Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media

Laura Plaza[1], Jorge Carrillo-de-Albornoz[1], Iván Arcos[2], María Aloy-Mayo[2], Paolo Rosso[2,3], Enrique García-Arias[1], Damiano Spina[4]

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
lplaza,jcalbornoz,e.garcia@lsi.uned.es
[2] Universitat Politècnica de València (UPV), 46022 Valencia, Spain
iarcgab@etsinf.upv.es, malomay@upvnet.upv.es, prosso@dsic.upv.es
[3] ValgrAI – Valencian Graduate School and Research Network of Artificial Intelligence, 46022 Valencia, Spain
[4] RMIT University, 3000 Melbourne, Australia
damiano.spina@rmit.edu.au

Abstract. This paper presents the EXIST 2026 Lab on sexism detection and categorization in social media, held at the CLEF 2026 conference, and marks the sixth edition of the EXIST Shared Task. EXIST 2026 further expands the study of online sexism by focusing on complex multimedia formats where sexist meaning may emerge from the interaction between visual, textual, temporal, and contextual cues. The lab comprises six subtasks in two languages, English and Spanish, organized around three core objectives: sexism identification, source intention detection, and sexism categorization. These objectives are applied across two multimedia formats: images, in the form of memes, and videos, in the form of TikToks. In contrast to previous editions, EXIST 2026 focuses exclusively on multimedia content and introduces a Human-Centered AI perspective by enriching the data with sensor-based information, including eye-tracking, heart-rate, and EEG signals collected from subjects exposed to potentially sexist content. As in previous editions, EXIST 2026 adopts the "Learning With Disagreement" paradigm, using annotations from multiple annotators to represent diverse and sometimes conflicting perceptions of sexism. This edition extends that paradigm by exploring how conscious labels and unconscious physiological or behavioral responses can jointly contribute to the development and evaluation of sexism detection systems. This overview describes the task design, datasets, evaluation methodology, participating systems, and results of EXIST 2026. The lab registered 247 teams from 26 countries, with 122 teams from 16 countries submitting runs, and a total of 1,303 runs processed.
Warning: Some of the examples included in this paper may contain offensive language and explicit descriptions of sexist behavior, which may be disturbing to the reader.

Keywords: sexism identification · sexism categorization · learning with disagreements · tweets · memes · TikTok videos · human-centric AI

1 Introduction

Sexism remains a pervasive form of social discrimination, manifesting itself in many areas of society, including sexual violence, economic inequality, and online harassment. Recent reports show that women account for the vast majority of sexual violence victims and continue to face significant gender pay gaps across many countries [6, 10, 11, 17]. In the digital sphere, they are also disproportionately exposed to harassment and discriminatory behaviour, with reported rates ranging from 16% in the USA to 41% in Australia, compared to 5–26% for men [7, 12]. Detecting and understanding sexist content is therefore an important step toward promoting safer and more inclusive online environments. Automatic methods for identifying and categorizing sexism can help researchers and practitioners better understand how prejudice is expressed and disseminated through digital communication.

In this context, the development of AI systems capable of detecting sexism on social media presents a particularly relevant challenge. The perception of what constitutes sexist behavior or expression involves a certain degree of subjectivity, as it may be influenced by cultural norms, personal experiences, and emotional reactions that cannot be fully captured through linguistic data alone. Despite significant advances in computational modeling, the mechanisms underlying human decision-making remain only partially understood. Empirical evidence suggests that human judgments are shaped not only by conscious factors—such as socio-demographic background, prior experiences, and explicit beliefs—but also by unconscious cues, including emotions, physiological states, and sensory responses that subtly guide perception and evaluation.

Current AI models, largely trained on textual or visual data, lack access to these deeper layers of cognitive and affective information, limiting their ability to replicate or interpret complex social phenomena. To bridge this gap, it becomes essential to explore new training paradigms that integrate human-centered and sensor-based data – such as eye tracking, heart rate, and EEG signals – to provide richer insights into how individuals consciously and unconsciously perceive sexist content. Neurophysiological signals collected through wearable devices have already proven effective in characterizing a wide range of complex information access tasks [8, 9, 20]. Therefore, incorporating this multimodal perspective could lead to fairer and more context-aware AI systems, capable of understanding not only the explicit meaning of language but also the implicit affective and cognitive processes that drive human interpretations.

EXIST 2026[5] will mark the sixth edition of the sEXism Identification in Social neTworks challenge. EXIST is a series of shared tasks and scientific events aimed at detecting and characterizing sexism in a broad sense, ranging from explicit misogyny to more subtle and implicit sexist behaviors. The last three editions of the EXIST shared task were organized as labs within CLEF [13–15], while the first two were held in the IberLEF evaluation forum [18, 19]. Over the

[5] https://nlp.uned.es/exist2026

past five years, more than 200 teams from research institutions and companies have participated in the EXIST challenge, submitting over 1,500 runs.

The EXIST 2026 lab will continue to focus on detecting sexism in social networks while introducing a novel paradigm that integrates human-centered signals into the AI development pipeline. Specifically, we incorporate sensor-based data—EEG, heart rate, and eye-tracking—collected from human annotators while they are exposed to potentially sexist content, with the aim of capturing implicit and pre-conscious responses to sexism.[6] These physiological signals provide complementary information to explicit annotations, reflecting affective and cognitive processes that may not be accessible through self-report or conscious reasoning alone. Findings from previous editions indicate that individuals often register sexist content at an implicit level, exhibiting measurable physiological reactions even when they are unable to explicitly articulate why the content is sexist or to consciously justify their annotation decisions.

By registering these signals during the annotation process, we aim to enrich training data with human-centered cues that capture early attentional shifts, emotional arousal, and cognitive load associated with the perception of harmful content. Given their inherently multimodal nature, our analysis will primarily focus on memes and short videos—formats that combine textual and visual information and are particularly suitable for eliciting such responses. We hypothesise that incorporating physiological correlates of implicit human perception into machine learning models will enable automated systems to better approximate human sensitivity to subtle or context-dependent forms of sexism, particularly in cases where surface linguistic features are insufficient, as previous studies have shown [2, 3, 5]. In this way, physiological signals are not treated as direct labels, but as auxiliary supervisory signals that can guide models toward more nuanced, human-aligned representations of sexist content.

This human-in-the-loop approach acknowledges the diversity of human reactions to sexist content and opens new avenues for developing more robust, fair, and interpretable AI systems. By combining explicit annotations with implicit physiological responses, EXIST 2026 aims to provide a richer and more ethically grounded representation of how sexism is perceived across different platforms and modalities.

In addition, EXIST 2026 will continue to adopt the Learning with Disagreement (LeWiDi) paradigm for both dataset construction and system evaluation, following the successful experience of previous editions. LeWiDi promotes a human-centred approach to AI by preserving, rather than eliminating, annotation disagreements, recognising that the perception of sexism is often subjective and context-dependent [4]. Instead of assuming a single "correct" interpretation, the paradigm enables models to learn from the diversity of human judgments, capturing the ambiguity inherent in complex social phenomena. This not only helps reduce potential biases introduced by majority voting, but also encourages the development of more transparent, inclusive, and human-aligned AI systems.

[6] Data collection was approved by the Research Ethics Committee of the Universitat Politècnica de València (Project IDs: P01_31-10-2025 and P13_26-12-2025).

In the following sections, we describe the tasks, dataset, and evaluation methodology that will be adopted in the EXIST 2026 challenge at CLEF.

2 Tasks

We define six experimental tasks within this lab, corresponding to three distinct task types applied across two multimodal datasets: memes and videos. Subjects are asked to identify the presence of sexism and to characterize it according to the intention of the source and the type of sexism. Each instance (meme or TikTok video) was independently evaluated by multiple annotators, whose physiological and behavioral responses were captured through eye tracking, heart rate monitoring, and electroencephalography. The resulting multimodal physiological signals, along with the corresponding annotation labels, are linked to each dataset instance.

2.1 Sexism Identification

The first task is a binary classification where systems must decide whether or not a given meme or video contains sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). Figure 1 show examples of sexist and not sexist memes, respectively.

(a) Sexist meme (b) Not sexist meme

Fig. 1: Examples of sexist and not sexist memes.

2.2 Source Intention Detection

This task aims to categorize a meme or TikTok video according to the intention of the author. We propose a binary classification task: (i) direct sexist, and (ii) judgmental. The following categories are defined:

– **Direct** sexism: the intention was to create content that is sexist by itself or incites to be sexist. Figure 2 shows an example of a direct meme.

– **Judgemental** message: the intention is judgmental, since the meme or video describes sexist situations or behaviors with the aim of condemning them. Figure 2 shows an example of a judgemental meme.

Figure 2 show examples of judgemental and direct sexist memes, respectively.

(a) Sexist (direct) (b) Sexist (judgemental)

Fig. 2: Examples of direct and judgemental memes.

2.3 Sexism Categorization

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. In this task, each sexist meme/video must be categorized in one or more of the following categories:

- **Ideological and inequality**: this category includes memes/videos that discredit the feminist movement in order to defame the struggle of women in any aspect of their lives. It also includes contents that reject inequality between men and women, or present men as victims of gender-based oppression.
- **Stereotyping and dominance**: this category includes memes/videos that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks. It also includes any claim that implies that men are somehow superior to women.
- **Objectification**: It includes content that objectifies women or reinforces stereotypes about their physical appearance, such as the expectation to meet traditional beauty standards or criticisms of their bodies.
- **Sexual violence**: this category includes memes/videos where sexual suggestions or harassment of a sexual nature (rape or sexual assault) are made.
- **Misogyny and non sexual violence**: this category includes expressions of hatred and violence towards women.

Examples of sexist memes from the previously described categories are shown in Figure 3.

(a) Ideological & inequality

(b) Objectification

(c) Stereotyping & dominance

(d) Sexual violence

(e) Misogyny & non-sexual violence

Fig. 3: Examples of memes from the different sexist categories.

3 Dataset

The EXIST 2026 dataset comprises 8,294 multimedia instances, including memes and short videos, in both English and Spanish. This collection integrates and extends the datasets developed for EXIST 2024 (memes) and EXIST 2025 (TikTok videos), which have been reused within a novel experimental setting designed to explore how individuals perceive and interpret sexism in online media. In this new dataset, subjects' conscious judgments are reflected in the labels they assigned to each instance, while their unconscious or implicit reactions are captured through sensor data such as eye tracking, heart rate, and EEG activity. The sensor data captured during the annotation process will be provided to the participant to use (if desired) in the training process. Together, these complementary layers of information offer a unique perspective on both the explicit and implicit dimensions of how sexism is processed and understood by human observers.

Detailed descriptions of the annotation methodologies for the EXIST 2024 and 2025 datasets are available in [13, 15]. Moreover, Table 1 summarises the dataset statistics.

Table 1: Distribution of multimedia instances in EXIST 2026 by modality, split, and language.

Modality	Split	Spanish	English	Total
Memes	Train	1,979	2,005	3,984
	Test	540	513	1,053
	Total	**2,519**	**2,518**	**5,037**
Videos	Train	1,524	1,000	2,524
	Test	304	370	674
	Total	**1,828**	**1,370**	**3,198**

3.1 Physiological Data

Physiological signals were recorded for all meme and video stimuli during a controlled annotation experiment. Participants viewed each stimulus on a screen and submitted their annotation before proceeding to the next item, with a 3-second pause between consecutive stimuli to minimise carryover effects. Each session lasted approximately 45 minutes, during which participants annotated between 100 and 170 stimuli. All participants provided informed consent for the anonymous use of their data. In total, 13 participants annotated memes and 8 participants annotated videos, using language-specific identifiers such as `ES1` or `EN2`. Each stimulus was annotated by 2–4 participants, allowing the same item to be associated with multiple subject-level physiological responses.

Physiological data are provided through three complementary modalities. Eye-tracking (ET) includes 24 variables describing visual attention and response dynamics, including left and right pupil diameter statistics, blink count and duration, fixation count and duration, saccade count and duration, and reaction time. Heart rate (HR) provides 4 variables—mean, standard deviation, minimum, and maximum heart rate—recorded using a Garmin wearable device and used as aggregate indicators of autonomic arousal.

EEG provides 80 bandpower features computed over 16 channels, EXG Channels 0 to 15, across five frequency bands: Delta, Theta, Alpha, Beta, and Gamma. Overall, each subject-level physiological record contains up to 108 numerical features: 24 from ET, 4 from HR, and 80 from EEG, capturing complementary signals related to attention, autonomic arousal, and cognitive and affective processing during the interpretation of potentially sexist content.

4 Evaluation Methodology and Metrics

As in previous EXIST editions, we applied two evaluation strategies: a **soft evaluation** and a **hard evaluation**. The soft evaluation aligns with the LeWiDi

paradigm and aims to assess how well models reflect disagreement among annotators. The hard evaluation follows the conventional approach, where systems output a single label per instance, and performance is assessed against a majority-vote gold standard.

1. **Soft-soft Evaluation**. For systems that return a probability distribution over the classes, we compare these distributions with the ones derived from human annotators. The class probabilities per instance are calculated from the frequency of annotations and the number of annotators involved. The primary metric for this evaluation is ICM-Soft, an adaptation of the original Information Contrast Model (ICM) [1]. We also report results using the normalized version, ICM-Soft Norm. For a description of the ICM-Soft metrics, please refer to [15].
2. **Hard-hard Evaluation**. For systems providing discrete (hard) predictions, we compute the reference labels using a task-specific probabilistic threshold based on annotator agreement: for subtasks 2.1 and 3.1, the label selected by more than three annotators is used; for subtasks 2.2 and 3.2, the threshold is more than two annotators; and for multilabel subtasks 2.3 and 3.3, any label assigned by more than one annotator is included. Instances without a dominant label (i.e., no class surpassing the threshold) are excluded from this evaluation. The main metric is the original ICM as defined by Amigó and Delgado [1]. We additionally report a normalized ICM (ICM Norm) and F1.

5 Overview of Approaches

This section offers an overview of the methodological approaches submitted to EXIST 2026. A total of 122 teams submitted results, participating across the six tasks with both hard and soft evaluation outputs. Table 2 summarizes the participation across tasks and evaluation contexts. Note that the task numbering is intended to ensure consistency with the structure of the 2025 dataset and the organization of the EvALL repository.

Table 2: Runs submitted and teams participating in each EXIST 2026 task.

Task	Description	#Runs	#Teams*
2.1	Sexism identification in memes	351	100
2.2	Intent classification in memes	293	86
2.3	Multi-label categorization in memes	295	86
3.1	Sexism identification in videos	131	29
3.2	Intent classification in videos	116	27
3.3	Multi-label categorization in videos	117	27

*unique teams that submitted at least one run

Table 3 disaggregates submissions by evaluation type. Hard-label evaluations were significantly more popular across all tasks, attracting approximately two to three times more teams than soft-label evaluations. For meme tasks, 98 teams submitted hard outputs compared to 66 for soft outputs in TASK2_1, with similar ratios observed across TASK2_2 (84 hard vs 54 soft) and TASK2_3 (84 hard vs 53 soft). For the video tasks, the pattern was consistent: 29 teams submitted hard outputs compared to 24 for soft outputs in TASK3_1, while in TASK3_2 and TASK3_3 the corresponding numbers were 27 and 20 teams, respectively.

Table 3: Runs submitted and teams participating in each EXIST 2026 task (by evaluation type).

Task	**#Runs (Hard)**	**#Runs (Soft)**	**#Teams (Hard)**	**#Teams (Soft)**
2.1	212	139	98	66
2.2	181	112	84	54
2.3	182	113	84	53
3.1	73	58	29	24
3.2	68	48	27	20
3.3	67	50	27	20

5.1 Methodology

To assess the relationship between the use of specific techniques and the final rankings, we employed the Spearman rank correlation coefficient. Spearman correlation is calculated using two variables: the first indicates whether a team uses a specific technique (binary variable), and the second measures the team's ranking. In practice, this calculation is equivalent to comparing the average rank of teams that use the technique against those that do not use it. If the average rank of those who use the technique is much lower (indicating a better position in the ranking), the resulting correlation will be negative. Therefore, a negative correlation implies that the technique is beneficial.

The Spearman correlation coefficient ρ is computed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where n is the number of teams analyzed in the task, and d_i is the difference between the ranks of the two variables for the i-th team.

The p-value is the probability that, under the hypothesis that the technique has no real effect (pure chance), we would observe a correlation equal to or more extreme than the one calculated. If the p-value is less than 0.05, we reject the random chance hypothesis and therefore consider the result statistically

significant. To achieve this value, we calculate a test statistic t that adjusts the correlation by the sample size:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

This t statistic follows a t-distribution with $n-2$ degrees of freedom. The p-value is then defined as:

$$p = 2 \cdot P(T > |t|)$$

where T is a random variable following that t-distribution.

The Rank improvement shown in the tables is a complementary and more intuitive measure to the Spearman correlation and refers to the difference in average ranking between teams that did not use the technique and teams that did use it. It is calculated as:

$$\text{Rank Improvement} = \text{Rank}_{\text{without}} - \text{Rank}_{\text{with}}$$

5.2 Sexism Detection in Memes

For meme classification, transformer encoders were the predominant architectural choice, being adopted by over eighty percent of teams, with XLM-RoBERTa emerging as the most frequently used backbone. CLIP was the most commonly employed visual feature extractor, used by nearly half of the teams, followed by ViT and CNN-based models. Large language models and vision-language models were also incorporated by several teams, particularly among the higher-ranked systems. Teams employed LLMs both as feature extractors, generating embeddings or explanations, and as zero-shot classifiers through prompting.

Multimodal fusion strategies were widely used across submissions. Early fusion, which combines textual and visual embeddings before classification, was adopted by approximately forty-three percent of teams. Cross-attention mechanisms were also present in a substantial proportion of systems, reflecting the growing adoption of attention-based multimodal architectures.

Regarding optimization strategies, ensemble methods, threshold tuning, and class balancing were among the most frequently applied techniques. LoRA and adapter-based fine-tuning approaches were also utilized by a number of teams to enable efficient parameter adaptation.

Table 4 summarizes the most notable associations identified through the statistical analysis. CLIP was associated with improved performance on meme tasks under soft evaluation, particularly for intent classification and multi-label categorization, while no significant effects were observed under hard evaluation. Large language models showed limited association with performance under soft evaluation but were linked to stronger rankings across the three meme tasks under hard evaluation. Vision-language models exhibited similar trends, particularly for identification and categorization tasks. In contrast, class balancing was associated with lower rankings in multi-label categorization under hard evaluation.

Table 4: Most significant correlations between techniques and rankings in meme tasks (hard evaluation).

Technique	Task	Correlation	p-value	Rank Improvement
LLM	2.1 (identification)	−0.338	<0.001	+41 positions
LLM	2.2 (intent)	−0.358	<0.001	+39 positions
VLM	2.1 (identification)	−0.328	<0.001	+52 positions
CLIP*	2.2 (intent)	−0.324	0.017	+20 positions
Class balancing	2.3 (categorization)	0.263	0.016	–43 positions (harmful)

*Soft evaluation

5.3 Sexism Detection in TikTok Videos

Video classification required the integration of visual, textual, and temporal information. Transformer encoders were the predominant architectural choice, being adopted by over eighty percent of teams, followed by CLIP and large language models. Cross-attention mechanisms were also common, particularly among higher-ranked systems, appearing in approximately one quarter of submissions.

To model temporal information, teams employed a variety of strategies, including frame sampling to capture key moments, attention-based mechanisms over temporal sequences, and multimodal fusion approaches combining video frames, audio transcripts, and metadata.

Table 5 summarizes the main associations identified for the video tasks. Large language models were linked to improved rankings across all tasks under hard evaluation. Vision-language models and VLM-based captioning approaches were similarly associated with stronger performance in video categorization. In contrast, the use of physiological sensors, such as EEG, heart rate, and eye-tracking data, was associated with lower rankings for intent classification.

Unlike the results observed under hard evaluation, soft evaluation revealed no statistically significant associations for any video task. This suggests that hard evaluation may provide greater differentiation between competing technical approaches in this benchmark.

6 Results

In the next subsections, we report the results of the participants and the baseline systems for each task. We only show the results obtained by the top five teams for each task. For more detailed results, including disaggregated results per language, please refer to the Lab Working Notes [16] or the EXIST website: https://nlp.uned.es/exist2026/. The discussion of the techniques employed is based on the self-reported system descriptions provided by the teams

Table 5: Most significant correlations between techniques and rankings in video tasks (Hard evaluation).

Technique	Task	Correlation	p-value	Rank Improvement
LLM	3.2 (intent)	−0.535	0.004	+21 positions
VLM caption	3.3 (categorization)	−0.482	0.011	+20 positions
VLM	3.3 (categorization)	−0.468	0.014	+25 positions
Cross-attention	3.1 (identification)	−0.433	0.019	+19 positions
Phys. sensors	3.2 (intent)	0.444	0.020	−12 pos. (harmful)

when submitting their runs through the official submission form. Therefore, these methodological observations should be interpreted as descriptive evidence rather than as controlled ablation results.

6.1 Task 2.1: Sexism Identification in Memes

Task 2.1 focuses on the classification of memes as sexist or not sexist. Table 6 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top systems for this meme subtask are best characterized by vision-language modelling (AI Wizards, ELiRF-UPV, Ordantis), LLM or prompt-based processing (AI Wizards, DDAIUNICC, ELiRF-UPV), multimodal fusion (ELiRF-UPV, Ordantis, DreamTeam), and fine-tuning (AI Wizards, Ordantis, alBERT_319). This reinforces the interpretation that high-ranked meme systems benefited from combining visual, textual, and task-adapted signals, rather than from a single generic classifier. The method evidence is based on free-text system descriptions, so it should be read as qualitative context for the ranking rather than as an ablation study.

Regarding sensor-derived information, Task 2.1 provides a useful cautionary case. Ordantis ranked first in soft evaluation while explicitly reporting that eye-tracking and heart-rate sensors did not add value once textual meme descriptions were included. SPECTRA, whose description centres on physiological and gaze responses, and FourBytes, which uses sensorial metadata as auxiliary supervision, did not reach the leading group. The evidence therefore suggests that sensors were not the main factor behind the best Task 2.1 systems.

Soft Evaluation Soft-soft evaluation was led by `Ordantis_1` (Ordantis), with ICM-Soft Norm=0.6158. The winner showed a clear separation from the second-ranked unique team (7.85 percentage points in ICM-Soft Norm), making the top position comparatively more distinct. The selected top-five systems were all above the simple majority and minority baselines.

Table 6: Top 5 systems from different teams in EXIST 2026 Task 2.1. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
Test_gold	–	3.1107	1.0000	0.5852	0
Ordantis_1	Ordantis	0.7206	0.6158	0.8155	1
aiwizards_3	aiwizards	0.2323	0.5373	0.9090	4
alBERT-319_2	alBERT-319	0.0838	0.5135	0.8479	7
DreamTeam_1	DreamTeam	0.0689	0.5111	1.3606	8
Jow_1	Jow	0.0108	0.5017	0.9005	10
Test_majority-class	–	−2.3568	0.1212	4.4015	140
Test_minority-class	–	−3.5089	0.0000	5.5672	141
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**Macro F1**	**Rank**
Test_gold	–	0.9832	1.0000	1.0000	0
DDAIUNICC_1	DDAIUNICC	0.4693	0.7387	0.8076	1
Ordantis_2	Ordantis	0.4079	0.7074	0.8006	3
DreamTeam_1	DreamTeam	0.3542	0.6801	0.7886	7
ELiRF-UPV_3	ELiRF-UPV	0.3274	0.6665	0.7790	8
alBERT-319_3	alBERT-319	0.3134	0.6594	0.7571	9
Test_majority-class	–	−0.4038	0.2947	0.6821	202
Test_minority-class	–	−0.6468	0.1711	0.0000	214

Under soft-soft evaluation, the leaderboard for the classification of memes as sexist or not sexist included 139 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.1744 to 0.6158, with a mean of 0.407 and a standard deviation of 0.073. All submitted systems outperformed the strongest simple baseline (Test_majority-class), indicating that participant submissions were generally effective under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 11.41 percentage points, with the largest drop (7.85 percentage points) occurring from the first- to the second-ranked unique-team submission. This spread suggests remaining room for improvement and differences in how systems model soft labels for multimodal data.

Hard Evaluation Hard-hard evaluation was led by DDAIUNICC_1, with ICM-Hard Norm=0.7387. The winner had a modest advantage over the second-ranked unique team (3.13 percentage points in ICM-Hard Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Under hard-hard evaluation, the leaderboard for the classification of memes as sexist or not sexist included 212 submitted systems. The normalized ICM-

Hard Norm scores ranged from 0.1711 to 0.7387, with a mean of 0.488 and a standard deviation of 0.100. 201/212 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 7.93 percentage points, with the largest drop (3.13 percentage points) occurring from the first- to the second-ranked unique-team submission. This spread suggests remaining room for improvement and differences in how systems model discrete labels for multimodal data.

The leading team changed between evaluation protocols: ORDANTIS led soft evaluation, whereas DDAIUNICC led hard evaluation. This indicates that optimization for probabilistic outputs and discrete-label performance did not necessarily favor the same system.

6.2 Task 2.2: Source Intention in Memes

Task 2.2 focuses on the classification of the source intention conveyed by sexist memes. Table 7 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top systems for this meme subtask are best characterized by vision-language modelling (AI WIZARDS, ELIRF-UPV, ORDANTIS), LLM or prompt-based processing (AI WIZARDS, DDAIUNICC, ELIRF-UPV), multimodal fusion (ELIRF-UPV, ORDANTIS, MATH ADDICTS), and fine-tuning (AI WIZARDS, ORDANTIS, MATH ADDICTS). This reinforces the interpretation that high-ranked meme systems benefited from combining visual, textual, and task-adapted signals, rather than from a single generic classifier.

For Task 2.2, the strongest evidence again comes from ORDANTIS, which led both soft and hard evaluation while stating that eye-tracking and heart-rate signals were not useful after incorporating textual meme descriptions. Other teams using numerical sensor features, metadata, or physiological signals appear lower in the ranking. Thus, for source-intention classification in memes, sensor information appears secondary to strong textual, visual, and fusion-based representations.

Soft Evaluation Soft-soft evaluation was led by `Ordantis_1` (ORDANTIS), with ICM-Soft Norm=0.5012. The winner showed a clear separation from the second-ranked unique team (7.27 percentage points in ICM-Soft Norm), making the top position comparatively more distinct. The selected top-five systems were all above the simple majority and minority baselines.

Under soft-soft evaluation, the leaderboard for the classification of the source intention conveyed by sexist memes included 112 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.0000 to 0.5012, with a mean of 0.269 and a standard deviation of 0.102. All but one submitted system outperformed the strongest simple baseline (Test_majority-class), indicating that most

Table 7: Top 5 systems from different teams in EXIST 2026 Task 2.2. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
Test_gold	–	4.7018	1.0000	0.9325	0
Ordantis_1	Ordantis	0.0114	0.5012	1.3334	1
aiwizards_1	aiwizards	−0.6720	0.4285	1.4668	4
Os-Faixa-Preta-do-Data_1	Os-Faixa-Preta-do-Data	−0.8372	0.4110	1.4069	6
Jow_1	Jow	−0.9680	0.3971	1.4315	8
Elliot Thorel_1	Elliot Thorel	−1.0997	0.3831	1.7144	9
Test_majority-class	–	−5.0745	0.0000	5.5565	112
Test_minority-class	–	−18.9382	0.0000	8.0245	114
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**Macro F1**	**Rank**
Test_gold	–	1.4383	1.0000	1.0000	0
Ordantis_3	Ordantis	0.3709	0.6289	0.6158	1
DDAIUNICC_1	DDAIUNICC	0.3267	0.6136	0.6081	2
Math Addicts_1	Math Addicts	0.1110	0.5386	0.5469	6
DreamTeam_1	DreamTeam	0.0977	0.5340	0.5559	7
ELiRF-UPV_3	ELiRF-UPV	0.0879	0.5306	0.5501	8
Test_majority-class	–	−1.0445	0.1369	0.1839	167
Test_minority-class	–	−2.0637	0.0000	0.0697	183

submissions were effective under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 11.81 percentage points, with the largest drop (7.27 percentage points) occurring from the first- to the second-ranked unique-team submission. This spread suggests remaining room for improvement and differences in how systems model soft labels for multimodal data.

Hard Evaluation Hard-hard evaluation was led by Ordantis_3 (Ordantis), with ICM-Hard Norm=0.6289. The winner had a modest advantage over the second-ranked unique team (1.53 percentage points in ICM-Hard Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Under hard-hard evaluation, the leaderboard for the classification of the source intention conveyed by sexist memes included 181 submitted systems. The normalized ICM-Hard Norm scores ranged from 0.0281 to 0.6289, with a mean of 0.324 and a standard deviation of 0.113. 165/181 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 9.83 percentage points, with the largest drop (7.50 percentage points) occurring from the second- to the third-ranked unique-team

submission. This spread suggests remaining room for improvement and differences in how systems model discrete labels for multimodal data.

The same team, ORDANTIS, led both soft and hard evaluation, suggesting a robust performance across the two evaluation protocols for this subtask.

6.3 Task 2.3: Sexism Categorization in Memes

Task 2.3 focuses on categorizing memes according to the type of sexist they express. Table 8 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top systems for this meme subtask are best characterized by vision-language modelling (AI WIZARDS, ELIRF-UPV, ORDANTIS), LLM or prompt-based processing (AI WIZARDS, DDAIUNICC, ELIRF-UPV), multimodal fusion (ELIRF-UPV, ORDANTIS, MATH ADDICTS), and fine-tuning (AI WIZARDS, ORDANTIS, MATH ADDICTS). This reinforces the interpretation that high-ranked meme systems benefited from combining visual, textual, and task-adapted signals, rather than from a single generic classifier.

In Task 2.3, sensor evidence is also inconclusive. The hard-evaluation winner DDAIUNICC includes sensor-based extended variants, however the soft-evaluation winner AI WIZARDS is better characterized by LLM/vision-language adaptation than by sensors. ELIRF-UPV, which reports physiological signals in its multimodal architecture, is competitive in hard evaluation, but the available results do not isolate sensors from the rest of the visual-textual pipeline. Overall, sensors may help as auxiliary context, but the ranking does not show them as a consistent driver of categorization performance.

Soft Evaluation Soft-soft evaluation was led by `aiwizards_2` (AIWIZARDS), with ICM-Soft Norm=0.3469. The winner showed a clear separation from the second-ranked unique team (7.01 percentage points in ICM-Soft Norm), making the top position comparatively more distinct. The selected top-five systems were all above the simple majority and minority baselines.

Under soft-soft evaluation, the leaderboard for the categorization of sexist memes included 113 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.0000 to 0.3469, with a mean of 0.107 and a standard deviation of 0.100. 73/113 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 9.53 percentage points, with the largest drop (7.01 percentage points) occurring from the first- to the second-ranked unique-team submission.

Hard Evaluation Hard-hard evaluation was led by `DDAIUNICC_1`, with ICM-Hard Norm=0.5008. The winner had a modest advantage over the second-ranked

Table 8: Top 5 systems from different teams in EXIST 2026 Task 2.3. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Rank
Test_gold	–	9.4343	1.0000	0
aiwizards_2	AIWIZARDS	−2.8881	0.3469	1
Jow_2	JOW	−4.2110	0.2768	4
tai6le_2	TAI6LE	−4.3410	0.2699	6
Legleye-y-Clara_1	LEGLEYE-Y-CLARA	−4.6246	0.2549	8
Ordantis_1	ORDANTIS	−4.6874	0.2516	10
Test_majority-class	–	−9.8173	0.0000	74
Test_minority-class	–	−50.0353	0.0000	114

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	2.4100	1.0000	1.0000	0
DDAIUNICC_1	DDAIUNICC	0.0036	0.5008	0.5746	1
ELiRF-UPV_3	ELIRF-UPV	−0.1069	0.4778	0.5524	3
CYUT-Giantsshoulders_1	CYUT-GIANTSSHOULDERS	−0.1901	0.4606	0.5387	5
Math Addicts_1	MATH ADDICTS	−0.2203	0.4543	0.5413	7
aiwizards_3	AIWIZARDS	−0.3953	0.4180	0.4500	8
Test_majority-class	–	−2.0711	0.0703	0.0919	140
Test_minority-class	–	−3.3135	0.0000	0.0318	174

unique team (2.30 percentage points in ICM-Hard Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Under hard-hard evaluation, the leaderboard for the categorization of sexist memes included 182 submitted systems. The normalized ICM-Hard Norm scores ranged from 0.0000 to 0.5008, with a mean of 0.190 and a standard deviation of 0.131. 137/182 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 8.28 percentage points, with a notable drop of 3.63 percentage points from the fourth- to the fifth-ranked unique-team submission. This range of performance indicates that there is still room for improvement and that models differ significantly in how they handle discrete labels in multimodal data.

The leading team changed between evaluation protocols: AIWIZARDS led soft evaluation, whereas DDAIUNICC led hard evaluation. This indicates that optimization for probabilistic outputs and discrete-label performance did not necessarily favor the same system.

6.4 Task 3.1: Sexism Identification in Videos

Task 3.1 focuses on the classification of videos as sexist or not sexist. Table 9 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top video systems suggest a strong emphasis on transcript, ASR, OCR, or subtitle signals (AIDONNOW, ELIRF-UPV, GERMAN &BLAI), LLM or prompt-based processing (AIDONNOW, ELIRF-UPV, GERMAN &BLAI), transformer text encoders (AIDONNOW, GERMAN&BLAI, DOMINICAN PAPI), multimodal fusion (ELIRF-UPV, GERMAN&BLAI, AEGIS ATHENA), and classical classifiers over extracted features (AIDONNOW, GERMAN&BLAI, AEGIS ATHENA). This is consistent with the difficulty of short-video sexism detection, where the relevant evidence may be distributed across speech, visible text, frames, and context. As with the other subtasks, these observations come from participant descriptions and should be treated as qualitative evidence rather than controlled comparisons.

For Task 3.1, sensor-related evidence is mixed. GERMAN&BLAI led soft evaluation and describes configurations that include biometric data, while ELIRF-UPV led hard evaluation with a multimodal pipeline that also mentions physiological signals. At the same time, AIDONNOW's sensor-augmented soft run is competitive, but its hard-label systems explicitly avoid sensor features and remain strong. This suggests that sensors can be incorporated into competitive systems, but the largest gains appear to come from robust transcript/ASR, frame, and visual-language processing rather than from sensor channels alone.

Soft Evaluation Soft-soft evaluation was led by `German&Blai_2` (GERMAN &BLAI), with ICM-Soft Norm=0.5940. The winner had a modest advantage over the second-ranked unique team (1.35 percentage points in ICM-Soft Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Under soft-soft evaluation, the leaderboard for the classification of videos as sexist or not sexist included 58 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.2359 to 0.5940, with a mean of 0.454 and a standard deviation of 0.092. 56/58 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 4.83 percentage points, with the largest drop (3.22 percentage points) occurring from the third- to the fourth-ranked unique-team submission.

Hard Evaluation Hard-hard evaluation was led by `ELiRF-UPV_2` (ELIRF-UPV), with ICM-Hard Norm=0.6669. The leading systems were tightly clustered, with only 0.12 percentage points in ICM-Hard Norm separating the first- and second-ranked unique teams. The selected top-five systems were all above the simple majority and minority baselines.

Table 9: Top 5 systems from different teams in EXIST 2026 Task 3.1. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Test_gold	–	2.8488	1.0000	0.1962	0
German&Blai_2	German&Blai	0.5358	0.5940	1.2302	1
Dominican-Papi_2	Dominican-Papi	0.4586	0.5805	0.8305	2
AIdonnow_1	AIdonnow	0.4505	0.5791	0.8436	4
NeverChorizoInMyPaella_3	NeverChorizoInMyPaella	0.2675	0.5469	1.3183	7
Aegis-Athena_2	Aegis-Athena	0.2602	0.5457	0.8116	8
Test_majority-class	–	−1.2877	0.2740	4.4285	57
Test_minority-class	–	−2.0051	0.1481	5.5402	60

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	0.9907	1.0000	1.0000	0
ELiRF-UPV_2	ELiRF-UPV	0.3307	0.6669	0.7627	1
German&Blai_3	German&Blai	0.3284	0.6657	0.7487	2
Aegis-Athena_2	Aegis-Athena	0.3096	0.6563	0.7528	4
XORIK_2	XORIK	0.3007	0.6518	0.7543	5
Dominican-Papi_1	Dominican-Papi	0.2743	0.6384	0.7339	7
Test_majority-class	–	−0.4244	0.2858	0.0000	74
Test_minority-class	–	−0.6036	0.1954	0.6117	75

Under hard-hard evaluation, the leaderboard for the classification of videos as sexist or not sexist included 73 submitted systems. The normalized ICM-Hard Norm scores ranged from 0.2887 to 0.6669, with a mean of 0.561 and a standard deviation of 0.069. All submitted systems outperformed the strongest simple baseline (Test_majority-class), indicating that participant submissions were generally effective under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 2.85 percentage points, with a notable drop of 1.34 percentage points from the fourth- to the fifth-ranked unique-team submission. This suggests that there is still room for improvement.

The leading team changed between evaluation protocols: German&Blai led soft evaluation, whereas ELiRF-UPV led hard evaluation. This indicates that optimization for probabilistic outputs and discrete-label performance did not necessarily favor the same system.

6.5 Task 3.2: Source Intention in Videos

Task 3.2 focuses on the classification of the source intention conveyed by sexist videos. Table 10 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top video systems suggest a strong emphasis on transcript, ASR, OCR, or subtitle signals (AIdonnow, ELiRF-UPV, Aegis Athena),

LLM or prompt-based processing (AIdonnow, ELiRF-UPV, Aegis Athena), transformer text encoders (AIdonnow, Dominican Papi, Aegis Athena), multimodal fusion (ELiRF-UPV, Aegis Athena, XORIK), and classical classifiers over extracted features (AIdonnow, Aegis Athena, XORIK). This is consistent with the difficulty of short-video sexism detection, where the relevant evidence may be distributed across speech, visible text, frames, and context.

Task 3.2 gives the strongest positive but still qualified signal for sensors. AIdonnow's run with transcript, OCR, video descriptions, and all sensor features leads soft evaluation, while the same team also leads hard evaluation with systems described as transcript+OCR ensembles without sensor features. This within-team contrast suggests that sensors may be helpful for calibrated soft outputs in this subtask, but they are not necessary for top hard-label performance. The best interpretation is therefore protocol-dependent: sensors may add useful uncertainty-related information, but strong textual evidence remains sufficient for the hard setting.

Soft Evaluation Soft-soft evaluation was led by `AIdonnow_1` (AIdonnow), with ICM-Soft Norm=0.4590. The winner had a modest advantage over the second-ranked unique team (1.31 percentage points in ICM-Soft Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Table 10: Top 5 systems from different teams in EXIST 2026 Task 3.2. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
`Test_gold`	–	4.6948	1.0000	0.2550	0
`AIdonnow_1`	AIdonnow	−0.3847	0.4590	1.0930	1
`Dominican-Papi_2`	Dominican-Papi	−0.5080	0.4459	1.2472	2
`XORIK_3`	XORIK	−0.7931	0.4155	0.9988	5
`Aegis-Athena_3`	Aegis-Athena	−0.9847	0.3951	1.0202	7
`nodicus_2`	nodicus	−1.1479	0.3777	1.5681	10
`Test_majority-class`	–	−3.1337	0.1663	4.4354	46
`Test_minority-class`	–	−15.4368	0.0000	8.8286	50
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**Macro F1**	**Rank**
`Test_gold`	–	1.3244	1.0000	1.0000	0
`AIdonnow_2`	AIdonnow	0.3344	0.6262	0.7065	1
`XORIK_3`	XORIK	0.2876	0.6086	0.6966	3
`ELiRF-UPV_1`	ELiRF-UPV	0.2843	0.6073	0.6901	4
`Dominican-Papi_3`	Dominican-Papi	0.2751	0.6038	0.6749	5
`Ontinyent 1931_2`	Ontinyent 1931	0.2554	0.5964	0.6888	8
`Test_majority-class`	–	−0.7537	0.2155	0.2375	69
`Test_minority-class`	–	−2.4749	0.0000	0.0586	70

Under soft-soft evaluation, the leaderboard for the classification of the source intention conveyed by sexist videos included 48 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.0705 to 0.4590, with a mean of 0.295 and a standard deviation of 0.095. 45/48 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 8.13 percentage points, with the largest drop (3.04 percentage points) occurring from the second- to the third-ranked unique-team submission.

Hard Evaluation Hard-hard evaluation was led by `AIdonnow_2` (AIDONNOW), with ICM-Hard Norm=0.6262. The winner had a modest advantage over the second-ranked unique team (1.76 percentage points in ICM-Hard Norm), suggesting a competitive top of the leaderboard. The selected top-five systems were all above the simple majority and minority baselines.

Under hard-hard evaluation, the leaderboard for the classification of the source intention conveyed by sexist videos included 68 submitted systems. The normalized ICM-Hard Norm scores ranged from 0.2196 to 0.6262, with a mean of 0.481 and a standard deviation of 0.095. All submitted systems outperformed the strongest simple baseline (Test_majority-class), indicating that participant submissions were generally effective under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 2.98 percentage points, with the largest drop (1.76 percentage points) occurring from the first- to the second-ranked unique-team submission.

The same team, AIDONNOW, led both soft and hard evaluation, suggesting a robust performance across the two evaluation protocols for this subtask.

6.6 Task 3.3: Sexism Categorization in Videos

Task 3.3 focuses on the categorization of sexist videos. Table 11 presents the results for this task. The following analysis first summarizes the main methodological tendencies observed among the strongest systems and then discusses the results separately for soft and hard evaluation.

Methodologically, the top video systems suggest a strong emphasis on transcript, ASR, OCR, or subtitle signals (AIDONNOW, GERMAN&BLAI, THE BOCADILLOS), LLM or prompt-based processing (AIDONNOW, GERMAN&BLAI, THE BOCADILLOS), transformer text encoders (AIDONNOW, GERMAN&BLAI, AEGIS ATHENA), multimodal fusion (GERMAN&BLAI, THE BOCADILLOS, AEGIS ATHENA), and classical classifiers over extracted features (AIDONNOW, GERMAN&BLAI, AEGIS ATHENA). This is consistent with the difficulty of short-video sexism detection, where the relevant evidence may be distributed across speech, visible text, frames, and context. As with the other subtasks, these observations come from participant descriptions and should be treated as qualitative evidence rather than controlled comparisons.

In Task 3.3, the evidence does not support a sensor advantage. AIdonnow leads hard evaluation with a no-sensor hard-label ensemble, and within its soft submissions the non-sensor variant ranks above the sensor-augmented run. German&Blai includes biometric configurations and remains competitive in hard evaluation, but not enough to show a systematic benefit. For video sexism categorization, the ranking points more clearly to transcript/OCR/ASR, frame-level visual cues, and fusion strategies than to physiological or biometric sensors.

Soft Evaluation Soft-soft evaluation was led by `NeverChorizoInMyPaella_3` (NeverChorizoInMyPaella), with ICM-Soft Norm=0.1724. The leading systems were tightly clustered, with only 0.53 percentage points in ICM-Soft Norm separating the first- and second-ranked unique teams. The selected top-five systems were all above the simple majority and minority baselines.

Table 11: Top 5 systems from different teams in EXIST 2026 Task 3.3. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2026/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Rank
Test_gold	–	8.3833	1.0000	0
NeverChorizoInMyPaella_3	NeverChorizoInMyPaella	−5.4930	0.1724	1
thebocadillos_1	thebocadillos	−5.5814	0.1671	2
nodicus_3	nodicus	−5.5964	0.1662	3
tai6le_3	tai6le	−5.7292	0.1583	6
AIdonnow_2	AIdonnow	−5.8311	0.1522	7
Test_majority-class	–	−6.8222	0.0931	25
Test_minority-class	–	−11.6668	0.0000	41

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	1.5453	1.0000	1.0000	0
AIdonnow_1	AIdonnow	−0.1162	0.4624	0.3750	1
Aegis-Athena_2	Aegis-Athena	−0.1394	0.4549	0.4193	3
German&Blai_3	German&Blai	−0.1743	0.4436	0.4285	4
nodicus_2	nodicus	−0.1886	0.4390	0.4300	6
XORIK_1	XORIK	−0.2662	0.4139	0.3644	8
Test_majority-class	–	−0.9530	0.1916	0.1188	52
Test_minority-class	–	−6.7467	0.0000	0.0025	68

Under soft-soft evaluation, the leaderboard for the categorization of sexist videos included 50 submitted systems. The normalized ICM-Soft Norm scores ranged from 0.0000 to 0.1724, with a mean of 0.078 and a standard deviation of 0.058. 24/50 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 2.02 percentage points, and the selected top-five unique-team submissions were tightly clustered.

Hard Evaluation Hard-hard evaluation was led by `AIdonnow_1` (AIDONNOW), with ICM-Hard Norm=0.4624. The leading systems were tightly clustered, with only 0.75 percentage points in ICM-Hard Norm separating the first- and second-ranked unique teams. The selected top-five systems were all above the simple majority and minority baselines.

Under hard-hard evaluation, the leaderboard for the categorization of sexist videos included 67 submitted systems. The normalized ICM-Hard Norm scores ranged from 0.0000 to 0.4624, with a mean of 0.293 and a standard deviation of 0.145. 51/67 submitted systems outperformed the strongest simple baseline (Test_majority-class), showing a more heterogeneous level of effectiveness under this protocol. The absolute difference between the highest and lowest scores among the selected top-five submissions from different teams was 4.85 percentage points, with a notable drop of 2.51 percentage points from the fourth- to the fifth-ranked unique-team submission.

The leading team changed between evaluation protocols: NEVERCHORIZOINMYPAELLA led soft evaluation, whereas AIDONNOW led hard evaluation. This indicates that optimization for probabilistic outputs and discrete-label performance did not necessarily favor the same system.

7 Conclusions

Overall, EXIST 2026 confirmed that sexism detection in memes and videos remains an active and challenging research area, attracting 1,303 submissions from 122 teams. Participation was substantially higher in the meme track than in the video track, reflecting the increased complexity and computational cost of processing multimodal video data. Across tasks, binary sexism identification proved consistently easier than fine-grained sexism categorization, particularly for videos, where performance remained comparatively low.

The comparison between hard and soft evaluation protocols showed that they capture complementary aspects of system performance. While some teams achieved strong results under both protocols, different winners emerged in most tasks, highlighting the distinction between making accurate discrete predictions and producing well-calibrated probabilistic outputs that better reflect annotator disagreement. These findings reinforce the value of evaluating both hard and soft labels to obtain a more complete assessment of model capabilities.

From a methodological perspective, the strongest systems relied on multimodal approaches combining textual, visual, OCR, transcript, and frame-level information, typically using transformer-based models, vision-language encoders, and multimodal fusion strategies. Although several teams explored physiological and sensor-based signals, these did not provide a consistent advantage over standard multimodal representations. Future work should therefore focus on improving multimodal fusion, uncertainty calibration, and the modeling of annotator disagreement, particularly for fine-grained sexism categorization in videos.

Acknowledgments This work was supported by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31 and PID2024-156022OB-C32)) funded by MICIU/AEI/10.13039/501100011 033 and the European Social Fund Plus (ESF+), and by the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Amigó, E., Delgado, A.: Evaluating Extreme Hierarchical Multi-label Classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. vol. Volume 1: Long Papers, p. 5809–5819. ACL, Dublin, Ireland (2022)

[2] Arcos, I., Rosso, P.: Do physiological signals improve both generalization and personal perception of sexism in memes? In: Lecture Notes in Computer Science. Springer, Cham (2026)

[3] Arcos, I., Rosso, P.: Human-centered multimodal fusion for sexism detection in TikTok videos with eye-tracking, heart rate, and EEG signals. Procesamiento del Lenguaje Natural 77 (2026)

[4] Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We Need to Consider Disagreement in Evaluation. In: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. pp. 15–21. Association for Computational Linguistics, Online (2021)

[5] Divya Venkatesh, J., Jaiswal, A., Nanda, G.: Comparing Human Text Classification Performance and Explainability with Large Language and Machine Learning Models Using Eye-Tracking. Scientific Reports 14(1), 14295 (Jun 2024)

[6] Eurostat: Violence Experienced by Total Population. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Violence_experienced_by_total_population (2025), Last accessed 22 Dec 2025

[7] Gobierno de España: Brecha Digital de Género. https://www.inmujeres.gob.es/publicacioneselectronicas/documentacion/Documentos/DE2110.pdf (2021), Last accessed 22 Dec 2025

[8] Ji, K., Hettiachchi, D., Salim, F.D., Scholer, F., Spina, D.: Characterizing information seeking processes with multiple physiological signals. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1006–1017. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024)

[9] Ji, K., Hettiachchi, D., Scholer, F., Salim, F.D., Spina, D.: Senseseek dataset: Multimodal sensing to study information seeking behaviors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 9(3) (Sep 2025)

[10] Ministerio del Interior: Informe Delitos Contra la Libertad Sexual 2023. https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/dam/jcr:0a219f1b-f2f5-4e95-9553-aa8cf9ee5b3f (2023), Last accessed 22 Dec 2025

[11] National Association of Services Against Sexual Violence: Data on Sexual Violence Services in Australia. https://www.nasasv.org.au/data (2025), Last accessed 22 Dec 2025

[12] Pew Research Center: The State of Online Harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (2021), Last accessed 22 Dec 2025
[13] Plaza, L., Carrillo-de Albornoz, J., Arcos, I., Rosso, P., Spina, D., Amigó, E., Gonzalo, J., Morante, R.: Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos. In: Carrillo-de Albornoz, J., García Seco de Herrera, A., Gonzalo, J., Plaza, L., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025). Madrid, Spain (2025)
[14] Plaza, L., Carrillo-de Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Thessaloniki, Greece (2023)
[15] Plaza, L., Carrillo-de Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., Spina, D.: Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Di Nunzio, G.M., Soulier, L., Galuščáková, P., García Seco de Herrera, A., Faggioli, G., Ferro, N. (eds.) Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024). Grenoble, France (2024)
[16] Plaza, L., Carrillo-de Albornoz, J., Arcos, I., Aloy-Mayo, M., Rosso, P., García-Arias, E., Spina, D.: Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview). CEUR Workshop Proceedings, CEUR Workshop Proceedings (CEUR-WS.org), Jena, Germany (September 2026)
[17] RAINN: Statistics: Victims of Sexual Violence. https://rainn.org/facts-statistics-the-scope-of-the-problem/statistics-victims-of-sexual-violence/ (2025), Last accessed 22 Dec 2025
[18] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: Sexism identification in social networks. Procesamiento del Lenguaje Natural 67, 195–207 (2021)
[19] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Spina, D., Gonzalo, J., Rosso, P.: Overview of EXIST 2022: Sexism identification in social networks. Procesamiento del Lenguaje Natural 69, 229–240 (2022)
[20] Spina, D., Gwizdka, J., Ji, K., Moshfeghi, Y., Mostafa, J., Ruotsalo, T., Zhang, M., Ahmad, A., Lawati, S.F.D.A., Boonprakong, N., Fernando, N., He, J., Hoeber, O., Jayawardena, G., Lee, B.G., Liu, H., Pike, M., Pirmoradi, A., Nakisa, B., Rastgoo, M.N., Salim, F.D., Scott, F., Sun, S., Tang, H., Towey, D., Wilson, M.L.: Report on the 3rd Workshop on NeuroPhysiological Approaches for Interactive Information Retrieval (NeuroPhysIIR 2025) at SIGIR CHIIR 2025. SIGIR Forum 59(1), 1–43 (Oct 2025)