Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos

Laura Plaza[1], Jorge Carrillo-de-Albornoz[1], Iván Arcos[2], Paolo Rosso[2,3], Damiano Spina[4], Enrique Amigó[1], Julio Gonzalo[1], Roser Morante[1],

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

Abstract. This paper presents the EXIST 2025 Lab on sexism detection and categorization in social media, which took place at the CLEF 2025 conference and marks the fifth edition of the EXIST Shared Task. Building on the success of previous editions, EXIST 2025 addresses the growing concern over the spread of offensive and discriminatory content targeting women across online platforms, which significantly impacts women's well-being and freedom of expression. The lab comprises nine tasks in two languages (English and Spanish), organized around three core objectives: sexism identification, source intention detection, and sexism categorization. These tasks are applied across three media types—text (tweets), image (memes), and video (TikToks)—offering a multimodal perspective that allows for a deeper understanding of how sexism manifests across different formats and user interactions. As in previous editions, EXIST 2025 adopts the "Learning With Disagreement" paradigm, using annotations from multiple annotators that reflect diverse and at times conflicting viewpoints. This overview describes the task design, datasets, evaluation methodology, participating systems, and results of EXIST 2025, which has surpassed participation expectations with 244 registered teams from 38 countries, 114 teams from 23 countries submitting runs, a total of 873 runs processed, and 33 working notes published.
Warning: Some of the examples included in this paper may contain offensive language and explicit descriptions of sexist behavior, which may be disturbing to the reader.

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

1 Introduction

Sexism refers to prejudice or discrimination based on a person's sex or gender, often manifesting in the belief that one gender is superior to another. It can take many forms, from overt aggression and harassment to subtler behaviors and norms that reinforce inequality. While sexism affects individuals of all genders, it disproportionately impacts women, particularly in digital spaces.

In recent years, online platforms like Twitter and TikTok have become breeding grounds for the proliferation of sexist discourse. On Twitter, sexism often manifests through harassment, trolling, and misogynistic hashtags that normalize discriminatory narratives [13, 24]. TikTok, by contrast, poses unique challenges due to its algorithm-driven content promotion and its popularity among younger audiences. Its recommendation system can generate filter bubbles that reinforce sexist ideologies [23], while visual trends and content moderation disparities contribute to the hypersexualization and objectification of women [10, 15]. These dynamics not only perpetuate traditional gender stereotypes but can also shape the perceptions and behaviors of young users.

To tackle these challenges, the sEXism Identification in Social neTworks (EXIST) campaign was launched in 2021. EXIST is a series of shared tasks and scientific events aimed at identifying, analyzing, and mitigating sexist content on social networks. The first two editions were hosted under the IberLEF forum [32, 33], and focused on textual data. In 2023, EXIST became a CLEF Lab [31], introducing a third task centered on detecting the communicative intention behind sexist messages and adopting for the first time the Learning with Disagreement (LeWiDi) paradigm [35]. This paradigm acknowledges that disagreements among annotators are not noise, but valuable signals that reflect the subjectivity inherent to tasks like sexism detection. The fourth edition of EXIST (2024) expanded the challenge to multimodal data by introducing tasks involving memes. Memes, while often humorous, are increasingly used to spread prejudices under the guise of irony [5, 8, 20, 22]. Their blend of text and image makes them particularly insidious vectors for normalizing sexist stereotypes, especially when humor is used to reduce the perceived harm [12, 16].

EXIST 2025 marks the fifth edition of the challenge and represents its most ambitious iteration yet. Held again as a CLEF Lab,[5] it comprises nine tasks in total—covering three core objectives (sexism identification, source intention detection, and sexism categorization) across three modalities: tweets (text), memes (image), and TikToks (video). This multimodal and bilingual (English and Spanish) design aims to capture the varied ways in which sexism is expressed and interpreted online, enabling researchers to develop AI models that are sensitive to both linguistic and visual cues, as well as the platform-specific dynamics that influence sexist content dissemination.

Throughout its four previous editions, more than 100 teams from universities and companies around the world have participated in EXIST, developing and testing state-of-the-art models to address this pressing social issue. The 2025

[5] https://nlp.uned.es/exist2025

edition continues to foster international participation, with 244 registered teams from 38 countries. Of these, 114 teams from 23 countries submitted valid runs, resulting in a total of 873 system submissions.

In the following sections, we present a detailed overview of the tasks, datasets, annotation process, evaluation methodology, and system results for EXIST 2025.

2 Tasks

The 2025 edition of EXIST features nine tasks, which are described below. The languages addressed are English and Spanish and the datasets are collections of tweets, memes and TikTok videos. For the tasks on TikTok, all the partitions of the dataset are new, whereas for the tasks on tweets and memes we employ the EXIST 2023 and 2024 datasets.

2.1 Task 1.1: Sexism Identification in Tweets

This is a binary classification task where systems must decide whether or not a given tweet expresses sexist ideas because it is sexist itself, it describes a sexist situation, or it criticizes a sexist behavior. The following examples from the dataset show sexist and not sexist messages, respectively.

(1) **Sexist**. *It's less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely.*
(2) **Not sexist**. *Just saw a woman wearing a mask outside spank her very tightly leashed dog and I gotta say I love learning absolutely everything about a stranger in a single instant.*

2.2 Task 1.2: Source Intention in Tweets

This task aims to categorize the message according to the intention of the author. We propose the following ternary classification of tweets:

- **Direct** sexist message. The intention is to write a message that is sexist by itself or incites sexism, as in:
  (3) *A woman needs love, to fill the fridge, if a man can give this to her in return for her services, I don't see what else she needs.*
- **Reported** sexist message. The intention is to report and share a sexist situation suffered by a woman or women in first or third person, as in:
  (4) *I doze in the subway, I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig.*
- **Judgemental** message. The intention is to condemn sexist situations or behaviours, as in:
  (5) *As usual, the woman was the one quitting her job for the family's welfare...*

2.3 Task 1.3: Sexism Categorization in Tweets

Many facets of a woman's life may be the focus of sexist attitudes including domestic role, career opportunities, and sexual image, to name a few. According to this, each sexist tweet must be assigned one or more of the following categories:

- **Ideological and inequality**. It includes messages that discredit the feminist movement. It also includes messages that reject inequality between men and women, or present men as victims of gender-based oppression.
  (6) *#Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.*
- **Stereotyping and dominance**. It includes messages that suggest women are more suitable or inappropriate for certain tasks, and somehow inferior to men.
  (7) *Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.*
- **Objectification**. It includes messages where women are presented as objects apart from their dignity and personal aspects. We also include messages that assume or describe certain physical qualities that women must have in order to fulfill traditional gender roles.
  (8) *No offense but I've never seen an attractive african american hooker. Not a single one.*
- **Sexual violence**. It includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made.
  (9) *I wanna touch your tits..you can't imagine what I can do on your body.*
- **Misogyny and non sexual violence**. It includes expressions of hatred and violence towards women.
  (10) *Domestic abuse is never okay... Unless your wife is a bitch.*

2.4 Task 2.1: Sexism Identification in Memes

As in Task 1.1, this involves a binary classification consisting on deciding whether or not a meme is sexist, as in Figure 1.

2.5 Task 2.2: Source Intention in Memes

As in Task 1.2, this task aims to categorize the meme according to the intention of the author. However, in this task systems should only classify memes in two classes: direct or judgemental, as shown in Figure 2.

2.6 Task 2.3: Sexism Categorization in Memes

This task aims to classify sexist memes according to the categorization provided for Task 1.3. Figure 3 shows one meme of each sexist category.

(a) Sexist meme (b) Non sexist meme

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

(a) Direct (b) Judgemental

Fig. 2: Examples of direct and judgemental memes.

(a) Ideological & inequality (b) Objectification

(c) Stereotyping & dominance (d) Sexual violence (e) Misogyny & non-sexual violence

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

2.7 Task 3.1: Sexism Identification in TikToks

As in Tasks 1.1 and 2.1, systems must determine whether a given short video shared on TikTok is sexist.

2.8 Task 3.2: Source Intention in TikToks

As in Tasks 1.2 and 2.2, this task aims to categorize the short video according to the intention of the author, as direct or judgemental.

2.9 Task 3.3: Sexism Categorization in TikToks

As in Tasks 1.3 and 3.3, this task aims to categorize short videos according to the categorization provided for Task 1.3.

3 Dataset

The EXIST 2025 dataset comprises three types of data: the tweets from the EXIST 2023 dataset, the memes from the EXIST 2024 dataset and a new dataset of TikTok videos. Here, we briefly describe the process followed to curate the TikTok dataset. More details can be found in the extended overview [29]. Moreover, Plaza et al. [31] and [28] provide a detailed description of the tweet and meme datasets, respectively.

3.1 Data Sampling

The data was collected using the Apify's TikTok Hashtag Scraper tool[6], using a previously curated list of 185 Spanish hashtags and 61 English hashtags associated with potentially sexist content.

More than 3,500 videos in English and Spanish were downloaded from different TikTok accounts. Rigorous manual cleaning procedures were applied, ensuring the removal of noise such as ads and duplicates.

3.2 Labeling with Disagreements

The learning with disagreement paradigm was adopted, as in EXIST 2023 and 2024. The annotation was performed by trained annotators, rather than crowd workers as in the previous EXIST editions. The annotation was conducted using Servipoli's service,[7] with eight students organized in pairs consisting of one male and one female student, in order to avoid biases. Each pair was tasked with annotating 1,000 TikTok videos.

[6] https://apify.com/clockworks/tiktok-hashtag-scraper
[7] https://www.servipoli.es/

4 Evaluation Methodology and Metrics

As in EXIST 2024, 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 Measure (ICM) [3]. We also report results using the normalized version, ICM-Soft Norm. For a description of the ICM-Soft metrics, please refer to [28].
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 Tasks 1.1, 2.1 and 3.1, the label selected by more than three annotators is used; for Tasks 1.2, 2.2 and 3.2, the threshold is more than two annotators; and for multilabel Tasks 1.3, 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 [3]. We additionally report a normalized ICM (ICM Norm) and F1.

5 Overview of Approaches

In this section, we provide a concise overview of the approaches presented at EXIST 2025. For a comprehensive description of the systems, please refer to the participant papers and the extended overview [29]. Although 244 teams from 38 different countries registered for participation, the number of participants who finally submitted results were 114, submitting 873 runs. Teams were allowed to participate in any of the nine tasks and submit hard and/or soft outputs. Table 1 summarizes the participation in the different tasks and evaluation contexts.

The evaluation campaign started on February 10, 2025 with the release of the training set. The test set was made available on April 7. Participants were provided with the official evaluation script. Runs had to be submitted by May 23. Each team could submit up to three runs per task.

5.1 Sexism Detection in Tweets

Sexism detection in tweets was predominantly approached through Natural Language Processing (NLP) techniques and neural network-based models. The majority of teams relied on pre-trained large language models (LLMs), such as

Table 1: Runs submitted and teams participating on each EXIST 2025 task.

	Tweets			Memes			TikTok		
	T1.1	T1.2	T1.3	T2.1	T2.2	T2.3	T3.1	T3.2	T3.3
#Runs	223	192	181	26	20	20	75	65	71
#Teams	117	105	101	15	13	12	36	32	33

BERT, RoBERTa, and domain-specific variants like BERTweet or HateBERT, often fine-tuned on the EXIST datasets.

While transformer-based models dominated, a minority of teams used traditional machine learning techniques such as Support Vector Machines (SVM) or Random Forests with TF-IDF, as well as rule-based or lexicon-based methods.

Many teams applied data preprocessing techniques tailored to social media content, including emoji normalization, hashtag segmentation, and URL removal. Data augmentation methods, such as back-translation, synonym replacement, or oversampling of minority classes, were also employed to mitigate class imbalance and improve generalization.

5.2 Sexism Detection in Memes

For memes, the inherently multimodal nature of the data led teams to combine computer vision and text analysis methods. Convolutional Neural Networks (CNNs) and visual feature extractors such as CLIP and ResNet were used to process image data. Meanwhile, embedded text within memes was handled using transformer-based NLP models.

Teams used both early fusion (merging textual and visual embeddings before classification) and late fusion (aggregating predictions from separate pipelines). Although multimodal fusion was key, some teams focused primarily on one modality, revealing diverse strategic preferences.

5.3 Sexism Detection in TikTok Videos

Sexism detection in TikToks required integrating audio, visual, and textual information, making multimodal analysis indispensable.

Despite the complexity of the modality, the dominant methods remained rooted in NLP (particularly for transcript analysis), followed by computer vision models. Multimodal fusion strategies—especially late fusion—were key in top-performing systems, and some teams adopted zero-shot or prompt-based learning using general-purpose LLMs such as GPT-3.

Given TikTok's social dynamics, models were also designed to be sensitive to context, sometimes incorporating meta-information, such as hashtags or background music features.

6 Results

In the next subsections, we report the results of the participants and the baseline systems for each task. We only show the results obtained by the top five teams for each task. For more detailed results (including disaggregated results per language), please refer to the Lab Working Notes [30] or the EXIST website: https://nlp.uned.es/exist2025/.

6.1 Task 1.1: Sexism Identification in Tweets

Soft Evaluation Table 2 presents the results for the soft-soft evaluation for Task 1.1, which attracted a total of 65 participating systems (excluding the gold reference and two baselines). Across all systems, the normalized ICM-Soft scores ranged from nearly 0 to 0.6700, with a mean of 0.490 and a standard deviation of 0.160. A total of 63 systems outperformed the strongest baseline (*Test_majority-class*, all instances labeled as 'NO'), indicating widespread above-baseline performance. The top 5 systems from distinct teams showed a relative difference of only 7.7% between them, pointing to a tightly clustered upper tier.

Table 2: Top 5 systems from different teams in EXIST 2025 Task 1.1. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
Test_gold	–	3.1182	1.0000	0.5472	0
GrootWatch_1	GrootWatch [25]	1.0600	0.6700	0.8893	1
DaniReinon_1	DaniReinon	0.8332	0.6336	0.9726	4
fhstp_3	fhstp [21]	0.7852	0.6259	0.8416	6
BERT-Simpson_2	BERT-Simpson	0.7461	0.6196	1.0426	7
CLiC_1	CLiC [27]	0.7386	0.6184	0.8201	9
Test_majority-class	–	−2.3585	0.1218	4.6115	64
Test_minority-class	–	−3.0717	0.0075	5.3572	66
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**	**Rank**
Test_gold	–	0.9948	1.0000	1.0000	0
Mario_1	Mario [34]	0.6774	0.8405	0.8167	1
CIMAT-GTO_2	CIMAT-GTO [36]	0.6297	0.8165	0.7996	2
warwick_1	warwick [1]	0.6249	0.8141	0.7991	4
CIMAT-CS-NLP_3	CIMAT-CS-NLP [37]	0.6127	0.8079	0.7945	5
BERT-Simpson_1	BERT-Simpson	0.5832	0.7931	0.7832	8
Test_majority-class	–	−0.4413	0.2782	0.0000	154
Test_minority-class	–	−0.5742	0.2114	0.5698	157

Hard Evaluation Table 2 presents the results for the hard-hard evaluation. In this scenario, the annotations from the six annotators are combined into a

single label using the majority vote. 158 systems competed using hard labels. The normalized ICM-Hard scores, bounded between 0 and 1, had a mean of 0.678 and a standard deviation of 0.149 across participants. Nearly all systems (153 out of 158) outperformed the strongest hard-label baseline (*Test_ majority-class*, all instances labeled as 'NO'), confirming robust overall performance in this setting. As seen in Table 2, the spread between the top and fifth-ranked system was only 5.6%, showing strong consistency among leading submissions. Interestingly, two teams from the same institution (CIMAT) appear in the top five with tightly clustered results, while the top system, *Mario_ 1*, led by a modest yet consistent margin.

6.2 Task 1.2: Source Intention in Tweets

Soft Evaluation Table 3 presents the results for the soft-soft evaluation of Task 1.2, focused on identifying the author's intent behind sexist tweets. This task received 54 system submissions in the Soft–Soft evaluation setting. Across participants, scores ranged from 0.0000 to 0.4647, with a mean of 0.182 and a standard deviation of 0.158. A total of 36 systems outperformed the strongest baseline (*Test_ majority-class*, all instances labeled as 'NO'), confirming moderate differentiation among participant quality. The relative difference between the best and the fifth best teams was 15.7%, indicating relatively close performance among the leading systems. This compact spread suggests that most top teams converged on similar probabilistic modeling strategies, even though overall scores were lower than in other tasks due to the increased ambiguity of intent classification.

Hard Evaluation Table 3 presents the hard-hard evaluation results for Task 1.2. In the Hard–Hard setting, 138 systems participated. The normalized ICM-Hard scores, which assess agreement with the aggregated label, ranged from 0.0000 to 0.6623, with an average of 0.3881 and a standard deviation of 0.2278. Impressively, 105 systems outperformed the best hard-label baseline (*Test_ majority-class*, Norm = 0.1910), demonstrating broad effectiveness across submissions. The normalized scores in the group of the top-5 teams were tightly packed, with a maximum relative difference of only 6.4%.

6.3 Task 1.3: Sexism Categorization in Tweets

Soft Evaluation Table 4 displays the results of the soft-soft evaluation Task 1.3, which involves multi-label categorization of sexist content in tweets under the Soft–Soft evaluation paradigm. 51 systems participated, excluding the gold and baseline runs. The normalized ICM-Soft scores spanned from 0.0000 to 0.4417, with a mean of 0.144 and a standard deviation of 0.163. Notably, 25 systems outperformed the strongest baseline (*Test_ majority-class*, all instances labeled as 'NO'), indicating a moderate level of competitiveness. The percentage difference between the best and the fifth team was 22.3%, suggesting a wider

Table 3: Top 5 systems from different teams in EXIST 2025 Task 1.2. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
Test_gold	–	6.2057	1.0000	0.9128	0
GrootWatch_1	GrootWatch [25]	−0.4385	0.4647	1.7711	1
Dandys-de-BERTganim_2	Dandys-de-BERTganim [17]	−0.7261	0.4415	1.3820	4
Cyberpuffs_3	Cyberpuffs [19]	−1.0572	0.4148	2.0396	6
fhstp_2	fhstp [21]	−1.1866	0.4044	1.6566	7
NetGuardAI_1	NetGuardAI [9]	−1.3444	0.3917	1.5681	8
Test_majority-class	–	−5.4460	0.0612	4.6233	37
Test_minority-class	–	−32.9552	0.0000	8.8517	56
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**Macro F1**	**Rank**
Test_gold	–	1.5378	1.0000	1.0000	0
Mario_1	Mario [34]	0.4991	0.6623	0.5692	1
CIMAT-GTO_3	CIMAT-GTO [36]	0.4678	0.6521	0.5555	2
CIMAT-CS-NLP_2	CIMAT-CS-NLP [37]	0.4264	0.6386	0.5461	4
Dandys-de-BERTganim_2	Dandys-de-BERTganim [17]	0.3752	0.6220	0.5522	6
BERTin-Osborne_2	BERTin-Osborne	0.3677	0.6196	0.5453	7
Test_majority-class	–	−0.9504	0.1910	0.1603	106
Test_minority-class	–	−3.1545	0.0000	0.0280	139

performance spread among the leading systems than in other tasks. This reflects the intrinsic difficulty of the multi-label classification task in the soft evaluation setting, where label ambiguity and annotator disagreement must be faithfully captured.

Hard Evaluation For the Hard–Hard evaluation of Task 1.3, a total of 130 systems were submitted. The normalized ICM-Hard values ranged from 0.0000 to 0.6514, with an average of 0.353 and a standard deviation of 0.193. Remarkably, 106 systems surpassed the best baseline (*Test_ majority-class*), demonstrating high effectiveness in predicting the aggregated ground truth labels. The range between the top and fifth systems was only 9.1%, highlighting a tight cluster of top performances. This compact variation among the leaders suggests strong generalization in handling categorical distinctions of sexism in tweets when annotations are aggregated, though continued progress is needed to reach the reliability of the gold standard.

6.4 Task 2.1: Sexism Identification in Memes

Soft Evaluation Table 5 presents the results for the classification of memes as sexist or not sexist. A total of 8 systems participated in the Soft–Soft evaluation. The normalized scores ranged from 0.0650 to 0.5110, with a mean of 0.373 and a standard deviation of 0.149. All but one system outperformed the strongest

Table 4: Top 5 systems from different teams in EXIST 2025 Task 1.3. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Rank
Test_gold	–	9.4686	1.0000	0
GrootWatch_1	GrootWatch [25]	−1.1034	0.4417	1
UMUTeam_2	UMUTeam [26]	−1.6711	0.4118	4
Cyberpuffs_3	Cyberpuffs [19]	−2.4632	0.3699	5
DaniReinon_1	DaniReinon	−2.5655	0.3645	6
A-squared_1	A-squared	−2.9711	0.3431	7
Test_majority-class	–	−8.7089	0.0401	26
Test_minority-class	–	−46.1080	0.0000	53

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	2.1533	1.0000	1.0000	0
Mario_1	Mario [34]	0.6519	0.6514	0.6533	1
CIMAT-GTO_2	CIMAT-GTO [36]	0.5413	0.6257	0.6392	2
NLPDame_3	NLPDame [7]	0.4842	0.6124	0.6335	4
UMUTeam_2	UMUTeam [26]	0.4506	0.6046	0.6262	7
CIMAT-CS-NLP_2	CIMAT-CS-NLP [37]	0.3980	0.5924	0.6125	8
Test_majority-class	–	−1.5984	0.1289	0.1069	108
Test_minority-class	–	−3.1295	0.0000	0.0288	128

baseline (*Test_majority-class*), indicating that most submissions were effective under this probabilistic evaluation. The relative difference between the highest and lowest among the top five submissions from different teams was substantial (87.3%), with a notable drop from the fourth to fifth system. This wide spread suggests room for improvement and divergence in approaches to modeling soft labels in multimodal data.

Hard Evaluation Table 5 presents the results for the hard-hard evaluation of Task 2.1. This task received 18 valid system submissions. The normalized ICM-Hard values ranged from 0.1711 to 0.6877, with an average of 0.471 and a standard deviation of 0.145. Out of these, 16 systems outperformed the *Test_majority-class* baseline. The top five systems from distinct teams showed a moderate performance spread, with a 28.3% relative difference between the highest and lowest performers in this top group. Compared to Task 1.1, the distribution in Task 2.1 reflects greater difficulty in aligning with aggregated hard labels in multimodal settings, likely due to the inherent ambiguity and subjective interpretation of memes.

6.5 Task 2.2: Source Intention in Memes

Soft Evaluation Table 6 presents the results for the classification of memes according to the intention of the author, with the outputs provided as the

Table 5: Top 5 systems from different teams in EXIST 2025 Task 2.1. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Test_gold	–	3.1107	1.0000	0.5852	0
TrankilTwice_1	TrankilTwice [18]	0.0683	0.5110	1.0096	1
surrey-mm-group_1	surrey-mm-group	−0.7061	0.3865	0.9364	4
I2C-UHU-Altair_1	I2C-UHU-Altair [14]	−0.8558	0.3624	0.9469	5
UMUTeam_2	UMUTeam [26]	−0.9623	0.3453	1.0554	6
Test_majority-class	–	−2.3568	0.1212	4.4015	8
Nogroupnocry_1	Nogroupnocry	−2.7060	0.0650	0.6782	9
Test_minority-class	–	−3.5089	0.0000	5.5672	10

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	F1 YES	Rank
Test_gold	–	0.9832	1.0000	1.0000	0
CogniCIC_1	CogniCIC [2]	0.3691	0.6877	0.7810	1
GrootWatch_3	GrootWatch [25]	0.3589	0.6825	0.7740	2
ArcosGPT_1	ArcosGPT [4]	0.3200	0.6627	0.7571	3
TrankilTwice_2	TrankilTwice [18]	0.1667	0.5848	0.7508	5
I2C-UHU-Altair_2	I2C-UHU-Altair [14]	−0.0134	0.4932	0.7125	9
Test_majority-class	–	−0.4038	0.2947	0.6821	17
Test_minority-class	–	−0.6468	0.1711	0.0000	20

probabilities of the different classes. Only 5 systems participated in the Soft–Soft evaluation. The average normalized score across systems was 0.228, with a standard deviation of 0.101. All five systems surpassed the strongest baseline (*Test_ majority-class*). Taking into account the top ranked submissions from distinct teams, the relative difference between the best and the worst among this top-4 was 81.7%, indicating a wide spread in system quality.

Hard Evaluation Table 6 presents the results for the hard-hard evaluation of Task 2.2. We received 15 system submissions. The normalized ICM-Hard metric ranged from 0.0000 to 0.5784, with an average of 0.308 and a standard deviation of 0.169. Thirteen systems outperformed the *Test_ majority-class* baseline, reflecting strong participation despite the challenging nature of the task. Concerning the five best submissions from different teams, the top system outperformed the fifth by 52.7%, a considerable difference suggesting uneven performance across modeling strategies. Nonetheless, the narrow gap among the three leading systems (within 10%) points to the emergence of competitive approaches for intent recognition, even in the presence of aggregated hard annotations derived from subjectively interpreted multimodal inputs.

Table 6: Top 5 systems from different teams in EXIST 2025 Task 2.2. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Test_gold	–	4.7018	1.0000	0.9325	0
UMUTeam_1	UMUTeam [26]	−1.6327	0.3264	1.7316	1
I2C-UHU-Altair_1	I2C-UHU-Altair [14]	−2.0736	0.2795	1.5556	2
surrey-mm-group_1	surrey-mm-group	−2.4423	0.2403	2.0468	3
Nogroupnocry_1	Nogroupnocry	−4.1395	0.0598	0.3164	5
Test_majority-class	–	−5.0745	0.0000	5.5565	6
Test_minority-class	–	−18.9382	0.0000	8.0245	7

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	1.4383	1.0000	1.0000	0
CogniCIC_1	CogniCIC [2]	0.2254	0.5784	0.5634	1
GrootWatch_3	GrootWatch [25]	0.1868	0.5649	0.5513	2
ArcosGPT_1	ArcosGPT [4]	0.0597	0.5208	0.5109	3
NaturalThinker_1	NaturalThinker	−0.5429	0.3113	0.3762	6
I2C-UHU-Altair_1	I2C-UHU-Altair [14]	−0.6519	0.2734	0.2685	7
Test_majority-class	–	−1.0445	0.1369	0.1839	14
Test_minority-class	–	−2.0637	0.0000	0.0697	17

6.6 Task 2.3: Sexism Categorization in Memes

Soft Evaluation Table 7 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as class probabilities. This task received submissions from 6 different teams. Among these, 5 systems outperformed the majority-class baseline. The average normalized ICM-Soft score was 0.151 with a standard deviation of 0.100, indicating a moderately dispersed distribution. The difference in normalized ICM-Soft between the top and bottom systems was 74.8%, showing a meaningful variation even within the upper ranks. Interestingly, all these systems clearly surpassed the worst-performing baseline (`Test_minority-class`), while four significantly exceeded the `Test_majority-class` baseline.

Hard Evaluation Finally, Table 7 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as a single class prediction. A total of 14 systems participated (excluding the gold and baselines). Thirteen of them scored above the best baseline, with an average normalized ICM-Hard of 0.262 and a standard deviation of 0.158. The relative difference between the top and fifth-best system was 59.5%, indicating competitive but not saturated performance across top ranks. As in the soft setting, these systems clearly outperformed both `Test_majority-class` and `Test_minority-class`.

Table 7: Top 5 systems from different teams in EXIST 2025 Task 2.3. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Rank
Test_gold	–	9.4343	1.0000	0
UMUTeam_1	UMUTeam [26]	−4.7791	0.2467	1
I2C-UHU-Altair_1	I2C-UHU-Altair [14]	−5.8210	0.1915	3
surrey-mm-group_1	surrey-mm-group	−6.3848	0.1616	4
Nogroupnocry_2	Nogroupnocry	−8.2621	0.0621	5
Test_majority-class	–	−9.8173	0.0000	6
Test_minority-class	–	−50.0353	0.0000	8

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	2.4100	1.0000	1.0000	0
CogniCIC_1	CogniCIC [2]	0.0244	0.5051	0.5763	1
GrootWatch_3	GrootWatch [25]	−0.0798	0.4834	0.5472	2
ArcosGPT_1	ArcosGPT [4]	−0.4187	0.4131	0.5501	4
I2C-UHU-Altair_1	I2C-UHU-Altair [14]	−0.9958	0.2934	0.4223	6
CLTL_2	CLTL [6]	−1.4243	0.2045	0.3143	8
Test_majority-class	–	−2.0711	0.0703	0.0919	14
Test_minority-class	–	−3.3135	0.0000	0.0318	16

6.7 Task 3.1: Sexism Identification in Videos

Soft Evaluation Table 8 presents the results for classifying videos as sexist or not sexist. The Soft–Soft evaluation of Task 3.1 attracted 34 participating systems. The normalized ICM-Soft values, which reflect alignment with the probabilistic distribution of annotator labels, ranged from 0.1481 to 0.5590. The average normalized score was 0.3584, with a standard deviation of 0.174, indicating considerable variance in system quality. A total of 25 systems outperformed the strongest baseline (*Test_majority-class*, Norm = 0.2740). The difference between the best and worst among the top five teams was approximately 18.2%, reflecting a modest but meaningful spread. Interestingly, most high-scoring systems came from teams with distinct modeling pipelines, suggesting diverse yet effective approaches to handling annotator disagreement in the multimodal context of video classification.

Hard Evaluation Finally, Table 8 presents the results for classifying videos on sexism identification in a hard-hard context. For this task, 41 systems submitted valid runs. Normalized ICM-Hard scores spanned from 0.1954 to 0.6001, with a mean of 0.4913 and a standard deviation of 0.1033. Nearly all participants (39 out of 41) exceeded the majority-class baseline (*Test_majority-class*), showing strong global performance. The top five teams, as can be observed from Table 8, were closely matched, with only a 4.0% difference between the best and lowest performer among the top five.

Table 8: Top 5 systems from different teams in EXIST 2025 Task 3.1. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation					
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**	**Rank**
Test_gold	–	2.8488	1.0000	0.1962	0
LaVellaPremium_2	LaVellaPremium	0.3362	0.5590	1.5731	1
MIARFID ducks_2	MIARFID ducks	0.2968	0.5521	1.7725	3
YesWeEXIST_1	YesWeEXIST	0.2759	0.5484	1.9034	5
profLayton_1	profLayton	0.1779	0.5312	0.9870	7
EmbeddingGuards_1	EmbeddingGuards	−0.2429	0.4574	0.8413	10
Test_majority-class	–	−1.2877	0.2740	4.4285	26
Test_minority-class	–	−2.0051	0.1481	5.5402	31
Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**	**Rank**
Test_gold	–	0.9907	1.0000	1.0000	0
ECA-SIMM-UVa_3	ECA-SIMM-UVa [11]	0.1984	0.6001	0.6935	1
CogniCIC_1	CogniCIC [2]	0.1940	0.5979	0.6835	2
EmbeddingGuards_1	EmbeddingGuards	0.1761	0.5889	0.6841	4
LaVellaPremium_1	LaVellaPremium	0.1563	0.5789	0.6899	5
AIDONTTOKSEXISM_2	AIDONTTOKSEXISM	0.1509	0.5761	0.7013	6
Test_majority-class	–	−0.4244	0.2858	0.0000	40
Test_minority-class	–	−0.6036	0.1954	0.6117	42

6.8 Task 3.2: Source Intention in Videos

Soft Evaluation Table 9 presents the results for the classification of videos according to the intention of the author, with the outputs provided as the probabilities of the different classes. In this task, the 29 participating systems showed normalized ICM-Soft scores that ranged from 0.0000 to 0.3728, with a mean of 0.252 and a standard deviation of 0.084. A total of 26 systems surpassed the strongest baseline (*Test_ majority-class*), indicating a generally competitive field. The difference between the best and the fifth ranked systems from distinct teams was modest, at 12.0%, revealing a cluster of high-performing submissions.

Hard Evaluation Table 9 presents the results for the hard-hard evaluation of Task 3.2. The normalized ICM-Hard scores for the 36 systems submitted ranged from 0.0000 to 0.5018, with a mean of 0.375 and a standard deviation of 0.116. Most systems (33 out of 36) outperformed the *Test_ majority-class* baseline. The best systems from five different teams showed a relative difference between the highest and lowest normalized scores of only 4.3%, reflecting a tight performance range. Interestingly, while the average performance remains moderate, the consistency among top runs suggests that author intent in video—despite its multimodal complexity—can be reliably modeled when annotations are aggregated, albeit with room for improving discriminatory power across subtle categories.

Table 9: Top 5 systems from different teams in EXIST 2025 Task 3.2. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation

System	Team	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Test_gold	–	4.6948	1.0000	0.2550	0
MIARFID ducks_2	MIARFID ducks	−1.1940	0.3728	1.7731	1
EXISTencialCrisis_1	EXISTencialCrisis	−1.3535	0.3558	3.0998	3
profLayton_1	profLayton	−1.3821	0.3528	1.4974	4
YesWeEXIST_1	YesWeEXIST	−1.6151	0.3280	1.5712	6
LaVellaPremium_1	LaVellaPremium	−1.6159	0.3279	1.3258	7
Test_majority-class	–	−3.1337	0.1663	4.4354	27
Test_minority-class	–	−15.4368	0.0000	8.8286	31

Hard-hard Evaluation

System	Team	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Test_gold	–	1.3244	1.0000	1.0000	0
CogniCIC_1	CogniCIC [2]	0.0048	0.5018	0.5623	1
jdsanroj_2	jdsanroj	−0.0068	0.4974	0.5781	2
profLayton_1	profLayton	−0.0283	0.4893	0.5902	3
LaVellaPremium_1	LaVellaPremium	−0.0487	0.4816	0.5742	4
EmbeddingGuards_1	EmbeddingGuards	−0.0529	0.4800	0.5738	5
Test_majority-class	–	−0.7537	0.2155	0.2375	34
Test_minority-class	–	−2.4749	0.0000	0.0586	38

6.9 Task 3.3: Sexism Categorization in Videos

Soft Evaluation Table 10 presents the results for classifying videos based on the aspects of women being attacked, with outputs provided as class probabilities. A total of 34 participant systems were submitted for this task. The normalized ICM-Soft scores ranged from 0.0000 to 0.1593, with a mean of 0.051 and standard deviation of 0.052. The majority baseline achieved a normalized ICM score of 0.0931, and was outperformed by 4 systems, while the minority baseline was not surpassed by any system. The top 5 systems from different teams achieved normalized ICM-Soft scores between 0.1593 and 0.0931. The relative difference between the best and the fifth-ranked system within this top group was 41.6%. Despite the low overall values, a meaningful gap between systems can be observed, which underlines the difficulty of probabilistic categorization in multi-class scenarios over multimodal video content.

Hard Evaluation Finally, Table 7 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as a single class prediction. This task attracted 41 participant systems . Normalized ICM-Hard scores spanned from 0.0000 to 0.3765, with a mean of 0.243 and standard deviation of 0.116. A total of 30 systems outperformed the majority baseline, while 13 did better than the minority baseline. The top 5 systems from distinct teams achieved normalized ICM-Hard scores ranging from 0.3765

Table 10: Top 5 systems from different teams in EXIST 2025 Task 3.3. The complete leaderboard for the task is available on the EXIST website https://nlp.uned.es/exist2025/, in the 'Results' section.

Soft-soft Evaluation				
System	**Team**	**ICM-Soft**	**ICM-Soft Norm**	**Rank**
Test_gold	–	8.3833	1.0000	0
EXISTencialCrisis_1	EXISTencialCrisis	−5.7131	0.1593	1
AIDONTTOKSEXISM_1	AIDONTTOKSEXISM	−6.0447	0.1395	2
LaVellaPremium_1	LaVellaPremium	−6.2730	0.1259	4
biasedmodels_1	biasedmodels	−6.5149	0.1114	6
Test_majority-class	–	−6.8222	0.0931	9
profLayton_2	profLayton	−6.9313	0.0866	10
Test_minority-class	–	−11.6668	0.0000	28

Hard-hard Evaluation					
System	**Team**	**ICM-Hard**	**ICM-Hard Norm**	**Macro F1**	**Rank**
Test_gold	–	1.5453	1.0000	1.0000	0
YesWeEXIST_1	YesWeEXIST	−0.3816	0.3765	0.2667	1
profLayton_3	profLayton	−0.3849	0.3755	0.3648	2
KeTEAM_1	KeTEAM	−0.3869	0.3748	0.3031	3
ScalaR_1	ScalaR	−0.4102	0.3673	0.2533	5
jdsanroj_1	jdsanroj	−0.4373	0.3585	0.2516	7
Test_majority-class	–	−0.9530	0.1916	0.1188	31
Test_minority-class	–	−6.7467	0.0000	0.0025	41

to 0.3585, showing a very tight performance band with only a 4.78% relative difference between the highest and the lowest scoring among them.

6.10 Global View

Figure 4 shows the results of Cross Entropy (horizontal axes) and normalized ICM-Soft (vertical axes). All the plots include the gold standard with maximum score. The first row (Tasks 1.1, 2.1, and 3.1), corresponds to **sexism detection** tasks, i.e., binary single-label classification on texts, images and video, respectively. The baseline approaches consisting of labeling everything as the majority class or as the minority class are marked in blue and red, respectively.

In terms of both Cross Entropy and ICM-Soft, the results of these two baselines fall below those of the other participant runs, indicating that the proposed systems contribute some informative value. Only in the case of the video task (Task 3.1) are there some runs that fall below the baseline in terms of ICM. This may be due to the fact that ICM penalizes false information based on class frequency.

Another observation is that, while high ICM values imply high Cross Entropy values, the reverse is not true, with several runs accumulating good performance (low scores) according to Cross Entropy but low ICM scores. This may be due,

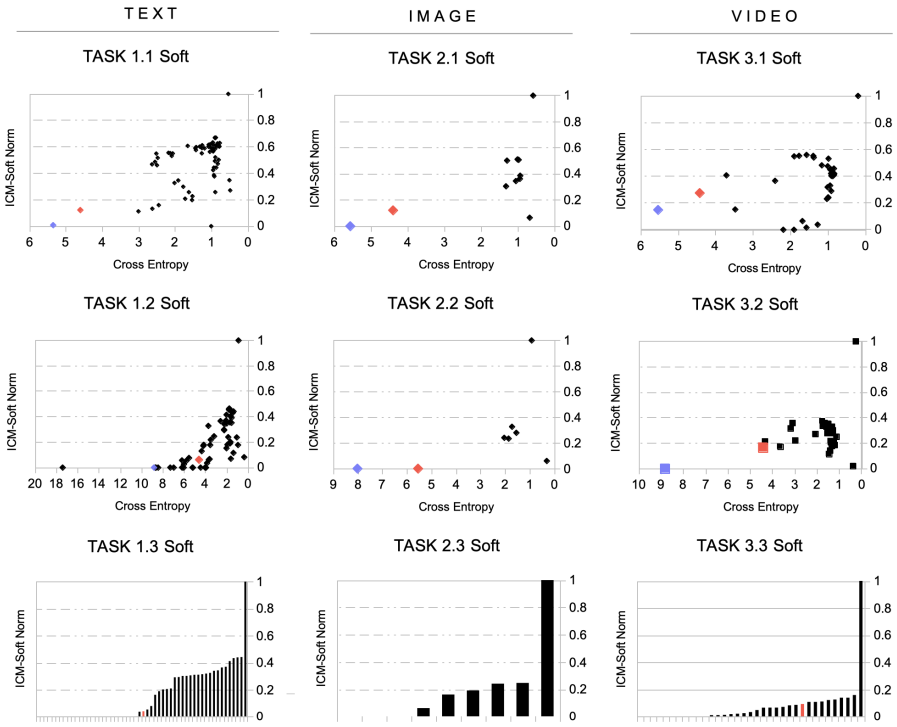

Fig. 4: ICM-Soft and Cross Entropy results across tasks.

among other factors, to the fact that ICM considers not only the similarity of the assigned values for each class, but also the distribution of classes throughout the corpus. In any case, in terms of ICM, there remains a significant gap between the best-performing systems and the perfect solution. The gap is notably larger for the image and video tasks (Tasks 1.2 and 1.3).

The second row corresponds to **intent detection** tasks. These are hierarchical classification tasks with an initial YES/NO decision and two or three sub-classes for the YES category. In this case, there is also an accumulation of runs with high performance in Cross Entropy but low ICM, suggesting that the second metric captures additional aspects. Most runs outperform the baselines, but the gap between the best run and the perfect output in terms of ICM is larger and in sexism detection, indicating a higher complexity of the task .

Finally, the third row corresponds to hierarchical multi-label classification tasks involving multiple **categories of sexism**. In this case, since the tasks are multi-label, the Cross Entropy metric is not applicable. The plots show system rankings ordered from lowest to highest ICM. An interesting finding is that, in this case, many of the runs—including the minority-class baseline—do not

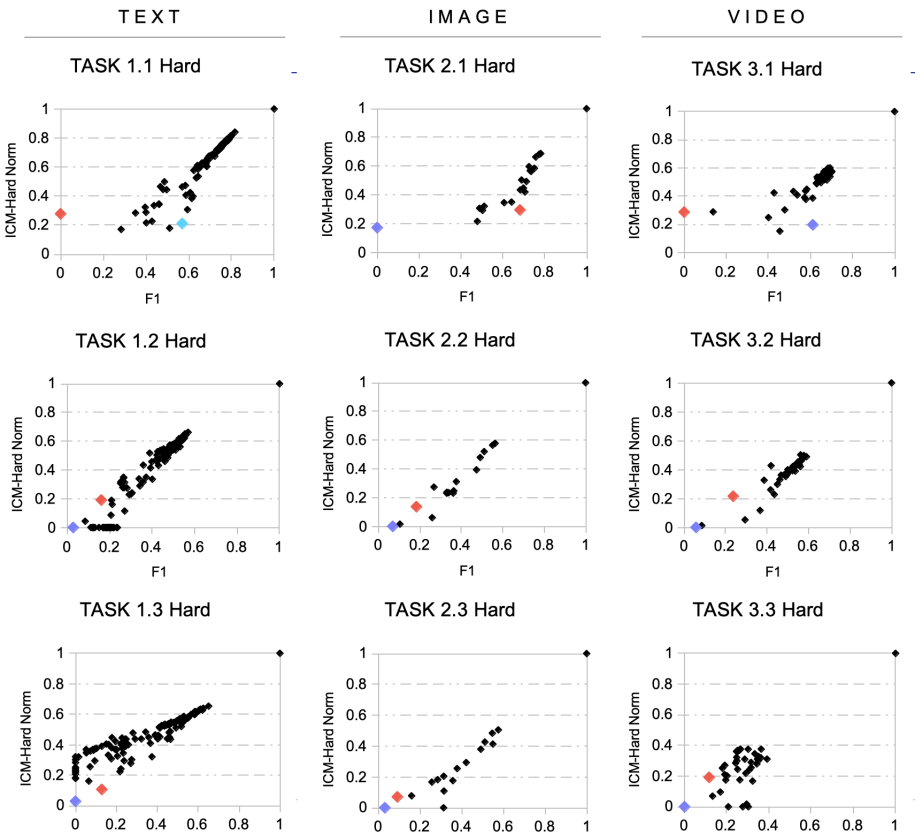

Fig. 5: ICM-Hard and F1 results across tasks.

surpass the zero threshold in normalized ICM. This suggests that the outputs, in terms of information content, do not outperform the empty output. In other words, the amount of noisy information exceeds the amount of useful information. As the number of categories increases and the task requires capturing annotation ambiguity (multi-label classification), the gap between the best run and the perfect output increases significantly compared to the previous tasks.

On the other hand, Figure 5 displays evaluation results for the **hard evaluation versions**, in which the assignment of items to classes depends on whether different thresholds of annotator agreement are met. The plot shows F1 scores for the positive class in the first row (sexism identification), and the average F1 score across all classes for the remaining tasks. The vertical axes show the results for ICM-Hard.

In general, a strong correlation between both metrics can be observed above a certain score threshold. This is because both F1 and ICM take class specificity or frequency within the corpus into account.

Again, most runs outperform the baselines. Moreover, by observing the gap between the best run and the ideal output, we can see that task difficulty increases as we move to setups with more classes, multi-labeling, or hierarchical structures (rows). An increase in task difficulty is also observed as we move from text-based tasks (first column), to image (second column), and video (third column).

7 Conclusions

The objective of the EXIST challenge is to encourage research on the automated detection and modeling of sexism in online environments, with a specific focus on social networks. The EXIST 2025 Lab, held as part of CLEF, attracted 114 participant teams, and received a total of 873 runs. Participants adopted a wide range of approaches, including vision transformer models, data augmentation through automatic translation, data duplication, utilization of data from past EXIST editions, multilingual language models, Twitter-specific language models, and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis. While many systems opted for the traditional approach of providing only hard labels as outputs, a significant number of systems leveraged the multiple annotations available in the dataset, and provided soft outputs, proving that there is an increasing interest by the research community in developing systems able to deal with disagreements and with different perspectives.

Acknowledgments. This work has been financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. It has also been financed by the Spanish Ministry of Science and Innovation (project FairTransNLP (PID2021-124361OB-C31 and PID2021-124361OB-C32)) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe, and by the Australian Research Council, ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005).

References

[1] Alajmi, A., Pergola, G.: Leveraging Model Confidence and Diversity: A Multi-Stage Framework for Sexism Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

[2] Alcantara, T., Garcia-Vazquez, O., Calvo, H., Valdez-Rodríguez, J.E.: CogniCIC at EXIST 2025: Identifying Sexist Content in Text and Visual Media using Transformers and Generative AI Models. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

[3] Amigó, E., Delgado, A.: Evaluating Extreme Hierarchical Multi-label Classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. vol. Volume 1: Long Papers, p. 5809–5819. ACL, Dublin, Ireland (2022)
[4] Arcos, I.: Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[5] Billig, M.: Humour and hatred: the racist jokes of the Ku Klux Klan. Discourse & Society 12(3), 267–289 (2014)
[6] Britez, A., Markov, I.: CLTL at EXIST 2025: Identifying Sexist Memes Using an Ensemble of Shallow and Transformer Models. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[7] Christodoulou, C.: NLPDame at EXIST: Sexism Categorization in Tweets via Multi-Head Multi-Task Models, LLM & RAG Voting Synergy. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[8] Chulvi, B., Fontanella, L., Labadie, R., P, R.: Social or Individual Disagreement? Perspectivism in the Annotation of Sexist Jokes. In: Proc. of NLPerspectives 2023: 2nd Workshop on Perspectivist Approaches to Disagreement in NLP, co-locotaed with ECAI-2023 (2023)
[9] Cotelin, M.D., Truică, C.O., Apostol, E.S.: NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[10] Davis, S.E.: Objectification, sexualization, and misrepresentation: Social media and the college experience. Social Media + Society 4(3) (2018)
[11] Fernández, D., Amigó, E., Cardeñoso, V.: ECA-SIMM-UVa at EXIST 2025: A Segmentation Oriented Approach to Sexism Detection in TikTok Videos Based on a "One Is Enough" Paradigm. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[12] Gasparini, F., Rizzi, G., Saibene, A., Fersini, E.: Benchmark Dataset of Memes with Text Transcriptions for Automatic Detection of Multi-modal Misogynistic Content. Data in Brief 44, 108526 (2022)
[13] Gil Bermejo, J.L., Martos Sánchez, C., Vázquez Aguado, O., García-Navarro, E.B.: Adolescents, ambivalent sexism and social networks, a conditioning factor in the healthcare of women. Healthcare (Basel) 9(6), 721 (Jun 2021)
[14] Guerrero-García, M., Carrillo García, F., Mata, J., Pachón-Álvarez, V.: I2C-UHU-Altair at EXIST2025: Multimodal Sexism Detection and Classification Using Advanced Vision-Language Models BLIP2 and Qwen, Large Language Models, and Learning with Disagreement Frameworks. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[15] Harriger, J., Thompson, J., Tiggemann, M.: TikTok, TikTok, the time is now: Future directions in social media and body image. ody Image B(44), 222–226 (2023)
[16] Hodson, G., Rush, J., MacInnis, C.C.: A Joke Is Just a Joke (except When It Isn't): Cavalier Humor Beliefs Facilitate the Expression of Group Dominance Motives. Journal of Personality and Social Psychology 99(4), 660–682 (2010)

[17] Hurtado, M., Tarrasó, A.: Dandys-de-BERTganim at EXSIST 2025: a Multi-task Learning Architecture for Sexism Identification. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[18] Italiani, P., Maqbool, F., Gimeno-Gómez, D., Fersini, E., Martínez-Hinarejos, C.D.: TrankilTwice at EXIST2025: Detecting Sexism in Memes under Multi-Lingual Settings. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[19] Khursheed, M.S., Hasan Abidi, S.R., Faisal Sikandar, S., Zahra, S., Alvi, F., Samad, A.: Sexism Identification in Tweets Using Ensembles & Augmentation: A Multilingual Approach. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[20] Labadie-Tamayo, R., Chulvi, B., Rosso, P.: Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter. In: Procesamiento del Lenguaje Natural (SEPLN). pp. 383–395. No. 71 (2023)
[21] Labadie-Tamayo, R., Böck, A.J., Slijepčević, D., Chen, X., Babic, A., Zeppelzauer, M.: FHSTP@EXIST 2025 Benchmark: Sexism Detection with Transparent Speech Concept Bottleneck Models. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[22] Mendiburo-Seguel, A., Ford, T.E.: The Effect of Disparagement Humor on the Acceptability of Prejudice. Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues pp. No Pagination Specified–No Pagination Specified (2019)
[23] Morales Rodríguez, G., Lopez-Figueroa, J.: The portrayal of women in media. Journal of Student Research 13(2) (2024)
[24] NewStatesman: Social media and the silencing effect: Why misogyny online is a human rights issue. https://bit.ly/3n3ox68 (nd), last accessed 18 Oct 2023
[25] Nowakowski, N., Calogiuri, L., Egyed-Zsigmond, E., Nurbakova, D., Erbani, J., Calabretto, S.: Automatic Sexism Detection on Social Networks: Classification of Tweets and Memes. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[26] Pan, R., Bernal Beltrán, T., García Díaz, J.A., Valencia-Garcia, R.: UMUTeam at EXIST 2025: Multimodal Transformer Architectures and Soft-Label Learning for Sexism Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[27] Pastells, P., Vázquez, M., Farrús, M., Taulé, M.: CLiC at EXIST 2025: Combining Fine-tuning and Prompting with Learning with Disagreement for Sexism Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[28] Plaza, L., Carrillo-de Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., Spina, D.a.: 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)

[29] Plaza, L., Carrillo-de Albornoz, J., Arcos, I., Rosso, P., Spina, D., Amigó, E., Gonzalo, J., Morante, R.: Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview). In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2025)
[30] Plaza, L., Carrillo-de Albornoz, J., Arcos, I., Rosso, P., Spina, D., Amigó, E., Gonzalo, J., Morante, R.: Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview). In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[31] Plaza, L., de Albornoz, J.C., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview). In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). vol. 497, pp. 813–854. CEUR Working Notes (2023)
[32] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: Sexism identification in social networks. Procesamiento del Lenguaje Natural 67, 195–207 (2021)
[33] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Spina, D., Gonzalo, J., Rosso, P.: Overview of EXIST 2022: Sexism identification in social networks. Procesamiento del Lenguaje Natural 69, 229–240 (2022)
[34] Tian, L., Trippas, J.R., Rizoiu, M.A.: Mario at EXIST 2025: A Simple Gateway to Effective Multilingual Sexism Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[35] Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., Poesio, M.: SemEval-2021 task 12: Learning with disagreements. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 338–347. Association for Computational Linguistics, Online (Aug 2021)
[36] Villarreal Haro, K., Sanchez-Vega, F., Pastor López Monroy, A.: Knowledge Expansion Guided by Justification for Improved Sexism Categorization. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
[37] Villarreal Haro, K., Segura Gómez, G., Tavarez Rodríguez, J., Sánchez Vega, F., Pastor López Monroy, A.: Leveraging Reasoning of Auto-Revealed Insights via Knowledge Injection and Evolutionary Prompting for Sexism Analysis. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)