Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview)*

Notebook for the EXIST Lab at CLEF 2026

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

[1]Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
[2]Universidad Politécnica de Valencia (UPV), 46022 Valencia, Spain
[3]Valencian Graduate School and Research Network Analysis of Artificial Analysis (ValgrAI), 46022 Valencia, Spain
[4]RMIT University, 3000 Melbourne, Australia

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, psexism 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 [1, 2, 3, 4]. 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 [5, 6]. 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.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
lplaza@lsi.uned.es (L. Plaza); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); iarcgab@etsinf.upv.es (I. Arcos); malomay@upvnet.upv.es (M. Aloy-Mayo); prosso@dsic.upv.es (P. Rosso); e.garcia@lsi.uned.es (E. García-Arias); damiano.spina@rmit.edu.au (D. Spina)

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 [7, 8, 9]. 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.

The EXIST Lab at CLEF 2026[1] [10, 11] marks 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 [12, 13, 14], while the first two were held in the IberLEF evaluation forum [15, 16]. Over the 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 lab at CLEF 2026 continues 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.[2] 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 [17, 18, 19]. 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

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

[2] 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).

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 continues 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 [20]. 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.

In the following sections, we describe the tasks, dataset, and evaluation methodology adopted in the EXIST 2026 lab at CLEF, as well as the results of the participants.

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

Figure 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)

Figure 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.

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 participants 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, 14]. Moreover, Table 1 summarises the dataset statistics.

(a) Ideological & inequality

(b) Objectification

(c) Stereotyping & dominance

(d) Sexual violence

(e) Misogyny & non-sexual violence

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

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. Multimodal Characterization of Sexist Data

In this section, we analyse the recurrent visual and linguistic patterns that contribute to the construction of sexist meanings in our dataset. Both memes and TikTok videos reflect broader social media narratives in which textual and visual resources interact to create multimodal artefacts that reinforce intentionally sexist messages. Memes typically combine images, embedded text, and emotionally charged emojis, whereas TikTok videos display greater multimodal complexity by integrating visual, textual, and audio resources into a single communicative product.

Focusing on the categories of memes and TikTok videos described in Section 2, we analyse some examples from the results previously presented in [21], where the most frequent category in both the English- and Spanish-language memes is Ideology and Inequality (39.7% of the Spanish memes and 42.3% of the English memes). This category encompasses explicitly sexist messages that promote ideological positions portraying women as inferior to men. Many of these memes ridicule feminist ideology, including feminist slogans, demonstrations, and activists, thereby reinforcing negative stereotypes

associated with the defense of women's rights. A recurrent representational strategy consists of portraying feminist women as lesbians, as women who "behave like men," or as individuals who endorse violence against men. As illustrated in Figure 4, the use of pink and rainbow colours further emphasizes these stereotypical representations. These visual resources interact with the embedded verbal text to intensify the ideological meaning of the memes, which frequently contain explicit statements such as "es lesbiana y se queja de que vivimos en una sociedad machista y falocéntrica" (*she is a lesbian and complains that we live in a sexist and phallocentric society*) or messages that justify or normalise violence against men (Figure 4).

(a) (b) (c)

Figure 4: Examples of memes from the Ideology and Inequality category.

The predominance of ideological content is closely followed by the Stereotyping and Dominance category, which represents 38.8% of the Spanish memes and 36.8% of the English memes. This category is also prominent in the TikTok dataset, accounting for 42.8% of the Spanish videos and 48.1% of the English videos. Across both formats, the multimodal discourse is characterised by visual scenes and verbal arguments that portray women as incapable of performing tasks traditionally associated with men while simultaneously reinforcing the notion of male superiority. Figure 5 illustrates several representative examples. One meme depicts a working-class man wearing a suit and displaying an expression of resignation, implicitly suggesting that women receive lower salaries because they deserve to be paid less. Another recurrent stereotype is that of the young blonde woman portrayed as intellectually incompetent and unable to understand everyday adult situations.

(a) (b) (c)

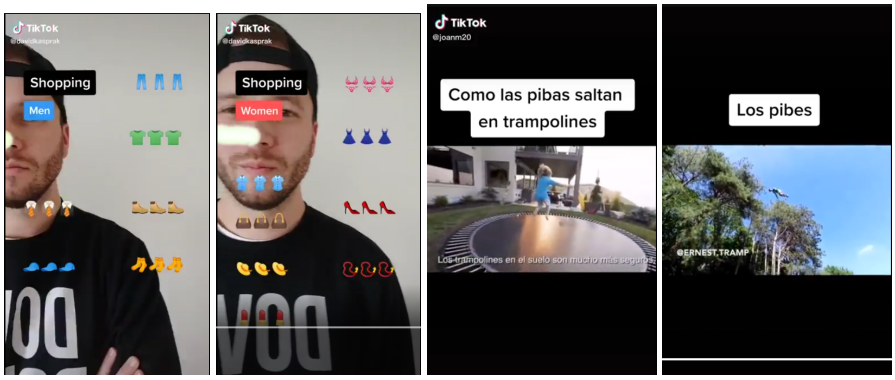
Figure 5: Examples of memes from the Stereotyping and Dominance category.

A particularly revealing example is the third meme (c), in which the semantic content of the image does not directly correspond to that of the accompanying text. The image depicts a scene of violence against a woman, whereas the embedded text explicitly orders her to "go back to the kitchen." Despite this apparent incongruity, the interaction between the verbal and visual modes constructs a coherent ideological message that normalizes male dominance and female subordination.

The representation of sexist attitudes grounded in gender stereotypes is equally prominent in the TikTok corpus. As illustrated in Figure 6, many videos focus on alleged differences between men and women, perpetuating stereotypical gender roles, behaviors, preferences, and social expectations. In

these videos, embedded text, along with emojis and other graphic elements, such as the highlighted colored text, reinforces the interpretation of the visual content and contributes to the construction of strongly ideological gender narratives. The interrelationship between verbal and visual resources is, therefore, central to the meaning-making process, as both modes operate either explicitly or implicitly to amplify the semantics of sexist messages. On the one hand, this is exemplified in frames (a) and (b), where shopping products are employed as symbolic resources to construct and reinforce stereotypical representations of masculine and feminine consumer practices, thereby reproducing conventional gender ideologies. This binary opposition is visually constructed through the use of emojis depicting products conventionally associated with each gender, thereby activating culturally shared assumptions about masculine and feminine consumption practices. The emojis serve as salient semiotic resources that reinforce the ideological contrast established by the visual composition.

On the other hand, frames (c) and (d) present a different type of TikTok video that similarly constructs gender differences. In this case, the narrative revolves around the use of a trampoline. The female in frame (c) is portrayed as a cautious and risk-averse kid, jumping only slightly while accompanied by the embedded text, "Ground-level trampolines are much safer." In contrast, the male in frame (d) is shown performing a considerably higher jump, thereby visually embodying confidence, bravery, and a greater willingness to take risks. The juxtaposition of these two scenes, together with the evaluative role of the embedded text, constructs a clear dichotomous representation of masculinity and femininity. Through this multimodal orchestration, the video reproduces essentialist gender stereotypes by portraying women as more innocent, inherently cautious and less adventurous than men, while simultaneously associating masculinity with courage, risk-taking, and physical confidence.



(a) (b) (c) (d)

Figure 6: Examples of TikToks from the Stereotyping and Dominance category.

Furthermore, the analysis of the embedded text also provides valuable insights into the linguistic representation of sexist discourse. The wordclouds presented in Figure 7 display the most frequent lexical items in the Spanish and English sexist meme corpora. In both languages, the most frequent terms are "women" and "men", highlighting the centrality of gender opposition in the construction of sexist discourse. Both terms are accompanied by words suggesting some kind of violence or words related to the feminist movement. In the Spanish corpus, the most frequent lexical items are “mujeres” (*women*) and “hombres” (*men*), followed by “género” (*gender*), “acoso” (*abuse*), and “violencia” (*violence*), "violación" (*rape*), "feminazis". Similarly, in the English corpus, the most frequent words are “women”, “men”, “girl/girls”, “bitches”, “feminism”, and “sex”.

Taken together, these findings reveal that sexist discourse is systematically organised around the binary opposition between women and men. The high frequency of lexical items referring to both genders, together with the recurrent use of multimodal representations of stereotypes and ideological sexism, demonstrates that antagonistic gender polarisation constitutes one of the principal semiotic mechanisms through which sexist narratives are constructed and legitimised across both memes and TikTok videos.

(a) Lexical items in Spanish memes (b) Lexical items in English memes

Figure 7: Most frequent lexical items in the Spanish and English corpora.

3.2. 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_Channel_0`–`EXG_Channel_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) [22]. We also report results using the normalized version, ICM-Soft Norm. For a description of the ICM-Soft metrics, please refer to [23].
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 [22]. 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-hard and Soft-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*
Task 2.1	Sexism identification in memes	351	100
Task 2.2	Intent classification in memes	293	86
Task 2.3	Multi-label categorization in memes	295	86
Task 3.1	Sexism identification in videos	131	29
Task 3.2	Intent classification in videos	116	27
Task 3.3	Multi-label categorization in videos	117	27

*Unique teams that submitted at least one run.

Table 3 shows a breakdown of 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 Task 2.1, with similar ratios observed across Task 2.2 (84 hard vs 54 soft) and Task 2.3 (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 Task 3.1, while in Task 3.2 and Task 3.3 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-hard)	#Runs (Soft-soft)	#Teams (Hard-hard)	#Teams (Soft-soft)
Task 2.1	212	139	98	66
Task 2.2	181	112	84	54
Task 2.3	182	113	84	53
Task 3.1	73	58	29	24
Task 3.2	68	48	27	20
Task 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 (LLMs) and vision-language models (VLMs) 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-soft evaluation, particularly for intent classification and multi-label categorization, while no significant effects were observed under hard-hard evaluation. LLMs showed limited association with performance under Soft-soft evaluation but were linked to stronger rankings across the three meme tasks under Hard-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-hard evaluation.

Table 4
Most significant correlations between techniques and rankings in meme tasks (Hard-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-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 LLMs. 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. LLMs were linked to improved rankings across all tasks under Hard-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-hard evaluation, Soft-soft evaluation revealed no statistically significant associations for any video task. This suggests that Hard-hard evaluation may provide greater differentiation between competing technical approaches in this benchmark.

Table 5
Most significant correlations between techniques and rankings in video tasks (Hard-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)

5.4. Summary of Approaches per Team

Next we provide a summary of the methodological approaches followed by the EXIST 2026 teams that submitted a description paper for the Working Notes. Unlike previous editions, EXIST 2026 focused exclusively on multimedia formats, comprising six subtasks applied to memes (Task 2) and TikTok videos (Task 3). A groundbreaking feature of this edition was the integration of Human-Centered AI principles, providing physiological and behavioral responses—such as eye tracking, heart rate, and EEG—which many teams explored alongside the Learning with Disagreement (LeWiDi) paradigm and the ICM-Soft metric. We start by the teams that participated only in Task 2 on processing memes.

UNAM_GIL Lab [24] participated exclusively in the meme subtasks (Task 2), evaluating three Qwen2.5-VL vision-language models: a 72B model fine-tuned using QLoRA, together with the 72B and 32B base

models. Their system relies on direct inference with task-specific prompts that simulate reasoning strategies such as a "panel of judges" or the "mirror test," integrating OCR-extracted text with the image and enforcing a JSON-formatted output. Unlike other teams, they deliberately discarded the physiological signals provided in the dataset. Their results reveal a discrepancy between evaluation metrics: the QLoRA fine-tuned model achieved the best official rankings under the ICM-Hard metric (top 10 in Subtask 2.1 and top 11 in Subtask 2.3), whereas the untuned 72B base model consistently outperformed the fine-tuned model in terms of the traditional F1 score, particularly on the English subset.

SafetyNLP [25] proposes an LLM-based (Qwen) OCR preprocessing pipeline to reconstruct meme text while preserving emojis and mentions. They combine text-only classifiers (XLM-R, Twitter-XLM-R, mDeBERTa) with multimodal models (CLIP, SigLIP) through late fusion and cross-attention. They employ a meta-ensemble for binary subtasks and LeWiDi soft-label strategies for multi-label classification.

FourBytes [26] participates in Subtask 2.1 with a dual-encoder architecture that fuses textual representations from XLM-T and visual features from SigLIP via a cross-attention mechanism. The system is trained directly on soft labels derived from annotator votes, although it includes an auxiliary head for physiological variables which was not activated in the final submission.

MRBLACK [27] proposes MemeSense, a multi-task cross-modal reasoning method for sexism detection in the English meme subset. Their approach employs a multimodal LLM using a two-stage Reason-Act prompting process to analyze image-text relations and infer communicative intent, applying schema-based output constraints to enforce hierarchical consistency. Additionally, they integrate physiological signals as auxiliary calibration information, fusing statistical features with the language model predictions at the decision level. Their system achieved competitive results, ranking first in Task 2.3 hard evaluation.

AI Wizards [28] addresses the meme subtasks by modeling them hierarchically as conditional soft-label prediction over empirical annotator distributions. Their system uses fixed Gemini Embedding 2 vision-language representations passed through a lightweight Gated MLP backbone, optimized using KL divergence and homoscedastic uncertainty weighting. Following statistical analysis showing limited EEG separability, physiological signals were excluded. Their submissions ranked first in Task 2.3 and fourth in Tasks 2.1 and 2.2 on the official Soft-Soft evaluation leaderboards.

SGF-Multimodal [29] participates in Task 2.1 with a framework that enhances input semantics and incorporates physiological signals. They use Qwen2.5-VL to generate visual descriptions, concatenated with OCR text and encoded using XLM-RoBERTa, alongside ViT image features. They propose a Sensor Gated Fusion (SGF) mechanism that uses aggregated physiological features to generate a gating vector, performing dimension-wise re-weighting on the joint text-image representation.

alBERT-319 [30] participates in Task 2.1 with an incremental approach designed to quantify the impact of physiological signals on classifier calibration. They propose a text-only baseline, a multimodal Qwen3-VL classifier, and a third system that adds a dynamic temperature scaling module. This module uses an MLP to compute a per-meme temperature parameter from physiological features. Their results show that physiological calibration improves hard evaluation by correcting pessimistic predictions, though it slightly degrades soft evaluation.

UNED-Annotate-Team [31] addresses Task 2.1 using a multimodal framework based on a Transformer Mean Pooling with Multi-Head Attention architecture. They enrich textual representations via early fusion with 13 visual features extracted using CLIP. To address annotator disagreement within the LeWiDi paradigm, they compare auxiliary multi-task learning heads against a contextual soft-token embedding strategy (SmartTokens) that injects sociodemographic information directly into the input layer.

Cloud-17 [32] participates in Task 2 with a multimodal architecture integrating text (mDeBERTa-v3), image (ALIGN), EEG, and eye-tracking signals. They introduce a Gated Fusion architecture where learnable scalar gates dynamically modulate the contribution of each non-textual modality. Their ablation reveals a stable modality hierarchy: image dominates, EEG provides a complementary neural signal, and eye-tracking is largely suppressed as redundant.

BioSentinel [33] addresses Sub-task 2.2 with a text-centric approach based on xlm-roberta-base, excluding physiological sensor signals due to a train-test feature mismatch. To handle the LeWiDi paradigm, they train using a composite loss combining KL divergence on soft distributions with weighted cross-entropy on hard labels. They introduce a non-differentiable hierarchy-aware penalty term for checkpoint selection and apply temperature scaling during inference to sharpen output distributions.

CYUT-Giantsshoulders_1 [34] participates in Tasks 2.1 and 2.2 leveraging GPT-4o for multimodal semantic augmentation. These generated features are concatenated with original text and processed using a language-aware weighted ensemble strategy, combining RoBERTa-large and XLM-RoBERTa-large for English, and XLM-RoBERTa-large with BERTIN for Spanish. They employ multi-seed probability averaging and hierarchical inference rules.

DDAI-UNICC [35] participates in Task 2 with a sensor-enhanced, feedback-driven prompt optimization framework. The system iteratively refines prompts using an Evaluator, Critic, Summarizer, and Rewriter. A distinguishing feature is the incorporation of physiological sensor data, encoded via a triplet-loss neural encoder, clustered, and injected as centroid-distance scores into the optimized prompts. They achieve first place in Task 2.1 across all official rankings and Task 2.3 in the global and Spanish rankings.

Bicocca_Unimib [36] participates in Task 2 with a human-centered multimodal framework integrating textual, visual, and physiological signals. They compare a Late Fusion architecture against a Cross-Attention model where physiological signals query the joint text-image representation. Both models are trained on soft labels using Automatic Weighted Loss for multi-task learning within the LeWiDi paradigm.

Ordantis [37] addresses Task 2 with the Gemini-Enriched Multimodal Fusion (GEMF) architecture, employing Gemini Flash 3.1 as a semantic mediator to translate visual content into structured text. This enriched text is concatenated with OCR and encoded using XLM-RoBERTa-base, while EEG signals are integrated via Set Attention Pooling. The models are trained on empirical annotator distributions, utilizing task-specific heads and exploring probability blending with Gemini's zero-shot outputs.

MathAddicts [38] participates in Task 2 with a heterogeneous ensemble of three Vision-Language Models (Qwen 3.5 4B, GLM-4.6V-Flash, and Qwen 3.5 9B). All models are fine-tuned using 4-bit quantized parameter-efficient methods to ensure diversity. For multi-label Subtask 2.3, they employ an "at-least-one" voting strategy, which significantly outperformed majority voting by preserving valid but infrequently predicted categories.

OCR-PragSex [39] participates in Sub-task 2.1 with a deliberately text-only approach to evaluate the limits of unimodal modeling. They submit a TF-IDF logistic regression classifier and a fine-tuned XLM-RoBERTa-base. Contrary to expectations, the classical TF-IDF system proved significantly more stable, and the team provides a detailed pragmatic error analysis identifying key failure categories such as visual dependence and implicit sexism.

Farabi University [40] participates in Sub-task 2.1 with a CLIP ViT-B/32 multimodal classification system. Their approach concatenates text and image embeddings passed through an MLP classification head. The authors report moderate results and a notable bias toward the negative class, identifying the lack of multilingual capacity in CLIP's text encoder for Spanish content as a primary limitation.

SPECTRA [41] participates in Sub-task 2.1 with a multimodal fusion framework integrating textual, visual, and physiological information. They introduce an asymmetric cross-attention mechanism where the textual representation acts as the primary semantic anchor. Their results demonstrate that text remains the dominant information source; the best-performing configuration relied on text, image, and caption features, while physiological signals did not consistently improve performance.

Semantica [42] participates in Sub-task 2.3 with a two-stage multimodal pipeline fusing XLM-RoBERTa with CLIP. A distinctive component is a handcrafted enriched bilingual lexicon providing 28-dimensional features, including category-specific keywords and regex-based syntactic patterns. Their ablation studies highlight that these lexicon features were essential for performance, particularly for detecting rare categories.

Aryan [43] participates in the three meme subtasks with a fully local, zero-shot pipeline. They propose progressively richer configurations using quantized LLMs with Chain-of-Thought reasoning,

and a novel framework using Gemma-4B that integrates physiological sensor fusion to post-process soft confidence scores using uncertainty-weighted arousal scores. Physiological sensor fusion achieved the best confidence alignment on identification and categorization.

VANGUARD [44] participates in Task 2 addressing all meme subtasks with a human-centered multimodal framework incorporating demographic and physiological characteristics. They fuse five input modalities through a cross-attention architecture conditioned via FiLM, and employ cross-lingual data augmentation using NLLB-200. A rigorous ablation revealed that no single architectural component is independently statistically significant, suggesting performance arises from the integration of all parts.

Sudharshan PS [45] participates in Sub-task 2.1 with a multimodal ensemble framework combining SentenceTransformer text embeddings, CLIP image embeddings, and physiological statistical descriptors. The fused representations are processed by an ensemble of traditional machine learning classifiers combined through weighted averaging, employing stratified cross-validation and class-aware learning strategies to handle subjectivity.

Next we present the approaches of the teams that participated only in the TikTok tasks (Tasks 3.X).

AIdonnow [46] approaches the TikTok tasks by converting videos into rich text representations. Their system employs a three-stage audio cleaning cascade followed by Whisper-large-v3, and a visual branch utilizing Qwen3.5 for OCR and scene descriptions. They compare classical TF-IDF classifiers with XLM-RoBERTa and Qwen3-8B via QLoRA. They find that multimodal text reconstruction is beneficial, whereas physiological features generally degrade performance. Their systems ranked first in Task 3.2 hard and soft evaluation, and first in Task 3.3 hard evaluation.

NeverChorizoInMyPaella [47] participates in the TikTok tasks with a multimodal framework based on Twitter-XLM-RoBERTa. They implement a Sequential Cascade Training pipeline to exploit hierarchical dependencies. Their system fuses OCR text with confidence-filtered ASR transcripts alongside SigLIP visual frames, employing a Visual Cross-Attention mechanism and integrating physiological signals via a shallow MLP. Their approach ranked 1st overall in Task 3.3 under Soft-Soft evaluation.

UniKo [48] participates in Task 3 on TikTok videos integrating physiological sensor signals as a dedicated fourth modality alongside text, audio, and visual frames. They encode per-annotator sensor responses to explicitly model inter-annotator disagreement. The system employs XLM-RoBERTa-large, wav2vec2, and CLIP ViT-L/14, fused through a Gated Cross-Modal Attention module. To address LeWiDi, they construct gender-weighted soft labels and utilize a combined loss function including agreement-weighted KL divergence and soft-label mixup.

UMUTeam [49] participates in Tasks 3.1 and 3.2 focusing on sexism detection in TikTok videos using a Stacking-based Late Fusion framework. They employ TF-IDF with character n-grams for text, MFCCs for audio, and frozen SigLIP aggregated temporally using an LSTM for video frames. Due to being trained on a reduced data subset and excluding physiological signals, the system obtained modest results, though it consistently outperformed the official baselines.

Aegis Athena [50] participates in the three TikTok video subtasks with a multimodal architecture fusing four streams through symmetric cross-attention: VLM scene descriptions, ASR transcripts, raw video features, and physiological signals. They train under the LeWiDi paradigm using soft labels augmented with Dirichlet and Beta perturbations. Their four-modality configuration ranks 4th on Task 3.1 and 3rd on Task 3.3 in the all-language hard evaluation, demonstrating that physiological signals provide consistent gains over strong content-based modalities.

A few teams participated on both meme and TikTok tasks (Tasks 2 and 3).

Team Marsad [51] participates in Tasks 2.1 and 3.1 using CLIP-based multimodal architectures. For memes, they evaluate models ranging from TF-IDF baselines to end-to-end fine-tuned CLIP ViT-L/14. For videos, they extract five frames per instance, encode them using CLIP with average pooling, and fuse them with text embeddings. Their results show that fine-tuned CLIP consistently outperforms baselines, though they observe significant performance degradation in Spanish compared to English.

AILS-NTUA [52] addresses both the meme and TikTok video tasks with a feature-engineered late-fusion pipeline built on gradient-boosted regression models. For memes, their system combines CLIP image embeddings, XLM-RoBERTa encodings, and higher-level semantic indicators extracted via LLM prompting. They implement hierarchical post-processing where the binary sexism prediction acts as a

probabilistic gatekeeper for downstream tasks.

Jow [53] addresses both Task 2 and Task 3 using deliberately lightweight, sub-billion-parameter architectures trained directly on soft label distributions. For memes, their system fuses XLM-RoBERTa-large with a mostly frozen CLIP ViT-Large/14, uniquely injecting annotator demographic count vectors into the fused representation. For videos, they incorporate VideoMAE-base adapted via LoRA, Whisper audio features, and biometric signals. Despite avoiding large-scale compute, their approach achieves highly competitive results, notably ranking 4th on Task 2.3.

ELiRF-UPV [54] participates in Tasks 2 and 3 with two complementary system families. The first is a trained multimodal architecture (PhysioMeme-Fusion) encoding enriched textual views and physiological branches via residual cross-attention. The second is a training-free, hierarchical-conditional few-shot pipeline prompting Qwen3.5-27B and Gemma-4-31B. Their results show that the training-free few-shot pipeline proves highly effective, achieving first place on video identification (T3.1) and second on meme categorization (T2.3) in the hard-hard setting.

Tai6le [55] participates in Subtasks 2.1, 2.3, 3.1, and 3.3 with systems built around XLM-RoBERTa text-only pipelines. They explore caption-augmented inputs, an XLM-RoBERTa-large plus CLIP late-fusion architecture, and calibrated variants applying threshold tuning for hard submissions and probability calibration for soft submissions optimized against ICM metrics. Their strongest results are achieved in the soft-soft categorization setting, ranking 6th in both Subtask 2.3 and Subtask 3.3, indicating that calibrated text-centered models can effectively capture annotator disagreement signals.

Finally, a few student teams did not submit a notebook paper, but provided additional information about their competitive runs.

Dream Team participates in Sbutasks 2.1, 2.2, and 2.3. Qwen 3.5 first generates rich English descriptions of each meme, and an XLM-RoBERTa-large model is then fine-tuned on the concatenation of these descriptions and the OCR-extracted text, using class-weighted losses and hierarchical inference.

Os Faixa Preta do Data participates in Subtasks 2.1, 2.2, and 2.3. Their proposed approach uses the OCR text of each meme enriched with a cached VLM representation. They concatenate the OCR text with a VLM-generated visual description and an explicit sexism-oriented analysis of the meme. This enriched text is then represented with word and character TF-IDF features. For Tasks 2.1 and 2.2 they use Logistic Regression with balanced class weights, while One-vs-Rest Logistic Regression is used for Task 2.3. Soft outputs correspond to the calibrated/probabilistic model outputs, while Hard outputs are obtained by argmax for Tasks 2.1 and 2.2, and by thresholding the multilabel probabilities for Task 2.3.

Legleye_y_clara participats in Sbutasks 2.1, 2.2, and 2.3. Their system is a text-hybrid model that combines cleaned meme text with a language tag, TF-IDF word n-grams, and multilingual MPNet sentence embeddings through feature concatenation. A balanced Logistic Regression classifier is used for Tasks 2.1 and 2.2, and a One-vs-Rest balanced Logistic Regression with a calibrated multilabel threshold for Task 2.3.

German&Blai participates in Subtasks 3.1, 3.2, and 3.3. Their runs are based on different combinations of ensembles: Qwen3-VL-Embeddings 8B, run both with and without physiological data, Reasoning-Aware Multimodal Fusion (RAMF) [56], and an encoder + Random Forest model over physiological signals that additionally uses the tasks' descriptions.

Dominican Papi particpates in Subtasks 3.1, 3.2, and 3.3. Their runs are based on multimodal late fusion using XLM-RoBERTa as the main textual backbone. The original post text is extended with audio transcripts extrcted using OpenAI's Whisper automatic speech recongizrs, and structured video summaries generated using Google DeepMind's Gemma – summaries were designed to capture gender dynamics, tone, intention, sociological tropes and other audiovisual cues that may not be explicit in the raw text. The textual representations were then fused with task-specific physiological features. The submitted runs explored different fusion approaches: (Run 1) single-query late fusion module; (Run 2) a set of learned fusion tokens instead of a single fusion query; and (Run 3) adds a stream-interaction layer, using a encoder, before the final fusion stage.

XORIK participates in Subtasks 3.1., 3.2, and 3.3. Their approach combines visual and textual information extracted from each video. A VLM (Qwen3.6 27B) first generates a textual description of the video content, which is encoded with XLM-RoBERTa Large to obtain both a global CLS representation and

contextual token embeddings. In parallel, the original video is processed with Qwen3-VL-Embeddings 8B to produce a single multimodal video embedding. This embedding serves as the query in a Transformer encoder layer, with the XLM-RoBERTa token embeddings acting as keys and values, allowing the visual representation to attend to relevant textual information. The contextualized video embedding is finally concatenated with the CLS token and passed to a linear classifier to produce the prediction.

Nodicus participates in Subtasks 3.1, 3.2, and 3.3. Their systems are based on mDeBERTa-v3-base and XLM-RoBERTa-large trained with an Annotator-KL loss, and complemented with Qwen2.5-VL-generated (stance-aware) captions. For some of the runs, a selection of the 12 most discriminative physiological features is applied.

6. Results

In the next subsections, we report the official all-language results of the participant submissions and the reference baseline systems for the EXIST 2026 multimedia tasks. We consider both evaluation protocols described in Section 4: soft-soft evaluation, which rewards systems that model annotator disagreement through probability distributions, and hard-hard evaluation, which evaluates discrete decisions. For each subtask, the complete leaderboard is reported in a dedicated long table. All leaderboards can also be consulted on the EXIST website: https://nlp.uned.es/exist2026/. The discussion combines these quantitative results with the methodological evidence available in the submitted working notes; when a high-ranked team did not submit a Working Note, we only use the shorter system description submitted with the runs and avoid attributing unsupported design choices to individual runs.

Overall, the results show that the binary identification subtasks are substantially easier than the fine-grained categorization subtasks. The best normalized ICM scores are obtained in Task 2.1 and Task 3.1, while Task 2.3 and especially Task 3.3 have lower best normalized ICM scores and many more systems close to, or below, the strongest simple baseline. The comparison between soft and hard evaluation also shows that the two protocols do not always reward the same behavior: source-intention classification in memes and videos is led by the same team under both protocols, whereas the winning teams change for sexism identification and categorization.

6.1. Task 2.1: Sexism Identification in Memes

Task 2.1 focuses on the binary classification of memes as sexist or not sexist. This is the most competitive meme subtask in absolute terms: the leading systems obtain the highest normalized ICM scores in Task 2, and most submissions clearly outperform the majority-class baseline.

6.1.1. Soft Evaluation

Table 6 presents the complete soft-soft evaluation results for Task 2.1. A total of 139 participant runs were submitted, in addition to the three reference systems. All participant runs outperformed the strongest non-informative baseline, the majority-class system, whose ICM-Soft Norm was 0.1212. The participant ICM-Soft Norm scores ranged from 0.6158 to 0.1744, with a mean of 0.407 and a standard deviation of 0.073.

Table 6
Complete leaderboard for EXIST 2026 Task 2.1, Sexism Identification in Memes, under the soft-soft evaluation protocol. The table reports all 142 evaluated runs, including 139 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	3.1107	1.0000	0.5852
Ordantis_1	1	0.7206	0.6158	0.8155
Ordantis_3	2	0.5175	0.5832	0.8227
Ordantis_2	3	0.5099	0.5820	0.8146
aiwizards_3	4	0.2323	0.5373	0.9090
aiwizards_1	5	0.2039	0.5328	0.8938
aiwizards_2	6	0.1846	0.5297	0.9118
alBERT-319_2	7	0.0838	0.5135	0.8479
DreamTeam_1	8	0.0689	0.5111	1.3606
alBERT-319_3	9	0.0334	0.5054	0.8485
Jow_1	10	0.0108	0.5017	0.9005
ROBERTA'S-SONS_2	11	−0.0258	0.4959	1.4023
ELiRF-UPV_2	12	−0.0340	0.4945	0.9125
ELiRF-UPV_1	13	−0.0359	0.4942	0.9065
SerranoTeam_1	14	−0.0801	0.4871	0.8969
Jow_2	15	−0.1061	0.4829	0.9558
Jow_3	16	−0.1118	0.4820	0.9526
Marsad_1	17	−0.1156	0.4814	0.9462
SafetyNLP_3	18	−0.1295	0.4792	0.9281
NaiveNeuron_2	19	−0.1557	0.4750	0.9665
Elliot Thorel_1	20	−0.1983	0.4681	1.0173
NaiveNeuron_1	21	−0.2048	0.4671	0.9730
alBERT-319_1	22	−0.2127	0.4658	0.9124
CJKTeam_1	23	−0.2230	0.4642	0.9932
Marsad_2	24	−0.2374	0.4618	0.9738
JBG_1	25	−0.2592	0.4583	0.9316
jell_3	26	−0.2617	0.4579	0.8962
Team1_1	27	−0.2659	0.4573	1.1618
KlaudiaBanasiewicz_1	28	−0.2707	0.4565	1.0213
BicoccaUnimib_2	29	−0.2804	0.4549	0.9078
Os-Faixa-Preta-do-Data_1	30	−0.2849	0.4542	0.8777
BicoccaUnimib_1	31	−0.2925	0.4530	0.8937
tai6le_1	32	−0.2956	0.4525	0.9275
SPSPH4572_2	33	−0.3042	0.4511	0.9356
BicoccaUnimib_3	34	−0.3043	0.4511	0.9336
CJKTeam_2	35	−0.3123	0.4498	1.0866
BraveNewWave_1	36	−0.3137	0.4496	0.9281
TokenTwins_1	37	−0.3241	0.4479	1.0041
SafetyNLP_2	38	−0.3292	0.4471	0.9253
EQUIPOACTIMEL_1	39	−0.3538	0.4431	1.0194
pgnamia-sfonpia-UPV_2	40	−0.3544	0.4430	2.0407
Legleye-y-Clara_1	41	−0.3619	0.4418	0.9198

Continued on next page

Table 6 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
SafetyNLP_1	42	−0.3696	0.4406	0.9647
MB1-UPV_1	43	−0.3925	0.4369	0.9315
JogaBonito_1	44	−0.3983	0.4360	0.9390
EQUIPOACTIMEL_2	45	−0.3985	0.4360	0.9142
JogaBonito_3	46	−0.4036	0.4351	0.9581
Senyu&Dani_1	47	−0.4046	0.4350	0.9522
JogaBonito_2	48	−0.4138	0.4335	0.9309
SPECTRA_1	49	−0.4185	0.4327	0.9702
Marsad_3	50	−0.4212	0.4323	0.9137
Cloud-17_2	51	−0.4301	0.4309	1.1128
EQUIPOACTIMEL_3	52	−0.4366	0.4298	0.9218
AndreayEmma_3	53	−0.4367	0.4298	0.9297
AndreayEmma_2	54	−0.4367	0.4298	0.9297
AndreayEmma_1	55	−0.4367	0.4298	0.9297
FourBytes_1	56	−0.4376	0.4297	0.9524
tai6le_3	57	−0.4415	0.4290	0.9042
teamJorge_1	58	−0.4448	0.4285	0.9362
AILS-NTUA_2	59	−0.4625	0.4257	0.9219
AILS-NTUA_1	60	−0.4675	0.4249	0.9282
AILS-NTUA_3	61	−0.4753	0.4236	0.9223
Legleye-y-Clara_3	62	−0.4805	0.4228	0.9959
teamJorge_2	63	−0.4828	0.4224	1.0094
jell_1	64	−0.4878	0.4216	0.9494
NACHUGO_1	65	−0.4913	0.4210	0.9986
NACHUGO_2	66	−0.5017	0.4194	0.9297
Marc&Pablo_1	67	−0.5049	0.4188	0.9550
SesgoCero_1	68	−0.5059	0.4187	0.9721
lonquets_1	69	−0.5083	0.4183	1.0226
pgnamia-sfonpia-UPV_1	70	−0.5152	0.4172	3.8940
Cloud-17_3	71	−0.5205	0.4163	1.1274
pgnamia-sfonpia-UPV_3	72	−0.5208	0.4163	2.7932
Castrillo_1	73	−0.5212	0.4162	1.2427
Cloud-17_1	74	−0.5363	0.4138	1.2091
VANGUARD_1	75	−0.5545	0.4109	0.9389
andreolo_1	76	−0.5968	0.4041	0.9524
Os-Faixa-Preta-do-Data_2	77	−0.6040	0.4029	0.9131
thebocadillos_1	78	−0.6157	0.4010	0.9962
Transformers_1	79	−0.6289	0.3989	0.9436
berlin_1	80	−0.6299	0.3987	1.3802
mandatory_1	81	−0.6420	0.3968	0.9806
CEVS_2	82	−0.6421	0.3968	1.3308
ANALYTICALGRAMMAR_2	83	−0.6424	0.3967	0.9531
JBG_2	84	−0.6434	0.3966	0.9604
ANALYTICALGRAMMAR_1	85	−0.6434	0.3966	0.9540
CEVS_1	86	−0.6455	0.3962	0.9537
lonquets_2	87	−0.6472	0.3960	1.1157
Legleye-y-Clara_2	88	−0.6473	0.3960	0.9462

Continued on next page

Table 6 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
maranda24_1	89	−0.6549	0.3947	1.0308
teamJorge_3	90	−0.6648	0.3931	0.9524
DuoDinamita_1	91	−0.7074	0.3863	0.9581
jell_2	92	−0.7079	0.3862	0.9310
Os-Faixa-Preta-do-Data_3	93	−0.7088	0.3861	0.9627
ELiRF-UPV_3	94	−0.7100	0.3859	1.0703
SPECTRA_2	95	−0.7218	0.3840	1.0303
lonquets_3	96	−0.7324	0.3823	1.0061
TheAntisexists_1	97	−0.7330	0.3822	0.9779
Vicente&Lorenzo_1	98	−0.7364	0.3816	0.9481
Los yets_1	99	−0.7389	0.3812	1.0088
APARAJITA-NITA-JH_1	100	−0.7461	0.3801	1.4517
AryanSomnathBanerjee_2	101	−0.7501	0.3794	1.4397
I2C-UHU-PERSEUS_1	102	−0.7513	0.3792	0.9291
AryanSomnathBanerjee_1	103	−0.7714	0.3760	1.4447
Marc&Pablo_3	104	−0.7866	0.3736	0.9527
Los yets_3	105	−0.7872	0.3735	0.9665
Meme-Hunters-Team_2	106	−0.7935	0.3724	1.2769
VANGUARD_3	107	−0.8020	0.3711	0.9512
run_1	108	−0.8194	0.3683	0.9400
ANALYTICALGRAMMAR_3	109	−0.8252	0.3674	0.9426
C-TEA_1	110	−0.8304	0.3665	0.9809
Meme-Hunters-Team_1	111	−0.8382	0.3653	1.3066
Meme-Hunters-Team_3	112	−0.8407	0.3649	1.2525
tai6le_2	113	−0.8586	0.3620	0.9470
AryanSomnathBanerjee_3	114	−0.8753	0.3593	1.3139
SPECTRA_3	115	−0.8861	0.3576	1.2022
NadalBQ_1	116	−0.9101	0.3537	1.1499
VANGUARD_2	117	−0.9177	0.3525	0.9704
TheAntisexists_3	118	−1.0002	0.3392	1.1580
Vicente&Lorenzo_2	119	−1.0823	0.3260	0.9730
iEXIST-UPV_2	120	−1.0898	0.3248	0.9842
iEXIST-UPV_3	121	−1.0898	0.3248	0.9842
iEXIST-UPV_1	122	−1.0898	0.3248	0.9842
Vicente&Lorenzo_3	123	−1.1591	0.3137	0.9846
Marc&Pablo_2	124	−1.1672	0.3124	0.9836
Los yets_2	125	−1.2177	0.3043	0.9984
FairMemers_1	126	−1.2203	0.3038	1.0000
PolskaGurom_1	127	−1.2782	0.2946	1.0242
TokenTwins_3	128	−1.3103	0.2894	1.9552
TheAntisexists_2	129	−1.3279	0.2866	1.0283
CEVS_3	130	−1.3887	0.2768	1.3800
berlin_2	131	−1.4461	0.2676	1.7005
Ctrl-Alt-Supr_2	132	−1.4770	0.2626	1.1765
Ctrl-Alt-Supr_3	133	−1.4770	0.2626	1.1765
Ctrl-Alt-Supr_1	134	−1.4770	0.2626	1.1765
MRBLACK_1	135	−1.5221	0.2553	3.3020

Continued on next page

Table 6 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
UNED-Annotate-Team_3	136	−1.5754	0.2468	1.0896
UNED-Annotate-Team_2	137	−1.6743	0.2309	1.1178
UNED-Annotate-Team_1	138	−1.6957	0.2274	1.1127
TokenTwins_2	139	−2.0258	0.1744	2.8539
Test_majority-class	140	−2.3568	0.1212	4.4015
Test_minority-class	141	−3.5089	0.0000	5.5672

The leading run was Ordantis_1, with ICM-Soft Norm=0.6158. The first position was relatively well separated from the rest of the unique-team top group: the difference in ICM-Soft Norm between the best and fifth-best unique-team systems was 11.41 percentage points, and most of that gap came from the distance between Ordantis and the next unique team. This pattern suggests that, for soft binary meme identification, probability calibration and the modelling of annotator distributions mattered in addition to simply choosing the correct hard label.

6.1.2. Hard Evaluation

Table 7 reports the hard-hard evaluation results for the same subtask. The hard setting received 212 participant runs. Of these, 201 outperformed the strongest simple baseline, again the majority-class system, whose ICM-Hard Norm was 0.2947. Participant ICM-Hard Norm scores ranged from 0.7387 to 0.1711, with a mean of 0.488 and a standard deviation of 0.100.

Table 7
Complete leaderboard for EXIST 2026 Task 2.1, Sexism Identification in Memes, under the hard-hard evaluation protocol. The table reports all 215 evaluated runs, including 212 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and F1 YES. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
Test_gold	0	0.9832	1.0000	1.0000
DDAIUNICC_1	1	0.4693	0.7387	0.8076
DDAIUNICC_2	2	0.4212	0.7142	0.7911
Ordantis_2	3	0.4079	0.7074	0.8006
Ordantis_1	4	0.4005	0.7037	0.8003
DDAIUNICC_3	5	0.3990	0.7029	0.7921
Ordantis_3	6	0.3943	0.7005	0.7969
DreamTeam_1	7	0.3542	0.6801	0.7886
ELiRF-UPV_3	8	0.3274	0.6665	0.7790
alBERT-319_3	9	0.3134	0.6594	0.7571
Math Addicts_1	10	0.3099	0.6576	0.7514
CYUT-Giantsshoulders_1	11	0.3003	0.6527	0.7590
alBERT-319_2	12	0.2870	0.6459	0.7488
GIL UNAM_2	13	0.2849	0.6449	0.7580
aiwizards_3	14	0.2673	0.6359	0.7509
GIL UNAM_1	15	0.2650	0.6348	0.7671

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
ELiRF-UPV_2	16	0.2598	0.6321	0.7763
ROBERTA'S-SONS_2	17	0.2316	0.6178	0.7596
aiwizards_1	18	0.2208	0.6123	0.7391
aiwizards_2	19	0.2188	0.6112	0.7618
Jow_1	20	0.2143	0.6090	0.7403
SerranoTeam_1	21	0.2053	0.6044	0.7389
PAULU_1	22	0.1873	0.5952	0.7348
pgnamia-sfonpia-UPV_3	23	0.1855	0.5943	0.7329
tai6le_3	24	0.1807	0.5919	0.7279
ELiRF-UPV_1	25	0.1777	0.5904	0.7423
Os-Faixa-Preta-do-Data_2	26	0.1728	0.5879	0.7352
datovalenciano_1	27	0.1716	0.5872	0.7366
pgnamia-sfonpia-UPV_1	28	0.1619	0.5824	0.7294
jell_3	29	0.1550	0.5788	0.7257
Os-Faixa-Preta-do-Data_1	30	0.1549	0.5788	0.7325
pgnamia-sfonpia-UPV_2	31	0.1451	0.5738	0.7101
BicoccaUnimib_1	32	0.1430	0.5727	0.7271
MICCAR_2	33	0.1354	0.5689	0.7256
BicoccaUnimib_2	34	0.1317	0.5669	0.7280
Team1_1	35	0.1316	0.5669	0.7014
SafetyNLP_3	36	0.1298	0.5660	0.7310
JBG_1	37	0.1278	0.5650	0.7222
EQUIPOACTIMEL_2	38	0.1252	0.5637	0.7252
EQUIPOACTIMEL_3	39	0.1221	0.5621	0.7296
NaiveNeuron_1	40	0.1219	0.5620	0.7325
Marsad_3	41	0.1218	0.5620	0.7219
EQUIPOACTIMEL_1	42	0.1217	0.5619	0.7240
UNED-Annotate-Team_2	43	0.1212	0.5616	0.7020
MemeMiner_1	44	0.1210	0.5615	0.6981
alBERT-319_1	45	0.1206	0.5613	0.7033
Legleye-y-Clara_1	46	0.1195	0.5608	0.7085
Marsad_1	47	0.1194	0.5607	0.6876
SerranoTeam_2	48	0.1173	0.5596	0.7183
Lidia_1	49	0.1142	0.5581	0.7138
CoEXIST_1	50	0.1062	0.5540	0.7154
Jow_3	51	0.1031	0.5524	0.7102
CJKTeam_1	52	0.1017	0.5517	0.7129
Jow_2	53	0.1013	0.5515	0.7052
MB1-UPV_1	54	0.0929	0.5472	0.7088
TokenTwins_1	55	0.0910	0.5463	0.7267
SPSPH4572_1	56	0.0890	0.5453	0.7210
SafetyNLP_2	57	0.0879	0.5447	0.7122
BraveNewWave_1	58	0.0837	0.5426	0.7001
CJKTeam_2	59	0.0835	0.5425	0.7155
KlaudiaBanasiewicz_1	60	0.0819	0.5417	0.7004
PAULU_2	61	0.0812	0.5413	0.7064
NaiveNeuron_2	62	0.0744	0.5378	0.6922

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
BicoccaUnimib_3	63	0.0700	0.5356	0.6903
Lorena_3	64	0.0660	0.5335	0.6865
GIL UNAM_3	65	0.0612	0.5311	0.6794
Marsad_2	66	0.0611	0.5311	0.7153
AILS-NTUA_3	67	0.0605	0.5308	0.7089
tai6le_1	68	0.0591	0.5301	0.6945
Daien and Camilla_1	69	0.0545	0.5277	0.7050
Senyu&Dani_1	70	0.0530	0.5269	0.6889
AILS-NTUA_2	71	0.0513	0.5261	0.7078
teamJorge_1	72	0.0413	0.5210	0.6945
CoEXIST_2	73	0.0413	0.5210	0.6963
CLUPV_1	74	0.0386	0.5197	0.7000
AILS-NTUA_1	75	0.0354	0.5180	0.7011
LNR-Erasmus_1	76	0.0348	0.5177	0.6994
Atipicas_3	77	0.0335	0.5170	0.6912
HECTOR-ALONSO_3	78	0.0334	0.5170	0.6974
ParabaTeam_1	79	0.0301	0.5153	0.6692
Women_3	80	0.0297	0.5151	0.7207
berlin_1	81	0.0284	0.5145	0.7298
Lorena_1	82	0.0256	0.5130	0.6978
UNED-Annotate-Team_1	83	0.0244	0.5124	0.6863
CoEXIST_3	84	0.0240	0.5122	0.6939
Lorena_2	85	0.0233	0.5119	0.6753
lonquets_1	86	0.0213	0.5108	0.7039
Castrillo_1	87	0.0191	0.5097	0.6776
AryanSomnathBanerjee_1	88	0.0180	0.5092	0.6728
AryanSomnathBanerjee_2	89	0.0180	0.5092	0.6728
OCRPragSex_1	90	0.0158	0.5080	0.6864
JogaBonito_2	91	0.0139	0.5071	0.7113
HECTOR-ALONSO_2	92	0.0114	0.5058	0.6910
Los yets_1	93	0.0093	0.5047	0.6679
Women_1	94	0.0091	0.5046	0.7098
M&M_3	95	0.0088	0.5045	0.6816
JogaBonito_3	96	0.0085	0.5043	0.7126
HECTOR-ALONSO_1	97	0.0073	0.5037	0.6941
Transformers_1	98	0.0066	0.5034	0.6616
I2C-UHU-PERSEUS_1	99	0.0055	0.5028	0.7229
JogaBonito_1	100	0.0008	0.5004	0.7046
Aurora_1	101	0.0005	0.5003	0.7045
Cloud-17_2	102	−0.0092	0.4953	0.7145
SesgoCero_1	103	−0.0094	0.4952	0.6833
Marc&Pablo_1	104	−0.0103	0.4948	0.6685
Cloud-17_3	105	−0.0104	0.4947	0.7145
jell_1	106	−0.0111	0.4944	0.7024
jell_2	107	−0.0120	0.4939	0.7062
Aitaners_2	108	−0.0174	0.4911	0.6782
SafetyNLP_1	109	−0.0179	0.4909	0.7003

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
FourBytes_1	110	−0.0188	0.4904	0.7323
Legleye-y-Clara_3	111	−0.0207	0.4895	0.6624
NACHUGO_2	112	−0.0230	0.4883	0.7080
ParabaTeam_3	113	−0.0235	0.4881	0.6715
tai6le_2	114	−0.0237	0.4879	0.6966
ANALYTICALGRAMMAR_3	115	−0.0253	0.4871	0.6811
teamJorge_2	116	−0.0269	0.4863	0.6747
VANGUARD_1	117	−0.0274	0.4861	0.6920
Daien and Camilla_2	118	−0.0289	0.4853	0.7007
ANALYTICALGRAMMAR_2	119	−0.0302	0.4846	0.6667
MariayLuna_2	120	−0.0317	0.4839	0.7063
CEVS_2	121	−0.0328	0.4833	0.6742
NACHUGO_1	122	−0.0340	0.4827	0.6708
Cloud-17_1	123	−0.0341	0.4827	0.7071
Alejandro_2	124	−0.0355	0.4820	0.6808
ANALYTICALGRAMMAR_1	125	−0.0360	0.4817	0.6690
run_1	126	−0.0390	0.4802	0.7084
MiguelSalva_2	127	−0.0417	0.4788	0.7186
Aitaners_1	128	−0.0511	0.4740	0.6713
VANGUARD_3	129	−0.0567	0.4712	0.6924
RUBLO_1	130	−0.0574	0.4708	0.6922
Legleye-y-Clara_2	131	−0.0591	0.4699	0.6684
DuoDinamita_1	132	−0.0600	0.4695	0.6755
thebocadillos_1	133	−0.0600	0.4695	0.6475
MiguelSalva_3	134	−0.0602	0.4694	0.7126
Os-Faixa-Preta-do-Data_3	135	−0.0606	0.4692	0.6565
RUBLO_2	136	−0.0607	0.4691	0.6864
Atipicas_1	137	−0.0621	0.4684	0.6655
KazNU-ISS_1	138	−0.0647	0.4671	0.6528
ParabaTeam_2	139	−0.0652	0.4669	0.6582
Alejandro_1	140	−0.0670	0.4659	0.6883
VANGUARD_2	141	−0.0674	0.4657	0.6766
AryanSomnathBanerjee_3	142	−0.0676	0.4656	0.6424
DataModel_2	143	−0.0685	0.4652	0.6895
Marc&Pablo_3	144	−0.0718	0.4635	0.6776
OCRPragSex_2	145	−0.0735	0.4626	0.6717
SPECTRA_1	146	−0.0737	0.4625	0.6462
MiguelSalva_1	147	−0.0747	0.4620	0.6782
teamJorge_3	148	−0.0824	0.4581	0.6818
MRBLACK_1	149	−0.0840	0.4573	0.5668
CEVS_1	150	−0.0861	0.4562	0.6944
Los yets_2	151	−0.0882	0.4552	0.6440
lonquets_2	152	−0.0884	0.4550	0.6791
PL Team_1	153	−0.0893	0.4546	0.6780
Atipicas_2	154	−0.0940	0.4522	0.6467
JBG_2	155	−0.0964	0.4510	0.6505
PL Team_3	156	−0.0983	0.4500	0.6748

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
MemeMiner_2	157	−0.0986	0.4498	0.6404
Meme-Hunters-Team_1	158	−0.1001	0.4491	0.6683
maranda24_1	159	−0.1011	0.4486	0.6736
Vicente&Lorenzo_1	160	−0.1043	0.4470	0.6732
Meme-Hunters-Team_2	161	−0.1113	0.4434	0.6788
MariayLuna_3	162	−0.1150	0.4415	0.6488
UNED-Annotate-Team_3	163	−0.1206	0.4387	0.5767
MemeMiner_3	164	−0.1231	0.4374	0.6006
haoxiang-lang_1	165	−0.1248	0.4365	0.6327
APARAJITA-NITA-JH_1	166	−0.1248	0.4365	0.6333
C-TEA_1	167	−0.1303	0.4337	0.6087
M&M_1	168	−0.1311	0.4333	0.6373
Alejandro_3	169	−0.1372	0.4302	0.6963
mandatory_1	170	−0.1374	0.4301	0.6916
TheAntisexists_1	171	−0.1417	0.4279	0.6373
SPECTRA_2	172	−0.1488	0.4243	0.6057
andreolo_1	173	−0.1632	0.4170	0.6878
GuiGil_1	174	−0.1693	0.4139	0.6789
Meme-Hunters-Team_3	175	−0.1708	0.4132	0.6677
RUBLO_3	176	−0.1742	0.4114	0.6493
Los yets_3	177	−0.1795	0.4087	0.6463
MariayLuna_1	178	−0.1872	0.4048	0.6870
AndreayEmma_1	179	−0.1876	0.4046	0.6956
AndreayEmma_2	180	−0.1876	0.4046	0.6956
AndreayEmma_3	181	−0.1876	0.4046	0.6956
Aitaners_3	182	−0.1879	0.4044	0.6130
M&M_2	183	−0.2161	0.3901	0.6196
iEXIST-UPV_3	184	−0.2176	0.3893	0.6773
iEXIST-UPV_2	185	−0.2176	0.3893	0.6773
iEXIST-UPV_1	186	−0.2176	0.3893	0.6773
TheAntisexists_3	187	−0.2179	0.3892	0.6767
PAULU_3	188	−0.2357	0.3801	0.6089
Vicente&Lorenzo_2	189	−0.2404	0.3777	0.6037
TokenTwins_3	190	−0.2640	0.3657	0.5987
SPECTRA_3	191	−0.2680	0.3637	0.5551
Alba & Sergio team_2	192	−0.2899	0.3526	0.6548
Alba & Sergio team_3	193	−0.3010	0.3469	0.6731
Vicente&Lorenzo_3	194	−0.3190	0.3378	0.5971
Women_2	195	−0.3438	0.3252	0.6868
DataModel_1	196	−0.3458	0.3242	0.5804
berlin_2	197	−0.3684	0.3126	0.6824
MICCAR_1	198	−0.3724	0.3106	0.6841
Marc&Pablo_2	199	−0.3730	0.3103	0.6491
lonquets_3	200	−0.3761	0.3087	0.6860
Alba & Sergio team_1	201	−0.3878	0.3028	0.6178
Test_majority-class	202	−0.4038	0.2947	0.6821
CEVS_3	203	−0.4176	0.2877	0.5156

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
IreLucIA_1	204	−0.4350	0.2788	0.3101
PL Team_2	205	−0.4497	0.2713	0.4028
TokenTwins_2	206	−0.4694	0.2613	0.4289
PolskaGurom_1	207	−0.5131	0.2391	0.5088
Ctrl-Alt-Supr_2	208	−0.5278	0.2316	0.1798
Ctrl-Alt-Supr_1	209	−0.5278	0.2316	0.1798
Ctrl-Alt-Supr_3	210	−0.5278	0.2316	0.1798
BEST-team_1	211	−0.5290	0.2310	0.5023
TheAntisexists_2	212	−0.6468	0.1711	0.0000
FairMemers_1	213	−0.6468	0.1711	0.0000
Test_minority-class	214	−0.6468	0.1711	0.0000

The best hard run was DDAIUNICC_1, with ICM-Hard Norm=0.7387. The top of the hard leaderboard was less separated than in soft evaluation: the difference in ICM-Hard Norm between the best and fifth-best unique-team systems was 7.93 percentage points. The leading team changed from ORDANTIS in soft evaluation to DDAIUNICC in hard evaluation, showing that strong probability distributions and strong discrete decisions were related but not identical objectives.

System Behavior and Methods. The strongest Task 2.1 systems with working notes tend to rely on either explicit image-text modelling, task-specific adaptation, or both. ORDANTIS describes a Gemini-enriched multimodal probabilistic model in which visual content is converted into structured textual information, combined with OCR, and then used to predict empirical annotator distributions [37]. DDAIUNICC reports a feedback-driven prompt-optimization framework and sensor-aware extensions, but the Working Note does not provide enough evidence to isolate the contribution of sensors from the prompt-optimization pipeline in a specific submitted run [35]. Other competitive working notes point in related directions: AIWIZARDS uses fixed Gemini vision-language embeddings and hierarchical soft-label learning [28], while ALBERT-319 explicitly studies physiological calibration and reports that it helps hard evaluation but can slightly hurt soft evaluation [30]. Taken together, the supported conclusion is not that a single modality dominated, but that the best systems were those that made the image-text relation or the task-specific decision process usable for either calibrated soft prediction or robust hard decisions.

6.2. Task 2.2: Source Intention in Memes

Task 2.2 evaluates the classification of the source intention conveyed by sexist memes. Compared with Task 2.1, the normalized ICM score distribution is lower and wider, which indicates that identifying intent is more difficult than binary sexism detection even when the input format is the same.

6.2.1. Soft Evaluation

Table 8 shows the soft-soft results for Task 2.2. There were 112 participant submissions and three reference systems. A total of 111 participant runs outperformed the strongest simple baseline, the majority-class baseline, which obtained ICM-Soft Norm=0.0000. Participant ICM-Soft Norm scores ranged from 0.5012 to 0.0000, with a mean of 0.269 and a standard deviation of 0.102.

Table 8
Complete leaderboard for EXIST 2026 Task 2.2, Source Intention in Memes, under the soft-soft evaluation protocol. The table reports all 115 evaluated runs, including 112 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	4.7018	1.0000	0.9325
Ordantis_1	1	0.0114	0.5012	1.3334
Ordantis_2	2	−0.5650	0.4399	1.5248
Ordantis_3	3	−0.5685	0.4395	1.4874
aiwizards_1	4	−0.6720	0.4285	1.4668
aiwizards_3	5	−0.7623	0.4189	1.4859
Os-Faixa-Preta-do-Data_1	6	−0.8372	0.4110	1.4069
aiwizards_2	7	−0.9496	0.3990	1.4902
Jow_1	8	−0.9680	0.3971	1.4315
Elliot Thorel_1	9	−1.0997	0.3831	1.7144
SerranoTeam_1	10	−1.1095	0.3820	1.4263
Jow_3	11	−1.1401	0.3788	1.5640
BicoccaUnimib_2	12	−1.1752	0.3750	1.4233
BicoccaUnimib_1	13	−1.1770	0.3748	1.4036
SerranoTeam_2	14	−1.1826	0.3742	1.5158
BicoccaUnimib_3	15	−1.2031	0.3721	1.4580
Os-Faixa-Preta-do-Data_2	16	−1.2159	0.3707	1.4423
Legleye-y-Clara_1	17	−1.2547	0.3666	1.4580
Jow_2	18	−1.2604	0.3660	1.5631
ELiRF-UPV_2	19	−1.3160	0.3601	1.4908
AILS-NTUA_3	20	−1.3266	0.3589	1.4405
AILS-NTUA_1	21	−1.3450	0.3570	1.4486
AILS-NTUA_2	22	−1.3525	0.3562	1.4343
ELiRF-UPV_1	23	−1.3594	0.3554	1.4758
ROBERTA'S-SONS_2	24	−1.3656	0.3548	2.4548
EQUIPOACTIMEL_1	25	−1.3700	0.3543	1.5866
Legleye-y-Clara_3	26	−1.3966	0.3515	1.6180
SafetyNLP_3	27	−1.4512	0.3457	1.6665
SafetyNLP_1	28	−1.4973	0.3408	1.5530
VANGUARD_1	29	−1.5152	0.3389	1.4575
TokenTwins_1	30	−1.5220	0.3381	1.4842
Legleye-y-Clara_2	31	−1.5581	0.3343	1.4672
VANGUARD_2	32	−1.5757	0.3324	1.4613
teamJorge_2	33	−1.5931	0.3306	1.4662
Castrillo_1	34	−1.6002	0.3298	1.8327
Cloud-17_2	35	−1.6095	0.3288	1.6527
VANGUARD_3	36	−1.6112	0.3287	1.4580
Os-Faixa-Preta-do-Data_3	37	−1.6271	0.3270	1.5058
ANALYTICALGRAMMAR_2	38	−1.6507	0.3245	1.4896
DuoDinamita_1	39	−1.6631	0.3231	1.4881
BioSentinal_1	40	−1.6655	0.3229	1.4629
run_1	41	−1.6678	0.3226	1.4684

Continued on next page

Table 8 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Cloud-17_3	42	−1.6826	0.3211	1.6652
teamJorge_3	43	−1.6865	0.3207	1.5020
teamJorge_1	44	−1.6866	0.3206	1.4680
andreolo_1	45	−1.6994	0.3193	1.5100
Senyu&Dani_1	46	−1.7245	0.3166	1.5219
ANALYTICALGRAMMAR_1	47	−1.7276	0.3163	1.5294
SafetyNLP_2	48	−1.7407	0.3149	1.5148
KlaudiaBanasiewicz_1	49	−1.7567	0.3132	1.5004
Cloud-17_1	50	−1.7775	0.3110	1.7582
Marc&Pablo_3	51	−1.7810	0.3106	1.5099
ANALYTICALGRAMMAR_3	52	−1.7981	0.3088	1.4758
ELiRF-UPV_3	53	−1.8847	0.2996	1.6244
lonquets_3	54	−1.8931	0.2987	2.3959
TheAntisexists_1	55	−1.9211	0.2957	1.5500
thebocadillos_1	56	−1.9742	0.2901	2.3850
Marc&Pablo_2	57	−1.9748	0.2900	1.5324
Ctrl-Alt-Supr_1	58	−2.0287	0.2843	1.5297
Ctrl-Alt-Supr_2	59	−2.0287	0.2843	1.5297
Ctrl-Alt-Supr_3	60	−2.0287	0.2843	1.5297
EQUIPOACTIMEL_2	61	−2.0399	0.2831	1.6730
C-TEA_1	62	−2.0412	0.2829	1.5816
mandatory_1	63	−2.0766	0.2792	1.5390
DreamTeam_1	64	−2.1532	0.2710	2.0674
Los yets_3	65	−2.1866	0.2675	1.5522
EQUIPOACTIMEL_3	66	−2.1958	0.2665	1.7061
MB1-UPV_1	67	−2.2602	0.2596	3.9056
TheAntisexists_2	68	−2.3037	0.2550	1.5425
Marc&Pablo_1	69	−2.3176	0.2535	1.5756
Los yets_2	70	−2.3541	0.2497	1.5847
FairMemers_1	71	−2.3581	0.2492	1.5804
I2C-UHU-PERSEUS_1	72	−2.4680	0.2375	4.0572
NACHUGO_2	73	−2.5393	0.2300	4.1787
lonquets_2	74	−2.5489	0.2289	2.2004
iEXIST-UPV_3	75	−2.6131	0.2221	1.7333
iEXIST-UPV_2	76	−2.6131	0.2221	1.7333
iEXIST-UPV_1	77	−2.6131	0.2221	1.7333
SesgoCero_1	78	−2.6286	0.2205	4.1159
CEVS_1	79	−2.6347	0.2198	4.3016
Meme-Hunters-Team_2	80	−2.6715	0.2159	2.4070
Meme-Hunters-Team_3	81	−2.6810	0.2149	2.4157
Meme-Hunters-Team_1	82	−2.6845	0.2145	2.4597
maranda24_1	83	−2.6885	0.2141	1.8881
APARAJITA-NITA-JH_1	84	−2.7795	0.2044	4.0169
AryanSomnathBanerjee_2	85	−2.7905	0.2033	4.0678
pgnamia-sfonpia-UPV_2	86	−3.0321	0.1776	5.8780
Transformers_1	87	−3.0590	0.1747	4.9298
AndreayEmma_1	88	−3.0763	0.1729	4.9912

Continued on next page

Table 8 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
AndreayEmma_3	89	−3.0763	0.1729	4.9912
AndreayEmma_2	90	−3.0763	0.1729	4.9912
pgnamia-sfonpia-UPV_3	91	−3.1324	0.1669	4.6230
AryanSomnathBanerjee_1	92	−3.1586	0.1641	2.7707
lonquets_1	93	−3.1610	0.1639	3.5146
Vicente&Lorenzo_1	94	−3.1818	0.1616	4.3067
Team1_1	95	−3.2461	0.1548	4.0341
JogaBonito_3	96	−3.3428	0.1445	6.0130
NACHUGO_1	97	−3.3611	0.1426	4.2734
CEVS_2	98	−3.4124	0.1371	4.0831
MRBLACK_1	99	−3.4838	0.1295	4.4012
PolskaGurom_1	100	−3.5024	0.1275	3.4483
TheAntisexists_3	101	−3.5091	0.1268	2.1592
JogaBonito_1	102	−3.6108	0.1160	5.8509
TokenTwins_3	103	−3.6333	0.1136	3.0310
Vicente&Lorenzo_2	104	−3.6391	0.1130	4.5796
JogaBonito_2	105	−3.6520	0.1116	5.8343
Vicente&Lorenzo_3	106	−3.9377	0.0813	4.7717
CEVS_3	107	−3.9899	0.0757	5.0466
Los yets_1	108	−4.0320	0.0712	4.9900
pgnamia-sfonpia-UPV_1	109	−4.0924	0.0648	10.5252
NadalBQ_1	110	−4.3418	0.0383	3.3063
TokenTwins_2	111	−4.4053	0.0315	3.5764
Test_majority-class	112	−5.0745	0.0000	5.5565
AryanSomnathBanerjee_3	113	−5.2134	0.0000	2.9626
Test_minority-class	114	−18.9382	0.0000	8.0245

Ordantis_1 ranked first with ICM-Soft Norm=0.5012. The best-to-fifth unique-team gap in ICM-Soft Norm was 11.81 percentage points, indicating a clearer separation among the leading soft systems than in the hard setting. This suggests that only a subset of systems managed to produce probability distributions that captured disagreement about source intention rather than merely approximating a majority label.

6.2.2. Hard Evaluation

Table 9 presents the hard-hard results. The task received 181 participant runs, of which 165 outperformed the strongest simple baseline. Participant ICM-Hard Norm scores ranged from 0.6289 to 0.0281, with a mean of 0.324 and a standard deviation of 0.113.

Table 9
Complete leaderboard for EXIST 2026 Task 2.2, Source Intention in Memes, under the hard-hard evaluation protocol. The table reports all 184 evaluated runs, including 181 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and Macro F1. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Test_gold	0	1.4383	1.0000	1.0000
Ordantis_3	1	0.3709	0.6289	0.6158
DDAIUNICC_1	2	0.3267	0.6136	0.6081
DDAIUNICC_2	3	0.2444	0.5850	0.5858
DDAIUNICC_3	4	0.2127	0.5739	0.5829
Ordantis_2	5	0.1790	0.5622	0.5453
Math Addicts_1	6	0.1110	0.5386	0.5469
DreamTeam_1	7	0.0977	0.5340	0.5559
ELiRF-UPV_3	8	0.0879	0.5306	0.5501
ELiRF-UPV_2	9	0.0728	0.5253	0.5519
CYUT-Giantsshoulders_1	10	0.0664	0.5231	0.5041
Ordantis_1	11	0.0659	0.5229	0.5368
aiwizards_3	12	0.0020	0.5007	0.5204
aiwizards_1	13	−0.0353	0.4877	0.5200
datovalenciano_1	14	−0.0615	0.4786	0.4391
ROBERTA'S-SONS_2	15	−0.0929	0.4677	0.4984
Os-Faixa-Preta-do-Data_1	16	−0.1193	0.4585	0.5008
aiwizards_2	17	−0.1248	0.4566	0.5182
MICCAR_2	18	−0.1338	0.4535	0.4234
Lidia_1	19	−0.1380	0.4520	0.4622
Jow_1	20	−0.1395	0.4515	0.4900
pgnamia-sfonpia-UPV_2	21	−0.1511	0.4475	0.4689
pgnamia-sfonpia-UPV_1	22	−0.1844	0.4359	0.4508
Team1_1	23	−0.2003	0.4304	0.4705
MemeMiner_1	24	−0.2120	0.4263	0.4810
Daien and Camilla_3	25	−0.2313	0.4196	0.4101
SerranoTeam_3	26	−0.2319	0.4194	0.4430
SerranoTeam_1	27	−0.2499	0.4131	0.4676
BicoccaUnimib_2	28	−0.2532	0.4120	0.4332
Legleye-y-Clara_1	29	−0.2553	0.4112	0.4542
SerranoTeam_2	30	−0.2601	0.4096	0.4660
BicoccaUnimib_1	31	−0.2604	0.4095	0.4110
Jow_3	32	−0.2628	0.4086	0.4643
BicoccaUnimib_3	33	−0.2710	0.4058	0.4363
Os-Faixa-Preta-do-Data_2	34	−0.2777	0.4035	0.4513
Aurora_1	35	−0.2898	0.3993	0.3740
pgnamia-sfonpia-UPV_3	36	−0.2912	0.3988	0.4621
ELiRF-UPV_1	37	−0.2916	0.3986	0.4554
Castrillo_1	38	−0.2997	0.3958	0.4530
AILS-NTUA_2	39	−0.3082	0.3929	0.4033
Jow_2	40	−0.3124	0.3914	0.4473
TokenTwins_1	41	−0.3126	0.3913	0.3791

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Women_1	42	−0.3179	0.3895	0.4342
AILS-NTUA_3	43	−0.3272	0.3863	0.3945
SafetyNLP_3	44	−0.3277	0.3861	0.4290
NACHUGO_2	45	−0.3305	0.3851	0.3629
MB1-UPV_1	46	−0.3312	0.3849	0.4497
MiguelSalva_1	47	−0.3397	0.3819	0.4436
AILS-NTUA_1	48	−0.3446	0.3802	0.3913
BioSentinal_1	49	−0.3516	0.3778	0.4236
MiguelSalva_2	50	−0.3533	0.3772	0.4095
AryanSomnathBanerjee_2	51	−0.3548	0.3767	0.4283
AryanSomnathBanerjee_1	52	−0.3548	0.3767	0.4283
ANALYTICALGRAMMAR_3	53	−0.3551	0.3765	0.4074
MRBLACK_1	54	−0.3608	0.3746	0.4379
SafetyNLP_1	55	−0.3634	0.3737	0.4508
Women_3	56	−0.3651	0.3731	0.4388
teamJorge_2	57	−0.3675	0.3722	0.4326
I2C-UHU-PERSEUS_1	58	−0.3692	0.3717	0.4186
MiguelSalva_3	59	−0.3698	0.3715	0.4414
lonquets_3	60	−0.3723	0.3706	0.4342
CEVS_1	61	−0.3786	0.3684	0.3509
PAULU_2	62	−0.3801	0.3679	0.4512
PAULU_3	63	−0.3801	0.3679	0.4512
Senyu&Dani_1	64	−0.3843	0.3664	0.4452
Legleye-y-Clara_2	65	−0.3845	0.3664	0.4102
run_1	66	−0.3876	0.3652	0.4276
Os-Faixa-Preta-do-Data_3	67	−0.3927	0.3635	0.4162
CEVS_2	68	−0.3972	0.3619	0.4212
ANALYTICALGRAMMAR_2	69	−0.3973	0.3619	0.3869
Meme-Hunters-Team_1	70	−0.3977	0.3618	0.4071
VANGUARD_1	71	−0.3992	0.3612	0.3719
LNR-Erasmus_1	72	−0.4004	0.3608	0.4343
Lorena_1	73	−0.4040	0.3595	0.4308
VANGUARD_3	74	−0.4043	0.3595	0.3692
ANALYTICALGRAMMAR_1	75	−0.4147	0.3558	0.4204
teamJorge_1	76	−0.4166	0.3552	0.4214
Legleye-y-Clara_3	77	−0.4175	0.3549	0.4295
NACHUGO_1	78	−0.4209	0.3537	0.4245
Alejandro_3	79	−0.4215	0.3535	0.4324
Atipicas_3	80	−0.4326	0.3496	0.3639
Meme-Hunters-Team_2	81	−0.4380	0.3477	0.3685
SesgoCero_1	82	−0.4382	0.3477	0.4117
Marc&Pablo_3	83	−0.4392	0.3473	0.4237
CoEXIST_2	84	−0.4443	0.3455	0.4197
ParabaTeam_1	85	−0.4460	0.3450	0.4327
ParabaTeam_3	86	−0.4460	0.3450	0.4327
SafetyNLP_2	87	−0.4490	0.3439	0.4585
RUBLO_2	88	−0.4507	0.3433	0.4060

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
MemeMiner_2	89	−0.4520	0.3429	0.4406
CoEXIST_3	90	−0.4643	0.3386	0.4162
Atipicas_1	91	−0.4655	0.3382	0.4167
mandatory_1	92	−0.4680	0.3373	0.3271
VANGUARD_2	93	−0.4688	0.3370	0.3576
Lorena_2	94	−0.4695	0.3368	0.4208
RUBLO_1	95	−0.4701	0.3366	0.3963
teamJorge_3	96	−0.4758	0.3346	0.4184
GIL UNAM_2	97	−0.4785	0.3337	0.3078
Cloud-17_3	98	−0.4828	0.3322	0.4407
Lorena_3	99	−0.4847	0.3315	0.4215
CoEXIST_1	100	−0.4962	0.3275	0.4015
Meme-Hunters-Team_3	101	−0.4987	0.3266	0.3644
Daien and Camilla_2	102	−0.5052	0.3244	0.4075
Aitaners_1	103	−0.5067	0.3239	0.4040
Los yets_1	104	−0.5082	0.3233	0.4222
CLUPV_1	105	−0.5099	0.3227	0.4261
ParabaTeam_2	106	−0.5210	0.3189	0.3958
KlaudiaBanasiewicz_1	107	−0.5218	0.3186	0.4196
APARAJITA-NITA-JH_1	108	−0.5253	0.3174	0.4036
GIL UNAM_1	109	−0.5259	0.3172	0.4312
DataModel_1	110	−0.5297	0.3159	0.3427
Cloud-17_2	111	−0.5301	0.3157	0.4298
DataModel_2	112	−0.5338	0.3144	0.4115
Cloud-17_1	113	−0.5408	0.3120	0.4280
MariayLuna_2	114	−0.5574	0.3062	0.4197
lonquets_2	115	−0.5604	0.3052	0.4179
EQUIPOACTIMEL_1	116	−0.5669	0.3029	0.4185
Women_2	117	−0.5694	0.3021	0.3915
DuoDinamita_1	118	−0.5734	0.3007	0.4284
Daien and Camilla_1	119	−0.5781	0.2991	0.4155
maranda24_1	120	−0.5797	0.2985	0.4051
MemeMiner_3	121	−0.5813	0.2979	0.4008
MariayLuna_3	122	−0.5842	0.2969	0.4060
thebocadillos_1	123	−0.5924	0.2941	0.3273
Aitaners_2	124	−0.5942	0.2934	0.4109
Atipicas_2	125	−0.5981	0.2921	0.4031
Ctrl-Alt-Supr_1	126	−0.6295	0.2812	0.3322
Ctrl-Alt-Supr_3	127	−0.6295	0.2812	0.3322
Ctrl-Alt-Supr_2	128	−0.6295	0.2812	0.3322
MariayLuna_1	129	−0.6296	0.2811	0.4030
PAULU_1	130	−0.6300	0.2810	0.4002
GIL UNAM_3	131	−0.6316	0.2804	0.4151
C-TEA_1	132	−0.6343	0.2795	0.3968
PL Team_3	133	−0.6420	0.2768	0.3462
PL Team_1	134	−0.6447	0.2759	0.3419
TheAntisexists_1	135	−0.7075	0.2541	0.3676

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
AryanSomnathBanerjee_3	136	−0.7099	0.2532	0.3892
RUBLO_3	137	−0.7120	0.2525	0.3923
Alba & Sergio team_3	138	−0.7226	0.2488	0.2963
lonquets_1	139	−0.7231	0.2486	0.3857
Los yets_2	140	−0.7338	0.2449	0.3441
MICCAR_1	141	−0.7390	0.2431	0.2001
EQUIPOACTIMEL_2	142	−0.7454	0.2409	0.3056
andreolo_1	143	−0.7463	0.2406	0.3805
CEVS_3	144	−0.7504	0.2391	0.3354
Marc&Pablo_2	145	−0.7702	0.2322	0.3461
JogaBonito_2	146	−0.7714	0.2318	0.1751
Transformers_1	147	−0.7732	0.2312	0.1753
Marc&Pablo_1	148	−0.7839	0.2275	0.1949
AndreayEmma_1	149	−0.7908	0.2251	0.1919
AndreayEmma_3	150	−0.7908	0.2251	0.1919
AndreayEmma_2	151	−0.7908	0.2251	0.1919
JogaBonito_1	152	−0.7912	0.2250	0.1839
GuiGil_1	153	−0.8200	0.2149	0.2201
Alba & Sergio team_1	154	−0.8524	0.2037	0.3224
Vicente&Lorenzo_1	155	−0.8592	0.2013	0.3643
Aitaners_3	156	−0.8699	0.1976	0.3525
Los yets_3	157	−0.8703	0.1975	0.2409
JogaBonito_3	158	−0.8721	0.1968	0.2468
EQUIPOACTIMEL_3	159	−0.8789	0.1945	0.2935
TokenTwins_3	160	−0.9109	0.1833	0.3589
Alejandro_1	161	−0.9273	0.1777	0.3574
haoxiang-lang_1	162	−0.9294	0.1769	0.2294
Alejandro_2	163	−0.9957	0.1539	0.3469
M&M_2	164	−1.0102	0.1488	0.2604
HECTOR-ALONSO_1	165	−1.0347	0.1403	0.2460
TheAntisexists_2	166	−1.0445	0.1369	0.1839
Test_majority-class	167	−1.0445	0.1369	0.1839
FairMemers_1	168	−1.0445	0.1369	0.1839
PolskaGurom_1	169	−1.0445	0.1369	0.1839
Vicente&Lorenzo_3	170	−1.0512	0.1346	0.3294
Vicente&Lorenzo_2	171	−1.0712	0.1276	0.3041
M&M_1	172	−1.0767	0.1257	0.2356
M&M_3	173	−1.0977	0.1184	0.2335
Alba & Sergio team_2	174	−1.1205	0.1105	0.3187
BEST-team_1	175	−1.1355	0.1053	0.1999
iEXIST-UPV_1	176	−1.1503	0.1001	0.2541
iEXIST-UPV_3	177	−1.1503	0.1001	0.2541
iEXIST-UPV_2	178	−1.1503	0.1001	0.2541
TheAntisexists_3	179	−1.2528	0.0645	0.2894
TokenTwins_2	180	−1.2835	0.0538	0.2981
PL Team_2	181	−1.3429	0.0332	0.2898
IreLucIA_1	182	−1.3574	0.0281	0.2100

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Test_minority-class	183	−2.0637	0.0000	0.0697

Ordantis_3 led the hard evaluation with ICM-Hard Norm=0.6289. Unlike several other subtasks, the same team led both soft and hard protocols, although with different runs. This makes Task 2.2 one of the clearest cases in which a team achieved stable performance across the two evaluation views.

System Behavior and Methods. The methodological evidence for Task 2.2 points to the usefulness of enriched textual representations of memes. Ordantis' Working Note supports this interpretation through its use of structured visual-to-text descriptions, OCR and probabilistic modelling [37]. SafetyNLP also focuses on improving OCR text before combining multilingual text models with CLIP/SigLIP-style multimodal models and LeWiDi target modelling [25]. BioSentinel, a Working Note focused specifically on source intention in memes, takes a more text-centric route with XLM-RoBERTa and explicitly excludes physiological signals because of a train-test feature mismatch [33]. These reports support a cautious reading of the leaderboard: source-intention performance seems to depend heavily on how well the system reconstructs the communicative meaning of the meme, while the available evidence does not show a general advantage for sensor features in this subtask.

6.3. Task 2.3: Sexism Categorization in Memes

Task 2.3 is the fine-grained meme categorization task. It is harder than Task 2.1 and Task 2.2: the best normalized ICM scores are lower, the mean normalized ICM scores are much lower, and a substantial fraction of systems fail to outperform the majority-class baseline.

6.3.1. Soft Evaluation

Table 10 reports the soft-soft results for Task 2.3. The leaderboard contains 113 participant runs and three reference systems. Only 73 participant runs outperformed the strongest simple baseline, whose ICM-Soft Norm was 0.0000. Participant ICM-Soft Norm scores ranged from 0.3469 to 0.0000, with a mean of 0.107 and a standard deviation of 0.100.

Table 10
Complete leaderboard for EXIST 2026 Task 2.3, Sexism Categorization in Memes, under the soft-soft evaluation protocol. The table reports all 116 evaluated runs, including 113 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	9.4343	1.0000	1.3468
aiwizards_2	1	−2.8881	0.3469	2.0785
aiwizards_3	2	−2.9507	0.3436	2.0321
aiwizards_1	3	−3.2537	0.3276	2.0847
Jow_2	4	−4.2110	0.2768	2.1026
Jow_3	5	−4.2168	0.2765	2.0904
tai6le_2	6	−4.3410	0.2699	2.1495

Continued on next page

Table 10 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
tai6le_3	7	−4.6064	0.2559	2.1570
Legleye-y-Clara_1	8	−4.6246	0.2549	2.2071
Jow_1	9	−4.6331	0.2545	2.0812
Ordantis_1	10	−4.6874	0.2516	5.0881
SerranoTeam_1	11	−4.8477	0.2431	2.1461
BicoccaUnimib_3	12	−4.8823	0.2412	2.1328
BicoccaUnimib_2	13	−4.9132	0.2396	2.1032
tai6le_1	14	−4.9452	0.2379	2.1129
ELiRF-UPV_1	15	−4.9489	0.2377	2.1502
Legleye-y-Clara_3	16	−5.0781	0.2309	2.4112
BicoccaUnimib_1	17	−5.1125	0.2290	2.0878
ELiRF-UPV_2	18	−5.2403	0.2223	2.1861
thebocadillos_1	19	−5.3580	0.2160	3.1141
Legleye-y-Clara_2	20	−5.3742	0.2152	2.2101
AILS-NTUA_2	21	−5.4766	0.2098	2.1286
AILS-NTUA_1	22	−5.4961	0.2087	2.1313
AILS-NTUA_3	23	−5.5438	0.2062	2.1295
Semantica_1	24	−5.6317	0.2015	2.3849
teamJorge_3	25	−5.6862	0.1986	3.0769
Ordantis_2	26	−5.7804	0.1937	5.1219
ELiRF-UPV_3	27	−5.7889	0.1932	2.2556
NACHUGO_2	28	−5.8662	0.1891	4.6305
Castrillo_1	29	−5.8714	0.1888	2.5006
teamJorge_2	30	−5.8867	0.1880	3.0864
TokenTwins_1	31	−5.9040	0.1871	2.3646
run_1	32	−5.9889	0.1826	2.1731
lonquets_3	33	−6.1165	0.1758	2.8263
DuoDinamita_1	34	−6.1591	0.1736	2.2513
Os-Faixa-Preta-do-Data_1	35	−6.1603	0.1735	2.1588
teamJorge_1	36	−6.1941	0.1717	3.1005
KlaudiaBanasiewicz_1	37	−6.2638	0.1680	2.2987
Ordantis_3	38	−6.3589	0.1630	5.1316
Semantica_2	39	−6.3802	0.1619	2.2732
andreolo_1	40	−6.4147	0.1600	2.2557
VANGUARD_1	41	−6.4755	0.1568	2.2329
Meme-Hunters-Team_1	42	−6.6326	0.1485	3.1166
Transformers_1	43	−6.6736	0.1463	5.3005
Meme-Hunters-Team_2	44	−6.8561	0.1366	3.0770
Meme-Hunters-Team_3	45	−6.9090	0.1338	3.0567
CEVS_1	46	−6.9324	0.1326	2.9753
mandatory_1	47	−7.0024	0.1289	2.3194
Los yets_3	48	−7.0842	0.1245	2.3638
Vicente&Lorenzo_1	49	−7.0920	0.1241	4.6550
pgnamia-sfonpia-UPV_2	50	−7.2563	0.1154	3.1014
TheAntisexists_1	51	−7.3299	0.1115	2.3086
Ctrl-Alt-Supr_3	52	−7.3631	0.1098	2.4718
Ctrl-Alt-Supr_2	53	−7.3631	0.1098	2.4718

Continued on next page

Table 10 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Ctrl-Alt-Supr_1	54	−7.3631	0.1098	2.4718
Os-Faixa-Preta-do-Data_2	55	−7.4047	0.1076	2.2703
FairMemers_1	56	−7.4495	0.1052	2.5685
SafetyNLP_2	57	−7.5252	0.1012	2.5756
SafetyNLP_1	58	−7.5253	0.1012	2.5756
SafetyNLP_3	59	−7.5337	0.1007	2.5758
pgnamia-sfonpia-UPV_3	60	−7.5441	0.1002	2.7477
lonquets_2	61	−7.6396	0.0951	3.0067
MRBLACK_1	62	−7.8172	0.0857	5.0641
pgnamia-sfonpia-UPV_1	63	−7.8854	0.0821	2.6821
Vicente&Lorenzo_2	64	−7.8970	0.0815	4.9738
Vicente&Lorenzo_3	65	−7.9392	0.0792	5.1894
C-TEA_1	66	−8.0271	0.0746	2.5346
Marc&Pablo_1	67	−8.2372	0.0634	2.9163
CEVS_2	68	−8.2665	0.0619	4.7538
TheAntisexists_3	69	−8.5789	0.0453	3.3251
Marc&Pablo_3	70	−8.6251	0.0429	2.8643
VANGUARD_2	71	−8.9001	0.0283	2.3267
Los yets_1	72	−9.1019	0.0176	2.5677
VANGUARD_3	73	−9.1174	0.0168	2.3290
Test_majority-class	74	−9.8173	0.0000	5.8836
ANALYTICALGRAMMAR_3	75	−9.9556	0.0000	2.4067
AryanSomnathBanerjee_3	76	−10.1289	0.0000	4.5492
AryanSomnathBanerjee_2	77	−10.1289	0.0000	4.5492
Team1_1	78	−10.3491	0.0000	3.5433
AryanSomnathBanerjee_1	79	−10.5257	0.0000	5.7936
ANALYTICALGRAMMAR_2	80	−10.8377	0.0000	2.3890
ANALYTICALGRAMMAR_1	81	−10.9566	0.0000	2.3922
EQUIPOACTIMEL_3	82	−11.1202	0.0000	2.3591
lonquets_1	83	−12.2530	0.0000	4.4695
EQUIPOACTIMEL_2	84	−12.2951	0.0000	2.4410
SesgoCero_1	85	−12.8016	0.0000	4.8015
Os-Faixa-Preta-do-Data_3	86	−12.8688	0.0000	2.4096
TheAntisexists_2	87	−13.1001	0.0000	2.4674
NACHUGO_1	88	−13.2280	0.0000	4.8396
CEVS_3	89	−13.5311	0.0000	5.8431
ROBERTA'S-SONS_2	90	−13.7236	0.0000	3.1943
Los yets_2	91	−13.9292	0.0000	2.4649
Cloud-17_2	92	−14.5195	0.0000	2.7173
EQUIPOACTIMEL_1	93	−14.6025	0.0000	2.4783
Cloud-17_1	94	−14.6147	0.0000	2.7850
Cloud-17_3	95	−14.6769	0.0000	2.7148
maranda24_1	96	−14.9718	0.0000	5.1208
PolskaGurom_1	97	−15.2494	0.0000	5.8488
Marc&Pablo_2	98	−16.0434	0.0000	3.6264
Senyu&Dani_1	99	−16.1653	0.0000	2.3707
DreamTeam_1	100	−16.2643	0.0000	2.6017

Continued on next page

Table 10 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Women_1	101	−16.7877	0.0000	2.4075
Women_2	102	−16.9637	0.0000	2.4034
Women_3	103	−17.7118	0.0000	2.4959
NadalBQ_1	104	−18.7546	0.0000	8.5347
JogaBonito_3	105	−21.2101	0.0000	5.0774
JogaBonito_1	106	−22.6613	0.0000	5.0101
JogaBonito_2	107	−23.3986	0.0000	4.9933
TokenTwins_3	108	−25.4116	0.0000	4.9449
I2C-UHU-PERSEUS_1	109	−25.8863	0.0000	4.8896
iEXIST-UPV_1	110	−29.3657	0.0000	2.8470
iEXIST-UPV_2	111	−29.3657	0.0000	2.8470
iEXIST-UPV_3	112	−29.3657	0.0000	2.8470
MB1-UPV_1	113	−31.6494	0.0000	4.8496
Test_minority-class	114	−50.0353	0.0000	9.3270
TokenTwins_2	115	−54.0051	0.0000	11.5854

The best soft run was aiwizards_2, with ICM-Soft Norm=0.3469. The best-to-fifth unique-team spread in ICM-Soft Norm was 9.53 percentage points. The gap between the first and second unique teams was especially large, indicating that the winning system modelled the soft categorization distributions more effectively than most competitors.

6.3.2. Hard Evaluation

Table 11 presents the hard-hard results. There were 182 participant runs, and 137 of them outperformed the majority-class baseline, which obtained ICM-Hard Norm=0.0703. Participant ICM-Hard Norm scores ranged from 0.5008 to 0.0000, with a mean of 0.190 and a standard deviation of 0.131.

Table 11
Complete leaderboard for EXIST 2026 Task 2.3, Sexism Categorization in Memes, under the hard-hard evaluation protocol. The table reports all 185 evaluated runs, including 182 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and Macro F1. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Test_gold	0	2.4100	1.0000	1.0000
DDAIUNICC_1	1	0.0036	0.5008	0.5746
DDAIUNICC_2	2	−0.0819	0.4830	0.5679
ELiRF-UPV_3	3	−0.1069	0.4778	0.5524
DDAIUNICC_3	4	−0.1566	0.4675	0.5417
CYUT-Giantsshoulders_1	5	−0.1901	0.4606	0.5387
ELiRF-UPV_2	6	−0.1914	0.4603	0.5411
Math Addicts_1	7	−0.2203	0.4543	0.5413
aiwizards_3	8	−0.3953	0.4180	0.4500
aiwizards_2	9	−0.4224	0.4124	0.4566

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Os-Faixa-Preta-do-Data_1	10	−0.4528	0.4060	0.5080
GIL UNAM_2	11	−0.4622	0.4041	0.4689
pgnamia-sfonpia-UPV_2	12	−0.5304	0.3900	0.4718
aiwizards_1	13	−0.5316	0.3897	0.4167
pgnamia-sfonpia-UPV_1	14	−0.6049	0.3745	0.4768
Os-Faixa-Preta-do-Data_2	15	−0.6115	0.3731	0.4650
SerranoTeam_2	16	−0.6375	0.3677	0.4712
datovalenciano_1	17	−0.6636	0.3623	0.4130
Team1_1	18	−0.6646	0.3621	0.4595
GIL UNAM_1	19	−0.6686	0.3613	0.4903
PAULU_1	20	−0.7245	0.3497	0.3414
ELiRF-UPV_1	21	−0.7347	0.3476	0.4461
MICCAR_2	22	−0.7408	0.3463	0.4105
GIL UNAM_3	23	−0.7502	0.3444	0.4859
Lidia_1	24	−0.7682	0.3406	0.4060
SerranoTeam_1	25	−0.7713	0.3400	0.4977
pgnamia-sfonpia-UPV_3	26	−0.8357	0.3266	0.4902
tai6le_1	27	−0.8413	0.3255	0.4069
Jow_2	28	−0.8579	0.3220	0.3849
tai6le_3	29	−0.8721	0.3191	0.3415
MemeMiner_1	30	−0.8760	0.3183	0.3296
AILS-NTUA_3	31	−0.9125	0.3107	0.3482
Jow_3	32	−0.9193	0.3093	0.3733
AILS-NTUA_2	33	−0.9207	0.3090	0.3502
lonquets_3	34	−0.9277	0.3075	0.3935
Jow_1	35	−0.9386	0.3053	0.3227
KlaudiaBanasiewicz_1	36	−0.9443	0.3041	0.4421
Atipicas_3	37	−0.9518	0.3025	0.3267
AILS-NTUA_1	38	−0.9530	0.3023	0.3501
RUBLO_2	39	−0.9537	0.3021	0.3942
RUBLO_1	40	−0.9559	0.3017	0.3911
MRBLACK_1	41	−0.9779	0.2971	0.3944
CEVS_1	42	−1.0303	0.2863	0.3713
Alejandro_2	43	−1.0367	0.2849	0.4221
run_1	44	−1.0483	0.2825	0.3172
DreamTeam_1	45	−1.0604	0.2800	0.5124
teamJorge_2	46	−1.0678	0.2785	0.4401
ParabaTeam_1	47	−1.0705	0.2779	0.4308
Senyu&Dani_1	48	−1.0879	0.2743	0.4567
AryanSomnathBanerjee_3	49	−1.0981	0.2722	0.3335
AryanSomnathBanerjee_2	50	−1.0981	0.2722	0.3335
MariayLuna_2	51	−1.1170	0.2683	0.3890
Os-Faixa-Preta-do-Data_3	52	−1.1371	0.2641	0.4365
EQUIPOACTIMEL_3	53	−1.1392	0.2636	0.4595
Cloud-17_2	54	−1.1424	0.2630	0.3371
Alejandro_1	55	−1.1473	0.2620	0.4196
CLUPV_1	56	−1.1475	0.2619	0.4484

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Alejandro_3	57	−1.1482	0.2618	0.4127
NACHUGO_1	58	−1.1557	0.2602	0.4163
ANALYTICALGRAMMAR_1	59	−1.1573	0.2599	0.4399
ParabaTeam_3	60	−1.1587	0.2596	0.4308
CEVS_2	61	−1.1603	0.2593	0.3628
Castrillo_1	62	−1.1661	0.2581	0.4631
ANALYTICALGRAMMAR_2	63	−1.1741	0.2564	0.4348
Aitaners_1	64	−1.1781	0.2556	0.4083
teamJorge_1	65	−1.1794	0.2553	0.4391
Cloud-17_3	66	−1.1828	0.2546	0.3322
ANALYTICALGRAMMAR_3	67	−1.2094	0.2491	0.4393
AryanSomnathBanerjee_1	68	−1.2102	0.2489	0.2413
Atipicas_1	69	−1.2116	0.2486	0.4110
Cloud-17_1	70	−1.2256	0.2457	0.3236
lonquets_2	71	−1.2617	0.2382	0.3481
Legleye-y-Clara_1	72	−1.2641	0.2377	0.4443
ParabaTeam_2	73	−1.2679	0.2370	0.3581
MemeMiner_2	74	−1.2790	0.2346	0.2818
Meme-Hunters-Team_1	75	−1.2919	0.2320	0.3756
teamJorge_3	76	−1.2932	0.2317	0.4393
MariayLuna_3	77	−1.2934	0.2317	0.3335
EQUIPOACTIMEL_2	78	−1.2977	0.2308	0.4450
Vicente&Lorenzo_1	79	−1.3109	0.2280	0.2606
EQUIPOACTIMEL_1	80	−1.3170	0.2268	0.4510
MariayLuna_1	81	−1.3395	0.2221	0.3673
MiguelSalva_1	82	−1.3584	0.2182	0.4312
PAULU_2	83	−1.4044	0.2086	0.3324
Meme-Hunters-Team_2	84	−1.4062	0.2083	0.3794
CoEXIST_2	85	−1.4179	0.2058	0.4131
thebocadillos_1	86	−1.4218	0.2050	0.2501
CoEXIST_1	87	−1.4264	0.2041	0.4128
PAULU_3	88	−1.4359	0.2021	0.3350
MiguelSalva_3	89	−1.4373	0.2018	0.4087
Aitaners_2	90	−1.4419	0.2009	0.3928
Lorena_2	91	−1.4521	0.1987	0.3792
Meme-Hunters-Team_3	92	−1.4572	0.1977	0.3705
Los yets_1	93	−1.4589	0.1973	0.3276
Women_2	94	−1.4699	0.1950	0.4404
PL Team_3	95	−1.4730	0.1944	0.3315
lonquets_1	96	−1.4770	0.1936	0.3304
Atipicas_2	97	−1.4881	0.1913	0.3852
PL Team_1	98	−1.5033	0.1881	0.3318
CoEXIST_3	99	−1.5125	0.1862	0.4018
RUBLO_3	100	−1.5572	0.1769	0.3512
DuoDinamita_1	101	−1.5642	0.1755	0.4128
Legleye-y-Clara_2	102	−1.5872	0.1707	0.3967
TokenTwins_1	103	−1.5874	0.1707	0.2110

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Daien and Camilla_1	104	−1.6052	0.1670	0.4194
Lorena_3	105	−1.6140	0.1652	0.3738
NACHUGO_2	106	−1.6165	0.1646	0.2130
Lorena_1	107	−1.6260	0.1627	0.4263
Women_1	108	−1.6376	0.1602	0.4455
tai6le_2	109	−1.6398	0.1598	0.1866
TheAntisexists_1	110	−1.6616	0.1553	0.3925
C-TEA_1	111	−1.6627	0.1550	0.2605
SesgoCero_1	112	−1.6641	0.1547	0.4202
BicoccaUnimib_3	113	−1.6827	0.1509	0.1934
Vicente&Lorenzo_2	114	−1.6896	0.1495	0.1617
ROBERTA'S-SONS_2	115	−1.7083	0.1456	0.4693
TheAntisexists_3	116	−1.7565	0.1356	0.2887
Legleye-y-Clara_3	117	−1.7745	0.1318	0.3893
HECTOR-ALONSO_1	118	−1.7778	0.1312	0.2847
Transformers_1	119	−1.7900	0.1286	0.2553
Vicente&Lorenzo_3	120	−1.8083	0.1248	0.2620
Daien and Camilla_2	121	−1.8275	0.1209	0.3950
BicoccaUnimib_2	122	−1.8306	0.1202	0.1588
BicoccaUnimib_1	123	−1.8993	0.1059	0.1389
haoxiang-lang_1	124	−1.9044	0.1049	0.3233
MemeMiner_3	125	−1.9080	0.1042	0.2281
Aitaners_3	126	−1.9190	0.1019	0.3204
Ctrl-Alt-Supr_3	127	−1.9248	0.1007	0.1293
Ctrl-Alt-Supr_2	128	−1.9248	0.1007	0.1293
Ctrl-Alt-Supr_1	129	−1.9248	0.1007	0.1293
Aurora_1	130	−1.9368	0.0982	0.4127
DataModel_1	131	−1.9418	0.0971	0.1486
Ordantis_1	132	−1.9465	0.0962	0.3790
MICCAR_1	133	−1.9473	0.0960	0.1297
Women_3	134	−1.9585	0.0937	0.3978
LNR-Erasmus_1	135	−1.9627	0.0928	0.4300
Ordantis_2	136	−1.9631	0.0927	0.3786
CEVS_3	137	−2.0506	0.0746	0.2165
VANGUARD_1	138	−2.0711	0.0703	0.0919
FairMemers_1	139	−2.0711	0.0703	0.0919
Test_majority-class	140	−2.0711	0.0703	0.0919
Los yets_3	141	−2.0711	0.0703	0.0919
Marc&Pablo_1	142	−2.0740	0.0697	0.2598
Ordantis_3	143	−2.0800	0.0685	0.3731
DataModel_2	144	−2.0898	0.0664	0.3727
GuiGil_1	145	−2.1360	0.0568	0.3066
Marc&Pablo_3	146	−2.2034	0.0429	0.3230
JogaBonito_2	147	−2.2086	0.0418	0.3136
JogaBonito_3	148	−2.2403	0.0352	0.3215
Marc&Pablo_2	149	−2.2453	0.0342	0.3176
SafetyNLP_1	150	−2.2541	0.0323	0.1745

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
SafetyNLP_2	151	−2.2549	0.0322	0.1744
SafetyNLP_3	152	−2.2590	0.0313	0.1760
JogaBonito_1	153	−2.2621	0.0307	0.3223
MiguelSalva_2	154	−2.2818	0.0266	0.3186
IreLucIA_1	155	−2.3259	0.0174	0.3019
Semantica_1	156	−2.3508	0.0123	0.3854
MB1-UPV_1	157	−2.3529	0.0119	0.4171
Semantica_2	158	−2.3638	0.0096	0.3392
PL Team_2	159	−2.3705	0.0082	0.2583
Alba & Sergio team_1	160	−2.3872	0.0047	0.2899
Alba & Sergio team_3	161	−2.3872	0.0047	0.2899
maranda24_1	162	−2.4706	0.0000	0.3748
andreolo_1	163	−2.4812	0.0000	0.3869
SPSPH4572_1	164	−2.5852	0.0000	0.3295
M&M_2	165	−2.6229	0.0000	0.2990
TokenTwins_3	166	−2.6514	0.0000	0.2733
AndreayEmma_1	167	−2.7530	0.0000	0.3021
AndreayEmma_2	168	−2.7530	0.0000	0.3021
AndreayEmma_3	169	−2.7530	0.0000	0.3021
BEST-team_1	170	−2.8232	0.0000	0.1780
M&M_1	171	−2.8388	0.0000	0.2882
M&M_3	172	−2.8560	0.0000	0.3034
Los yets_2	173	−3.1829	0.0000	0.3106
Test_minority-class	174	−3.3135	0.0000	0.0318
VANGUARD_3	175	−3.4351	0.0000	0.2629
VANGUARD_2	176	−3.5368	0.0000	0.2683
PolskaGurom_1	177	−3.5794	0.0000	0.2943
mandatory_1	178	−3.6499	0.0000	0.3595
TokenTwins_2	179	−3.8765	0.0000	0.2561
Alba & Sergio team_2	180	−4.2003	0.0000	0.2475
iEXIST-UPV_3	181	−4.3617	0.0000	0.2554
iEXIST-UPV_2	182	−4.3617	0.0000	0.2554
iEXIST-UPV_1	183	−4.3617	0.0000	0.2554
TheAntisexists_2	184	−4.7287	0.0000	0.3595

DDAIUNICC_1 ranked first with ICM-Hard Norm=0.5008. The winning team changed between protocols: AIWIZARDS led soft evaluation, while DDAIUNICC led hard evaluation. This difference is consistent with the nature of the task: fine-grained categories are sparse and overlapping, so predicting calibrated distributions and selecting final labels can favour different modelling choices.

System Behavior and Methods. The working notes suggest several plausible reasons for the difficulty of Task 2.3. AIWIZARDS frames the meme tasks as hierarchical conditional soft-label prediction, which is well aligned with the official soft evaluation and with the dependency between sexism identification, intention and categorization [28]. DDAIUNICC's prompt-optimization framework is reported to perform strongly in Task 2.3, but the paper does not justify attributing the result to sensors alone [35]. MATH ADDICTS focuses on heterogeneous vision-language ensembles and reports an "at-least-one" voting strategy for multi-label categorization, a design choice directly motivated by rare

categories [38]. SEMANTICA follows a two-stage multimodal pipeline and uses a handcrafted bilingual lexicon for category-specific cues, also targeting rare or subtle categories [42]. These papers support the interpretation that Task 2.3 rewards hierarchy-aware and category-aware systems, not just stronger generic encoders.

6.4. Task 3.1: Sexism Identification in Videos

Task 3.1 evaluates binary sexism identification in TikTok videos. The video setting introduces spoken language, OCR, frames and temporal context, but the binary nature of the task keeps the leaderboard more compact than in video categorization.

6.4.1. Soft Evaluation

Table 12 presents the soft-soft results for Task 3.1. A total of 58 participant runs were submitted. Of these, 56 outperformed the majority-class baseline, whose ICM-Soft Norm was 0.2740. Participant ICM-Soft Norm scores ranged from 0.5940 to 0.2359, with a mean of 0.454 and a standard deviation of 0.092.

Table 12
Complete leaderboard for EXIST 2026 Task 3.1, Sexism Identification in Videos, under the soft-soft evaluation protocol. The table reports all 61 evaluated runs, including 58 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	2.8488	1.0000	0.1962
German&Blai_2	1	0.5358	0.5940	1.2302
Dominican-Papi_2	2	0.4586	0.5805	0.8305
German&Blai_3	3	0.4549	0.5798	0.7784
AIdonnow_1	4	0.4505	0.5791	0.8436
Dominican-Papi_3	5	0.4219	0.5740	0.8401
Dominican-Papi_1	6	0.3331	0.5585	0.9701
NeverChorizoInMyPaella_3	7	0.2675	0.5469	1.3183
Aegis-Athena_2	8	0.2602	0.5457	0.8116
Aegis-Athena_1	9	0.2554	0.5448	0.8051
NeverChorizoInMyPaella_2	10	0.2271	0.5399	1.3819
nodicus_1	11	0.2058	0.5361	0.9823
NeverChorizoInMyPaella_1	12	0.2058	0.5361	1.1717
nodicus_2	13	0.2047	0.5359	0.9655
Aegis-Athena_3	14	0.1877	0.5329	0.7643
nodicus_3	15	0.1284	0.5225	0.8499
German&Blai_1	16	0.1242	0.5218	0.7996
CarmenMachi_1	17	0.1234	0.5217	2.1138
XORIK_1	18	0.1085	0.5190	0.7450
tai6le_2	19	0.1074	0.5188	1.1215
CarmenMachi_2	20	0.0562	0.5099	2.1578
AILS-NTUA_3	21	0.0349	0.5061	0.9374
XORIK_2	22	0.0087	0.5015	0.8152
Jow_3	23	0.0013	0.5002	0.9858

Continued on next page

Table 12 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
AILS-NTUA_2	24	−0.0093	0.4984	0.9429
AILS-NTUA_1	25	−0.0201	0.4965	0.9314
DanCar_1	26	−0.0245	0.4957	1.2998
TurboAbuelas_3	27	−0.0251	0.4956	0.8214
GXX_2	28	−0.1024	0.4820	0.9278
Jow_1	29	−0.1258	0.4779	1.0718
CEVS_2	30	−0.1789	0.4686	1.1226
Jow_2	31	−0.1874	0.4671	0.8422
thebocadillos_1	32	−0.2764	0.4515	1.0563
TurboAbuelas_2	33	−0.2794	0.4510	0.7936
Thinkpol_1	34	−0.2796	0.4509	0.8127
Marsad_2	35	−0.3836	0.4327	0.8819
TurboAbuelas_1	36	−0.4183	0.4266	0.8104
CarmenMachi_3	37	−0.4364	0.4234	0.7916
GXX_1	38	−0.4527	0.4205	0.8363
UniKo_1	39	−0.4559	0.4200	0.8343
tai6le_3	40	−0.4728	0.4170	0.8594
UniKo_3	41	−0.5028	0.4117	0.8463
tai6le_1	42	−0.5232	0.4082	0.8279
LosMejores_1	43	−0.5647	0.4009	1.0620
UniKo_2	44	−0.5923	0.3960	0.8452
GXX_3	45	−0.5935	0.3958	0.8203
Castrillo_1	46	−0.7335	0.3713	0.9417
Mapaches-Hiperactivos_2	47	−0.7712	0.3646	0.8970
Mapaches-Hiperactivos_3	48	−0.8003	0.3595	0.8952
Mapaches-Hiperactivos_1	49	−0.8686	0.3475	0.8954
Marsad_1	50	−0.9274	0.3372	0.9019
ELiRF-UPV_1	51	−0.9707	0.3296	1.1172
ELiRF-UPV_2	52	−0.9824	0.3276	1.0598
AIdonnow_3	53	−1.0528	0.3152	0.8938
CEVS_1	54	−1.0591	0.3141	0.9132
Marsad_3	55	−1.1539	0.2975	0.9255
AIdonnow_2	56	−1.2271	0.2846	0.9496
Test_majority-class	57	−1.2877	0.2740	4.4285
CEVS_3	58	−1.4889	0.2387	0.9939
TeamOne_1	59	−1.5046	0.2359	1.0001
Test_minority-class	60	−2.0051	0.1481	5.5402

German&Blai_2 ranked first with ICM-Soft Norm=0.5940. This team did not have a Working Note in the available set, so its methodological discussion must rely on the submitted system description rather than on a paper. The top unique-team spread in ICM-Soft Norm was only 4.83 percentage points, making this one of the most competitive soft leaderboards.

6.4.2. Hard Evaluation

Table 13 reports the hard-hard results. The task received 73 participant runs, and all of them outperformed the strongest simple baseline, whose ICM-Hard Norm was 0.2858. Participant ICM-Hard Norm scores ranged from 0.6669 to 0.2887, with a mean of 0.561 and a standard deviation of 0.069.

Table 13
Complete leaderboard for EXIST 2026 Task 3.1, Sexism Identification in Videos, under the hard-hard evaluation protocol. The table reports all 76 evaluated runs, including 73 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and F1 YES. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
Test_gold	0	0.9907	1.0000	1.0000
ELiRF-UPV_2	1	0.3307	0.6669	0.7627
German&Blai_3	2	0.3284	0.6657	0.7487
ELiRF-UPV_1	3	0.3195	0.6613	0.7504
Aegis-Athena_2	4	0.3096	0.6563	0.7528
XORIK_2	5	0.3007	0.6518	0.7543
Aegis-Athena_1	6	0.2864	0.6446	0.7484
Dominican-Papi_1	7	0.2743	0.6384	0.7339
German&Blai_1	8	0.2632	0.6328	0.7210
Aegis-Athena_3	9	0.2582	0.6303	0.7308
Dominican-Papi_3	10	0.2569	0.6296	0.7360
nodicus_3	11	0.2538	0.6281	0.7239
German&Blai_2	12	0.2536	0.6280	0.7165
1931_2	13	0.2531	0.6277	0.7357
AIdonnow_1	14	0.2456	0.6239	0.7150
Dominican-Papi_2	15	0.2390	0.6206	0.7153
AIdonnow_3	16	0.2185	0.6103	0.7047
TurboAbuelas_3	17	0.2130	0.6075	0.7157
TurboAbuelas_2	18	0.2129	0.6074	0.7083
XORIK_1	19	0.2110	0.6065	0.7303
2Tazas_1	20	0.2092	0.6056	0.7154
AIdonnow_2	21	0.1994	0.6006	0.6957
nodicus_1	22	0.1991	0.6005	0.7027
UniKo_3	23	0.1932	0.5975	0.7185
LosMejores_1	24	0.1907	0.5962	0.7068
2Tazas_3	25	0.1907	0.5962	0.7080
TurboAbuelas_1	26	0.1872	0.5945	0.7012
nodicus_2	27	0.1854	0.5936	0.7038
CarmenMachi_1	28	0.1772	0.5894	0.6946
UniKo_1	29	0.1755	0.5886	0.7072
GXX_1	30	0.1586	0.5801	0.6869
Jow_2	31	0.1562	0.5788	0.6815
Thinkpol_1	32	0.1560	0.5788	0.6726
tai6le_3	33	0.1460	0.5737	0.6679
GXX_3	34	0.1432	0.5723	0.6690
CarmenMachi_2	35	0.1410	0.5712	0.6812
CarmenMachi_3	36	0.1394	0.5703	0.6933
tai6le_2	37	0.1327	0.5670	0.6619
NeverChorizoInMyPaella_3	38	0.1312	0.5662	0.6852
LosMejores_3	39	0.1304	0.5658	0.6655
Jow_3	40	0.1261	0.5637	0.6643
thebocadillos_1	41	0.1200	0.5605	0.6906

Continued on next page

Table 13 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
Castrillo_1	42	0.1104	0.5557	0.6536
UniKo_2	43	0.1072	0.5541	0.7151
NeverChorizoInMyPaella_1	44	0.1064	0.5537	0.6722
NeverChorizoInMyPaella_2	45	0.1055	0.5533	0.6786
Ontinyent 1931_1	46	0.1033	0.5521	0.6573
Mapaches-Hiperactivos_3	47	0.1033	0.5521	0.6573
2Tazas_2	48	0.0935	0.5472	0.6678
tai6le_1	49	0.0930	0.5470	0.6667
CEVS_2	50	0.0862	0.5435	0.6516
AILS-NTUA_2	51	0.0850	0.5429	0.6645
DanCar_1	52	0.0831	0.5419	0.6335
AILS-NTUA_3	53	0.0788	0.5397	0.6589
Marsad_2	54	0.0774	0.5391	0.6656
AILS-NTUA_1	55	0.0652	0.5329	0.6708
PAULU_3	56	0.0597	0.5301	0.6521
BeingChilling_2	57	0.0597	0.5301	0.6510
PAULU_2	58	0.0584	0.5295	0.6579
PAULU_1	59	0.0554	0.5280	0.6510
Jow_1	60	0.0492	0.5248	0.6441
BeingChilling_1	61	0.0400	0.5202	0.6305
LosMejores_2	62	0.0365	0.5184	0.6514
DanCar_2	63	0.0242	0.5122	0.6979
BeingChilling_3	64	0.0181	0.5091	0.6508
Mapaches-Hiperactivos_2	65	−0.0666	0.4664	0.6246
CEVS_1	66	−0.0836	0.4578	0.5911
TeamOne_1	67	−0.0854	0.4569	0.5603
Marsad_1	68	−0.0898	0.4547	0.6227
Marsad_3	69	−0.0958	0.4517	0.6007
UMUTeam_1	70	−0.1237	0.4376	0.5644
Mapaches-Hiperactivos_1	71	−0.1828	0.4077	0.6065
GXX_2	72	−0.2881	0.3546	0.5226
CEVS_3	73	−0.4187	0.2887	0.0199
Test_majority-class	74	−0.4244	0.2858	0.0000
Test_minority-class	75	−0.6036	0.1954	0.6117

ELiRF-UPV_2 led the hard evaluation with ICM-Hard Norm=0.6669. The top of the ranking was very tight: the best-to-fifth unique-team spread in ICM-Hard Norm was only 2.85 percentage points. This suggests that several different video-processing strategies can reach similar hard-label performance in binary video identification.

System Behavior and Methods. The working notes support a strong trend toward reducing videos to robust textual and visual evidence. ELiRF-UPV reports a combination of multimodal fusion and training-free few-shot LLM prompting over video frames and ASR-derived information, and its Working Note identifies the few-shot pipeline as particularly effective for video identification [54]. AIDONNOW reconstructs videos into rich text using cleaned ASR, OCR and frame descriptions generated by a vision-language model; its paper concludes that this text reconstruction is consistently useful, while

physiological fusion requires more careful modelling [46]. Aegis-Athena uses structured VLM descriptions, transcripts, raw video features and physiological signals in a cross-attention architecture, and reports strong hard-evaluation ranks for Task 3.1 [50]. Marsad, which also addresses Task 3.1, reports a notable English-Spanish gap for CLIP-based video models [51]. These reports indicate that strong video systems relied less on raw video alone and more on extracting stable cues from speech, OCR, frames and generated scene descriptions.

6.5. Task 3.2: Source Intention in Videos

Task 3.2 evaluates source-intention classification in sexist TikTok videos. Among the video subtasks, this is the clearest case of cross-protocol stability, since the same team leads both soft and hard evaluation.

6.5.1. Soft Evaluation

Table 14 presents the soft-soft results for Task 3.2. There were 48 participant runs and three reference systems. A total of 45 runs outperformed the majority-class baseline, whose ICM-Soft Norm was 0.1663. Participant ICM-Soft Norm scores ranged from 0.4590 to 0.0705, with a mean of 0.295 and a standard deviation of 0.095.

Table 14
Complete leaderboard for EXIST 2026 Task 3.2, Source Intention in Videos, under the soft-soft evaluation protocol. The table reports all 51 evaluated runs, including 48 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	4.6948	1.0000	0.2550
AIdonnow_1	1	−0.3847	0.4590	1.0930
Dominican-Papi_2	2	−0.5080	0.4459	1.2472
Dominican-Papi_3	3	−0.6813	0.4274	1.1624
Dominican-Papi_1	4	−0.7860	0.4163	1.2824
XORIK_3	5	−0.7931	0.4155	0.9988
XORIK_1	6	−0.8585	0.4086	0.9959
Aegis-Athena_3	7	−0.9847	0.3951	1.0202
Aegis-Athena_2	8	−1.0841	0.3845	1.1268
XORIK_2	9	−1.1473	0.3778	1.0061
nodicus_2	10	−1.1479	0.3777	1.5681
nodicus_1	11	−1.1713	0.3753	1.4215
CarmenMachi_3	12	−1.2418	0.3677	2.9512
Aegis-Athena_1	13	−1.2488	0.3670	1.0477
CarmenMachi_2	14	−1.2875	0.3629	3.1706
nodicus_3	15	−1.3116	0.3603	1.4193
CarmenMachi_1	16	−1.3221	0.3592	3.1211
NeverChorizoInMyPaella_2	17	−1.3753	0.3535	2.6836
NeverChorizoInMyPaella_1	18	−1.4342	0.3473	2.3084
NeverChorizoInMyPaella_3	19	−1.4777	0.3426	2.1615
AILS-NTUA_3	20	−1.5763	0.3321	1.2847
AILS-NTUA_1	21	−1.7564	0.3129	1.2992
CEVS_2	22	−1.7617	0.3124	3.0839

Continued on next page

Table 14 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Jow_3	23	−1.7634	0.3122	1.3396
Thinkpol_1	24	−1.7910	0.3093	1.2380
AILS-NTUA_2	25	−1.8038	0.3079	1.3146
Jow_1	26	−1.8075	0.3075	1.4644
TurboAbuelas_3	27	−1.9132	0.2962	1.1402
TurboAbuelas_2	28	−1.9553	0.2918	1.0823
TurboAbuelas_1	29	−2.0626	0.2803	1.0810
Jow_2	30	−2.0952	0.2769	1.2130
AIdonnow_3	31	−2.2360	0.2619	1.3135
thebocadillos_1	32	−2.2440	0.2610	2.2986
GXX_2	33	−2.3477	0.2500	1.3226
AIdonnow_2	34	−2.4218	0.2421	1.4578
UniKo_1	35	−2.5466	0.2288	1.2049
Castrillo_1	36	−2.6550	0.2172	1.2949
CEVS_1	37	−2.6844	0.2141	3.7889
UniKo_3	38	−2.7470	0.2074	1.2328
UniKo_2	39	−2.7964	0.2022	1.2280
ELiRF-UPV_1	40	−2.8787	0.1934	3.6428
GXX_3	41	−2.8869	0.1925	1.1880
ELiRF-UPV_2	42	−2.9173	0.1893	3.5028
TeamOne_1	43	−3.0156	0.1788	2.8040
CEVS_3	44	−3.1162	0.1681	4.4321
German&Blai_3	45	−3.1294	0.1667	1.5871
Test_majority-class	46	−3.1337	0.1663	4.4354
LosMejores_1	47	−3.5975	0.1169	1.4928
German&Blai_2	48	−3.7766	0.0978	4.0591
German&Blai_1	49	−4.0329	0.0705	4.6903
Test_minority-class	50	−15.4368	0.0000	8.8286

AIdonnow_1 ranked first with ICM-Soft Norm=0.4590. The best-to-fifth unique-team spread in ICM-Soft Norm was 8.13 percentage points, so the top soft systems were more separated than in Task 3.1. The result suggests that modelling disagreement about communicative intent in videos is more difficult than binary identification, but still tractable when systems extract reliable textual and visual cues.

6.5.2. Hard Evaluation

Table 15 reports the hard-hard results. The hard setting received 68 participant runs, all of which outperformed the majority-class baseline, whose ICM-Hard Norm was 0.2155. Participant ICM-Hard Norm scores ranged from 0.6262 to 0.2196, with a mean of 0.481 and a standard deviation of 0.095.

Table 15
Complete leaderboard for EXIST 2026 Task 3.2, Source Intention in Videos, under the hard-hard evaluation protocol. The table reports all 71 evaluated runs, including 68 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and Macro F1. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Test_gold	0	1.3244	1.0000	1.0000
AIdonnow_2	1	0.3344	0.6262	0.7065
AIdonnow_3	2	0.3165	0.6195	0.7001
XORIK_3	3	0.2876	0.6086	0.6966
ELiRF-UPV_1	4	0.2843	0.6073	0.6901
Dominican-Papi_3	5	0.2751	0.6038	0.6749
XORIK_2	6	0.2730	0.6031	0.6889
ELiRF-UPV_2	7	0.2675	0.6010	0.6789
Ontinyent 1931_2	8	0.2554	0.5964	0.6888
XORIK_1	9	0.2488	0.5939	0.6825
German&Blai_3	10	0.2445	0.5923	0.6726
Ontinyent 1931_1	11	0.2194	0.5828	0.6832
Dominican-Papi_2	12	0.2012	0.5760	0.6508
German&Blai_1	13	0.1840	0.5695	0.6598
AIdonnow_1	14	0.1809	0.5683	0.6385
German&Blai_2	15	0.1754	0.5662	0.6570
Dominican-Papi_1	16	0.1687	0.5637	0.6494
Aegis-Athena_1	17	0.1603	0.5605	0.6483
TurboAbuelas_2	18	0.1576	0.5595	0.6471
nodicus_1	19	0.1563	0.5590	0.6309
Aegis-Athena_3	20	0.1553	0.5586	0.6512
TurboAbuelas_1	21	0.1271	0.5480	0.6320
2Tazas_1	22	0.1218	0.5460	0.6530
Aegis-Athena_2	23	0.1191	0.5450	0.6364
nodicus_3	24	0.1104	0.5417	0.6193
nodicus_2	25	0.0621	0.5235	0.6045
CarmenMachi_2	26	0.0425	0.5160	0.6240
TurboAbuelas_3	27	0.0343	0.5130	0.6078
UniKo_3	28	0.0239	0.5090	0.6035
Thinkpol_1	29	0.0047	0.5018	0.5781
UniKo_1	30	−0.0017	0.4994	0.5924
CarmenMachi_3	31	−0.0046	0.4983	0.6149
2Tazas_3	32	−0.0126	0.4952	0.5935
BeingChilling_1	33	−0.0128	0.4952	0.6085
CarmenMachi_1	34	−0.0131	0.4951	0.6099
DanCar_2	35	−0.0201	0.4924	0.5798
AILS-NTUA_3	36	−0.0276	0.4896	0.5781
NeverChorizoInMyPaella_2	37	−0.0470	0.4823	0.5746
AILS-NTUA_1	38	−0.0588	0.4778	0.5751
NeverChorizoInMyPaella_3	39	−0.0687	0.4741	0.5729
AILS-NTUA_2	40	−0.0696	0.4737	0.5752
UniKo_2	41	−0.0759	0.4713	0.5828

Continued on next page

Table 15 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Jow_3	42	−0.0789	0.4702	0.5699
BeingChilling_2	43	−0.0938	0.4646	0.5746
NeverChorizoInMyPaella_1	44	−0.0940	0.4645	0.5680
Jow_2	45	−0.0971	0.4633	0.5567
Jow_1	46	−0.1366	0.4484	0.5571
2Tazas_2	47	−0.1412	0.4467	0.5632
CEVS_2	48	−0.1622	0.4388	0.5379
BeingChilling_3	49	−0.2083	0.4214	0.5260
GXX_3	50	−0.2125	0.4198	0.5074
Castrillo_1	51	−0.2208	0.4166	0.5280
Mapaches-Hiperactivos_3	52	−0.2355	0.4111	0.5059
thebocadillos_1	53	−0.2643	0.4002	0.5135
PAULU_3	54	−0.2948	0.3887	0.5208
PAULU_1	55	−0.2970	0.3879	0.5199
DanCar_1	56	−0.3212	0.3788	0.5271
PAULU_2	57	−0.3409	0.3713	0.5135
LosMejores_3	58	−0.3486	0.3684	0.5005
TeamOne_1	59	−0.3604	0.3639	0.4976
GXX_1	60	−0.3626	0.3631	0.4733
LosMejores_1	61	−0.3753	0.3583	0.4876
LosMejores_2	62	−0.3876	0.3537	0.4878
CEVS_1	63	−0.4167	0.3427	0.4744
UMUTeam_1	64	−0.4232	0.3402	0.5024
Mapaches-Hiperactivos_2	65	−0.4275	0.3386	0.4560
Mapaches-Hiperactivos_1	66	−0.5071	0.3085	0.4466
GXX_2	67	−0.6441	0.2568	0.4702
CEVS_3	68	−0.7428	0.2196	0.2532
Test_majority-class	69	−0.7537	0.2155	0.2375
Test_minority-class	70	−2.4749	0.0000	0.0586

AIdonnow_2 led the hard evaluation with ICM-Hard Norm=0.6262. The top unique-team spread in ICM-Hard Norm was only 2.98 percentage points. The same team led both protocols, but not with the same run, which suggests that the team's general video-to-text pipeline was robust while the best configuration still depended on the evaluation target.

System Behavior and Methods. AIdonnow's Working Note is the central methodological evidence for this subtask. It compares TF-IDF classifiers, XLM-RoBERTa and Qwen3-8B with QLoRA on video representations built from ASR, OCR and generated frame descriptions, and reports first place in Task 3.2 under both official protocols [46]. Importantly, the paper supports the claim that multimodal text reconstruction is useful, but it does not support a simple claim that physiological sensors are the decisive factor. The available system description further indicates that the hard systems used transcript and OCR text without sensor features. Aegis-Athena and ELiRF-UPV also provide working notes describing multimodal video systems with transcripts, generated visual descriptions, frames and, in some configurations, physiological signals [50, 54]. The supported conclusion is therefore protocol-sensitive: sensor channels were explored, but the strongest evidence in Task 3.2 points to high-quality text reconstruction from video and speech as the most stable component.

6.6. Task 3.3: Sexism Categorization in Videos

Task 3.3 is the fine-grained categorization task for TikTok videos. It is the most difficult subtask in the benchmark: both the best soft score and the mean soft score are the lowest among all subtasks, and many systems do not outperform the majority-class baseline.

6.6.1. Soft Evaluation

Table 16 shows the soft-soft results for Task 3.3. There were 50 participant runs. Only 24 outperformed the majority-class baseline, which obtained ICM-Soft Norm=0.0931. Participant ICM-Soft Norm scores ranged from 0.1724 to 0.0000, with a mean of 0.078 and a standard deviation of 0.058.

Table 16
Complete leaderboard for EXIST 2026 Task 3.3, Sexism Categorization in Videos, under the soft-soft evaluation protocol. The table reports all 53 evaluated runs, including 50 participant submissions, ordered by the official rank and showing ICM-Soft, ICM-Soft Norm, and Cross Entropy. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Test_gold	0	8.3833	1.0000	0.5062
NeverChorizoInMyPaella_3	1	−5.4930	0.1724	2.2205
thebocadillos_1	2	−5.5814	0.1671	2.4749
nodicus_3	3	−5.5964	0.1662	1.7565
NeverChorizoInMyPaella_2	4	−5.6924	0.1605	2.5509
NeverChorizoInMyPaella_1	5	−5.6935	0.1604	2.3661
tai6le_3	6	−5.7292	0.1583	1.7919
AIdonnow_2	7	−5.8311	0.1522	1.7218
nodicus_1	8	−5.9308	0.1463	1.7479
AIdonnow_3	9	−6.0975	0.1363	1.7528
AIdonnow_1	10	−6.1562	0.1328	1.7031
Aegis-Athena_2	11	−6.2080	0.1297	1.6874
nodicus_2	12	−6.2676	0.1262	1.6272
CEVS_2	13	−6.3699	0.1201	3.5335
CarmenMachi_2	14	−6.4127	0.1175	3.4689
AILS-NTUA_3	15	−6.5497	0.1094	1.7547
AILS-NTUA_1	16	−6.5768	0.1077	1.8364
AILS-NTUA_2	17	−6.5801	0.1075	1.8386
XORIK_3	18	−6.6244	0.1049	1.5629
XORIK_1	19	−6.6244	0.1049	1.5629
tai6le_1	20	−6.6733	0.1020	1.6297
Aegis-Athena_3	21	−6.6937	0.1008	1.5934
ELiRF-UPV_1	22	−6.7240	0.0990	3.7955
CEVS_1	23	−6.7487	0.0975	2.7991
XORIK_2	24	−6.7780	0.0957	1.5499
Test_majority-class	25	−6.8222	0.0931	4.5052
ELiRF-UPV_2	26	−6.9242	0.0870	3.6934
CEVS_3	27	−6.9481	0.0856	4.5139
German&Blai_3	28	−6.9605	0.0849	1.5266
Thinkpol_1	29	−6.9658	0.0845	1.5619

Continued on next page

Table 16 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
CarmenMachi_3	30	−6.9720	0.0842	3.8940
Jow_3	31	−7.1330	0.0746	1.7253
Jow_1	32	−7.1499	0.0736	1.7975
CarmenMachi_1	33	−7.2390	0.0682	3.2934
tai6le_2	34	−7.3379	0.0624	1.5853
Aegis-Athena_1	35	−7.3978	0.0588	1.5550
Jow_2	36	−7.4666	0.0547	1.6077
German&Blai_2	37	−8.1729	0.0125	1.8027
German&Blai_1	38	−8.2224	0.0096	1.8592
Dominican-Papi_1	39	−10.8079	0.0000	1.5977
TeamOne_1	40	−11.0928	0.0000	3.4721
Test_minority-class	41	−11.6668	0.0000	9.7964
Castrillo_1	42	−11.9906	0.0000	1.7828
Castrillo_2	43	−11.9906	0.0000	1.7828
Dominican-Papi_2	44	−13.6352	0.0000	1.6971
Dominican-Papi_3	45	−14.3178	0.0000	1.7271
UniKo_3	46	−18.1032	0.0000	2.0311
UniKo_1	47	−20.1994	0.0000	2.0732
UniKo_2	48	−20.7637	0.0000	2.1039
LosMejores_1	49	−21.8543	0.0000	2.1835
TurboAbuelas_1	50	−22.6328	0.0000	1.9248
TurboAbuelas_2	51	−24.7430	0.0000	1.9653
TurboAbuelas_3	52	−30.9358	0.0000	1.9824

NeverChorizoInMyPaella_3 ranked first with ICM-Soft Norm=0.1724. Although the best ICM-Soft Norm score is low in absolute terms, the top systems were tightly clustered: the best-to-fifth unique-team spread in ICM-Soft Norm was only 2.02 percentage points. This suggests that the main difficulty was not only choosing the best model family, but the inherent ambiguity and sparsity of fine-grained sexist categories in short videos.

6.6.2. Hard Evaluation

Table 17 presents the hard-hard results. The task received 67 participant submissions, of which 51 outperformed the majority-class baseline, whose ICM-Hard Norm was 0.1916. Participant ICM-Hard Norm scores ranged from 0.4624 to 0.0000, with a mean of 0.293 and a standard deviation of 0.145.

Table 17
Complete leaderboard for EXIST 2026 Task 3.3, Sexism Categorization in Videos, under the hard-hard evaluation protocol. The table reports all 70 evaluated runs, including 67 participant submissions, ordered by the official rank and showing ICM-Hard, ICM-Hard Norm, and Macro F1. The table also includes an official gold-standard reference run, which establishes the maximum value obtained by the ICM-based metrics. It also includes two non-informative baselines (majority-class and minority-class).

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Test_gold	0	1.5453	1.0000	1.0000

Continued on next page

Table 17 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
AIdonnow_1	1	−0.1162	0.4624	0.3750
AIdonnow_2	2	−0.1392	0.4549	0.3999
Aegis-Athena_2	3	−0.1394	0.4549	0.4193
German&Blai_3	4	−0.1743	0.4436	0.4285
AIdonnow_3	5	−0.1817	0.4412	0.3228
nodicus_2	6	−0.1886	0.4390	0.4300
nodicus_3	7	−0.2421	0.4217	0.3493
XORIK_1	8	−0.2662	0.4139	0.3644
XORIK_2	9	−0.2762	0.4106	0.3707
Ontinyent 1931_1	10	−0.2879	0.4068	0.3740
nodicus_1	11	−0.2952	0.4045	0.4214
German&Blai_2	12	−0.2971	0.4039	0.4026
XORIK_3	13	−0.3230	0.3955	0.3411
Dominican-Papi_2	14	−0.3276	0.3940	0.3370
Dominican-Papi_3	15	−0.3280	0.3939	0.3154
Aegis-Athena_1	16	−0.3299	0.3933	0.3449
Ontinyent 1931_2	17	−0.3303	0.3931	0.3879
German&Blai_1	18	−0.3338	0.3920	0.4158
ELiRF-UPV_1	19	−0.3389	0.3903	0.4162
TurboAbuelas_1	20	−0.3404	0.3899	0.4195
NeverChorizoInMyPaella_3	21	−0.3523	0.3860	0.4097
tai6le_2	22	−0.3630	0.3825	0.2656
AILS-NTUA_1	23	−0.3731	0.3793	0.2978
Jow_1	24	−0.3758	0.3784	0.3787
AILS-NTUA_3	25	−0.3812	0.3766	0.2866
tai6le_3	26	−0.3832	0.3760	0.3072
AILS-NTUA_2	27	−0.3877	0.3746	0.2902
Dominican-Papi_1	28	−0.4070	0.3683	0.3727
Jow_3	29	−0.4137	0.3661	0.3485
thebocadillos_1	30	−0.4181	0.3647	0.3362
2Tazas_1	31	−0.4323	0.3601	0.4152
Castrillo_2	32	−0.4530	0.3534	0.2486
Castrillo_1	33	−0.4530	0.3534	0.2486
NeverChorizoInMyPaella_2	34	−0.4552	0.3527	0.3821
ELiRF-UPV_2	35	−0.4625	0.3504	0.4101
Aegis-Athena_3	36	−0.4748	0.3464	0.4100
DanCar_1	37	−0.5031	0.3372	0.3703
UniKo_3	38	−0.5169	0.3328	0.2926
CEVS_2	39	−0.5197	0.3318	0.2477
Jow_2	40	−0.5205	0.3316	0.3084
BeingChilling_2	41	−0.5233	0.3307	0.3007
NeverChorizoInMyPaella_1	42	−0.5281	0.3291	0.3783
UniKo_2	43	−0.5792	0.3126	0.2734
TurboAbuelas_2	44	−0.5794	0.3125	0.4113
CarmenMachi_3	45	−0.6454	0.2912	0.3247
TurboAbuelas_3	46	−0.6637	0.2852	0.3781
CarmenMachi_2	47	−0.6642	0.2851	0.3547

Continued on next page

Table 17 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
BeingChilling_3	48	−0.6785	0.2805	0.3336
BeingChilling_1	49	−0.6938	0.2755	0.2920
UniKo_1	50	−0.7596	0.2542	0.2896
CarmenMachi_1	51	−0.7907	0.2442	0.3374
Test_majority-class	52	−0.9530	0.1916	0.1188
CEVS_3	53	−0.9610	0.1890	0.1187
CEVS_1	54	−1.0029	0.1755	0.2133
GXX_1	55	−1.1204	0.1375	0.2635
PAULU_1	56	−1.1821	0.1175	0.2831
PAULU_3	57	−1.2224	0.1045	0.2829
tai6le_1	58	−1.2690	0.0894	0.3670
PAULU_2	59	−1.3220	0.0723	0.2965
Thinkpol_1	60	−1.4807	0.0209	0.3485
TeamOne_1	61	−1.7926	0.0000	0.2214
2Tazas_3	62	−2.0130	0.0000	0.2920
LosMejores_2	63	−2.1216	0.0000	0.2877
LosMejores_1	64	−2.1216	0.0000	0.2877
2Tazas_2	65	−2.1790	0.0000	0.2951
Mapaches-Hiperactivos_1	66	−2.9765	0.0000	0.1707
Mapaches-Hiperactivos_2	67	−6.6100	0.0000	0.1612
Test_minority-class	68	−6.7467	0.0000	0.0025
Mapaches-Hiperactivos_3	69	−8.0795	0.0000	0.1451

AIdonnow_1 led the hard evaluation with ICM-Hard Norm=0.4624. The best-to-fifth unique-team spread in ICM-Hard Norm was 4.85 percentage points. As in Task 2.3, the soft and hard winners differ, reinforcing the idea that calibrated multi-label distributions and final hard category decisions are not equivalent optimization targets.

System Behavior and Methods. The working notes for Task 3.3 emphasize hierarchy, multimodal evidence and calibration. NeverChorizoInMyPaella describes a sequential cascade training strategy for TikTok videos, fusing OCR text, ASR transcripts and SigLIP visual frames with visual cross-attention, and its paper reports first place in Task 3.3 under soft-soft evaluation [47]. AIdonnow reports first place in Task 3.3 hard evaluation and attributes its overall video performance mainly to multimodal text reconstruction and validation-optimized ensembles, while noting that soft calibration and physiological fusion remain difficult [46]. Aegis-Athena reports a four-stream cross-attention architecture and obtains a strong hard rank in this subtask [50]. Finally, tai6le reports that calibrated text-centred systems can be competitive in soft categorization, even without relying on full video modelling in every submitted variant [55]. These papers support a cautious interpretation: Task 3.3 benefits from combining speech, visible text and visual cues, but the low scores indicate that fine-grained categorization in videos remains weakly solved.

6.7. Cross-task Observations

Across the six subtasks, the strongest evidence points to three broad conclusions. First, the systems that performed well usually transformed multimodal content into a representation where textual meaning, visual context and label uncertainty could interact. For memes, this often meant OCR plus generated descriptions or vision-language embeddings; for videos, it often meant ASR, OCR and frame

descriptions. This pattern is supported by working notes from Ordantis, SafetyNLP, aiwizards, ELiRF-UPV, AIdonnow, Aegis-Athena and NeverChorizoInMyPaella [37, 25, 28, 54, 46, 50, 47].

Second, the results do not support a single global conclusion about physiological or behavioral signals. Several working notes report using EEG, eye-tracking, heart-rate or demographic information, and some report positive uses in particular configurations [30, 32, 36, 50, 44]. Others explicitly exclude sensors, leave them for future work, or report that they require more careful modelling [25, 33, 46]. Therefore, the supported conclusion is mixed: sensor data can provide useful human-centred context, but the leaderboard does not by itself prove a systematic advantage for sensor-based systems across tasks.

Third, the hardest cases are the categorization subtasks, especially Task 3.3. In soft evaluation, only 24 of 50 Task 3.3 participant runs outperformed the strongest simple baseline, and the best ICM-Soft Norm was only 0.1724. In Task 2.3, 73 of 113 soft runs outperformed the strongest baseline, with a best ICM-Soft Norm of 0.3469. These numbers show that modelling fine-grained sexist content and annotator disagreement remains substantially more difficult than binary identification. The working notes suggest promising directions–hierarchical prediction, cascade training, category-aware voting, calibrated probability outputs and better multimodal text reconstruction–but the current results leave considerable room for improvement.

7. Conclusions

Overall, EXIST 2026 confirms that sexism detection in multimedia content remains an active, technically diverse, and still challenging evaluation context. The lab processed 1303 submitted runs from 122 teams. Participation was broader in the meme track (939 runs from 102 teams) than in the video track (364 runs from 29 teams), which is consistent with the higher cost of processing videos and with the additional difficulty introduced by temporal, spoken, visual, and OCR-derived evidence. This difference in participation should not be read only as a matter of dataset size: the working notes show that video systems often required additional preprocessing steps, including ASR, OCR, frame sampling, visual description generation, or multimodal feature fusion [46, 54, 50, 47].

The results also point to clear differences in task difficulty. Under Soft-soft Evaluation, the highest best ICM-Soft Norm score was obtained in Sexism Identification in Memes (0.6158), whereas Sexism Categorization in Videos obtained the lowest best ICM-Soft Norm score (0.1724); under Hard-hard Evaluation, the highest best ICM-Hard Norm score was obtained in Sexism Identification in Memes (0.7387), whereas Sexism Categorization in Videos obtained the lowest best ICM-Hard Norm score (0.4624). In these comparisons, the strongest simple baseline is the best of the two non-informative reference systems: the majority-class baseline, which always predicts the most frequent class, and the minority-class baseline, which always predicts the least frequent class. In Soft-soft Evaluation, 30/30 selected top-five systems and 448/520 submitted systems outperformed this strongest non-informative baseline; in Hard-hard Evaluation, 30/30 selected top-five systems and 695/783 submitted systems outperformed it. The top of the leaderboards was usually well above these non-informative baselines, but the full ranking remained heterogeneous, especially for the categorization subtasks, where lower averages and wider score dispersion suggest that fine-grained sexist-content modelling is still substantially harder than binary identification.

This difficulty is particularly visible in Task 3.3. Only 24/50 submitted soft runs outperformed the strongest non-informative baseline in video categorization, which in this case was the majority-class baseline. The best ICM-Soft Norm score was 0.1724. The corresponding hard evaluation was stronger, with 51/67 runs above the strongest non-informative baseline, again the majority-class baseline, but still obtained the lowest best ICM-Hard Norm among all hard subtasks. The working notes support this quantitative picture: competitive systems for categorization tend to introduce hierarchy-aware modelling, cascade training, category-aware voting, calibration, or richer video-to-text reconstruction rather than relying on a single generic classifier [28, 38, 42, 47, 46, 55]. These results indicate that modelling fine-grained sexist categories, especially under disagreement-aware soft evaluation, remains

the main open challenge.

The comparison between hard and soft evaluation confirms that the two evaluation contexts capture related but not identical system behaviours. The source-intention subtasks were the most stable across contexts, with the same teams leading the meme and video leaderboards in both soft and hard evaluation. By contrast, the winning teams changed in all identification and categorization subtasks, which suggests that these tasks are more sensitive to the way uncertainty is represented and then converted into final labels. Hard evaluation rewards robust discrete decisions, whereas soft evaluation gives more weight to calibrated probabilistic outputs and to the ability to model annotator disagreement. This distinction is also reflected in the working notes: several teams explicitly adapted losses, calibration, target distributions, or hierarchical constraints to the Learning with Disagreement setting [25, 28, 33, 53, 55]. The practical implication is that future systems should not treat hard and soft outputs as interchangeable products of the same classifier, especially for the more ambiguous categorization tasks.

From a methodological perspective, the submitted descriptions show that the most widely used approaches were multimodal fusion, transformer-based text encoders, lexical or handcrafted features, CLIP-style vision-language encoders, and classical machine-learning classifiers. Strong submissions generally relied on combinations of text, visual, OCR, transcript, or frame-level evidence rather than on a single modality. For memes, this means aligning the image with embedded text and pragmatic context; for videos, it means combining what is said, what appears on screen, and what can be inferred from selected frames. LLM prompting, vision-language encoders, transformer text models, classical classifiers over fused features, and ensembling all appear in competitive systems, suggesting that careful feature extraction and fusion remain as important as model scale.

The working notes add more detail to this methodological picture. For memes, strong systems often made the image-text relation more explicit through OCR improvement, generated visual descriptions, vision-language embeddings, prompt optimization, or probabilistic modelling of annotator distributions [25, 37, 35, 28]. For videos, several competitive systems converted the original video into richer textual evidence using ASR, OCR, scene descriptions, and frame-level visual information before classification [46, 54, 50, 47]. These findings support a practical conclusion: in this edition, performance gains often came from making noisy multimedia content easier for downstream classifiers to reason over, rather than from treating the raw image or video stream as sufficient on its own.

A final observation concerns auxiliary human-centred information. Several teams explored sensor signals, such as eye-tracking, heart rate, or EEG, while others also considered demographic information from annotators. The rankings do not show a systematic advantage for sensor-driven systems across tasks. The evidence is therefore best interpreted cautiously: these signals and metadata may add useful complementary information in some settings, particularly for uncertainty-aware soft prediction, but they do not replace strong multimodal text-image or text-video modelling. This interpretation is consistent with recent studies on memes and TikTok videos showing that physiological and behavioural signals can provide complementary information for modelling sexism perception, although their contribution depends on the task, modality, and fusion strategy [18, 19]. This mixed picture is supported by the working notes. Some teams report positive or promising uses of physiological or human-centred features in specific configurations [30, 32, 36, 50, 44], while others exclude them, leave them for future work, or report that they require more careful modelling [25, 33, 46]. Future work should include more explicit ablations, stronger calibration strategies for soft labels, and analyses of disagreement patterns to better understand when auxiliary human-response signals and participant metadata contribute beyond standard multimodal representations.

Taken together, EXIST 2026 suggests that future work should focus on improving how systems identify the signals that support each sexist-content category, especially in videos. The results show that current models still struggle to connect subtle visual, textual, pragmatic, and temporal evidence with fine-grained categories. The soft evaluation context is particularly important here: by following the Learning with Disagreement paradigm, it better represents the subjective nature of sexism perception and the fact that different annotators may legitimately assign different degrees of evidence to the same item. Hard evaluation remains useful for measuring final discrete decisions, but soft evaluation should guide the development of systems that model uncertainty, ambiguity, and disagreement more

faithfully. The leaderboard results and working notes therefore point in the same direction: robust multimedia sexism detection requires not only stronger encoders, but also better representations of category-specific evidence, calibrated modelling of disagreement, and transparent ablation studies that identify which modalities and human-centred signals genuinely contribute to performance.

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/10.13039/501100011033 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).

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] D. Spina, J. Gwizdka, K. Ji, Y. Moshfeghi, J. Mostafa, T. Ruotsalo, M. Zhang, A. Ahmad, S. F. D. A. Lawati, N. Boonprakong, N. Fernando, J. He, O. Hoeber, G. Jayawardena, B.-G. Lee, H. Liu, M. Pike, A. Pirmoradi, B. Nakisa, M. N. Rastgoo, F. D. Salim, F. Scott, S. Sun, H. Tang, D. Towey, M. L. Wilson, Report on the 3rd Workshop on NeuroPhysiological Approaches for Interactive Information Retrieval (NeuroPhysIIR 2025) at SIGIR CHIIR 2025, SIGIR Forum 59 (2025) 1–43. doi:10.1145/3769733.3769740.

[8] K. Ji, D. Hettiachchi, F. D. Salim, F. Scholer, D. Spina, Characterizing Information Seeking Processes with Multiple Physiological Signals, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1006–1017. doi:10.1145/3626772.3657793.

[9] K. Ji, D. Hettiachchi, F. Scholer, F. D. Salim, D. Spina, SenseSeek Dataset: Multimodal Sensing to Study Information Seeking Behaviors, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 9 (2025). doi:10.1145/3749501.

[10] L. Plaza, J. Carrillo-de Albornoz, E. Gomis-Vicent, I. Arcos, M. Aloy-Mayo, P. Rosso, D. Spina, EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media, in:

Advances in Information Retrieval, ECIR '26, Springer Nature Switzerland, Cham, 2026. doi:10.1007/978-3-032-21321-1_40.

[11] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vincent, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media, Lecture Notes in Computer Science (LNCS), Springer, Cham, 2026.

[12] 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: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Thessaloniki, Greece, 2023.

[13] 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: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Grenoble, France, 2024.

[14] 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: J. Carrillo-de Albornoz, A. García Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Madrid, Spain, 2025.

[15] 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.

[16] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, D. Spina, J. Gonzalo, P. Rosso, Overview of EXIST 2022: Sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[17] J. Divya Venkatesh, A. Jaiswal, G. Nanda, Comparing Human Text Classification Performance and Explainability with Large Language and Machine Learning Models Using Eye-Tracking, Scientific Reports 14 (2024) 14295. doi:10.1038/s41598-024-65080-7.

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

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

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

[21] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, 2026. URL: https://arxiv.org/abs/2602.23862. arXiv:2602.23862.

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

[23] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. a. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[24] L. A. H. Miranda, E. M. Arriaga Santana, A. Flores Reyes, L. E. Quintero Cuevas, F. F. López-Ponce, G. Sierra, Detecting Sexism in Memes with Vision-Language Models for EXIST 2026, in: [57], 2026.

[25] B. Tamba, A. Petrescu, C.-O. Truică, SafetyNLP at EXIST 2026: Multimodal Sexism Detection in Memes using LLM-Driven OCR Enhancement and LeWiDi Target Modeling, in: [57], 2026.

[26] S. N. John, T. Ravialagappan, S. Thota, V. Muralidharan, S. K. Shivaanee, FourBytes at EXIST 2026: Cross-Attention Multimodal Fusion for Sexism Detection in Memes, in: [57], 2026.
[27] C. Sun, Y. Lu, L. Kong, Multi-Task Cross-Modal Reasoning with Physiological-Aware Fusion for Sexism Detection in Memes, in: [57], 2026.
[28] M. Fasulo, A. Gravina, L. Tedeschini, L. Babboni, AI Wizards at EXIST 2026: Hierarchical Soft-Label Learning for Multimodal Sexism Identification in Memes, in: [57], 2026.
[29] Y. Chen, L. Kong, A Multimodal Method for Sexism Identification Incorporating Visual Descriptions and Physiological Signals, in: [57], 2026.
[30] A. Rey, alBERT-319 at EXIST 2026: Calibrating a Multimodal Sexism Classifier with Physiological Signals, in: [57], 2026.
[31] E. García-Arias, L. Plaza, J. Carrillo-de Albornoz, UNED-Annotate-Team at EXIST 2026: A Multimodal Multi-Task Approach to Sexism Identification in Memes, in: [57], 2026.
[32] D. Zuniga, Gated Multimodal Fusion with Neurophysiological Signals for Sexism Detection in Memes, in: [57], 2026.
[33] C. Munisamy, K. Raguveer, A. Kuila, BioSentinel at EXIST 2026: Soft-label optimization with xlm-roberta for sexism intent classification in memes, in: [57], 2026.
[34] S.-H. Wu, Y.-F. Huang, CYUT at EXIST 2026: GPT-Enhanced Multilingual Ensemble for Sexism Detection in Memes, in: [57], 2026.
[35] A. Llinares, M. Franco-Salvador, DDAI-UNICC at EXIST-2026 Task 2: Human-Cognitive and Feedback-Driven Prompt Optimization for Sexism Detection, in: [57], 2026.
[36] M. Borgia, M. Viviani, Bicocca_Unimib at EXIST 2026: Human-Centered Multimodal Sexism Detection in Memes with Physiological Signals, in: [57], 2026.
[37] S. Ortiz Montesinos, F. Martínez Gómez, Gemini-Enriched Probabilistic Modeling for Multimodal Sexism Characterization in Memes, in: [57], 2026.
[38] A. names omitted for review, MathAddicts at EXIST 2026: Heterogeneous Vision-Language Ensembles with Diverse PEFT Adapters for Multimodal Sexism Detection, in: [57], 2026.
[39] A. Mier González, OCR-PragSex at EXIST 2026: Text-Based Sexism Identification in Memes Using OCR and Pragmatic Error Analysis, in: [57], 2026.
[40] M. Yermukhamed, B. Milana, CLIP-Based Multimodal Approach for Sexism Identification in Memes: Farabi University at EXIST 2026 Task 2.1, in: [57], 2026.
[41] S. Mandelli, G. Rizzi, A. Saibene, F. Gasparini, E. Fersini, SPECTRA at EXIST2026: Sexism Prediction and Evaluation of Multimedia Content Through Physiological and Gaze Responses Analysis, in: [57], 2026.
[42] M. Bessghaier, W. Zaghouani, Semantica at EXIST 2026: A Two-Stage Multimodal Approach for Sexism Categorization in Memes, in: [57], 2026.
[43] Aryan, S. Banerjee, LLM-Based Zero-Shot Pipeline with Physiological Sensor Fusion for Sexism Detection in Memes, in: [57], 2026.
[44] A.-M. L. Mocanu, S. Mocanu, C.-O. Truică, E.-S. Apostol, Through the Eyes of the Beholder: Biometric and Demographic Conditioning for Multimodal Sexism Detection, in: [57], 2026.
[45] P. Sudharshan, Subjective Sexism Identification in Social Media Memes Using Multimodal Embedding Ensembles, in: [57], 2026.
[46] O. Meneses Albalá, J. Martínez Llinares, AIdonnow at EXIST 2026: Multimodal Text Reconstruction and Sensor Fusion for TikTok Sexism Detection, in: [57], 2026.
[47] G. Di-Palma, R. Hernández-López, NeverChorizoInMyPaella at EXIST 2026: Multimodal Sexism Identification and Categorisation in TikTok Videos via Sequential Cascade Training and Visual Cross-Attention, in: [57], 2026.
[48] C. Marade, M. Ernst, F. Hopfgartner, UniKo at EXIST 2026: Integrating physiological sensor signals into a multimodal fusion architecture for sexism detection in tiktok videos, in: [57], 2026.
[49] A. de Rojas-Ortuño, R. Pan, R. Valencia-García, UMUTeam at EXIST2026: Multimodal Stacking Architecture for Sexism Detection in Short Videos, in: [57], 2026.
[50] R. Pylypiv, M. Szabóová, Aegis Athena at EXIST 2026: Multimodal Sexism Detection in Short Videos via Cross-Attention Fusion, in: [57], 2026.

[51] K. K. Aldous, W. Zaghouani, Team Marsad at EXIST 2026: End-to-End Vision-Language Fine-Tuning for Sexism Identification in Multilingual Memes and TikTok Videos, in: [57], 2026.
[52] K. Chaviaras, M. Lymperaiou, A. Voulodimos, Multimodal Sexism Identification and Characterization using Large Language Models and Gradient Boosting, in: [57], 2026.
[53] J. Santos da Silva, R. P. Lopes, A. G. A. Ali Ibrahim, Lightweight Multimodal Sexism Detection Across Memes and Videos with Demographic-Aware Soft-Label Learning for EXIST 2026, in: [57], 2026.
[54] J. A. García-Real, V. Ahuir, M. J. Castro-Bleda, ELiRF-UPV at EXIST 2026: Multimodal Fusion and Few-Shot LLMs for Sexism Detection in Memes and Videos, in: [57], 2026.
[55] J. Wang, S. Chen, H. Yang, H. Dong, T. Du, tai6le at EXIST 2026: System Description, in: [57], 2026.
[56] S. Yang, T. Chen, J. Yue, G. Cheng, J. Jiao, Z. FU, Reasoning-Aware Multimodal Fusion for Hateful Video Detection, Transactions on Machine Learning Research (2026). URL: https://openreview.net/forum?id=U9KnNiuMu1.
[57] A. Barrón-Cedeño, A. García Seco de Herrera, E. Sánchez Salido, P. Schaer, E. Zangerle, S. MacAvaney (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR Workshop Proceedings (CEUR-WS.org), Jena, Germany, 2026.