UNED-Annotate-Team at EXIST 2026: A Multimodal Multi-Task Approach to Sexism Identification in Memes*

Notebook for the EXIST Lab at CLEF 2026

Enrique García-Arias[1,*,†], Laura Plaza[1,†] and Jorge Carrillo-de-Albornoz[1,†]

[1]Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain

Abstract

The harmful effects of sexism on social media change over time and can affect different groups of people in different ways. This is especially true for memes, where meaning is often conveyed through a combination of text and images, making sexist content harder to identify. As a result, people may interpret the same meme differently and disagree on whether it is harmful or sexist, a challenge that the Learning with Disagreement (LeWiDi) paradigm addresses precisely by preserving the variability in those annotations rather than forcing a single consensus label.

In this paper, we present the work of the UNED-Annotate-Team for Task 2.1 of the EXIST (sEXism Identification in Social neTworks) lab at CLEF 2026, which focuses on spotting sexism in Spanish and English memes. We propose a multimodal approach that integrates visual features with annotator demographic information (gender and age), exploring two distinct strategies: auxiliary multitask learning and contextual soft-token embeddings. We further enrich the textual representation with CLIP-derived visual semantic features and adopt a disagreement-aware annotation aggregation strategy aligned with the LeWiDi paradigm.

The results reveal the importance of prioritizing the proper generation of the visual semantic representation. We leveraged an efficient CLIP-based preprocessing strategy combined with early fusion, which significantly improved our detection of the minority (sexist) class. In the Soft assessment environment, we compared multitask learning with a novel soft token embedding strategy, demonstrating that integrating sociodemographic information as contextual tokens—rather than discrete auxiliary tasks—produces better alignment with the annotator's subjectivity, achieving an F_1 score of 72.36%.

Keywords

Sexism Detection, Memes, Multimodal Learning, Demographic Profiling, Multitask, LeWiDi

1. Introduction

The problem of sexism is amplified by its widespread presence on social media, whose manifestations range from explicit harassment to subtle trivialization, contributing to the normalization of discriminatory discourse and having profound repercussions for vulnerable populations[1, 2].

In this context, memes have emerged as potent vehicles for sexist meaning, their capacity for virality, combined with the interplay between text and imagery and the use of irony as a masking mechanism, complicates the identification of misogynistic content [3]. These characteristics highlight the difficulty of detecting sexism in memes, which must be addressed as a multimodal process with implicit meanings shaped by cultural factors.

Addressing this complexity requires considering many elements that are close to subjectivity; it is not a uniform phenomenon, and the interpretation of such content additionally varies depending on sociodemographic characteristics and individuals' experience. As a result, annotator disagreement in labeling tasks should not be treated as mere noise, but as a valuable signal of the diverse social positions regarding sexist discourse [4, 5]. The integration of annotators' sociodemographic information provides the essential context to influence content interpretation and categorization [6].

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

e.garcia@lsi.uned.es (E. García-Arias); lplaza@lsi.uned.es (L. Plaza); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz)

https://uned.es/ (E. García-Arias); https://lplaza@lsi.uned.es/ (L. Plaza); http://uned.es/ (J. Carrillo-de-Albornoz)

0009-0006-6274-3880 (E. García-Arias)

Across its different editions, the sEXism Identification in Social neTworks (EXIST) shared task has consolidated the study of sexism in social media under the Learning with Disagreement paradigm. The 2024 and 2025 editions expanded the identification and characterization of sexism in tweets, memes, and short videos, incorporating multiple annotations and annotators' sociodemographic profiles [7, 8]. This work addresses EXIST 2026 subtask 2.1 by modeling sexism not as a conventional binary classification but as a nuanced, multimodal, and subjectively grounded phenomenon.

We explore two primary hypotheses regarding the role of annotator sociodemographic information. First, we hypothesize that these attributes are important for capturing the the inherently subjective nature of sexist content. Different people may interpret the same message in different ways, and these differences are often linked to their personal experiences and perspectives. Second, we hypothesize that multitask learning (MTL) can improve model performance by training the system not only to detect sexism, but also to predict annotator characteristics. By evaluating these hypotheses, we aim to uncover how annotator background modulates the model's internal representation of discriminatory discourse.

The rest of this paper is organized as follows: Section 2 reviews related work; Section 3 details the methodology; Section 4 presents experiments and results; Section 5 discusses the results; Section 6 outlines the limitations and threats to validity; and Section 7 concludes the paper and presents directions for future work.

2. Related Work

The EXIST Lab and Learning with Disagreement

The EXIST (sEXism Identification in Social neTworks) lab has evolved substantially since its first edition in 2021. In the early editions the focus has been on detecting sexism in tweets [9, 10, 11], and in the more recent editions the scope has been expanded to multimodal and video-based content, including memes in 2024 [7] and TikTok videos in 2025 [8].

The adoption of the Learning with Disagreement (LeWiDi) paradigm [12] has been a core component in all editions, allowing for the explicit modeling of disagreement among annotators rather than treating it as noise in the annotations. This perspective moves away from the assumption of a single ground-truth label and instead acknowledges subjectivity in sexism perception. As a result, datasets provide soft labels reflecting annotator distributions, and evaluation is conducted using both hard metrics (majority vote) and soft metrics such as ICM-Soft [17], which measure alignment with the full annotation distribution.

The 2025 edition attracted substantial participation, with 244 registered teams from 38 countries and 114 teams submitting results, yielding 873 runs in total [8]. Despite the maturity of the evaluation framework, most systems have primarily focused on optimizing predictive performance under these metrics, while little attention has been given to analysing the structure or sources of annotator disagreement itself.

Multimodal Approaches for Meme Classification

Approaches to meme classification in EXIST have evolved rapidly, particularly with the increasing availability of vision-language models. A common strategy has been to rely on multimodal fusion architectures combining visual encoders with text-based transformers. For instance, ArcosGPT [13] integrates BLIP-generated captions, OCR-extracted text, and GPT-4o descriptions within a fusion model that combines vision transformers with RoBERTa. Similarly, NaturalThinkers [14] combines BLIP, BERT, and vision transformers with attention-based fusion, followed by an MLP classifier.

Other systems move away from dense fusion architectures. TrankilTwice [15] explores a combination of LLM-based prompting and graph-based modeling at the meme level, aiming to capture structural relationships between samples. In contrast, CLTL [16] adopts a simpler ensemble strategy, combining Swin Transformer V2 with RoBERTa/BERT models and supplementing them with SVM baselines based on stylometric and emotion-related features. I2C-UHU-Altair [17] leverages vision-language models

such as BLIP2 and Qwen within the LeWiDi framework, explicitly addressing annotation ambiguity in subjective labeling settings.

While these approaches achieve strong performance, they largely rely on either heavy multimodal fusion or ensemble strategies. In contrast, we introduce a concept-based CLIP scoring mechanism that computes a continuous sexism signal by comparing meme embeddings against curated sets of sexist and non-sexist concepts in English and Spanish. This provides a lightweight alternative that avoids additional LLM inference while still capturing semantic alignment with sexist content.

CLIP for Multimodal Feature Extraction

CLIP (Contrastive Language-Image Pre-training) [18] has become a standard model for aligning visual and textual representations in a shared embedding space. Its ability to perform zero-shot classification and general-purpose feature extraction has made it widely applicable across multimodal tasks.

Previous work has successfully applied CLIP-based representations to hate speech and sexism detection in memes [19, 20], typically leveraging its embeddings as inputs to downstream classifiers. Building on this line of work, we derive a concept-based sexism score by comparing meme embeddings against curated sets of sexist and non-sexist concepts in both English and Spanish.

The score combines visual and textual signals through an adaptive weighting scheme, where the contribution of the textual modality depends on the availability of OCR-extracted text. This produces a continuous feature that complements the raw CLIP embeddings used elsewhere in the model, rather than replacing them.

Multi-Task Learning for Sexism Detection

Multi-task learning has been widely studied as a mechanism to improve generalisation by encouraging shared representations across related tasks [19]. In sexism detection, auxiliary tasks based on annotator demographics such as gender and age have been proposed as a form of regularisation, helping reduce overfitting to dataset-specific biases [21].

In our setting, we investigate the use of annotator demographic prediction as auxiliary tasks within a multi-task framework. Unlike prior approaches that incorporate demographic information directly as input features, we predict these attributes from meme content itself. This design encourages the model to learn representations that are informative both for sexism classification and for capturing correlates of annotator characteristics.

Several recent studies have explored the use of demographic signals in multimodal and textual models for sexism detection, although their primary focus is typically on improving predictive performance rather than analysing disagreement phenomena.

One line of work integrates demographic metadata directly into the input representation. Mumul03 [22] concatenates annotator attributes such as gender, ethnicity, and age with the original text using separator tokens, and trains annotator-specific models whose predictions are later aggregated. While this approach improves soft-label accuracy, it remains centred on reproducing annotator distributions. The authors also note reduced stability in cases with high disagreement, suggesting that modelling disagreement structure itself remains challenging.

Other approaches align more closely with the LeWiDi framework. Dandys-de-BERTganim [23] introduce a multi-task architecture combined with a soft-label reader that preserves full annotator distributions, yielding well-calibrated predictions under disagreement. However, their analysis remains focused on predictive performance, without further examining how disagreement varies across types of sexist content or semantic categories.

NYCU-NLP [24] extends demographic modelling by embedding annotator attributes and aggregating them through attention mechanisms, alongside bilingual fusion to improve robustness across languages. Despite the architectural complexity, their evaluation is restricted to standard classification metrics, leaving open the question of how demographic variation interacts with specific disagreement patterns.

Across these approaches, demographic information is primarily used as an additional signal for improving prediction quality. Less attention has been given to using it as a tool for understanding disagreement structure, which is the direction explored in this work.

We further distinguish our approach by treating demographic prediction as an auxiliary task rather than an input feature. This induces a form of shared representation learning that must jointly support sexism detection and demographic prediction. To assess the impact of this design choice, we compare single-task and multi-task configurations, observing consistent gains in recall for the sexist class under the multi-task setting (64.03% $\rightarrow$ 71.54%). This suggests that auxiliary demographic supervision provides a useful inductive bias in subjective classification settings.

3. Methodology

Next, we describe the main components of our system. The process consists of three components: (i) corpus preprocessing, including text cleaning and visual feature extraction using CLIP; (ii) a unified multimodal framework based on the Transformer Mean Pooling with Multi-Head Attention (Transp-Mean MHA) architecture that integrates auxiliary data features with the primary textual representation through a two-stage process: first, the transformer encoder output is compressed via a mean-max pooling strategy to capture semantic nuances; second, auxiliary scalar and categorical features are processed by a dedicated encoder using a Multi-Head Attention (MHA) mechanism. The number of attention heads in the encoder is dynamically set to capture interactions between functions and dependencies in data before fusion and (iii) multi-task learning that incorporates auxiliary heads corresponding to the gender and age of the annotators. Figure 1 provides an overview of the complete pipeline.

3.1. Preprocessing Phase

In this phase, we process the EXIST 2026 JSON corpus to generate two datasets with the structure required for training. Figure 2 illustrates the three components of this phase.

Data preparation. The texts extracted by OCR from the original images and provided as part of the corpus metadata in both English and Spanish are cleaned. This cleaning process consists of URL removal, hashtag and user mention normalization, removal of residual special characters, and conversion to lowercase.

Multimodal Feature Extraction with CLIP. To process memes, we complement the textual signal with image representations derived from the Contrastive Language–Image Pretraining (CLIP) model [18]. Specifically, we use the `openai/clip-vit-base-patch32` checkpoint [25]. Below we detail the process performed in the feature extraction:

- *Image embeddings:* Each meme image is loaded and converted to RGB format, then processed through CLIP's vision encoder to obtain a 512-dimensional ℓ_2-normalised feature vector. Images that cannot be located on disk are assigned zero vectors and flagged as missing.
- *Dimensionality reduction.* The full 512-dimensional CLIP vectors were projected to a 64-dimensional subspace via Principal Component Analysis (PCA) fitted on the training split. From this decomposition, we selected the top 10 components capturing the highest variance, discarding the remaining components to reduce noise. The selected components $\{\texttt{img\_pca}_i\}_{i=0}^{63}$ are stored alongside each record.
- *Visual cluster assignment.* To capture discrete visual patterns, we apply K-Means clustering ($K = 8$) on the original 512-dimensional space. Each record is annotated with a cluster identifier $\texttt{img\_cluster} \in \{0, \ldots, 7\}$.
- *Multimodal Sexism Score.* We have implemented a MemeSexismScorer that, using the visual representation and the OCR text, generates a continuous score $\texttt{img\_score} \in [0, 1]$ to indicate the estimated probability that a meme conveys sexist content.

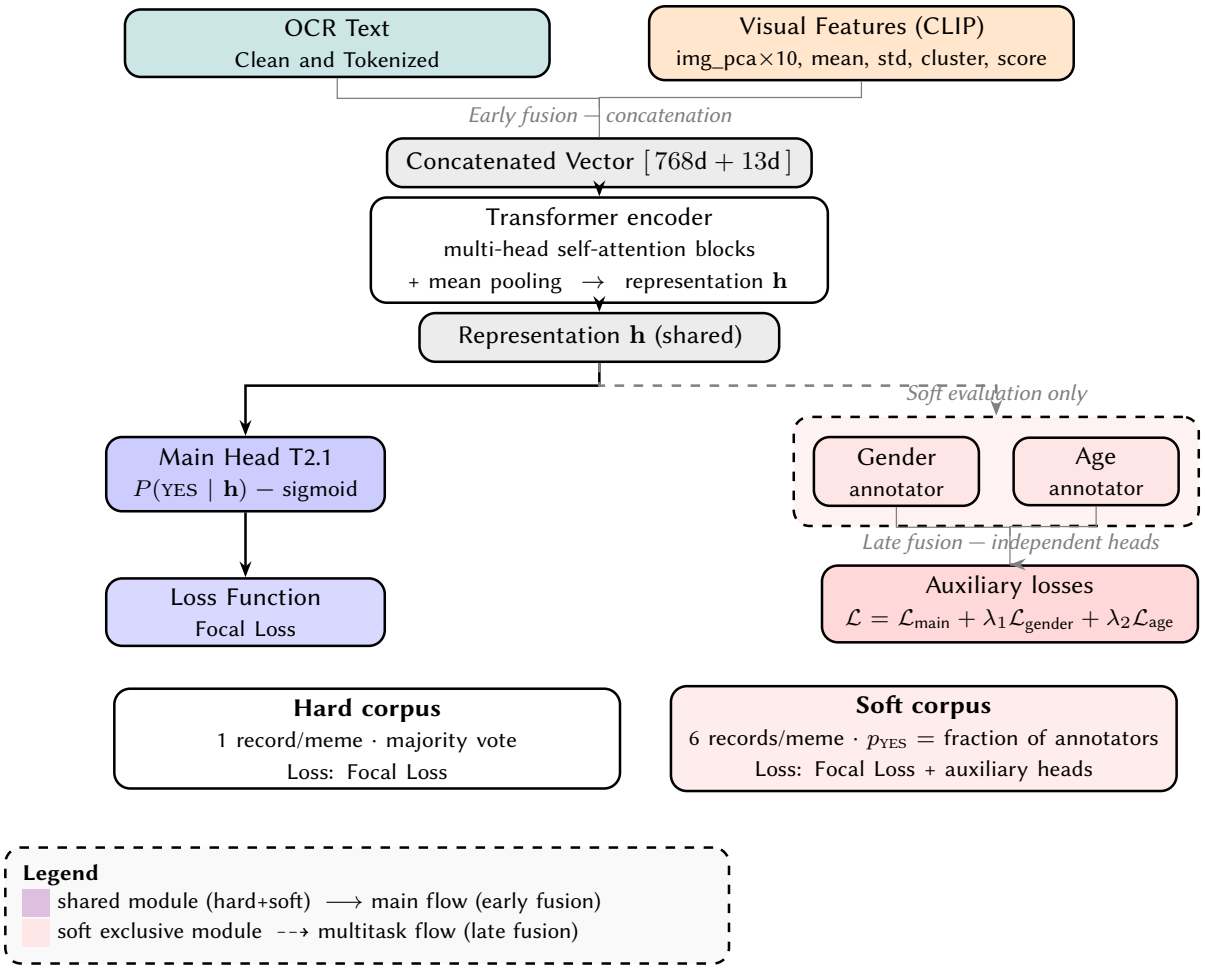

Figure 1: Model architecture for detecting sexism in memes in EXIST 2026. The *early fusion* concatenates the textual embeddings (768 d) and the numerical features derived from CLIP (13 d) before the encoder transformer with *mean pooling*. The *late fusion* adds, exclusively in the *soft* evaluation regime, independent auxiliary heads to predict the annotator's gender and age, whose losses are combined weighted with the main loss.

For each meme, a fused representation is calculated as a weighted combination (with $\alpha = 0.6$ for the text modality) of the normalized OCR text representation and the image representation. The sexism score is then obtained from the maximum cosine similarity between this fused representation and two predefined sets of concepts: sexist concepts and non-sexist concepts.

The aggregate image statistics (`img_mean`, `img_std`) are also appended as scalar features to each record.

3.2. Multihead attention framework

Our system is built around a unified architecture called `TransformerMeanPoolingBase`. It handles three types of input: meme text, CLIP-based visual features, and annotator metadata. The architecture supports two operating modes: single-task (used for hard evaluation) and multi-task (used for soft evaluation).

The core encoder is a multilingual transformer (XLM-RoBERTa-base, DistilBERT-base-mc, or DeBERTa-v3-base). After the transformer, we apply a pooling layer (mean, CLS, or mean+max) to obtain a text vector of dimension $d_{\text{text}} = 768$. The 13 numeric features from CLIP (10 PCA components, mean, std, and the sexism score) go through an `ExtraFeatureEncoder`. This encoder includes a multi-head attention layer that lets the model learn interactions among the visual features. The number of attention heads is set to $\max(1, \lfloor d_{\text{proj}}/64 \rfloor)$, where $d_{\text{proj}} = 384$.

The projected visual representation is combined with the pooled textual embedding to form a shared

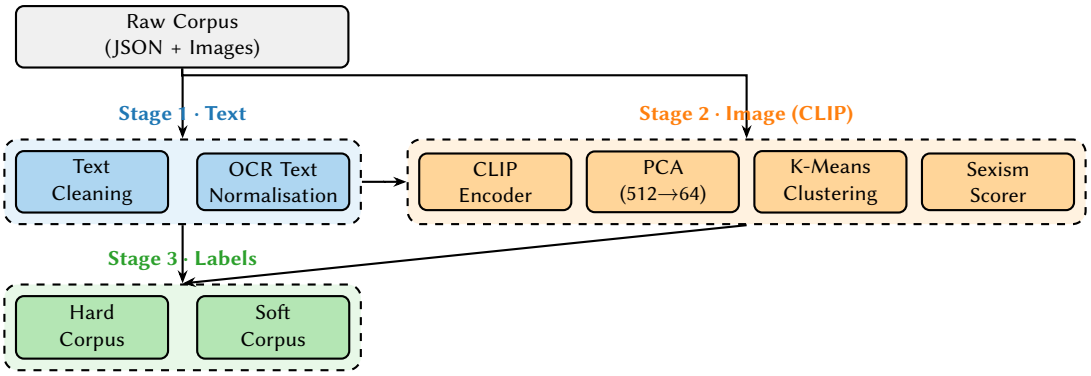

Figure 2: Overview of the three-stage preprocessing pipeline. *Stage 1* handles text cleaning and OCR normalisation; *Stage 2* extracts multimodal image features via CLIP; *Stage 3* aggregates annotations into the hard and soft evaluation corpora.

multimodal representation of dimension $d_{\text{fusion}} = d_{\text{text}} + d_{\text{proj}}$. The subsequent bidirectional LSTM layers capture higher-order dependencies between semantic textual information and CLIP-derived visual features prior to classification.

We use Focal Loss [26] for the main binary classification task (sexist or not) with $\alpha = 0.45$ and $\gamma = 2.0$ to reduce the impact of class imbalance.

We use AdamW with an initial learning rate of either 1.5×10^{-5} or 2×10^{-5} depending on the run and weight decay between 0.03 and 0.05. The learning rate follows a cosine annealing schedule with a warmup period of 3 epochs ($\texttt{pct\_start} = 0.2$). Early stopping monitors either macro F1 or balanced accuracy with a patience of 8 epochs. The parameter `freezeLayers` (values from 0 to 4) controls how many transformer layers are kept frozen. Training uses mixed precision (AMP) and gradient accumulation (4 steps), giving an effective batch size of 32.

We apply two text augmentation techniques: synonym replacement using WordNet, and random word swapping within the sentence. Depending on the configuration, we generate either one or two augmented versions per sample. Augmentation is applied only to the target classes (YES, or both YES and NO, depending on the run). The system's versatility in regularization and optimization options allowed us to perform systematic comparisons between different configurations.

In the inference process, hard predictions are obtained by applying a threshold τ (typically between 0.35 and 0.45) to the output probability of the main head. For soft evaluation, we directly use those same probabilities as $P(\text{YES} \mid \text{meme})$.

3.3. Multitask

The sociodemographic characteristics of the annotators incorporated into the corpus determine the subjective nature and basis of disagreement in the perception of sexism. To efficiently capture this valuable information, we have designed a multitasking learning architecture that we apply exclusively in the soft evaluation regime. The fundamental criterion is that we do not treat this sociodemographic data, specifically gender and age, as simple input characteristics, but rather they act as auxiliary monitoring tasks.

Architecture and geometric regularization: Our model employs a shared encoder (as described in Section 3.2) that projects the multimodal input into a high-dimensional latent space. From this shared representation, the architecture branches into three independent classification heads:

- Main head: Responsible for the identification of sexism (YES/NO) using a focal loss function.
- Auxiliary head 1 (gender): Predicts the annotator's gender (binary classification) using cross-entropy.
- Auxiliary head 2 (age): Predicts the scorer's age range (categorical: 18-22, 23-45, 46+) using cross-entropy with label smoothing ($\varepsilon = 0.05$) to improve generalization.

The core of this approach lies in the backward step and during training, the total loss is defined as a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sex}}\mathcal{L}_{\text{sex}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{age}}\mathcal{L}_{\text{age}}$$

with weights $\lambda_{\text{sex}} = 1.2$, $\lambda_{\text{gen}} = 0.4$ and $\lambda_{\text{age}} = 0.4$.

These heads of the multitasking process operate independently during inference, but what is very important is that the communication between tasks occurs dynamically during backpropagation so that the shared encoder reorganizes the latent space motivated by the regularization mechanism generated by the gradients derived from the tasks in the multitask. By processing the sexism prediction concurrently with the demographic prediction tasks, we forced the encoder to capture subtle signals that correlate with the annotators' profiles; these signals would likely be dismissed as noise in a single-task model.

Multitasking design is not limited to adding auxiliary monitoring; it also structures discrepancies between annotators in the shared encoder. Gender and age are not arbitrary labels, but rather sociodemographic characteristics on which perceptions of sexism differ within the corpus. This is accounted for in the architecture through the gradients of λ_{gen} and λ_{age}, which reflect this disagreement structure. While single-task models would dismiss this information as label noise, our model encodes it, capturing and processing a real annotative phenomenon: that the perception of a meme as sexist is clearly influenced by who annotates it.

In the implementation, and to ensure stability across different loss landscapes, we use a UnifiedLossFactory. The sexism identification task is handled by a stabilized Focal Loss ($\gamma = 2.0, \alpha = 0.45$), which addresses the class imbalance characteristic of sexism datasets by focusing the training process on "hard-to-classify" instances. Auxiliary tasks employ standard cross-entropy formulations, and all losses are masked for ignored samples (index -100) to ensure consistency in batch processing.

4. Experimental evaluation

4.1. Setup

The experiments were conducted on the EXIST 2026 meme dataset, which includes 3984 memes for training and 1053 for testing, evenly distributed between Spanish and English. All training processes were executed on a workstation equipped with 30 GB of RAM and an NVIDIA RTX 4060Ti GPU. The split process was configured with 70% for training and 15% for validation and (internal) testing.

For hard labels, we followed the official majority voting threshold: a meme is labeled **YES** if at least 4 out of 6 annotators agree; otherwise, it is labeled **NO**. Consequently, the hard corpus contains one registry per meme. In contrast, the soft corpus maintains the individual annotator perspective, we used the raw proportion of YES votes as the target probability distribution.

All experiments share the same base architecture (*(TranspMean MHA)*) with a multilingual Transformer encoder, bidirectional LSTM, multi-head attention, and early fusion of 13 CLIP-based visual features (`img_pca_0`…`img_pca_9`, `img_mean`, `img_std`, `img_score`). Hyperparameters were configured using a JSON configuration file. Training was performed using AdamW (initial learning rate of 1.5×10^{-5} or 2×10^{-5}), cosine warm-up annealing (3 epochs), gradient accumulation (steps 2–4), and mixed precision. Early stopping was applied with a patience ranging from 4 to 8 epochs.

We present three runs for the soft (probabilistic) configuration and three for the strict (binary) configuration, all targeting Subtask 2.1, using the same CLIP feature extraction pipeline.

4.2. Hyperparameter configurations

We evaluated our models under two different settings: *soft runs*, which produce a probability distribution over the $\{\text{YES}, \text{NO}\}$ classes to better capture annotation variability, and *hard runs*, which generate a binary decision.

For the *soft* configuration, all models were trained using the full annotator distribution (without applying majority voting) and optimized via Focal Loss ($\gamma = 2$). Table 1 details the hyperparameters

for these runs, where we experimented with different architectural backbones (XLM-RoBERTa and DistilBERT) and multi-task learning settings.

The *hard* runs were trained to predict a single label, with hyperparameters fine-tuned to balance class-imbalance sensitivity and convergence speed (see Table 2). We utilized a range of models, including DeBERTa-v3, and incorporated specific regularization strategies such as layer freezing and early stopping to prevent overfitting.

Table 1
Hyperparameter configuration for soft-label runs.

Parameter	Team_1	Team_2	Team_3
Model	XLM-RoBERTa-base	XLM-RoBERTa-base	DistilBERT-base-mc
Multi-task	Yes	No	Yes
Freeze layers	2	2	2
Dropout	0.3	0.3	0.3
Threshold	0.45	0.45	0.40
Learning rate	1.5×10^{-5}	1.5×10^{-5}	2.0×10^{-5}
Batch size	8	8	32
Accumulation steps	4	4	2
Focal α	0.45	0.45	0.75
Focal γ	2.0	2.0	2.0
Class weights (NO/YES)	1.1 / 1.0	1.1 / 1.0	1.0 / 1.0
Augmentation	synonym	synonym + swap	synonym

Table 2
Hyperparameter configuration for hard-label runs.

Parameter	Team_1	Team_2	Team_3
Model	XLM-RoBERTa-base	DistilBERT-base-mc	DeBERTa-v3-base
Multi-task	No	No	No
Freeze layers	0	1	0
Dropout	0.4	0.3	0.4
Threshold	0.35	0.40	0.38
Learning rate	1.5×10^{-5}	1.0×10^{-5}	1.5×10^{-5}
Batch size	8	32	12
Focal α	0.45	0.30	0.70
Focal γ	2.0	2.5	2.5
Class weights (YES/NO)	1.1 / 1.0	1.0 / 1.0	1.0 / 1.0
Augmentation	synonym	swap + synonym	synonym
Early stopping patience	4	8	3

4.3. Results

The results achieved by the UNED-Annotate-Team in Task 2.1 (Memes), as reported by the EXIST 2026 organizers [27], highlight the effectiveness of our multitask architecture and the contribution of semantic visual information to sexism detection. As shown in Tables 3, 4, and 5, corresponding respectively to the results for all instances, Spanish instances, and English instances, the proposed approach performs consistently across evaluation settings. In particular, it achieves competitive results under both Hard and Soft evaluation metrics, suggesting that it is able to capture not only the dominant label assigned to each meme but also the diversity of annotator perceptions reflected in the soft labels.

In the English (EN) task, the UNED-Annotate-Team_1 system stands out with an F_1-score of 0.7306 in Hard Labels, suggesting that integrating CLIP-derived visual features with demographic information is particularly effective at capturing the nuances of sexism in English-language memes.

We also want to highlight the system's capabilities in a multilingual context; when comparing results between Spanish (ES) and English (EN), UNED-Annotate-Team_2 achieved a notable F_1-score of 0.7069 in Spanish, validating that our multitask learning approach is transferable across different linguistic and cultural structures.

Regarding our ability to handle disagreement among annotators, the analysis of the Soft results (ICM-Soft and its normalized form) indicates that our model achieves better alignment with the variability of original annotations compared to traditional approaches. The divergence observed in the ICM values reflects the ability of the UNED-Annotate-Team model to capture dissent in the annotators' judgments in dealing with the complex content of sexist memes.

In conclusion, these results, obtained by combining multitasking learning using demographic attributes as auxiliary variables with semantic visual enrichment, indicate that our system has the capacity to manage the ambiguity inherent in the multimodal content of sexist memes.

Table 3
Results for Subtask 2.1 for both English and Spanish instances.

	Hard Labels		Soft Labels	
System	F_1	ICM-Norm	CrossEntropy	ICM-Norm
UNED-Annotate-Team_1	0.6863	0.5124	1.1127	0.2274
UNED-Annotate-Team_2	0.7020	0.5616	1.1178	0.2309
UNED-Annotate-Team_3	0.5767	0.4387	1.0896	0.2468

Table 4
Results for Subtask 2.1 for Spanish-only instances.

	Hard Labels		Soft Labels	
System	F_1	ICM-Norm	CrossEntropy	ICM-Norm
UNED-Annotate-Team_1	0.6431	0.4380	1.1083	0.2303
UNED-Annotate-Team_2	0.7069	0.5361	1.1129	0.2374
UNED-Annotate-Team_3	0.6113	0.4569	1.0758	0.2560

Table 5
Results for Subtask 2.1 for English-only instances.

	Hard Labels		Soft Labels	
System	F_1	ICM-Norm	CrossEntropy	ICM-Norm
UNED-Annotate-Team_1	0.7306	0.5873	1.1174	0.2245
UNED-Annotate-Team_2	0.6962	0.5864	1.1229	0.2239
UNED-Annotate-Team_3	0.5365	0.4197	1.1042	0.2371

To contextualize our results in the EXIST 2026 competition, tables 6 and 7 show a comparison of our three runs (UNED-Annotate-Team1, 2, and 3) with the best-performing proposal (DDAIUNICC_1 / Ordantis_1) and the EXIST 2025 benchmarks.

In the Hard Comp results (Table 6), our approach outperforms the benchmark of the majority class. While DDAIUNICC_1 is the best-performing system, our Team_2 is among the most efficient solutions, considering the optimization of the resources used.

Regarding the Soft results (Table 7),while our approach maintains a stable baseline, it does not achieve the expected alignment with the variability of original annotator judgments compared to the best-performing systems in the competition. As shown in the comparison tables, our models exhibit higher Cross-Entropy and lower alignment scores in the Soft setting.

Table 6
Comparative performance of the UNED-Annotate-Team systems against top-ranking entries and EXIST2025 baselines on Hard Labels.

	ALL		English		Spanish	
System	ICM	F_1	ICM	F_1	ICM	F_1
EXIST2025-test_gold	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DDAIUNICC_1	0.7387	0.8076	0.7680	0.8318	0.7092	0.7855
UNED-Annotate-Team_2	0.5616	0.7020	0.5864	0.6962	0.5361	0.7069
UNED-Annotate-Team_1	0.5124	0.6863	0.5873	0.7306	0.4380	0.6431
UNED-Annotate-Team_3	0.4387	0.5767	0.4197	0.5365	0.4569	0.6113
EXIST2025-test_majority-class	0.2947	0.6821	0.2931	0.6880	0.2962	0.6765
EXIST2025-test_minority-class	0.1711	0.0000	0.1761	0.0000	0.1660	0.0000

Table 7
Comparative performance of the UNED-Annotate-Team systems against top-ranking entries and EXIST2025 baselines on Soft Labels.

	ALL		English		Spanish	
System	ICM	CE	ICM	CE	ICM	CE
EXIST2025-test_gold	1.0000	0.5852	1.0000	0.5528	1.0000	0.6160
Ordantis_1	0.6158	0.8155	0.6380	0.7823	0.5912	0.8470
UNED-Annotate-Team_3	0.2468	1.0896	0.2371	1.1042	0.2560	1.0758
UNED-Annotate-Team_2	0.2309	1.1178	0.2239	1.1229	0.2374	1.1129
UNED-Annotate-Team_1	0.2274	1.1127	0.2245	1.1174	0.2303	1.1083
EXIST2025-test_majority-class	0.1212	4.4015	0.1390	4.4798	0.1014	4.3270
EXIST2025-test_minority-class	0.0000	5.5672	0.0000	5.4888	0.0000	5.6416

*Note: CE stands for Cross-Entropy.

4.4. Ablation Study

Ablation Study on Features Extra

The first ablation study focuses on the contribution of extra features directly related to the meme image. These features are the 10 PCA components, `img_score`, `img_mean`, and `img_std`. All configurations start from the same base setup described in Section 4.2 (run"hard") and four configurations were evaluated: Configuration A, which includes all features and is the one presented in the final run; Configuration B, which includes only img_score, img_mean, and img_std; Configuration C, which includes only the PCA components; and Configuration D, which includes no features.

In the results presented in Table A, the configuration A used in our work achieves the best performance across all metrics, with an accuracy of 0.7177 and a weighted F-score of 0.7049. The contribution of the multimodal features is confirmed when their removal results in a 4.77% reduction in accuracy. A more detailed analysis reveals that the statistics (score, mean, and standard deviation) contribute slightly more than the PCA components; their removal causes a greater drop in performance. They also contribute to maintain a stable balanced accuracy (0.6813), positively impacting class balance.

Table 8
Performance comparison across different feature configurations

Experiment	**Features**	**Accuracy**	**F_1-weighted**	**Balanced Acc**	**MCC**
A (Complete)	10 PCA + 3 stats (13)	0.7177	0.7049	0.6813	0.3978
B (Score+Mean+Std)	3 stats	0.6918	0.6922	0.6811	0.3614
C (Only PCA)	10 PCA	0.6938	0.6919	0.6764	0.3575
D (No features)	None	0.6700	0.6713	0.6620	0.3213

Ablation Study on Auxiliary Tasks and Smart Tokens

To evaluate how sociodemographic information modulates sexism detection, we performed an ablation study on the soft corpus. Our base architecture (XLMRoBERTa-TranspMean MHA with multimodal features) was subjected to different strategies: first, using auxiliary multitask heads for gender and age prediction; second, an experimental configuration (SmartTokens) where demographic attributes are injected as specialized input-level tokens. The goal was to determine whether treating annotator background as an auxiliary task or as contextual modulation yields more robust representations. Table 9 summarizes these results.

Table 9
Ablation study on auxiliary tasks and smart tokens for sexism detection.

Configuration	Acc.	F_1-macro	F_1-YES	F_1-NO	MCC	Time (s)
A. Multi-task (Gender + Age)	0.6833	0.6792	0.6431	0.7154	0.3585	6980
B. Multi-task (Gender only)	0.6735	0.6690	0.6305	0.7075	0.3380	6115
C. Multi-task (Age only)	0.6710	0.6650	0.6205	0.7096	0.3307	8137
D. SmartTokens	0.6918	0.6877	0.7236	0.6517	0.3754	4815

The results presented in Table 9 identify the benefits of these different strategies. Configuration D (SmartTokens) exhibits the highest overall accuracy (0.6918) and MCC (0.3754), and a significant improvement in minority class detection, achieving an F_1-YES value of 0.7236, compared to the range of 0.6205–0.6431 for the multitasking configurations (A, B, and C), which also show a bias toward the majority class (F_1-NO). Furthermore, the SmartTokens approach proves to be computationally more efficient, reducing training time to 4815 seconds, a reduction that can be attributed to the simplified architecture achieved by avoiding optimization goals in secondary tasks and potential gradient conflicts.

5. Discussion of Results

In this work, we have highlighted the complexity of detecting sexist content in memes. Its multimodal nature means that model performance is strongly influenced by the synergy between visual semantics and textual cues. Our experiments suggest that the fidelity and richness of the representation of the input data have a greater impact on the detection outcome than the implemented architecture.

A key finding in our ablation studies relates to the integration of the annotator's subjectivity. Initially, we hypothesized that multitask learning (MTL), by requiring the encoder to learn demographic patterns alongside sexism labels, would act as a powerful regularizer. However, our results indicate that the MTL architecture generates a competing process where the demographic auxiliary tasks create noisy gradients, which can destabilize the primary semantic learning performed from the CLIP features and the OCR text of the meme. This approach treats sociodemographic cues as decoupled, watertight compartments (Late Fusion), rather than modulating them in the input space.

Conversely, better performance is obtained with input-level injection (SmartTokens), suggesting that sociodemographic attributes have a more positive influence on training when directly interconnected with the context. This indicates that, for meme-based moderation, the model performs better when the annotator's context is incorporated at the beginning of the transformation process, rather than being treated as a target to be predicted. These results reveal an architectural limitation: current MTL implementations may not be detailed enough to capture the nuances of human subjectivity in multimodal content. Future work will investigate whether this can be further refined through dynamic fine-tuning of visual encoders (e.g., CLIP-LoRA) or via specialized sensor-based gating mechanisms that adjust the influence of annotator background based on the inherent ambiguity of the meme.

Furthermore, the robustness of our results, particularly in detecting the (sexist) minority class (achieving an F_1-score of 72.37% in our best configuration), confirms the choice of our CLIP-based preprocessing strategy combined with early fusion. Consequently, our findings demonstrate that,

for multimodal content moderation, model performance is optimized when architectures prioritize integrating contextual features into the input layer, allowing the model to properly integrate the ambiguity characteristic of sexist memes.

6. Limitations of the Study

Despite the performance improvements offered by the proposed architecture, our approach has several limitations that warrant analysis:

Static multimodal representation versus end-to-end fine-tuning: Our decision to use CLIP as a static feature extractor—during the preprocessing stage—was based on maintaining computational efficiency. However, this approach may restrict the model's ability to more seamlessly adapt visual representations to the nuances of sexist content during training. Future work should explore fine-tuning with efficient parameters (e.g., LoRA) to enable dynamic multimodal adjustment during training.

Dependence on OCR-provided text: Our model relies on OCR-generated transcriptions. Since memes often contain text embedded in the image with different fonts, sizes, and orientations, we must consider the possibility of OCR errors. Any noise introduced at this stage propagates throughout the entire process, potentially leading to misalignment between the text representation and the visual context.

Dimensionality reduction using PCA: We opted to reduce the CLIP embeddings from 512 dimensions to 10 principal components. While our experiments indicated that increasing this number introduced excessive noise, this significant compression implies a potential loss of detailed visual information. Further investigation into nonlinear dimensionality reduction (e.g., UMAP or autoencoders) could preserve more complex visual structures without sacrificing stability.

Manual definition of semantic concept sets: The Multimodal Sexism Evaluator relies on a manually defined lexicon of sexist and non-sexist concepts in both English and Spanish, which may limit its scope and introduce biases inherent in the chosen terms. This approach has provided an effective semantic foundation, but it is subject to improvement, where large-scale language models (LLMs) could certainly strengthen this scoring mechanism.

7. Conclusions and Future Work

Our experimental results lead us to two main conclusions:

First, the complexity of moderating multimodal content cannot be addressed solely by increasing the depth of the architecture or treating the annotator's subjectivity as a secondary objective. We have shown that the Multitasking Learning (MTL) framework, while theoretically attractive, often introduces gradient interference that hinders the model's ability to learn robust representations.

Second, integrating contextual sociodemographic information is more effective when done at the model's input level. Experiments using input-level token injection in conjunction with text indicate that allowing the Transformer to pay attention to the annotator's context, acting as a modulator—rather than as a prediction task—improves performance in detecting sexist minority classes. This, combined with our early fusion strategy based on CLIP, establishes a resource-efficient foundation that adequately addresses the inherent ambiguity of sexist memes.

Our work points to several avenues for future research:

- Advancing the use of smart tokens that inject sociodemographic information directly into the Transformer's input stream, allowing these attributes to be represented within the same latent space as the textual content. This enables the attention mechanism to incorporate annotator-related contextual information during both encoding and representation learning.
- Improving multimodal representations by incorporating parameter-efficient fine-tuning techniques (e.g., LoRA) into CLIP, enabling the visual encoder to adapt feature extraction to the specific nuances of sexism detection in multimodal content.

- Leveraging information provided by sensors and contextual signals. Future work could explore the integration of data collected from wearable devices, mobile phones, or interaction logs to model contextual factors that may influence how users perceive and interpret potentially sexist content. Such information could contribute to the development of more personalized and context-aware moderation systems capable of better capturing the diversity of human judgments.

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31) funded by MICIU/AEI/10.13039/501100011 033 and the European Social Fund Plus (ESF+).

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Martínez-Bacaicoa, N. Henry, E. Mateos-Pérez, M. Gámez-Guadix, Victimización online por violencia de género entre adultos: prevalencia, predictores y resultados psicológicos, Psicothema 36 (2024) 247–256.

[2] A. C. Salerno-Ferraro, C. Erentzen, R. A. Schuller, Young women's experiences with technology-facilitated sexual violence from male strangers, Journal of interpersonal violence 37 (2022) NP17860–NP17885.

[3] U. K. Schmid, Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes, New Media & Society 27 (2025) 1588–1606.

[4] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[5] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110.

[6] R. Wan, J. Kim, D. Kang, Everyone's voice matters: Quantifying annotation disagreement using demographic information, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 14523–14530.

[7] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.

[8] L. Plaza, J. Carrillo-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: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 266–289.

[9] 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.

[10] F. R. Sánchez, J. C. de Albornoz, L. P. Morales, A. M. Aragón, G. M. Remón, M. Makeienko, M. Plaza, J. G. Arroyo, D. Spina, P. Rosso, Overview of exist 2022:: sexism identification in social networks, Procesamiento del lenguaje natural (2022) 229–240.

[11] 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.
[12] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 338–347.
[13] I. Arcos, Identifying sexism in memes with multimodal deep learning: fusing text and visual cues, Working Notes of CLEF (2025).
[14] D. K. Jha, M. K. Mandal, A. K. Madasamy, Sexism identification using annotator ranking in memes: A multimodal approach using transformers, Working Notes of CLEF (2025).
[15] P. Italiani, F. Maqbool, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, Trankiltwice at exist2025: detecting sexism in memes under multi-lingual settings, Working Notes of CLEF (2025).
[16] A. Britez, I. Markov, Cltl at exist 2025: identifying sexist memes using an ensemble of shallow and transformer models, Working Notes of CLEF (2025).
[17] M. Guerrero-García, F. Carrillo García, J. Mata, V. Pachón-Álvarez, I2c-uhu-altair at exist2025: multimodal sexism detection and classification using advanced vision-language models blip2 and qwen, large language models, and learning with disagreement frameworks, Working Notes of CLEF (2025).
[18] 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, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.
[19] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content, Data in brief 44 (2022) 108526.
[20] 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: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.
[21] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).
[22] Q. Chen, L. Kong, Y. Chen, C. Sun, Modeling annotator subjectivity for sexism detection on social media, Working Notes of CLEF (2025).
[23] M. Hurtado, A. Tarrao, Dandys-de-bertganim at exist 2025: a multi-task learning architecture for sexism identification, CLEF 2025 Working Notes (2025).
[24] J. C. Prajogo, L.-H. Lee, H. I. Lin, Nycu-nlp at exist 2025: An empirical study of annotator-aware two-stage pipeline for sexism detection in tweets, CLEF 2025 Working Notes (2025).
[25] OpenAI, CLIP ViT-B/32: openai/clip-vit-base-patch32, Hugging Face, 2021. URL: https://huggingface.co/openai/clip-vit-base-patch32, accessed: 2026-05-20.
[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.
[27] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, D. Spina, Overview of exist 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.
[28] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, D. Spina, Overview of exist 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.