaches, we investigated three complementary strategies:

- **Caption-Based Classification**: representation of meme images as textual captions and classification of the captions using a fine-tuned text model.
- **Frozen Multimodal Classification**: usage of pretrained VLMs in zero-shot and few-shot settings without fine-tuning.
- **Fine-Tuned Multimodal Classification**: fine-tuning of medium-to-large VLMs on labeled sexist and not sexist memes for task-specific performance.

4.2.3. Caption-based Classification

In this text-based classification approach, represented in Figure 5, meme images were first transformed into textual descriptions using Qwen 2.5 VL 32B (1) and these captions (2), jointly with their respective ground truths (3), were then used as input for fine-tuning XLM-RoBERTa (4). This two-stage pipeline was designed to exploit the visual understanding of vision-language models and the adaptability of multilingual transformers.

To analyse the impact of visual description granularity, we generated two types of captions:

- **Short Captions**: concise descriptions capturing minimal visual content.
- **Detailed Captions**: rich, context-aware descriptions reflecting nuanced or subtle cues in the image.

Figure 6 shows how the textual representation can differ from the same meme. The prompts employed for the generation of captions are disclosed in Appendices L to N.

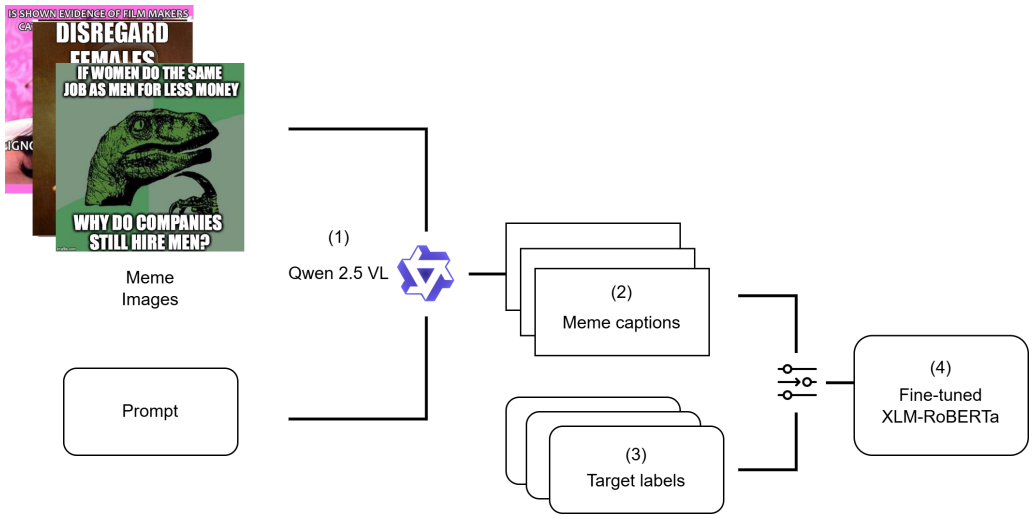

Figure 5: Pipeline for caption-based classification

Short caption
"The image is a meme featuring a close-up of a person with the text "A LAS MUJERES NO HAY QUE ENTENDERLAS HAY QUE AMARLAS" overlaid, which translates to "Women don't need to be understood; they need to be loved." The meme generator's watermark is also visible at the bottom right corner"

Detailed caption
"The image is a meme featuring a close-up of a person with their hand covering part of their face, appearing to cry or express distress. The text overlaid on the image reads: "A LAS MUJERES NO HAY QUE ENTENDERLAS HAY QUE AMARLAS," which translates from Spanish to English as "With women, you don't need to understand them; you just need to love them." The meme includes a watermark at the bottom right corner that says "memegenerator.es "

Figure 6: Comparison of two captions for the same meme, generated with Qwen 2.5 VL 32B on a different level of detail.

4.2.4. Frozen Multimodal Classification

This approach used frozen vision-language models in zero-shot and few-shot scenarios without task-specific fine-tuning, in order to simulate realistic low-resource classification settings. We evaluated the following VLMs:

- **Qwen-VL 2.5** in its 7B, 32B, and 72B variants (zero-shot and few-shot)
- **Aya Vision 8B** (zero-shot)
- **Mistral Small 3.1 24B** (zero-shot)

In the zero-shot setting, models were given only the meme image and a minimal classification prompt (showed in the Appendices G to K), with no prior examples. The employed prompts were significantly based on the guidelines provided to annotators for meme labelling across the three subtasks. We also tested variants where the model received either only the image, image plus superimposed text, or only the superimposed text. These variations aimed to quantify the importance of superimposed textual content over the final prediction. To evaluate few-shot performance, we included six example memes in the prompt using two different sampling strategies:

- **Random Few-Shot Sampling**: six random examples from the training set, imposing a balanced extraction between sexist and not sexist memes
- **Polarised Few-Shot Sampling**: three clearly sexist and three clearly non-sexist memes (i.e., with $\geq$5 of 6 annotators in agreement).

Images were resized to a maximum of 262,144 pixels (e.g., the size of a 512×512 image) maintaining their original proportions to fit within GPU memory constraints.

4.2.5. Fine-Tuned Multimodal Classification

Finally, we tested the effectiveness of fine-tuning a set of VLMs for both sexism identification and classification. Specifically:

- For subtask 2.1, we fine-tuned **Florence 2** and **Qwen 2.5 VL** (7B and 32B). The dataset used for fine-tuning was gathered from the available ground truths for the task, for a total 3,420 meme image-label pairs.
- For subtasks 2.2 and 2.3, only **Qwen 2.5 VL** 32B was fine-tuned. The data curation criteria was slightly different with respect to subtask 2.1, since we excluded the memes labelled as not sexist from the ground truths of the sexism identification task. Eventually, the number of considered records was 1,815 (over the 3,197 available ground truths) for the source intention classification and 2,868 (over 4,250 available ground truths) for sexism categorisation.

The experimental setup and the fine-tuning hyperparameters for both Florence-2 and Qwen 2.5 VL are presented in detail in Appendix O. In contrast to previously proposed methods for meme analysis, our proposed LLM-based solution offers a lightweight yet effective approach to detecting and classifying sexism in memes while incorporating the entire visual content into the classification pipeline. By avoiding high inference costs and proprietary APIs, this approach ensures compatibility with low-to-mid-tier hardware and promotes reproducibility by reducing computational requirements.

5. Results

5.1. System setting

All experiments were conducted using PyTorch 2.5.1, the Hugging Face Transformer 4.50 library, and the Smolagents 1.4 library for AI agent development. The computational environment consisted of two GPUs with the following specifications:

- NVIDIA A40 (46 GiB), driver version 555.42.06, CUDA 12.5
- NVIDIA A100 (40 GiB), driver version 555.42.02, CUDA 12.5

Additionally, VLMs with more than 7 billion parameters were loaded using Bitsandbytes 4-bit quantisation technique, which reduces the size of the model and computational costs by representing weights and activations with just 16 discrete levels [15]. This technique significantly reduces memory usage and accelerates inference while having minimal impact on model accuracy.

5.2. Development Phase

In this section, we present the performance metrics of our proposed methods across all the three tasks. For tweets, evaluations were conducted under both soft and hard contexts, whereas meme-based methods were assessed under the hard evaluation setting. Our model was trained on the provided training dataset and evaluated on the corresponding validation dataset for tweets. For memes, as mentioned before, we split the training dataset to create a validation dataset, thereby enabling us to assess the model's generalisation capabilities.

Regarding the sexism identification task in memes, the main results presented in Table 10 (full results in the Appendix F) indicate that models incorporating multimodal inputs generally have better performance on the task. Effectively, since the creation of memes and their virality across communities is based on a strong correlation between textual and visual elements, the analysis of the textual content alone could result in partial or impractical comprehension of the content.

Table 8
Tweets Hard label results for development dataset

Subtask 1.1		Subtask 1.2		Subtask 1.3	
ICM-Hard norm	**Macro F1**	**ICM-Hard norm**	**Macro F1**	**ICM-Hard norm**	**Macro F1**
0.81	0.88	0.59	0.53	0.56	0.60

Table 9
Tweets Soft label results for development dataset

	Subtask 1.1	**Subtask 1.2**	**Subtask 1.3**
ICM-Soft norm	0.68	0.47	0.44

Table 10
Results for Subtask 2.1 in hard evaluation for development dataset

Methodology	**Input**	**Model**	**ICM-Hard Norm**	**F1(YES)**
Caption-based	Detailed captions	XLM-RoBERTa large	0.5053	0.6591
Few-shot	Polarised meme images	Qwen 2.5 VL 32B	0.5121	0.7482
Zero-shot	Superimposed text		0.5352	0.7643
Zero-shot	Meme images		0.6189	0.8086
Fine-Tuned VLM	Meme images		**0.6671**	**0.8424**

However, it is worth mentioning that zero-shot classification of superimposed text using Qwen 2.5 VL 32B achieves results that are relatively close to the best ICM values obtained, while outperforming other methods that leverage meme images in their pipeline. This suggests that, for this specific type of memes, text plays a significant role in the final prediction. This may be explained by the fact that the creators of the EXIST Meme dataset gathered images by curating a lexicon of 250 terms that were used as search queries on Google Images. Additionally, textless images were removed manually, centring the dataset on textual elements [23]. In the domain of text-based methods, the performance of caption-based classification with XLM-RoBERTa was found to be inferior to that of superimposed text-based prediction. This suggests that captions may be lacking in descriptive information relevant crucial for a proper classification.

The fine-tuned Qwen 2.5 VL 32B model achieved the best results across all metrics, showing a +7.8% points improvement in the ICM-Hard Norm metric compared to the zero-shot classification performed using the off-the-shelf version of the same model.

To gain a clearer view of the results obtained by the best-performing method on subtask 2.1 Hard, we calculated the proportion of misclassified memes for which the annotators gave unanimous answers (e.g. all YES or all NO answers). Only 11.33% of misclassifications fall into this category, indicating that the model is highly confident in predictions for which there is a human agreement. Considering memes for which there is only one disagree answer, they account for 34.13% of misclassifications. More than half of the misclassifications (54.54%) come from memes for which two annotators disagree with the others., i.e. situations in which the evaluation of content is inherently more intricate from a human perspective.

Table 11
Results for Subtask 2.2 in hard evaluation for development dataset

Methodology	**Input**	**Model**	**ICM-Hard Norm**	**Macro F1**
Fine-tuned VLM	Meme images	Qwen 2.5 VL 32B	0.5324	0.5253

Given the superior performance of fine-tuned Qwen 2.5 VL 32B on subtask 2.1, we adopted this method for subtasks 2.2 and 2.3. This decision allowed us to focus on the scope of the study and avoid

redundant evaluations. Additionally, we reduced the number of experiments to minimise computational cost and environmental impact, striking a balance between empirical validation and responsible resource usage. Tables 11 and 12 show the results for source intention classification in memes.
A thorough exploration into the sub-values of the F1 score indicates that the model demonstrates a high capacity in identifying memes that overtly promote sexist ideologies. Indeed, the relatively high F1 score for the DIRECT class indicates that this category of content is more easily identifiable by the model. Performance drops sharply for the JUDGEMENTAL class, as the low F1 score of 0.1413 suggests that the model has difficulty to identify contents that criticise sexism. This may be due to the complex nature of such memes, which often rely on sarcasm, as shown in Figure 7. Additionally, this degradation in performance may be correlated with the under-representation of this class, accounting for just 14.38% of all ground truths.

Figure 7: Example of a sarcastic meme for the JUDGEMENTAL class. It uses a exaggerated self-description to promote women's autonomy while rejecting societal judgement.

Table 12
Subvalues of F1 score for all the labels in Subtask 2.2 for development dataset

No	Direct	Judgemental
0.7299	0.7048	0.1413

Similar considerations could be applied to the results obtained in the sexism categorisation task displayed in Tables 13 and 14. Moderate F1 scores are observed among sexist categories for IDEOLOGICAL INEQUALITY, STEREOTYPING DOMINANCE and OBJECTIFICATION (between 0.56 and 0.58), each of which also appears in over a quarter of the ground truth data. However, the model struggles to identify the categories MISOGYNY NON SEXUAL VIOLENCE and SEXUAL VIOLENCE, which represent 11.65% and 14.18% of the ground truths, respectively. The observation that the two lowest F1 sub-values are associated with these classes, jointly with the considerations made on the results of subtask 2.2, suggests that a low statistical representation constitutes a strong learning limit for this model. In this field of research, the relationship between the volume of data available and the precision of classification has been already examined for other types of models [48]. Providing a larger number of examples could therefore improve the ability of fine-tuned Qwen 2.5 VL 32B to recognise more generalised patterns associated with sexism categorisation and identify these instances more precisely.

5.3. Evaluation phase

The present section is dedicated to the presentation of results that have been obtained by our team in the EXIST 2025 challenge on the given test data.

Table 13
Results for subtask 2.3 in hard evaluation for development dataset

Methodology	Input	Model	ICM-Hard Norm	Macro F1
Fine-tuned VLM	Meme images	Qwen 2.5 VL 32B	0.3914	0.5152

Table 14
Sub-values of F1 score for all the labels in subtask 2.3 for development dataset

Ideological Inequality	Misogyny Non Sexual Violence	No	Stereotyping Dominance	Objectification	Sexual Violence
0.5759	0.3789	0.7206	0.5567	0.5600	0.2987

Tweet Classification

We trained our tweet classification model with three different seeds (0, 1, and 42), resulting in three submissions: GrootWatch_1, GrootWatch_2, and GrootWatch_3. The performance of these models on the tweet test set is shown in Tables 15 to 17. Notably, our model consistently ranked first in the Soft-Soft category across all languages for subtasks 1.1, 1.2, and 1.3. In the more challenging Hard-Hard category, we always placed within the top 20 out of over 130 submissions.

Table 15
Competition results for Subtask 1.1

Run	Soft Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Hard Rank	ICM-Hard	ICM-Hard Norm	F1 YES
GrootWatch_1	1	**1.0600**	**0.6700**	**0.8893**	16	0.5727	0.7878	0.7802
GrootWatch_2	2	1.0538	0.6690	0.9171	15	**0.5732**	**0.7881**	**0.7773**
GrootWatch_3	3	1.0368	0.6662	0.9088	20	0.5560	0.7795	0.7763

Table 16
Competition results for Subtask 1.2

Run	Soft Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Hard Rank	ICM-Hard	ICM-Hard Norm	Macro F1
GrootWatch_1	1	**-0.4385**	**0.4647**	**1.7711**	13	0.3016	0.5981	0.5325
GrootWatch_2	2	-0.5066	0.4592	1.8176	10	0.3266	0.6062	0.5421
GrootWatch_3	3	-0.5655	0.4544	1.8088	9	**0.3434**	**0.6117**	**0.5384**

Table 17
Competition results for Subtask 1.3

Run	Soft Rank	ICM-Soft	ICM-Soft Norm	Hard Rank	ICM-Hard	ICM-Hard Norm	Macro F1
GrootWatch_1	1	**-1.1034**	**0.4417**	11	0.3623	0.5841	0.6175
GrootWatch_2	3	-1.2566	0.4336	16	0.2829	0.5657	0.5986
GrootWatch_3	2	-1.1495	0.4393	9	**0.3809**	**0.5884**	**0.6171**

Meme Classification

Based on the results on memes for development dataset, we submitted our runs using the following methods:

- **GrootWatch_1**: Zero-shot classification of the superimposed text with Qwen 2.5 VL 32B
- **GrootWatch_2**: Zero-shot classification of the meme images with Qwen 2.5 VL 32B
- **GrootWatch_3**: Classification with fine-tuned Qwen 2.5 VL 32B

For subtasks 2.2 and 2.3, we used the fine-tuned Qwen 2.5 VL 32B model based on the YES predictions from the three distinct submissions on subtask 2.1. The results for meme classification in the hard evaluation setting are shown in Tables 18 to 20. Our methods demonstrated remarkable strength, with eight out of nine submissions achieving a top five ranking. The predictions obtained by fine-tuning Qwen 2.5 VL 32B consistently ranked second across all subtasks, achieving first place in subtask 2.1 on the Spanish instances.

Table 18
Competition results for subtask 2.1 in hard evaluation

Run	Language	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
GrootWatch_3	Spanish	1	0.3552	0.6810	0.7681
GrootWatch_3	All	2	**0.3589**	**0.6825**	**0.7740**
GrootWatch_2	All	4	0.1898	0.5965	0.7253
GrootWatch_1	All	8	0.0062	0.5032	0.6898

Table 19
Competition results for subtask 2.2 in hard evaluation

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
GrootWatch_3	2	**0.1868**	**0.5649**	**0.5513**
GrootWatch_2	4	-0.0588	0.4796	0.4917
GrootWatch_1	5	-0.3055	0.3938	0.4738

Table 20
Competition results for subtask 2.3 in hard evaluation

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
GrootWatch_3	2	**-0.0798**	**0.4834**	**0.5472**
GrootWatch_2	3	-0.3550	0.4263	0.5119
GrootWatch_1	5	-0.5812	0.3794	0.4921

6. Conclusion and Future Work

Our sexism detection approach achieved state-of-the-art performance in Soft-Soft classification for tweet analysis. The combination of contextual information search, annotator profile integration, soft label learning, and multi-task architecture proved particularly effective in this category. However, the Hard-Hard category remains a challenging task to overcome. Notably, our results revealed that simply using the soft probabilities to infer the hard label is not a sufficient strategy for tackling this challenge. One potential avenue for future research lies in optimising the inference time for context retrieval with AI agents. Currently, this process is relatively slow compared to traditional language models like LLM or BERT. To address this limitation, a possible solution could be the development of a shared dictionary or database of contexts that can be efficiently queried and retrieved. In cases where the desired context is not already present in the database, the system could be designed to search for it online and then store it in the database for future reference. This approach has the potential to significantly reduce inference times, enabling more efficient and scalable AI-powered language understanding.
Furthermore, despite the promise of incorporating context into language models, our experiments

suggest that fine-tuning LLM with context actually degrades performance. A possible explanation for this phenomenon is the concept of *context hijacking* [42], where the model overemphasises contextual cues and loses focus on the primary task. Further research is needed to verify this hypothesis and uncover the underlying causes of this performance drop, which will be crucial in unlocking the full potential of context-aware language models.

With respect to the best results obtained on meme classification in the past edition of EXIST, which very mostly based on textual elements, the results obtained by our team in the current edition confirmed that a full integration of meme images into the classification pipeline leads to better performance. Despite the top-tier results achieved, the proposed approaches present some limitations:

- **Multi-task learning**: Qwen 2.5 VL and Florence-2 were fine-tuned using the available ground truths for the three subtasks to minimise Cross-Entropy Loss. However, introducing a specific loss function that captures the interaction between subtasks could help the model to leverage the full potential of the given data and achieve better performance
- **Meme Dataset split**: The dataset was split 80/20 for training and testing. Despite the significant computational time required for repeated VLM fine-tuning, future work may consider cross-validation to obtain a more comprehensive assessment of model generalisation

Using optimal transport theory and the principle of maximum entropy, Erbani et al. [49] proposed the extended confusion matrix (TCM), which applies to single-label, multi-label, and soft-label classification tasks. TCM keeps the familiar structure of a standard confusion matrix: a square matrix sized by the number of classes, with diagonal entries representing correct predictions and off-diagonal entries showing confusions.

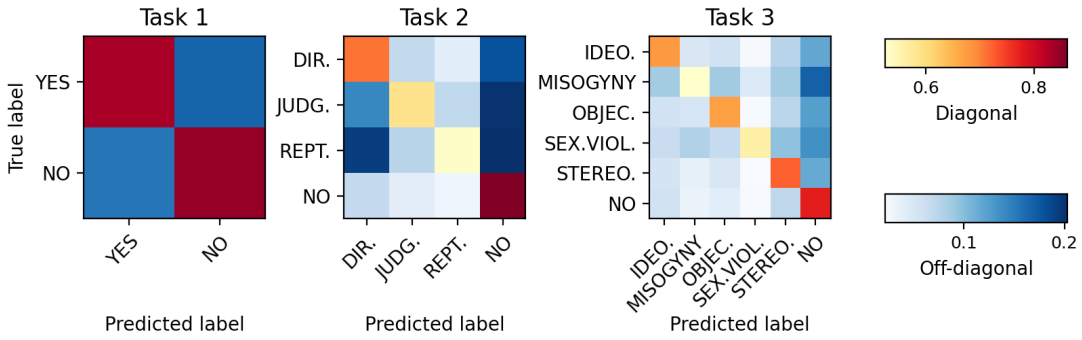

Figure 8: Row-normalised confusion matrix (TCM) from [49] with a weighting factor of 1 (i.e., all predicted label pairs contribute equally to TCM), showing model performance on soft subtasks 1.1, 1.2, and 1.3. Because diagonal values are much higher than off-diagonal ones, two colour scales—shared across matrices—are used to highlight variations within diagonal and off-diagonal entries.

Figure 8 reveals:

- **Subtask 1.1**: The confusion matrix shows a strong diagonal, indicating strong performance.
- **Subtask 1.2**: The diagonal entries are higher than off-diagonal ones, showing good model accuracy. Classes DIRECT and NO have the highest diagonal values but also strong column values, suggesting the model over-predicts these classes. This is especially true for NO, which shows the lightest row and the darkest column. JUDGEMENTAL and REPORTED have lower diagonal values and are often confused with DIRECT and NO, especially REPORTED.
- **Subtask 1.3**: Again, diagonal values are higher than others, confirming good model behaviour. The class NO has the lowest row and highest column values, indicating over-prediction that harms other classes. Notable confusions include MISOGYNY-NON-SEXUAL-VIOLENCE being misclassified as NO, and SEXUAL-VIOLENCE being confused with MISOGYNY-NON-SEXUAL-VIOLENCE, STEREOTYPING-DOMINANCE, or NO.

Future work could build on this analysis to reduce current misclassifications and enhance our method.

Acknowledgments

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organisations (see https://www.grid5000.fr).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and DeepL Write in order to: grammar and spelling corrections, rewriting of unnatural phrases, tone improving. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. I. Toluna, Baromètre Sexisme, Etude 4, Haut Conseil à l'Egalité entre les Femmes et les Hommes, 2024. URL: https://www.haut-conseil-egalite.gouv.fr/IMG/pdf/rapport_toluna_harris_-_baromc_tre_sexisme_vague_4_-_2024_dgcs-hce_-_avec_note_vf.pdf.

[2] Meta, More speech and fewer mistakes, 2025. URL: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

[3] Amnesty International, Les nouvelles politiques de Meta en matière de contenus risquent d'alimenter davantage de violences de masse et de génocides, 2025. URL: https://www.amnesty.org/fr/latest/news/2025/02/metas-new-content-policies-risk-fueling-more-mass-violence-and-genocide/.

[4] J. L. Gil Bermejo, C. Martos Sánchez, O. Vázquez Aguado, E. B. García-Navarro, Adolescents, ambivalent sexism and social networks, a conditioning factor in the healthcare of women, in: Healthcare, volume 9, MDPI, 2021, p. 721.

[5] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[6] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[9] A. Chhabra, D. K. Vishwakarma, A literature survey on multimodal and multilingual automatic hate speech identification, Multimedia Systems 29 (2023) 1203–1230.

[10] A. Vetagiri, P. Pakray, A. Das, A deep dive into automated sexism detection using fine-tuned deep learning and large language models, Engineering Applications of Artificial Intelligence 145 (2025) 110167.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[12] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[13] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, Nycu-nlp at exist 2024–leveraging transformers with diverse annotations for sexism identification in social networks, Working Notes of CLEF (2024).
[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.
[16] L. M. Quan, D. V. Thin, Sexism identification in social networks with generation-based language models, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271856112.
[17] F. Belbachir, T. Roustan, A. Soukane, Detecting online sexism: Integrating sentiment analysis with contextual language models, AI 5 (2024) 2852–2863.
[18] A. Debnath, S. Sumukh, N. Bhakt, K. Garg, Sexist Stereotype Classification on Instagram Data, 2020. URL: https://github.com/djinn-anthrope/Sexist_Stereotype_Classification.
[19] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, arXiv preprint arXiv:2303.04222 (2023).
[20] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.
[21] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.
[22] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.
[23] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, Exist 2024: sexism identification in social networks and memes, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 498–504. URL: https://doi.org/10.1007/978-3-031-56069-9_68. doi:10.1007/978-3-031-56069-9_68.
[24] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), 2023. URL: https://arxiv.org/abs/2304.14803. arXiv:2304.14803.
[25] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74/. doi:10.18653/v1/2022.semeval-1.74.
[26] A. Singh, D. Sharma, V. K. Singh, Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2024). URL: https://doi.org/10.1145/3656169. doi:10.1145/3656169, just Accepted.
[27] S. Deng, S. Belongie, P. E. Christensen, Large vision-language models for knowledge-grounded data annotation of memes, 2025. URL: https://arxiv.org/abs/2501.13851. arXiv:2501.13851.
[28] M.-H. Van, X. Wu, Detecting and correcting hate speech in multimodal memes with large visual language model, 2023. URL: https://arxiv.org/abs/2311.06737. arXiv:2311.06737.
[29] B. Zhao, A. Zhang, B. Watson, G. Kearney, I. Dale, A review of vision-language models and their performance on the hateful memes challenge, 2023. URL: https://arxiv.org/abs/2305.06159. arXiv:2305.06159.
[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin,

J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[31] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a unified representation for a variety of vision tasks, 2023. URL: https://arxiv.org/abs/2311.06242. arXiv:2311.06242.

[32] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[33] M. AI, Mistral Small 3.1, https://mistral.ai/news/mistral-small-3-1, 2025. [Online; accessed 27-May-2025].

[34] S. Dash, Y. Nan, J. Dang, A. Ahmadian, S. Singh, M. Smith, B. Venkitesh, V. Shmyhlo, V. Aryabumi, W. Beller-Morales, J. Pekmez, J. Ozuzu, P. Richemond, A. Locatelli, N. Frosst, P. Blunsom, A. Gomez, I. Zhang, M. Fadaee, M. Govindassamy, S. Roy, M. Gallé, B. Ermis, A. Üstün, S. Hooker, Aya vision: Advancing the frontier of multilingual multimodality, 2025. URL: https://arxiv.org/abs/2505.08751. arXiv:2505.08751.

[35] E. Valavi, J. Hestness, N. Ardalani, M. Iansiti, Time and the value of data, arXiv preprint arXiv:2203.09118 (2022).

[36] A. Menárguez-Box, D. Torres-Bertomeu, Ditana-pv at sexism identification in social networks (exist) tasks 4 and 6: The effect of translation in sexism identification, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271844312.

[37] V. Ruiz, J. Carrillo-de Albornoz, L. Plaza, Concatenated transformer models based on levels of agreements for sexism detection, Working Notes of CLEF (2024).

[38] J. Ma, R. Li, Rojing-cl at exist 2024: Leveraging large language models for multimodal sexism detection in memes, in: G. Faggioli, N. F. 0001, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1080–1090. URL: https://ceur-ws.org/Vol-3740/paper-100.pdf.

[39] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.

[40] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al., A survey on large language model based autonomous agents, Frontiers of Computer Science 18 (2024) 186345.

[41] Z. Dong, J. Ni, D. Bikel, E. Alfonseca, Y. Wang, C. Qu, I. Zitouni, Exploring dual encoder architectures for question answering, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9414–9419. URL: https://aclanthology.org/2022.emnlp-main.640/. doi:10.18653/v1/2022.emnlp-main.640.

[42] J. Jeong, Hijacking context in large multi-modal models, arXiv preprint arXiv:2312.07553 (2023).

[43] C. Sanchez, Z. Zhang, The effects of in-domain corpus size on pre-training bert, arXiv preprint arXiv:2212.07914 (2022).

[44] S. de Vries, D. Thierens, Learning with confidence: Training better classifiers from soft labels, arXiv preprint arXiv:2409.16071 (2024).

[45] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[46] A. C. Stickland, I. Murray, BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International

Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 5986–5995. URL: https://proceedings.mlr.press/v97/stickland19a.html.
[47] Y. Peng, Q. Chen, Z. Lu, An empirical study of multi-task learning on bert for biomedical text mining, arXiv preprint arXiv:2005.02799 (2020).
[48] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.
[49] J. Erbani, P.-É. Portier, E. Egyed-Zsigmond, D. Nurbakova, Confusion Matrices: A Unified Theory, IEEE Access 12 (2024) 1–1. URL: https://hal.science/hal-04820752. doi:10.1109/ACCESS.2024.3507199.

A. Hyperparameter Settings and Prompts

To achieve the results, we fine-tuned BERT with the following hyperparameters:

- Learning rate: 1e-5
- Batch size: 64
- Weight decay: 0.05
- Number of epochs: 5

For LLM fine-tuning using LoRA, we employed the following configuration:

- Batch size (training and evaluation): 32
- Gradient accumulation steps: 4
- Optimizer: `PagedAdamW_8bit`
- Learning rate: 5e-5
- Precision format: bf16
- Warm-up ratio: 0.1
- LoRA's matrix decomposition rank (r): 4
- Alpha LoRA: 16
- Targeted modules: `self_attn.q_proj`, `self_attn.k_proj`, `self_attn.v_proj`, `self_attn.o_proj`, `mlp.gate_proj`, `mlp.up_proj`, and `mlp.down_proj`

The zero-shot and fine-tune prompts used in our experiments are:

> **Task**: Classify tweets as YES (sexist) or NO (not sexist).
>
> **YES**: Explicit sexism, descriptions of sexist situations, or criticism of sexism (even implied).
>
> **NO**: Neutral content. Ignore non-sexist vulgarity. Use societal context.
>
> Answer: (Only YES or NO)
>
> **Tweet:** {tweet}

B. User Prompt Subtask 1.1 - AI agent direct classification

> **Task**: Determine whether a tweet is sexist. Categories: **YES**: The tweet is inherently sexist, describes a sexist situation, or criticises sexist behaviour. Examples:
>
> - "Women are too emotional to hold leadership positions."
> - "At the meeting, all my ideas were ignored until a male colleague repeated them."
> - "Catcalling is not a compliment; it's harassment."
>
> **NO**: The tweet does not contain sexist content, nor does it describe or criticise sexist situations or behaviours. Examples:

- “Looking forward to the weekend!”
- “Really looking forward to today’s ‘women in web3’ lunch meetup! If you’re in the la area and want to join, send me a dm!! See you ladies soon.”
- “Wow! Trouble making witches unite!”

Additional Guidelines:

- Ambiguous Language: If the tweet’s sexism is implied rather than explicit, classify it as ’YES.’ If context is insufficient, classify it as ’NO.’
- Strong or Vulgar Language: Classify based on content relevance to sexism, not on the presence of strong language alone.
- Contextual Understanding: Consider societal norms and the broader conversation when evaluating the tweet.

Your final answer will be YES or NO.

Tweet: {tweet}

C. User Prompt for AI agent context retrieval

Task: Retrieve concise external context to clarify ambiguous tweets or cultural references for sexism classification. Do NOT classify the tweet—only provide context that would help a downstream model to decide.

When to retrieve context:

- The tweet references events, lyrics, memes, or cultural artefacts unfamiliar to a general audience.
- The language is ambiguous (e.g., sarcasm, coded terms, or terms with dual meanings).
- The tweet hints at a broader societal debate or news story.

Guidelines:

1. No classification: Never output YES/NO. Your role is purely contextual.
2. Conciseness: Summarise external context in ≤ 100 tokens.
3. Relevance: Only include context directly tied to potential sexism (e.g., explain a referenced event’s sexist controversy, not general info).
4. No context? Output ”No external context needed.”

Output Format: [Summary of context, or ”No external context needed.”]

Examples:

1. **Tweet**: ”Ugh, not another ‘Boss Babe’ anthem. . . ”
   Output: ”The term ’Boss Babe’ is associated with MLM schemes targeting women, often criticised for exploiting feminist rhetoric. Some view it as empowering, others as patronising.”
2. **Tweet**: ”This is why we need more #NotAllMen energy.”
   Output: ”#NotAllMen is a hashtag used to critique men who derail conversations about sexism by insisting ’not all men’ are problematic. Often cited in debates about systemic misogyny.”
3. **Tweet**: ”Finally got tickets to the concert!”
   Output: ”No external context needed.

D. Context Analysis

We conducted a preliminary assessment of the generated contexts to explore at their quality, relevance and accuracy. Our aim was to explore how well the generated contexts align with the original tweets.
Methodology
We randomly selected 30 context samples from each dataset (train, dev, test) and evaluated them based on three criteria:

- **Relevance**: How well did the generated context align with the original tweet? (Score: 1-5)
- **Accuracy**: Did the generated context provide correct information or insights? (Score: 1-5)
- **Quality**: Was the generated context coherent, well-structured, and easy to understand? (Score: 1-5)
- In case of 'No external context needed.': Was it appropriate not to generate external context for the given tweet? (Score: 1-5)

Results
The small-scale study reveals that the generated contexts consistently achieve perfect scores in terms of relevance (100%) and quality. The accuracy, however, is satisfactory but not outstanding, with an average score of 3.7/5. Notably, our model demonstrates 100% capability in identifying when no additional context is required. Neither did we observe any hallucinations in the generated texts.
To delve deeper into context accuracy, we stratified the results according to the agreement rate of the six annotators on the binary sexist classification of the tweet (only applicable for training and development datasets, as test dataset results are not available).

Table 21
Context Accuracy Analysis according to Annotator Agreement Rate

Annotator Agreement Rate	Context Accuracy
100%	3
83%	4.4
66%	4.3
50%	4.5

As shown in Table 21, we observe that accuracy is less satisfactory when there is a high annotator agreement rate for subtask 1.1. However, with lower agreement rates, accuracy tends to improve. While this limited analysis provides an encouraging initial look at the generated contexts, we acknowledge that more samples and evaluators are necessary to draw more robust conclusions.

E. User Prompt Subtask 1.1 - LLM classification with context

Task: Classify tweets as YES (sexist) or NO (not sexist).

YES: Explicit sexism, descriptions of sexist situations, or criticism of sexism (even implied).

NO: Neutral content. Ignore non-sexist vulgarity. Use societal context.

Answer: (Only YES or NO)

Tweet: {tweet}

Context: {context}

F. Full Results Subtask 2.1 and Further Discussion

The full set of experiments conducted for subtask 2.1 Hard is showed in Table 22. With regard to multimodal models, it is interesting to note that an increase in model size does not necessarily lead

to improved performance. For instance, zero-shot classification using models from the Qwen 2.5 VL family with the same prompt shows that Qwen 2.5 VL 32B outperforms the 7B release, while the 72B version performs worse than the medium-sized model. Interestingly, despite differences in architecture and parameter count, the results obtained from Aya Vision 8B, Mistral Small 3.1 24B, and the fine-tuned Florence 2 model exhibit notable similarity. This convergence in performance suggests that these models have achieved a similar level of representational capability for sexism detection in memes. Fine-tuned Qwen 2.5 VL 7B has shown that an increase of +8.7% in the ICM-Hard Norm metric when compared to zero-shot classification with the same model. This percentage increase is comparable to the value computed for the 32B release.

Table 22
Results for subtask 2.1 in hard evaluation

Methodology	Input(s)	Model	ICM-Hard Norm	F1(YES)
Caption-based classification	Short captions	XLM-RoBERTa base	0.4647	0.7163
	Short captions	XLM-RoBERTa large	0.3575	0.7377
	Detailed captions	XLM-RoBERTa base	0.5081	0.7497
	Detailed captions	XLM-RoBERTa large	0.5121	0.7482
Zero-shot classification	Meme images	Qwen 2.5 VL 7B	0.5551	0.6786
	Meme images	Qwen 2.5 VL 32B	0.6189	0.8088
	Meme images	Qwen 2.5 VL 72B	0.6004	0.7576
	Meme images	Aya Vision 8B	0.4354	**0.8924**
	Meme images	Mistral Small 3.1 24B	0.4458	0.6916
	Meme images, superimposed text	Qwen 2.5 VL 32B	0.5774	0.8212
	Superimposed text	Qwen 2.5 VL 32B	0.5352	0.7643
Random few-shot	Meme images	Qwen 2.5 VL 7B	0.4828	0.8228
Polarised few-shot	Meme images	Qwen 2.5 VL 32B	0.5053	0.8591
Fine-tuned VLM	Meme images	Florence-2 Large	0.5015	0.7580
	Meme images	Qwen 2.5 VL 7B	0.6038	0.7693
	Meme images	Qwen 2.5 VL 32B	**0.6817**	0.8412

G. System Prompt Subtasks 2.1, 2.2 and 2.3

"You are an Artificial Intelligence for sexism detection and classification in social media contents."

H. User Prompt Subtask 2.1 - Multimodal classification

"Is this meme sexist towards women? Answer just YES if you judge the meme as sexist towards women. Answer just NO if you judge the meme as not sexist towards women."

I. User Prompt Subtask 2.1 - Text-based classification

"Given the superimposed text of a meme '[superimposed_text]', is this meme sexist towards women? Answer just YES if you judge the meme as sexist towards women. Answer just NO if you judge the meme as not sexist towards women."

J. User Prompt Subtask 2.2

"Answer just DIRECT if the meme promotes or encourages sexism. Answer just JUDGEMENTAL if the meme criticises or condemns sexist behaviour."

K. User Prompt Subtask 2.3

"Classify the given meme into one or more of these categories (multi-label allowed):

- IDEOLOGICAL-INEQUALITY if it rejects feminism or denies gender inequality.
- STEREOTYPING-DOMINANCE if it promotes traditional gender roles or male superiority.
- OBJECTIFICATION if it reduces women to appearance or sexualises them.
- SEXUAL-VIOLENCE if it contains sexual harassment or assault references.
- MISOGYNY-NON-SEXUAL-VIOLENCE if it expresses hatred or non-sexual violence toward women.

The answer is just and strictly a list of strings, as the following example:

["CATEGORY_1", "CATEGORY_4"] "

L. System Prompt for Meme Caption Generation

"You are an Artificial Intelligence for meme captioning."

M. User Prompt for Meme Caption Generation - Simple captions

"Generate a caption in plain text of this meme without expressing a judgement on it. Answer in 80 words maximum."

N. User Prompt for Meme Caption Generation - Detailed captions

"Generate a detailed caption in plain text of this meme without expressing a judgement on it."

O. Fine-tuning Setup for Florence-2 and Qwen 2.5 VL

On Florence-2, the experiments were conducted by freezing the DaViT vision encoder and using a batch size of 5. The training was conducted over 3 epochs using the AdamW optimiser with a linear learning rate scheduler and no warm up steps. The model was optimised to minimise the cross-entropy loss between predicted and target YES/NO labels, performing the validation after each epoch.
For Qwen 2.5 VL 7B and 32B, due to a larger size of the models the fine-tuning strategy was different, in order to keep reasonable training times. We applied Low-Rank Adaptation (LoRA) to the query and value projection layers using a rank of 8, a scaling factor of 16, and a dropout rate of 0.05. Only the low-rank adaptor weights were updated during training, resulting in a significant reduction in the number of trainable parameters. For the 7B model, the trainable parameters were 2,523,136 (0.0304% of the total), while for the 32B model the number of trainable parameters was 8,388,608 (0.0251 % of the total). The models were fine-tuned on 3 epochs by scaling the image resolution up to 262,144 pixels, using a batch size of 5. As for Florence-2, the loss function to minimise was Cross Entropy Loss.