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

Notebook for the EXIST Lab at CLEF 2025

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

[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 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 disagreement, 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 [1, 2]. 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 [3], while visual trends and content moderation disparities contribute to the hypersexualization and objectification of

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
lplaza@lsi.uned.es (L. Plaza); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); iarcgab@etsinf.upv.es (I. Arcos); prosso@dsic.upv.es (P. Rosso); damiano.spina@rmit.edu.au (D. Spina); enrique@lsi.uned.es (E. Amigó); julio@lsi.uned.es (J. Gonzalo); r.morant@lsi.uned.es (R. Morante)

women [4, 5]. 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 [6, 7], and focused on textual data. In 2023, EXIST became a CLEF Lab [8], 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 [9]. 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 [10, 11, 12, 13]. 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 [14, 15].

EXIST 2025 marks the fifth edition of the challenge and represents its most ambitious iteration yet. Held again as a CLEF Lab,[1] 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 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 (see Section 3). For the tasks on TikTok videos, all the partitions of the dataset are new, whereas for the tasks on tweets and memes we employ the EXIST 2023 and 2024 datasets, respectively.

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:

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

- **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.

(a) Sexist meme (b) Non sexist meme

Figure 1: Examples of sexist and not sexist memes.

(a) Direct (b) Judgemental

Figure 2: Examples of direct and judgemental memes.

(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.

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.

2.7. Task 3.1: Sexism Identification in TikToks

As in Tasks 1.1 and 2.1, systems must determine whether short videos shared on TikTok are sexist.

2.8. Task 3.2: Source Intention in TikToks

As in Tasks 1.2 and 2.2, this task aims to categorize TikTok short videos 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 sexism categories 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. Plaza et al. [8] and [16] provide a detailed description of the tweets and memes datasets, respectively. Here we provide a summarized description of the three datasets.

3.1. Data Sampling

3.1.1. EXIST 2023 Tweets Dataset

We first collected different popular expressions and terms, both in English and Spanish, commonly used to underestimate the role of women in our society. These expressions were later used as seeds to retrieve Twitter data. To mitigate the **seed bias**, we have also gathered other common hashtags and expressions less frequently used in sexist contexts to ensure a balanced distribution between sexist/not sexist expressions. This first set of seeds contains more than 400 expressions.

The set of seeds was then used to extract tweets in English and Spanish (more than 8,000,000 tweets were downloaded). The crawling was performed during the period from the September 1, 2021 till September 30, 2022. 100 tweets were downloaded for each seed per day (no retweets and promotional tweets were included). To ensure an appropriate balance between seeds, we removed those with less than 60 tweets. The final set of seeds contains 183 seeds for Spanish and 163 seeds for English.

To mitigate the **terminology and temporal bias**, the final sets of tweets were selected as follows: for each seed, approximately 20 tweets were randomly selected within the period from 1st September 1, February 28, 2022 for the training set, taking into account a representative temporal distribution among tweets of the same seed. Similarly, 3 tweets per seed were selected for the development set within the period from 1st to 31st May of 2022, and 6 tweets per seed within the period from August 1, 2022 to September, 30 2022 were selected for the test set. Only one tweet per author was included in the final selection to avoid **author bias**. Finally, tweets containing less than 5 words were removed. As a result, we have more than 3,200 tweets per language for the training set, around 500 per language for the development set, and nearly 1,000 tweets per language for the test set.

3.1.2. EXIST 2024 Memes Dataset

We first curated a lexicon of terms and expressions leading to sexist memes. The set of seeds encompasses diverse topics and contains 250 terms, with 112 in English and 138 in Spanish. The terms were used as search queries on Google Images to obtain the top 100 images. Rigorous manual cleaning procedures were applied, defining memes and ensuring the removal of noise such as textless images, text-only images, ads, and duplicates. The final set consists of more than 3,000 memes per language.

Since the proportion of memes per term was heterogeneous, we discarded the most unbalanced seeds and made sure that all seeds have at least five memes. To avoid introducing **selection bias**, we randomly selected memes, ensuring the appropriate distribution per seed. As a result, we have 2,000 memes per language for the training set and 500 memes per language for the test set.

3.1.3. TikTok Dataset

The data was collected with the Apify's TikTok Hashtag Scraper tool.[2] 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.

The collected TikTok videos were divided into training and test sets following a chronological and author-based partitioning strategy. This approach ensured temporal coherence while preventing data leakage. To achieve this, authors present in the training set were excluded from the test set, preventing the model from learning author-specific patterns and enhancing its generalization capabilities. Additionally, each hashtag (seed) was required to contribute a minimum number of videos, ensuring a more uniform distribution across the dataset. The final selection of videos was conducted randomly but maintained a temporal distribution to ensure diversity and avoid overrepresentation of any specific time period.

3.2. Datasets Size

3.2.1. EXIST 2023 Tweets Dataset

The dataset consists of three partitions per language. The distribution of tweets per partition and language is shown in Table 1.

Table 1
Number of **tweets** in the EXIST 2023 dataset per partition and language.

	Training	Development	Test	Total
Spanish	3,660	549	1,098	5,307
English	3,260	489	978	7,727
Total	6,920	1,038	2,076	10,034

3.2.2. EXIST 2024 Memes Dataset

The memes dataset is provided in two partitions per language, training and test. The distribution per partition and language is shown in Table 2.

3.2.3. TikTok Dataset

The TikTok dataset consists of three partitions per language. The distribution of tweets per partitions is shown in Table 3.

[2] https://apify.com/clockworks/tiktok-hashtag-scraper

Table 2
Number of **memes** in the EXIST 2024 dataset per partition and language.

	Training	Test	Total
Spanish	2,034	540	2,573
English	2,010	513	2,523
Total	4,044	1,044	5,096

Table 3
Number of **TikTok videos** in the EXIST 2025 dataset per partition and language.

	Training	Test	Total
Spanish	1,524	304	1,828
English	1,000	370	1,370
Total	2,524	674	3,198

3.3. Labeling with Disagreements

The LeWiDi paradigm was adopted to label the TikTok videos, in the same way that it was adopted to label the tweets and memes datasets for EXIST 2023 and 2024, respectively. Differently from previous EXIST editions, the annotation was performed by trained annotators, instead of crowd workers. The annotation was conducted using Servipoli's service,[3] 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.

4. Evaluation Methodology and Metrics

As in EXIST 2023 and 2024, we have carried out a **soft evaluation** and a **hard evaluation**. The soft evaluation relates to the LeWiDi paradigm and is intended to measure the ability of the model to capture disagreements, by considering the probability distribution of labels in the output as a soft label and comparing it with the probability distribution of the annotations. The hard evaluation is the standard paradigm and assumes that a single label is provided by the systems for every instance in the dataset.

From the point of view of evaluation metrics, the tasks can be described as follows:

- Tasks 1 and 4 (sexism identification): binary classification, monolabel.
- Tasks 2 and 5 (source intention): multiclass hierarchical classification, monolabel. The hierarchy of classes has a first level with two categories, sexist/not sexist, and a second level for the sexist category with three mutually-exclusive subcategories: direct/reported/judgemental. A suitable evaluation metric must reflect the fact that a confusion between not sexist and a sexist category is more severe than a confusion between two sexist subcategories.
- Tasks 3 and 6 (sexism categorization): multiclass hierarchical classification, multilabel. Again the first level is a binary distinction between sexist/not sexist, and there is a second level for the sexist category that includes five subcategories: ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence. These classes are not mutually exclusive: a tweet may belong to several subcategories at the same time.

The LeWiDi paradigm can be considered in both sides of the evaluation process:

[3] https://www.servipoli.es/

- **The ground truth.** In a hard evaluation setting, the variability in the human annotations is reduced by selecting one and only one gold category per instance, the hard label. In a soft evaluation setting, the gold standard label for one instance is the set of all the human annotations existing for that instance. Therefore, the evaluation metric incorporates the proportion of human annotators that have selected each category (soft labels). Note that in Tasks 1, 2, 4 and 5, which are monolabel problems, the sum of the probabilities of each class must be one. But in Task 3, which is multilabel, each annotator may select more than one category for a single instance. Therefore, the sum of probabilities of each class may be larger than one.
- **The system output.** In a hard, traditional setting, the system predicts one or more categories for each instance. In a soft setting, the system predicts a probability for each category, for each instance. The evaluation score is maximized when the probabilities predicted match the actual probabilities in a soft ground truth.

In EXIST 2025, for each of the tasks, two types of evaluation have been performed:

1. **Soft-soft evaluation**. For systems that provide probabilities for each category, we perform a soft-soft evaluation that compares the probabilities assigned by the system with the probabilities assigned by the set of human annotators. The probabilities of the classes for each instance are calculated according to the distribution of labels and the number of annotators for that instance. We use a modification of the original ICM metric (Information Contrast Measure [17]), ICM-Soft (see details below), as the official evaluation metric in this variant and we also provide results for the normalized version of ICM-Soft (ICM-Soft Norm).
2. **Hard-hard evaluation**. For systems that provide a hard, conventional output, we perform a hard-hard evaluation. To derive the hard labels in the ground truth from the different annotators' labels, we use a probabilistic threshold computed for each task. As a result, for Tasks 1 and 4, the class annotated by more than 3 annotators is selected; for Tasks 2 and 5, the class annotated by more than 2 annotators is selected; and for Tasks 3 ad 6 (multilabel), the classes annotated by more than 1 annotator are selected. The instances for which there is no majority class (i.e., no class receives more probability than the threshold) are removed from this evaluation scheme. The official metric for this task is the original ICM, as defined by [17]. We also report a normalized version of ICM (ICM Norm) and F1 (F1_{YES}). In Tasks 1 and 4, we use F1 for the positive class. In Tasks 2, 3, 5 and 6, we use the macro-average of F1 for all classes (Macro F1). Note, however, that F1 is not ideal in our experimental setting: although it can handle multilabel situations, it does not take into account the relationships between classes. In particular, a confusion between not sexist and any of the sexist subclasses, and a confusion between two of the sexist subclasses, are penalized equally.

ICM is a similarity function that generalizes Pointwise Mutual Information (PMI), and can be used to evaluate outputs in classification problems by computing their similarity to the ground truth. The general definition of ICM is:

$$\text{ICM}(A, B) = \alpha_1 IC(A) + \alpha_2 IC(B) - \beta IC(A \cup B)$$

Where $IC(A)$ is the Information Content of the instance represented by the set of features A. ICM maps into PMI when all parameters take a value of 1. The general definition of ICM by [17] is applied to cases where categories have a hierarchical structure and instances may belong to more than one category. The resulting evaluation metric is proved to be analytically superior to the alternatives in the state of the art. The definition of ICM in this context is:

$$\text{ICM}(s(d), g(d)) = 2IC(s(d)) + 2IC(g(d)) - 3IC(s(d) \cup g(d))$$

Where $IC()$ stands for Information Content, $s(d)$ is the set of categories assigned to document d by system s, and $g(d)$ the set of categories assigned to document d in the gold standard. The score for

a perfect output ($s(d) = g(d)$) is the gold standard Information Content ($IC(g(d))$). The score for a zero-information system (no category assignment) is $-IC(g(d))$. We use these two boundaries for normalisation purposes, truncating to 0 the scores lower than $-IC(g(d))$.

As there is not, to the best of our knowledge, any current metric that fits hierarchical multilabel classification problems in a LeWiDi scenario, we have defined an extension of ICM (ICM-soft) that accepts both soft system outputs and soft ground truth assignments. ICM-soft works as follows: first, we define the Information Content of a single assignment of a category c with an agreement v to a given instance as the probability of instances in the gold standard to exceed the agrement level v for the category c:

$$IC(\{\langle c, v\rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\})$$

In order to estimate IC, we compute the mean and deviation of the agreement levels for each class across instances, and applying the cumulative probability over the inferred normal distribution. In the case of zero variance, we must consider that the probability for values equals or below the mean is 1 (zero IC) and the probability for values above the mean must be smoothed. But this is not the case of the EXIST datasets.

Due to the multi-label and hierarchical nature of the classification task,for each classification instance, the gold standard, the system output and their unions ($IC(s(d))$ $IC(g(d))$ and $IC(s(d))Ug(d))$ are sets of category assignments. The union of the assignments (i.e. $s(d))Ug(d))$ is calculated as fuzzy sets, i.e. the maximum values., in order to estimate information content, we apply a recursive function similar to the one described by Amigó and Delgado [17] for assignment sets and avoid the redundant information of parent categories.

$$\begin{aligned} IC\left(\bigcup_{i=1}^{n}\{\langle c_i, v_i\rangle\}\right) &= IC(\langle c_1, v_1\rangle) + IC\left(\bigcup_{i=2}^{n}\{\langle c_i, v_i\rangle\}\right) \\ &- IC\left(\bigcup_{i=2}^{n}\{\langle \text{lca}(c_1, c_i), min(v_1, v_i)\rangle\}\right) \end{aligned} \tag{11}$$

where $\text{lca}(a, b)$ is the lowest common ancestor of categories a and b.

5. Overview of Approaches

This section offers an overview of the methodological approaches submitted to EXIST 2025.

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 4 summarizes the participation in the different tasks and evaluation contexts.

Table 4
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	223.0	192.0	181.0	26.0	20.0	20.0	75.0	65.0	71.0
#Teams	117.0	105.0	101.0	15.0	13.0	12.0	36.0	32.0	33.0

Table 5 summarizes the number of system submissions and participating teams for each of the EXIST 2025 tasks, disaggregated by evaluation type (hard vs. soft labels). Overall, hard-label evaluations were more popular across all tasks, and the distribution of teams remained relatively balanced between evaluation types, particularly in the TikTok subtasks.

Next, we provide a brief overview of the approaches submitted by participants, organized by task groups according to the type of input they target—tweets, memes, or videos. This structure allows

Table 5
Runs submitted and teams participating in each EXIST 2025 task (by evaluation type).

	Tweets			Memes			TikTok		
	T1.1	T1.2	T1.3	T2.1	T2.2	T2.3	T3.1	T3.2	T3.3
#Runs (hard)	158.0	138.0	130.0	18.0	15.0	14.0	41.0	36.0	39.0
#Runs (soft)	65.0	54.0	51.0	8.0	5.0	6.0	34.0	29.0	32.0
#Teams (hard)	87.0	79.0	76.0	10.0	9.0	8.0	20.0	17.0	18.0
#Teams (soft)	30.0	26.0	25.0	5.0	4.0	4.0	16.0	15.0	15.0

for a clearer comparison of modeling strategies across different modalities and highlights trends and innovations specific to each content type.

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 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.

5.4. Summary of Approaches per Team

Next we provide a summary of the methodological approaches followed by the EXIST 2025 teams that submitted a description paper for the Working Notes. We start by the teams that participated only in some or all subtasks of Task 1 on processing **tweets**.

ANLP-Uniso [18] uses the mT5 model for contextual embeddings and a system that integrates several machine learning and deep learning classifiers, including both traditional models (Logistic Regression, SVM) and neural networks (RNN, GRU, hybrid FNN+GRU). To enhance classification accuracy, they apply extensive preprocessing, feature normalization, dimensionality reduction via PCA, and data balancing techniques such as SMOTE and class weighting.

NLPDame [19] addresses Sub-task 1.3 with a methodology that includes fine-tuning twelve transformer LLMs within a tailored multi-head and multi-task model architecture that employs CLS, mean, and max pooling for multi-label text classification. The multi-head architecture is chosen to deal with multilinguality, while the multi-task architecture incorporates sentiment analysis to enhance the multi-label classification process. The methodology also involves utilizing the open-source multilingual LLM Llama-3.2-3B-Instruct and prompt engineering to classify tweets. Additionally, a method incorporating RAG (Retrieval Augmented Generation), chain-of-thought reasoning and annotators' profiles was used to provide contextual information within the LLM prompt engineering framework. A majority voting system was submitted that includes the predictions from (i) the twelve Transformer models with LLM prompt engineering, and (ii) the twelve transformer models with LLM prompt engineering, including chain-of-thought and annotators' profiles, along with RAG. Various loss functions and thresholds were applied, as well as the use of class positive weights to tackle class imbalance.

ECORBI-UPV [20] leverages semantic embeddings generated using pre-trained models from Google's Generative AI suite, evaluated on both frozen and fine-tuned forms. For classification they use traditional machine learning models, such as Random Forest, SVM, and MLP.

Mumul03 [21] employs ModernBERT-large and incorporates demographic information from the annotator such as gender, ethnicity, age, and other attributes into the model input. By modeling individual annotator perspectives and aggregating predictions across submodels, they aim at capturing the subjectivity in annotations.

Fosu-students [22] reformulates the binary classification problem of Task 1 into a seven-class task. They implement ModernBERT-large with layered learning rate decay for hierarchical feature optimization. The model is enhanced with Supervised Contrastive Learning (SCL) to improve discrimination of nuanced sexism expressions through metric learning. Their architecture incorporates: (1) Task reformulation from binary to fine-grained seven-class prediction, (2) ModernBERT's memory-efficient attention mechanisms for long-context understanding, and (3) Hybrid CE+SCL loss (λ=0.9) for robust representation learning.

Warwick [23] develops a hybrid detection framework that integrates the outputs of multiple neural language models, each encoding different perspectives on the task. Their system combines fine-tuned monolingual transformers (BERTweet for English, RoBERTuito for Spanish) with instruction-tuned LLMs such as Claude 3 Sonnet and LLaMA3-70B-Instruct. These models are combined within a confidence-based multi-stage pipeline: high-confidence predictions from task-specialized models are preserved, while uncertain instances are routed to general-purpose LLMs for zero-shot classification. This dynamic strategy combines high-confidence predictions from specialized models with broader judgments from instruction-tuned LLMs.

CLiC [24] employs BERT fine-tuning for Task 1.1 and DSPy-based prompt optimization for Tasks 1.2 and 1.3. They explore BERT-based methods for Task 1.1 and contrasting prompt-based methods, including variants with annotator information and RAG, for the subsequent tasks.

NetGuardAI [25] experiments with several transformer-based models, including DeBERTa, mDeBERTa, XLM-RoBERTa, Detoxify, and HateBERT, alongside three levels of text preprocessing: Light, Classic, and Aggressive Cleaning. Although they tested various data augmentation strategies, such as translation-based augmentation using Meta AI's NLLB model and pseudo-labeling with the EDOS dataset, the final submitted system does not include these enhancements.

EquityExplorer-2.0 [26] proposes a pipeline that combines label-aware translation, domain-adaptive pre-training, and ensemble learning. A central component of their system is a prompt-based Spanish-to-English translation step, designed to preserve the tone and task-relevant semantics of the original message, selectively incorporating label cues during training. They aimed at enabling the use of high-performance monolingual models, while maintaining semantic fidelity across languages. They further

adapt DeBERTa-v3-Large and RoBERTa-Large using 2 million unlabeled posts from the EDOS dataset and fine-tune them individually and in a fused configuration (DTFN). Final predictions are generated via majority voting, with a tie-handling rule that improves robustness.

Exist@CeDRI [27] uses a combination of multiple text augmentation strategies, including AEDA (punctuation-based), synonym replacement, back-translation, and light code-switching via round-trip translation, in order to enhance model reliability and deal with data sparsity. Their architecture builds on XLM-RoBERTa-large, fine-tuned for three subtasks: binary sexism detection, source classification, and sexism categorization. Both soft and hard label strategies are incorporated to account for annotation disagreement and label smoothing and class-weighted loss functions are applied to manage class imbalance.

Awakened [28] employs an adaptive Mixture of Transformers architecture. The system combines nine transformer-based models—spanning both English-specific and multilingual variants—each specialized by language, platform, or task. A dynamic weighting mechanism automatically adjusts the contribution of each model in the ensemble, based on the detected language and performance metrics, in order to enable robust and context-aware classification across diverse linguistic settings.

Dandys-de-BERTganim [29] adopts a multi-task learning architecture with language-specific transformers for English and Spanish, integrating demographic information from annotators as contextual signals. They enhance model generalization through data augmentation techniques such as back-translation and a punctuation-based augmentation method. Furthermore, they introduce a soft-labeling data reader to better reflect annotation disagreement, aligning with the LeWiDi paradigm.

DuthThrace [30] develops a transformer-based multilingual architecture, fine-tuned with techniques such as oversampling, class weighting, and soft-label learning to account for class imbalance and annotator disagreement.

CIMAT-CS-NLP [31] proposes a method based on a single multitask query to LLMs, designing a query that first generates chain-of-thought justifications and then requests answers for all tasks simultaneously. To automate query refinement, they apply evolutionary computation, optimizing the F1-macro on a development subset. Experiments are performed with DeepSeek-R1-Distill-Llama-8B and Gemini-1.5-Flash. They fine-tune a BERT-like model with the LLM-generated justifications, with DeepSeek achieving similar performance to the Gemini-based justifications despite the reduced model size.

UC3M-LI [32] develops a variety of systems for Task 1.1 and Task 1.2, combining traditional machine learning models, Transformer-based architectures, ensemble methods, and hybrid CNN-BERT approaches. Their approach incorporates data augmentation, and multilingual modeling strategies to address challenges such as label disagreement and language variation.

Cyberpuffs [33] uses several LLMs, prominently multilingual BERT and XLM_Roberta, combined with an ensemble learning approach to process tweets. They employ data augmentation techniques such as cross-translation, EASE, and AEDA, and develop separate models for English and Spanish to optimize language-specific predictions. Model evaluation is conducted using hard labels, derived through majority annotator voting, and soft labels, derived from class probability distributions.

COMFOR [34] approaches the tasks with an SVM based on a comprehensive feature representation, including embeddings and lexical features. For the third subtask, this classifier was used as the basis for a classifier chain.

CIMAT-GTO [35] uses a hybrid setting aimed at taking advantage of the reasoning produced by generative LLMs using justification-guided knowledge expansion when fine-tuning a smaller transformer-based model for classification.

Mario [36] applies hierarchical Low-Rank Adaptation (LoRA) of Llama 3.1 8B. Their method introduces conditional adapter routing that explicitly models label dependencies across the three hierarchically structured subtasks. Unlike conventional LoRA applications that target only attention layers, they apply adaptation to all linear transformations, enhancing the model's capacity to capture task-specific patterns. They train separate LoRA adapters (rank=16, QLoRA 4-bit) for each subtask using unified multilingual training that leverages Llama 3.1's native bilingual capabilities. The method requires minimal preprocessing and uses standard supervised learning.

FHSTP [37] proposes three machine learning models to address these tasks, including Speech Concept Bottleneck Model (SCBM), Speech Concept Bottleneck Model with Transformer (SCBMT) and a fine-tuned XLM-RoBERTa transformer model that serves as baseline. SCBM uses descriptive adjectives as human-interpretable bottleneck concepts. SCBM leverages LLMs to map input texts to an abstract adjective-based representation, which is then utilized to train a light-weight classifier for downstream tasks. SCBMT extends this approach by fusing transformer-based contextual embeddings with the adjective-based representation, aiming to balance interpretability and classification performance.

NYCU-NLP [38] integrates annotator demographics and leverages bilingual fusion by combining original and cross-translated tweets. They implement a hierarchical pipeline and compare three distinct modeling strategies: a fine-tuned transformer-based dual-encoder architecture with early and late fusion, a zero-shot auto-regressive LLM, and a zero-shot diffusion-based LLM. The transformer-based approach consistently achieves the highest performance across most metrics.

Next we present the approaches of teams that participated only in Task 2 on processing **memes**.

TrankilTwice [39] participates in Task 2.1 with an end-to-end system integrating LLM-based prompting strategies, cross-modal language encoding, and graph-based modeling at meme level, obtaining performance gaps across languages.

NaturalThinkers [40] integrates visual and textual feature extraction using BLIP (Bootstrapping Language-Image Pretraining), BERT and ViT (Vision transformer) followed by a fusion mechanism employing attention-based. Then they use multi-layer perceptron (MLP) for the final classification, with a Gradio-based user interface.

ArcosGPT [41] adds BLIP-generated image captions to OCR text. By further including a GPT-4o description of the memes they obtain an increae of 8.2 points. They obtain the best overall performance with a ViT+RoBERTa fusion model.

CLTL [42] follows a hard majority voting ensemble strategy to process memes, where the component models included a multimodal model that combines the representations of Swin Transformer V2 and a pre-trained language model (RoBERTa or BERT), and the text-only model that uses meme text and image captions as input. The text-only approaches included pre-trained transformer models (RoBERTa, BERT, and a BERTweet model fine-tuned for sexism detection) and a conventional machine learning approach, namely an SVM with stylometric and emotion-based features.

I2C-UHU-Altair [43] uses LLMs and vision-language models (VLMs) to process both textual and visual information in memes. To enhance model robustness, they adopt the LeWiDi framework, as an attempt to allow the system to benefit from divergent annotations that reflect the inherent ambiguity and subjectivity in sociolinguistic tasks.

GrootWatch [44] participates in both the **tweets and memes** tasks. For tweet classification, they used a multi-task headed BERT model enriched with relevant information surrounding the tweet, helping the model achieve a full understanding of the tweet and its context. For memes, they used a VLM-based application to detect and categorise sexism in different scenarios.

The following are the approaches of the teams that participated only in the **TikTok** tasks.

ECA-SIMM-UVa [45] follows a segmentation oriented approach, splitting TikTok videos into textual, audio, and video channels, driven by the hypothesis that sexism can manifest in spoken words, embedded text, speaker tone, or visual content (text, pictures or other images). They train individual deep learning classifiers for each channel and explore various prediction fusion mechanisms like One Is Enough (OIE), Majority Voting, and Probabilistic OIQ for hard evaluation, as well as Logistic Regression and Weighted Sum for soft evaluation, to combine predictions. Models using the textual channel show superior performance, specially when using the original text provided with each sample in the dataset. These models consistently outperform audio and video channels, indicating that textual information is the most informative source for sexism detection in this context.

DS@GT EXIST [46] implements a multimodal framework for automated sexism detection in short-form videos, incorporating audio, visual, and textual signals. They explore the use of transformer-based models including RoBERTa for text, VideoMAE for video, and CNN-LSTM pipelines for audio and they introduce a generative AI-enhanced pipeline using Gemini to produce video summaries and analyses, which are combined with traditional modalities.

Finally, a few teams participated on the three tasks, processing **tweets, memes and TikTok videos**.

UMUTeam [47] addresses all three subtasks with multilingual Transformer-based models, including XLM-RoBERTa (base and large versions) for text, ViT for image features, and VideoMAE for video input. They apply specialized preprocessing and label handling for each modality. Soft-label learning is implemented using mean squared error (MSE) loss for Subtasks 1 and 2, which involve binary and multiclass classification, respectively, and binary cross-entropy (BCE) loss for Subtask 3, which is a multilabel classification problem. In all cases, annotator votes are transformed into probability distributions to capture label uncertainty. For hard-label variants, discrete predictions are obtained by selecting the class or classes with the highest probability from the model's output during the evaluation stage.

CogniCIC [48] explores tailored methodologies to process tweets, memes, and TikTok videos. For subtask 1 they compare two approaches: the transformer-based HateBERT model and the generative Claude 3.7 model. HateBERT is optimized through tweet preprocessing, regularized training, and multitask learning, and Claude 3.7, which leverages advanced multimodal capabilities, integrating visual and textual cues for flexible and effective content interpretation. For Subtasks 2 and 3 they use Claude 3.7, which incorporates multimodal inputs, including visual frames from memes and videos, enabling nuanced distinctions, such as direct sexist expressions versus judgmental critiques.

Bergro [49] follows a generalizable BERT-based approach to identify and classify the source intent of sexism across different social network channels. This approach focuses on individual models trained on tweets that are then applied to both meme (image) and TikTok data using OCR and annotations, respectively. This is an example of single model fine-tuned on one media type and applied to multiple media types with minimal data preprocessing required.

BeatrizRuiz [50] uses three transformer-based models—DistilBERT, XLM-RoBERTa, and DistilGPT-2 to address all tasks. The results show that, while all models tend to overpredict sexist content and underutilize the non-sexist class in complex subtasks, DistilBERT demonstrates the most balanced performance in binary classification, XLM-RoBERTa shows robustness, but a propensity for overgeneralization, and DistilGPT-2 exhibits greater flexibility in multilabel assignments, despite its generative architecture.

6. Results

In the following subsections, we present the results of both, the participants and the baseline systems for each task, organized by evaluation mode (soft or hard).

6.1. Task 1.1: Sexism Identification in Tweets

6.1.1. Soft Evaluation

Table 6 presents the results of the soft-soft evaluation for Task 1.1, which received a total of 65 participating systems (excluding the gold reference and two baselines). The normalized ICM-Soft scores ranged from close to 0 up 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, *EXIST2025-test_majority-class*, which assigns the label 'NO' to all instances. Notably, all systems exceeded the performance of the minority-class baseline, confirming overall effectiveness in this evaluation setting.

Table 6
Leaderboard for EXIST 2025 Task 1.1 (sexism detection in tweets), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	3.1182	1.0000	0.5472	0
GrootWatch_1	GrootWatch	**1.0600**	**0.6700**	0.8893	1

(continued on next page)

(continued from previous page)

System	Team	ICM-S	ICM-S Nr	CE	Rank
GrootWatch_2	GrootWatch	1.0538	0.6690	0.9171	2
GrootWatch_3	GrootWatch	1.0368	0.6662	0.9088	3
DaniReinon_1	DaniReinon	0.8332	0.6336	0.9726	4
DaniReinon_3	DaniReinon	0.8058	0.6292	0.7815	5
fhstp_3	fhstp	0.7852	0.6259	0.8416	6
BERT-Simpson_2	BERT-Simpson	0.7461	0.6196	1.0426	7
DaniReinon_2	DaniReinon	0.7397	0.6186	0.9694	8
CLiC_1	CLiC	0.7386	0.6184	0.8201	9
BERT-Simpson_3	BERT-Simpson	0.7349	0.6178	0.9639	10
BERT-Simpson_1	BERT-Simpson	0.7203	0.6155	1.0872	11
A-squared_2	A-squared	0.6990	0.6121	1.2623	12
Cyberpuffs_3	Cyberpuffs	0.6767	0.6085	0.8894	13
NetGuardAI_1	NetGuardAI	0.6707	0.6076	0.8774	14
A-squared_1	A-squared	0.6686	0.6072	1.6623	15
Dandys-de-BERTganim_1	Dandys-de-BERTganim	0.6575	0.6054	0.7964	16
NYCU-NLP_1	NYCU-NLP	0.6569	0.6053	0.8302	17
bergro_1	bergro	0.6382	0.6023	0.7921	18
ArPa Project_3	ArPa Project	0.6339	0.6016	1.3316	19
GPTesla Smashers_1	GPTesla Smashers	0.6282	0.6007	1.0896	20
ArPa Project_2	ArPa Project	0.6122	0.5982	1.4289	21
Vallecas_1	Vallecas	0.5729	0.5919	1.0218	22
ArPa Project_1	ArPa Project	0.5637	0.5904	1.2723	23
fosu-student_2	fosu-student	0.5592	0.5897	1.4542	24
mumule03_3	mumule03	0.5527	0.5886	1.2003	25
Chai Cheers Chutney_1	Chai Cheers Chutney	0.5410	0.5868	0.9926	26
Awakened_3	Awakened	0.5328	0.5854	0.9084	27
Awakened_1	Awakened	0.5310	0.5851	0.9531	28
fosu-student_3	fosu-student	0.4714	0.5756	1.4373	29
GPTesla Smashers_3	GPTesla Smashers	0.4460	0.5715	0.9706	30
Awakened_2	Awakened	0.4074	0.5653	0.9645	31
mumule03_1	mumule03	0.3864	0.5620	1.0554	32
M&Ms_1	M&Ms	0.3406	0.5546	2.1901	33
M&Ms_2	M&Ms	0.3270	0.5524	2.1425	34
GPTesla Smashers_2	GPTesla Smashers	0.3167	0.5508	1.2758	35
fhstp_1	fhstp	0.3135	0.5503	2.0735	36
fhstp_2	fhstp	0.3039	0.5487	2.5240	37
DuthThace_1	DuthThace	0.1960	0.5314	2.1029	38
BeatrizRuiz_1	BeatrizRuiz	0.1285	0.5206	0.9037	39
M&Ms_3	M&Ms	0.1025	0.5164	2.4862	40
UMUTeam_2	UMUTeam	0.0138	0.5022	0.8119	41
BeatrizRuiz_2	BeatrizRuiz	−0.0483	0.4922	0.9148	42
CLiC_2	CLiC	−0.0849	0.4864	2.5585	43
UMUTeam_1	UMUTeam	−0.1729	0.4723	0.8362	44
CLiC_3	CLiC	−0.2035	0.4674	2.6423	45
CYUT_1	CYUT	−0.2420	0.4612	2.5070	46
Vallecas_2	Vallecas	−0.3542	0.4432	0.8795	47
BeatrizRuiz_3	BeatrizRuiz	−0.3562	0.4429	0.9420	48
JOW_1	JOW	−0.4744	0.4239	0.9640	49
NYCU-NLP_3	NYCU-NLP	−0.7152	0.3853	0.9415	50
NYCU-NLP_2	NYCU-NLP	−0.7700	0.3765	0.9325	51
Cyberpuffs_2	Cyberpuffs	−0.9514	0.3475	0.5168	52
GuerraMendez_1	GuerraMendez	−0.9604	0.3460	1.9215	53
GuerraMendez_2	GuerraMendez	−1.0803	0.3268	2.0251	54
AlbaandRita_1	AlbaandRita	−1.2573	0.2984	1.7845	55
Cyberpuffs_1	Cyberpuffs	−1.4295	0.2708	**0.4954**	56

(continued on next page)

(continued from previous page)

System	Team	ICM-S	ICM-S Nr	CE	Rank
mumule03_2	mumule03	−1.5131	0.2574	0.9052	57
AlbaandRita_2	AlbaandRita	−1.5147	0.2571	1.6312	58
exist@CeDRI_1	exist@CeDRI	−1.7048	0.2266	1.5323	59
IratxeCarla_1	IratxeCarla	−1.8245	0.2074	1.7326	60
ecorbi-upv_2	ecorbi-upv	−1.8858	0.1976	1.5434	61
SalaPlanes_1	SalaPlanes	−2.1261	0.1591	2.4581	62
ecorbi-upv_1	ecorbi-upv	−2.2943	0.1321	2.6337	63
EXIST2025-test_majority-class	–	−2.3585	0.1218	4.6115	64
ecorbi-upv_3	ecorbi-upv	−2.4186	0.1122	3.0160	65
EXIST2025-test_minority-class	–	−3.0717	0.0075	5.3572	66
Nogroupnocry_1	Nogroupnocry	−3.1726	0.0000	1.0170	67

6.1.2. Hard Evaluation

Table 7 presents the results for the hard-hard evaluation. In this setting, the annotations from the six annotators were aggregated into a single label using majority voting. A total of 158 systems participated using hard-label predictions.

The normalized ICM-Hard scores, ranging from 0 to 1, had a mean of 0.678 and a standard deviation of 0.149 across participants. The best-performing system achieved a normalized score of 0.8405, while the lowest-scoring system obtained 0.1710.

Notably, 153 out of 158 systems outperformed the strongest hard-label baseline (*EXIST2025-test_majority-class*, which assigns the label 'NO' to all instances). Only three systems fell below the minority-class baseline, confirming a robust overall performance in this setting.

As shown in Table 7, the gap between the top and fifth-ranked systems was only 5.6%, demonstrating strong consistency among the 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.

Table 7
Leaderboard for EXIST 2025 Task 1.1 (sexism detection in tweets), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, F1Y = F1 YES.

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
EXIST2025-test_gold	–	0.9948	1.0000	1.0000	0
Mario_1	Mario	**0.6774**	**0.8405**	**0.8167**	1
CIMAT-GTO_2	CIMAT-GTO	0.6297	0.8165	0.7996	2
CIMAT-GTO_3	CIMAT-GTO	0.6256	0.8144	0.7968	3
warwick_1	warwick	0.6249	0.8141	0.7991	4
CIMAT-CS-NLP_3	CIMAT-CS-NLP	0.6127	0.8079	0.7945	5
CIMAT-CS-NLP_2	CIMAT-CS-NLP	0.6076	0.8054	0.7940	6
warwick_2	warwick	0.5834	0.7932	0.7892	7
BERT-Simpson_1	BERT-Simpson	0.5832	0.7931	0.7832	8
alBERTAberg_1	alBERTAberg	0.5814	0.7922	0.7803	9
Equity-Explorer-2.0_1	Equity-Explorer-2.0	0.5806	0.7918	0.7837	10
UMUTeam_2	UMUTeam	0.5799	0.7915	0.7824	11
warwick_3	warwick	0.5793	0.7912	0.7888	12
Equity-Explorer-2.0_2	Equity-Explorer-2.0	0.5779	0.7904	0.7827	13
BERT-Simpson_2	BERT-Simpson	0.5757	0.7894	0.7804	14
GrootWatch_2	GrootWatch	0.5732	0.7881	0.7773	15
GrootWatch_1	GrootWatch	0.5727	0.7878	0.7802	16
ArPa Project_2	ArPa Project	0.5623	0.7826	0.7748	17
fhstp_3	fhstp	0.5610	0.7819	0.7839	18

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
BERT-Simpson_3	BERT-Simpson	0.5580	0.7805	0.7741	19
GrootWatch_3	GrootWatch	0.5560	0.7795	0.7763	20
fhstp_2	fhstp	0.5545	0.7787	0.7767	21
BERTin-Osborne_1	BERTin-Osborne	0.5505	0.7767	0.7734	22
ArPa Project_1	ArPa Project	0.5448	0.7738	0.7721	23
PabloyFede_1	PabloyFede	0.5431	0.7730	0.7727	24
pau-rus_1	pau-rus	0.5425	0.7727	0.7717	25
samuel-sergei_2	samuel-sergei	0.5424	0.7726	0.7683	26
BERTinators_1	BERTinators	0.5396	0.7712	0.7718	27
samuel-sergei_3	samuel-sergei	0.5371	0.7699	0.7673	28
ArPa Project_3	ArPa Project	0.5341	0.7684	0.7640	29
Cachapas_1	Cachapas	0.5340	0.7684	0.7695	30
A-squared_2	A-squared	0.5319	0.7673	0.7672	31
A-squared_3	A-squared	0.5312	0.7670	0.7692	32
CS-GO_1	CS-GO	0.5263	0.7645	0.7684	33
CIMAT-GTO_1	CIMAT-GTO	0.5253	0.7640	0.7639	34
CIMAT-CS-NLP_1	CIMAT-CS-NLP	0.5246	0.7637	0.7652	35
moniclaudia_1	moniclaudia	0.5244	0.7636	0.7660	36
JosepYSergio_1	JosepYSergio	0.5210	0.7619	0.7641	37
yow_1	yow	0.5208	0.7617	0.7632	38
bergro_1	bergro	0.5194	0.7611	0.7654	39
PorTod@s_1	PorTod@s	0.5193	0.7610	0.7636	40
A-squared_1	A-squared	0.5186	0.7606	0.7666	41
NeuralNomads_1	NeuralNomads	0.5181	0.7604	0.7660	42
carlamiguel_1	carlamiguel	0.5178	0.7603	0.7633	43
Mouctar Diakhaby_1	Mouctar Diakhaby	0.5174	0.7600	0.7661	44
Team PCIC_1	Team PCIC	0.5158	0.7593	0.7629	45
UC3M-LI_1	UC3M-LI	0.5135	0.7581	0.7613	46
TheMagicToken_1	TheMagicToken	0.5083	0.7555	0.7610	47
TransformerHotspur_3	TransformerHotspur	0.5082	0.7554	0.7554	48
UMUTeam_1	UMUTeam	0.5064	0.7545	0.7595	49
CLiC_1	CLiC	0.5059	0.7543	0.7538	50
Dandys-de-BERTganim_2	Dandys-de-BERTganim	0.5026	0.7526	0.7548	51
BRAINSTORMERS_1	BRAINSTORMERS	0.5011	0.7519	0.7559	52
Lirili-Larila_1	Lirili-Larila	0.5010	0.7518	0.7535	53
Dandys-de-BERTganim_1	Dandys-de-BERTganim	0.5000	0.7513	0.7597	54
NetGuardAI_1	NetGuardAI	0.4966	0.7496	0.7553	55
Cyberpuffs_3	Cyberpuffs	0.4953	0.7490	0.7558	56
LolaClaudia_1	LolaClaudia	0.4940	0.7483	0.7542	57
samuel-sergei_1	samuel-sergei	0.4935	0.7480	0.7516	58
UC3M-LI_3	UC3M-LI	0.4931	0.7479	0.7595	59
NYCU-NLP_1	NYCU-NLP	0.4912	0.7469	0.7512	60
Lirili-Larila_3	Lirili-Larila	0.4905	0.7465	0.7523	61
Juanji&Jowi_1	Juanji&Jowi	0.4904	0.7465	0.7524	62
Lim-go-home_1	Lim-go-home	0.4848	0.7437	0.7556	63
DoubleA_1	DoubleA	0.4844	0.7434	0.7484	64
Alberto and Ángel_1	Alberto and Ángel	0.4831	0.7428	0.7539	65
sadiqlovers_1	sadiqlovers	0.4831	0.7428	0.7497	66
Güeypingüino_1	Güeypingüino	0.4819	0.7422	0.7542	67
GPTesla Smashers_1	GPTesla Smashers	0.4800	0.7412	0.7496	68
E.T._1	E.T.	0.4800	0.7412	0.7516	69
MakeTwitterGreatAgain_1	MakeTwitterGreatAgain	0.4778	0.7402	0.7513	70
SalaPlanes_1	SalaPlanes	0.4777	0.7401	0.7491	71
Joses_1	Joses	0.4776	0.7400	0.7524	72
BocadilloDelDia_1	BocadilloDelDia	0.4775	0.7400	0.7497	73

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
Minerva_2	Minerva	0.4755	0.7390	0.7479	74
CogniCIC_3	CogniCIC	0.4718	0.7371	0.7495	75
SalaPlanes_2	SalaPlanes	0.4703	0.7364	0.7459	76
Lirili-Larila_2	Lirili-Larila	0.4691	0.7358	0.7456	77
LNP-SP_1	LNP-SP	0.4685	0.7355	0.7480	78
BRAINSTORMERS_2	BRAINSTORMERS	0.4684	0.7354	0.7462	79
BERTin-Osborne_2	BERTin-Osborne	0.4673	0.7349	0.7495	80
Chai Cheers Chutney_1	Chai Cheers Chutney	0.4662	0.7343	0.7487	81
GarCha_1	GarCha	0.4654	0.7339	0.7501	82
fosu-student_2	fosu-student	0.4649	0.7337	0.7480	83
exist@CeDRI_1	exist@CeDRI	0.4645	0.7335	0.7471	84
UC3M-LI_2	UC3M-LI	0.4634	0.7329	0.7448	85
Vallecas_1	Vallecas	0.4633	0.7328	0.7503	86
Güeypingüino_2	Güeypingüino	0.4631	0.7328	0.7456	87
DuthThace_1	DuthThace	0.4628	0.7326	0.7432	88
sheila-vicente_1	sheila-vicente	0.4615	0.7320	0.7437	89
M&Ms_1	M&Ms	0.4613	0.7319	0.7493	90
fosu-student_3	fosu-student	0.4587	0.7305	0.7471	91
M&Ms_2	M&Ms	0.4584	0.7304	0.7511	92
Awakened_3	Awakened	0.4562	0.7293	0.7499	93
Awakened_1	Awakened	0.4550	0.7287	0.7485	94
fosu-student_1	fosu-student	0.4532	0.7278	0.7432	95
Niklas Team_1	Niklas Team	0.4519	0.7271	0.7460	96
carlamiguel_2	carlamiguel	0.4500	0.7262	0.7439	97
Token-Trouble_1	Token-Trouble	0.4486	0.7255	0.7432	98
carlamiguel_3	carlamiguel	0.4384	0.7203	0.7408	99
mumule03_1	mumule03	0.4375	0.7199	0.7359	100
mumule03_3	mumule03	0.4360	0.7191	0.7316	101
Awakened_2	Awakened	0.4350	0.7186	0.7390	102
BRAINSTORMERS_3	BRAINSTORMERS	0.4311	0.7167	0.7403	103
bergro_2	bergro	0.4289	0.7156	0.7350	104
fhstp_1	fhstp	0.4288	0.7155	0.7392	105
IratxeCarla_1	IratxeCarla	0.4269	0.7145	0.7378	106
JorgeyQuique_2	JorgeyQuique	0.4265	0.7143	0.7395	107
Tweetbusters_2	Tweetbusters	0.4248	0.7135	0.7318	108
GPTesla Smashers_3	GPTesla Smashers	0.4243	0.7133	0.7358	109
Tweetbusters_1	Tweetbusters	0.4240	0.7131	0.7301	110
GPTesla Smashers_2	GPTesla Smashers	0.4150	0.7086	0.7333	111
M&Ms_3	M&Ms	0.4050	0.7035	0.7311	112
Data-force_1	Data-force	0.3971	0.6996	0.7222	113
rubenmanu_1	rubenmanu	0.3903	0.6962	0.7202	114
Minerva_1	Minerva	0.3635	0.6827	0.7154	115
alBERTAberg_2	alBERTAberg	0.3611	0.6815	0.7055	116
CogniCIC_1	CogniCIC	0.3493	0.6756	0.6958	117
CogniCIC_2	CogniCIC	0.3465	0.6742	0.7220	118
BeatrizRuiz_1	BeatrizRuiz	0.3263	0.6640	0.6907	119
CLiC_2	CLiC	0.3243	0.6630	0.6907	120
CLiC_3	CLiC	0.3146	0.6581	0.6978	121
Niklas Team_3	Niklas Team	0.3105	0.6561	0.6870	122
EXISTsValenciaWork_2	EXISTsValenciaWork	0.2837	0.6426	0.6868	123
TransformerHotspur_2	TransformerHotspur	0.2809	0.6412	0.6794	124
Niklas Team_2	Niklas Team	0.2555	0.6284	0.6626	125
megabeats_1	megabeats	0.2462	0.6238	0.6812	126
BeatrizRuiz_2	BeatrizRuiz	0.2200	0.6106	0.6407	127
CYUT_2	CYUT	0.2081	0.6046	0.6839	128

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
CYUT_1	CYUT	0.2075	0.6043	0.6843	129
Vallecas_2	Vallecas	0.2059	0.6035	0.6489	130
JENLCB_1	JENLCB	0.2034	0.6022	0.6553	131
Vectorious_1	Vectorious	0.2032	0.6021	0.6398	132
IratxeCarla_2	IratxeCarla	0.1695	0.5852	0.6439	133
BeatrizRuiz_3	BeatrizRuiz	0.1527	0.5767	0.6237	134
NYCU-NLP_2	NYCU-NLP	0.0706	0.5355	0.6444	135
NYCU-NLP_3	NYCU-NLP	0.0504	0.5253	0.6353	136
SergioAngel_1	SergioAngel	−0.0027	0.4987	0.4865	137
GuerraMendez_1	GuerraMendez	−0.0529	0.4734	0.5846	138
JOW_1	JOW	−0.0694	0.4651	0.4677	139
GuerraMendez_2	GuerraMendez	−0.0727	0.4634	0.5695	140
Sofia_1	Sofia	−0.1074	0.4460	0.4794	141
SQUADRAMP_1	SQUADRAMP	−0.1107	0.4443	0.4967	142
TransformerHotspur_1	TransformerHotspur	−0.1501	0.4246	0.6075	143
Cyberpuffs_2	Cyberpuffs	−0.1871	0.4060	0.5869	144
AlbaandRita_3	AlbaandRita	−0.2013	0.3988	0.6201	145
AlbaandRita_2	AlbaandRita	−0.2311	0.3839	0.6142	146
COMFOR_3	COMFOR	−0.3061	0.3461	0.4609	147
Cyberpuffs_1	Cyberpuffs	−0.3124	0.3430	0.4611	148
COMFOR_2	COMFOR	−0.3278	0.3352	0.4379	149
mumule03_2	mumule03	−0.3517	0.3233	0.3964	150
AlbaandRita_1	AlbaandRita	−0.3840	0.3070	0.5942	151
Nogroupnocry_1	Nogroupnocry	−0.4216	0.2881	0.4000	152
COMFOR_1	COMFOR	−0.4300	0.2839	0.3510	153
EXIST2025-test_majority-class	–	−0.4413	0.2782	0.0000	154
shm2025_1	shm2025	−0.5432	0.2270	0.4283	155
ANLP-UniSo_1	ANLP-UniSo	−0.5650	0.2160	0.4015	156
EXIST2025-test_minority-class	–	−0.5742	0.2114	0.5698	157
ecorbi-upv_2	ecorbi-upv	−0.6377	0.1795	0.5105	158
ecorbi-upv_1	ecorbi-upv	−0.6546	0.1710	0.2838	159
ecorbi-upv_3	ecorbi-upv	−0.6546	0.1710	0.2838	160

6.2. Task 1.2: Source Intention in Tweets

6.2.1. Soft Evaluation

Table 8 presents the results of the soft-soft evaluation for Task 1.2, which focuses on identifying the author's intent behind sexist tweets. This task received 54 system submissions under the Soft–Soft setting. Participant 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 (*EXIST2025-test_majority-class*, where all instances are labeled as 'NO'), indicating moderate variation in system effectiveness. All systems also outperformed the *EXIST2025-test_minority-class* baseline.

The relative difference between the best and fifth-best teams (*GrootWatch* and *NetGuardAI*) was 15.7%, suggesting relatively close performance among the top submissions. This narrow spread points to a convergence in probabilistic modeling strategies among leading participants, despite overall scores being lower than in other tasks—likely due to the increased ambiguity inherent in intent classification.

Table 8
Leaderboard for EXIST 2025 Task 1.2 (author intention analysis in tweets), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	6.2057	1.0000	0.9128	0
GrootWatch_1	GrootWatch	**-0.4385**	**0.4647**	1.7711	1
GrootWatch_2	GrootWatch	−0.5066	0.4592	1.8176	2
GrootWatch_3	GrootWatch	−0.5655	0.4544	1.8088	3
Dandys-de-BERTganim_2	Dandys-de-BERTganim	−0.7261	0.4415	1.3820	4
Dandys-de-BERTganim_1	Dandys-de-BERTganim	−0.8526	0.4313	1.4563	5
Cyberpuffs_3	Cyberpuffs	−1.0572	0.4148	2.0396	6
fhstp_2	fhstp	−1.1866	0.4044	1.6566	7
NetGuardAI_1	NetGuardAI	−1.3444	0.3917	1.5681	8
A-squared_2	A-squared	−1.6106	0.3702	2.0824	9
A-squared_1	A-squared	−1.6677	0.3656	2.6014	10
fhstp_3	fhstp	−1.7928	0.3556	1.7156	11
fhstp_1	fhstp	−1.8291	0.3526	1.5690	12
Chai Cheers Chutney_1	Chai Cheers Chutney	−1.8949	0.3473	2.0200	13
NYCU-NLP_1	NYCU-NLP	−1.9749	0.3409	2.2312	14
exist@CeDRI_1	exist@CeDRI	−2.1362	0.3279	3.7432	15
ArPa Project_2	ArPa Project	−2.5004	0.2985	2.2713	16
ArPa Project_3	ArPa Project	−2.5553	0.2941	2.2657	17
mumule03_2	mumule03	−3.1488	0.2463	3.2075	18
LNP-SP_1	LNP-SP	−3.2350	0.2394	1.8763	19
Cyberpuffs_2	Cyberpuffs	−3.2616	0.2372	1.0764	20
CLiC_2	CLiC	−3.5005	0.2180	3.5219	21
UMUTeam_2	UMUTeam	−3.6965	0.2022	1.6759	22
UMUTeam_1	UMUTeam	−3.8401	0.1906	1.7063	23
ecorbi-upv_2	ecorbi-upv	−3.9778	0.1795	1.9562	24
NYCU-NLP_2	NYCU-NLP	−3.9860	0.1788	4.1786	25
Cyberpuffs_1	Cyberpuffs	−4.0013	0.1776	0.9632	26
CLiC_3	CLiC	−4.0057	0.1773	3.5732	27
JOW_1	JOW	−4.0883	0.1706	4.2210	28
ecorbi-upv_3	ecorbi-upv	−4.3185	0.1521	2.0465	29
ecorbi-upv_1	ecorbi-upv	−4.3185	0.1521	2.0465	30
NYCU-NLP_3	NYCU-NLP	−4.6004	0.1293	4.3754	31
GuerraMendez_1	GuerraMendez	−4.7747	0.1153	1.3886	32
Nogroupnocry_1	Nogroupnocry	−5.1913	0.0817	**0.3806**	33
BeatrizRuiz_1	BeatrizRuiz	−5.2659	0.0757	5.5611	34
GuerraMendez_2	GuerraMendez	−5.3429	0.0695	1.6346	35
M&Ms_1	M&Ms	−5.3751	0.0669	3.6315	36
EXIST2025-test_majority-class	–	−5.4460	0.0612	4.6233	37
IratxeCarla_1	IratxeCarla	−5.4535	0.0606	3.7090	38
BeatrizRuiz_3	BeatrizRuiz	−5.5061	0.0564	6.2201	39
BeatrizRuiz_2	BeatrizRuiz	−5.5642	0.0517	5.8331	40
ArPa Project_1	ArPa Project	−5.7726	0.0349	3.8719	41
Vallecas_2	Vallecas	−5.8868	0.0257	6.0455	42
M&Ms_3	M&Ms	−6.5863	0.0000	3.9697	43
DaniReinon_1	DaniReinon	−6.8094	0.0000	5.1993	44
M&Ms_2	M&Ms	−6.8875	0.0000	4.0516	45
Vallecas_1	Vallecas	−6.9867	0.0000	6.0645	46
Awakened_2	Awakened	−7.8903	0.0000	6.9973	47
CLiC_1	CLiC	−8.5179	0.0000	4.4695	48
Awakened_1	Awakened	−9.0253	0.0000	5.2347	49
LNP-SP_2	LNP-SP	−9.4518	0.0000	6.2154	50
Awakened_3	Awakened	−18.4462	0.0000	17.4256	51

(continued on next page)

(continued from previous page)

System	Team	ICM-S	ICM-S Nr	CE	Rank
DuthThace_1	DuthThace	−18.5641	0.0000	7.3212	52
GPTesla Smashers_2	GPTesla Smashers	−20.5699	0.0000	8.4612	53
GPTesla Smashers_1	GPTesla Smashers	−21.3006	0.0000	8.7286	54
GPTesla Smashers_3	GPTesla Smashers	−21.4927	0.0000	8.6678	55
EXIST2025-test_minority-class	–	−32.9552	0.0000	8.8517	56

6.2.2. Hard Evaluation

Table 9 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. Remarcably, 105 systems outperformed the best hard-label baseline (*EXIST2025-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%.

Table 9
Leaderboard for EXIST 2025 Task 1.2 (author intention analysis in tweets), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-test_gold	–	1.5378	1.0000	1.0000	0
Mario_1	Mario	**0.4991**	**0.6623**	**0.5692**	1
CIMAT-GTO_3	CIMAT-GTO	0.4678	0.6521	0.5555	2
CIMAT-GTO_2	CIMAT-GTO	0.4392	0.6428	0.5582	3
CIMAT-CS-NLP_2	CIMAT-CS-NLP	0.4264	0.6386	0.5461	4
CIMAT-CS-NLP_3	CIMAT-CS-NLP	0.4118	0.6339	0.5482	5
Dandys-de-BERTganim_2	Dandys-de-BERTganim	0.3752	0.6220	0.5522	6
BERTin-Osborne_2	BERTin-Osborne	0.3677	0.6196	0.5453	7
CIMAT-CS-NLP_1	CIMAT-CS-NLP	0.3619	0.6177	0.5266	8
GrootWatch_3	GrootWatch	0.3434	0.6117	0.5384	9
GrootWatch_2	GrootWatch	0.3266	0.6062	0.5421	10
Dandys-de-BERTganim_1	Dandys-de-BERTganim	0.3217	0.6046	0.5425	11
UMUTeam_2	UMUTeam	0.3064	0.5996	0.5236	12
GrootWatch_1	GrootWatch	0.3016	0.5981	0.5325	13
BERT-Simpson_1	BERT-Simpson	0.2958	0.5962	0.5244	14
BERTin-Osborne_1	BERTin-Osborne	0.2889	0.5939	0.5307	15
fhstp_2	fhstp	0.2795	0.5909	0.5175	16
BERT-Simpson_2	BERT-Simpson	0.2711	0.5882	0.5111	17
UMUTeam_1	UMUTeam	0.2647	0.5860	0.5065	18
alBERTAberg_1	alBERTAberg	0.2461	0.5800	0.5167	19
BERTin-Osborne_3	BERTin-Osborne	0.2374	0.5772	0.4968	20
A-squared_3	A-squared	0.2286	0.5743	0.4837	21
BERT-Simpson_3	BERT-Simpson	0.2212	0.5719	0.5003	22
samuel-sergei_2	samuel-sergei	0.2176	0.5708	0.5092	23
Lirili-Larila_3	Lirili-Larila	0.2166	0.5704	0.5034	24
CIMAT-GTO_1	CIMAT-GTO	0.2069	0.5673	0.5268	25
Cyberpuffs_3	Cyberpuffs	0.1987	0.5646	0.5101	26
ArPa Project_1	ArPa Project	0.1971	0.5641	0.4935	27
carlamiguel_1	carlamiguel	0.1945	0.5632	0.4968	28
Lirili-Larila_1	Lirili-Larila	0.1943	0.5632	0.5049	29
CogniCIC_1	CogniCIC	0.1899	0.5617	0.5313	30
Mouctar Diakhaby_1	Mouctar Diakhaby	0.1773	0.5577	0.4794	31
Lirili-Larila_2	Lirili-Larila	0.1753	0.5570	0.5072	32

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	M F1	Rank
pau-rus_1	pau-rus	0.1566	0.5509	0.4824	33
A-squared_1	A-squared	0.1558	0.5507	0.4797	34
NYCU-NLP_1	NYCU-NLP	0.1529	0.5497	0.4798	35
M&Ms_1	M&Ms	0.1504	0.5489	0.4835	36
sadiqlovers_1	sadiqlovers	0.1485	0.5483	0.4943	37
LolaClaudia_2	LolaClaudia	0.1442	0.5469	0.4544	38
GarCha_1	GarCha	0.1377	0.5448	0.5030	39
MakeTwitterGreatAgain_1	MakeTwitterGreatAgain	0.1365	0.5444	0.4699	40
UC3M-LI_3	UC3M-LI	0.1350	0.5439	0.4749	41
exist@CeDRI_1	exist@CeDRI	0.1334	0.5434	0.4922	42
Niklas Team_1	Niklas Team	0.1333	0.5434	0.4787	43
A-squared_2	A-squared	0.1294	0.5421	0.4767	44
Chai Cheers Chutney_1	Chai Cheers Chutney	0.1271	0.5413	0.5000	45
sheila-vicente_1	sheila-vicente	0.1228	0.5399	0.4810	46
samuel-sergei_3	samuel-sergei	0.1215	0.5395	0.5106	47
alBERTAberg_2	alBERTAberg	0.1214	0.5395	0.4459	48
samuel-sergei_1	samuel-sergei	0.1169	0.5380	0.4873	49
Tweetbusters_2	Tweetbusters	0.1109	0.5361	0.4382	50
yow_1	yow	0.1098	0.5357	0.4410	51
Alberto and Ángel_1	Alberto and Ángel	0.1003	0.5326	0.4877	52
Tweetbusters_1	Tweetbusters	0.1000	0.5325	0.4570	53
NetGuardAI_1	NetGuardAI	0.0969	0.5315	0.4567	54
fhstp_3	fhstp	0.0953	0.5310	0.4888	55
Joses_1	Joses	0.0873	0.5284	0.4843	56
carlamiguel_3	carlamiguel	0.0870	0.5283	0.4263	57
carlamiguel_2	carlamiguel	0.0867	0.5282	0.4613	58
Vectorious_1	Vectorious	0.0856	0.5278	0.4546	59
Juanji&Jowi_1	Juanji&Jowi	0.0819	0.5266	0.4799	60
CogniCIC_3	CogniCIC	0.0770	0.5250	0.4851	61
BocadilloDelDia_1	BocadilloDelDia	0.0727	0.5236	0.4800	62
M&Ms_3	M&Ms	0.0669	0.5217	0.4801	63
ArPa Project_3	ArPa Project	0.0628	0.5204	0.4748	64
ArPa Project_2	ArPa Project	0.0628	0.5204	0.4748	65
CLiC_3	CLiC	0.0584	0.5190	0.4578	66
SalaPlanes_2	SalaPlanes	0.0582	0.5189	0.4659	67
CLiC_2	CLiC	0.0556	0.5181	0.4491	68
UC3M-LI_1	UC3M-LI	0.0539	0.5175	0.4583	69
TransformerHotspur_1	TransformerHotspur	0.0537	0.5175	0.3883	70
UC3M-LI_2	UC3M-LI	0.0496	0.5161	0.4627	71
fhstp_1	fhstp	0.0341	0.5111	0.4709	72
SalaPlanes_1	SalaPlanes	0.0262	0.5085	0.4729	73
M&Ms_2	M&Ms	0.0099	0.5032	0.4741	74
rubenmanu_1	rubenmanu	0.0093	0.5030	0.4225	75
E.T._1	E.T.	−0.0051	0.4983	0.4684	76
Niklas Team_2	Niklas Team	−0.0245	0.4920	0.4536	77
Data-force_1	Data-force	−0.0441	0.4857	0.4564	78
JorgeyQuique_2	JorgeyQuique	−0.0475	0.4846	0.4729	79
mumule03_2	mumule03	−0.0821	0.4733	0.4324	80
Lim-go-home_1	Lim-go-home	−0.0853	0.4723	0.4237	81
CLiC_1	CLiC	−0.0969	0.4685	0.4248	82
moniclaudia_1	moniclaudia	−0.1006	0.4673	0.4162	83
IratxeCarla_1	IratxeCarla	−0.1230	0.4600	0.4597	84
mumule03_1	mumule03	−0.1322	0.4570	0.3984	85
TransformerHotspur_2	TransformerHotspur	−0.2034	0.4339	0.3583	86
CogniCIC_2	CogniCIC	−0.2106	0.4315	0.4296	87

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	M F1	Rank
IratxeCarla_2	IratxeCarla	−0.2633	0.4144	0.3924	88
TransformerHotspur_3	TransformerHotspur	−0.4626	0.3496	0.2654	89
LNP-SP_1	LNP-SP	−0.4630	0.3495	0.3718	90
NYCU-NLP_2	NYCU-NLP	−0.4735	0.3461	0.2658	91
SergioAngel_1	SergioAngel	−0.4973	0.3383	0.3382	92
Cyberpuffs_2	Cyberpuffs	−0.5046	0.3359	0.4008	93
NYCU-NLP_3	NYCU-NLP	−0.5610	0.3176	0.2526	94
ecorbi-upv_2	ecorbi-upv	−0.5668	0.3157	0.3571	95
BRAINSTORMERS_1	BRAINSTORMERS	−0.5727	0.3138	0.2699	96
JOW_1	JOW	−0.5924	0.3074	0.2509	97
ecorbi-upv_3	ecorbi-upv	−0.6489	0.2890	0.3418	98
ecorbi-upv_1	ecorbi-upv	−0.6489	0.2890	0.3418	99
BRAINSTORMERS_2	BRAINSTORMERS	−0.6827	0.2780	0.2614	100
BRAINSTORMERS_3	BRAINSTORMERS	−0.6875	0.2765	0.2672	101
Token-Trouble_1	Token-Trouble	−0.6943	0.2743	0.2819	102
GuerraMendez_1	GuerraMendez	−0.7964	0.2411	0.3051	103
GuerraMendez_2	GuerraMendez	−0.8234	0.2323	0.3003	104
Cyberpuffs_1	Cyberpuffs	−0.8345	0.2287	0.2917	105
EXIST2025-test_majority-class	–	−0.9504	0.1910	0.1603	106
COMFOR_1	COMFOR	−0.9562	0.1891	0.2088	107
SQUADRAMP_1	SQUADRAMP	−1.0451	0.1602	0.2117	108
AlbaandRita_2	AlbaandRita	−1.1860	0.1144	0.2703	109
AlbaandRita_1	AlbaandRita	−1.2749	0.0855	0.2060	110
Nogroupnocry_1	Nogroupnocry	−1.4049	0.0432	0.0844	111
Cachapas_1	Cachapas	−1.5988	0.0000	0.2099	112
BeatrizRuiz_1	BeatrizRuiz	−1.6052	0.0000	0.1993	113
BERTinators_1	BERTinators	−1.6325	0.0000	0.2185	114
Sofia_1	Sofia	−1.6386	0.0000	0.1253	115
DoubleA_1	DoubleA	−1.6430	0.0000	0.2010	116
NeuralNomads_1	NeuralNomads	−1.6469	0.0000	0.2077	117
EXISTsValenciaWork_2	EXISTsValenciaWork	−1.6498	0.0000	0.1682	118
PabloyFede_2	PabloyFede	−1.6531	0.0000	0.1966	119
JosepYSergio_1	JosepYSergio	−1.6785	0.0000	0.1807	120
Vallecas_2	Vallecas	−1.6898	0.0000	0.1677	121
PorTod@s_1	PorTod@s	−1.7132	0.0000	0.2167	122
Vallecas_1	Vallecas	−1.7330	0.0000	0.1855	123
TheMagicToken_1	TheMagicToken	−1.7458	0.0000	0.2022	124
BeatrizRuiz_2	BeatrizRuiz	−1.7670	0.0000	0.1698	125
JENLCB_1	JENLCB	−1.7709	0.0000	0.1855	126
Güeypingüino_1	Güeypingüino	−1.8177	0.0000	0.1917	127
LNP-SP_2	LNP-SP	−1.8189	0.0000	0.1759	128
Güeypingüino_2	Güeypingüino	−1.8228	0.0000	0.1873	129
shm2025_1	shm2025	−1.8296	0.0000	0.1209	130
BeatrizRuiz_3	BeatrizRuiz	−1.8617	0.0000	0.1298	131
DuthThace_1	DuthThace	−1.8988	0.0000	0.1967	132
CS-GO_2	CS-GO	−1.9274	0.0000	0.1499	133
Awakened_3	Awakened	−2.1971	0.0000	0.2359	134
Awakened_1	Awakened	−2.2016	0.0000	0.2332	135
Awakened_2	Awakened	−2.2184	0.0000	0.2197	136
GPTesla Smashers_2	GPTesla Smashers	−3.0869	0.0000	0.1224	137
GPTesla Smashers_1	GPTesla Smashers	−3.1125	0.0000	0.1180	138
EXIST2025-test_minority-class	–	−3.1545	0.0000	0.0280	139
GPTesla Smashers_3	GPTesla Smashers	−3.2048	0.0000	0.1111	140

6.3. Task 1.3: Sexism Categorization in Tweets

6.3.1. Soft Evaluation

Table 10 displays the results of the soft-soft evaluation Task 1.3, which involves multi-label categorization of sexist content in tweets. 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 (*EXIST2025-test_majority-class*, all instances labeled as 'NO'), indicating a moderate level of competitiveness. The percentage difference between the best and the fifth team (*GrootWatch* and *A-Square*) was 22.3%, suggesting a wider 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 captured.

Table 10
Leaderboard for EXIST 2025 Task 1.3 (sexism categorization in tweets), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm.

System	Team	ICM-S	ICM-S Nr	Rank
EXIST2025-test_gold	–	9.4686	1.0000	0
GrootWatch_1	GrootWatch	**-1.1034**	**0.4417**	1
GrootWatch_3	GrootWatch	−1.1495	0.4393	2
GrootWatch_2	GrootWatch	−1.2566	0.4336	3
UMUTeam_2	UMUTeam	−1.6711	0.4118	4
Cyberpuffs_3	Cyberpuffs	−2.4632	0.3699	5
DaniReinon_1	DaniReinon	−2.5655	0.3645	6
A-squared_1	A-squared	−2.9711	0.3431	7
UMUTeam_1	UMUTeam	−3.0327	0.3399	8
A-squared_2	A-squared	−3.3214	0.3246	9
exist@CeDRI_1	exist@CeDRI	−3.4215	0.3193	10
GPTesla Smashers_1	GPTesla Smashers	−3.5062	0.3149	11
GPTesla Smashers_2	GPTesla Smashers	−3.5791	0.3110	12
GPTesla Smashers_3	GPTesla Smashers	−3.5882	0.3105	13
fhstp_2	fhstp	−3.6730	0.3060	14
Awakened_2	Awakened	−3.7444	0.3023	15
Awakened_1	Awakened	−3.7539	0.3018	16
NetGuardAI_1	NetGuardAI	−3.9061	0.2937	17
Awakened_3	Awakened	−3.9428	0.2918	18
CLiC_2	CLiC	−5.5147	0.2088	19
Cyberpuffs_2	Cyberpuffs	−5.5889	0.2049	20
CLiC_3	CLiC	−5.6447	0.2019	21
NYCU-NLP_1	NYCU-NLP	−5.8946	0.1887	22
Cyberpuffs_1	Cyberpuffs	−6.3964	0.1622	23
fhstp_1	fhstp	−7.9676	0.0793	24
Nogroupnocry_1	Nogroupnocry	−8.4635	0.0531	25
EXIST2025-test_majority-class	–	−8.7089	0.0401	26
Dandys-de-BERTganim_1	Dandys-de-BERTganim	−8.7671	0.0370	27
CLiC_1	CLiC	−9.6890	0.0000	28
COMFOR_2	COMFOR	−9.7944	0.0000	29
M&Ms_3	M&Ms	−10.1272	0.0000	30
fhstp_3	fhstp	−11.2121	0.0000	31
JOW_1	JOW	−11.5720	0.0000	32
NYCU-NLP_2	NYCU-NLP	−12.2392	0.0000	33
ecorbi-upv_3	ecorbi-upv	−12.7425	0.0000	34
LNP-SP_1	LNP-SP	−13.1963	0.0000	35
Vallecas_1	Vallecas	−13.5138	0.0000	36

(continued on next page)

(continued from previous page)

System	Team	ICM-S	ICM-S Nr	Rank
M&Ms_1	M&Ms	−14.0736	0.0000	37
M&Ms_2	M&Ms	−15.4497	0.0000	38
Vallecas_2	Vallecas	−17.7285	0.0000	39
ArPa Project_2	ArPa Project	−18.9609	0.0000	40
ArPa Project_1	ArPa Project	−18.9609	0.0000	41
ecorbi-upv_2	ecorbi-upv	−19.8913	0.0000	42
ArPa Project_3	ArPa Project	−19.9739	0.0000	43
ecorbi-upv_1	ecorbi-upv	−19.9740	0.0000	44
IratxeCarla_1	IratxeCarla	−24.0522	0.0000	45
GuerraMendez_1	GuerraMendez	−24.5888	0.0000	46
DuthThace_1	DuthThace	−25.9339	0.0000	47
BeatrizRuiz_2	BeatrizRuiz	−26.9696	0.0000	48
GuerraMendez_2	GuerraMendez	−28.4218	0.0000	49
NYCU-NLP_3	NYCU-NLP	−28.4446	0.0000	50
BeatrizRuiz_1	BeatrizRuiz	−31.8750	0.0000	51
BeatrizRuiz_3	BeatrizRuiz	−43.3581	0.0000	52
EXIST2025-test_minority-class	–	−46.1080	0.0000	53

6.3.2. Hard Evaluation

For the Hard–Hard evaluation of Task 1.3, a total of 130 systems were submitted (see Table 11). 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 (*EXIST2025-test_mayority-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. All except four systems achieved better results than the minority class baseline (all instances labeled as 'SEXUAL-VIOLENCE')

Table 11
Leaderboard for EXIST 2025 Task 1.3 (sexism categorization in tweets), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-test_gold	–	2.1533	1.0000	1.0000	0
Mario_1	Mario	**0.6519**	**0.6514**	**0.6533**	1
CIMAT-GTO_2	CIMAT-GTO	0.5413	0.6257	0.6392	2
CIMAT-GTO_3	CIMAT-GTO	0.5211	0.6210	0.6266	3
NLPDame_3	NLPDame	0.4842	0.6124	0.6335	4
NLPDame_1	NLPDame	0.4814	0.6118	0.6324	5
NLPDame_2	NLPDame	0.4515	0.6048	0.6272	6
UMUTeam_2	UMUTeam	0.4506	0.6046	0.6262	7
CIMAT-CS-NLP_2	CIMAT-CS-NLP	0.3980	0.5924	0.6125	8
GrootWatch_3	GrootWatch	0.3809	0.5884	0.6171	9
BERT-Simpson_3	BERT-Simpson	0.3751	0.5871	0.6038	10
GrootWatch_1	GrootWatch	0.3623	0.5841	0.6175	11
BERT-Simpson_1	BERT-Simpson	0.3586	0.5833	0.6006	12
BERT-Simpson_2	BERT-Simpson	0.3552	0.5825	0.5979	13
CIMAT-CS-NLP_1	CIMAT-CS-NLP	0.3353	0.5779	0.6039	14
UMUTeam_1	UMUTeam	0.3276	0.5761	0.6062	15
GrootWatch_2	GrootWatch	0.2829	0.5657	0.5986	16
Dandys-de-BERTganim_1	Dandys-de-BERTganim	0.2244	0.5521	0.5827	17
BERTin-Osborne_1	BERTin-Osborne	0.2227	0.5517	0.5764	18

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	M F1	Rank
fhstp_2	fhstp	0.1852	0.5430	0.5700	19
Awakened_1	Awakened	0.1491	0.5346	0.5629	20
M&Ms_1	M&Ms	0.1412	0.5328	0.5807	21
M&Ms_3	M&Ms	0.1395	0.5324	0.5754	22
CIMAT-CS-NLP_3	CIMAT-CS-NLP	0.1310	0.5304	0.5861	23
pau-rus_1	pau-rus	0.1205	0.5280	0.5738	24
Awakened_3	Awakened	0.1195	0.5277	0.5523	25
Awakened_2	Awakened	0.1078	0.5250	0.5525	26
CIMAT-GTO_1	CIMAT-GTO	0.0981	0.5228	0.5436	27
fhstp_3	fhstp	0.0978	0.5227	0.5728	28
alBERTAberg_2	alBERTAberg	0.0965	0.5224	0.5613	29
carlamiguel_1	carlamiguel	0.0805	0.5187	0.5721	30
samuel-sergei_3	samuel-sergei	0.0770	0.5179	0.5196	31
yow_1	yow	0.0600	0.5139	0.5667	32
M&Ms_2	M&Ms	0.0498	0.5116	0.5690	33
BERTin-Osborne_2	BERTin-Osborne	0.0493	0.5114	0.5478	34
sheila-vicente_1	sheila-vicente	0.0443	0.5103	0.5701	35
Juanji&Jowi_1	Juanji&Jowi	0.0379	0.5088	0.5637	36
carlamiguel_2	carlamiguel	0.0296	0.5069	0.5660	37
carlamiguel_3	carlamiguel	0.0290	0.5067	0.5658	38
samuel-sergei_1	samuel-sergei	0.0118	0.5027	0.5570	39
LolaClaudia_3	LolaClaudia	−0.0064	0.4985	0.5576	40
TransformerHotspur_3	TransformerHotspur	−0.0227	0.4947	0.5557	41
CLiC_3	CLiC	−0.0355	0.4918	0.5106	42
fhstp_1	fhstp	−0.0825	0.4809	0.5461	43
Niklas Team_2	Niklas Team	−0.0840	0.4805	0.5458	44
GarCha_1	GarCha	−0.1390	0.4677	0.5478	45
sadiqlovers_1	sadiqlovers	−0.1390	0.4677	0.5405	46
samuel-sergei_2	samuel-sergei	−0.1395	0.4676	0.5412	47
NetGuardAI_1	NetGuardAI	−0.1516	0.4648	0.4357	48
CLiC_1	CLiC	−0.1568	0.4636	0.4876	49
BocadilloDelDia_1	BocadilloDelDia	−0.1648	0.4617	0.5381	50
GPTesla Smashers_1	GPTesla Smashers	−0.1724	0.4600	0.4493	51
CLiC_2	CLiC	−0.1816	0.4578	0.4797	52
Niklas Team_1	Niklas Team	−0.1817	0.4578	0.5291	53
Alberto and Ángel_1	Alberto and Ángel	−0.1906	0.4557	0.5345	54
Tweetbusters_2	Tweetbusters	−0.1906	0.4557	0.5327	55
Token-Trouble_1	Token-Trouble	−0.1939	0.4550	0.5379	56
GPTesla Smashers_3	GPTesla Smashers	−0.2024	0.4530	0.4430	57
GPTesla Smashers_2	GPTesla Smashers	−0.2180	0.4494	0.4390	58
ArPa Project_3	ArPa Project	−0.2351	0.4454	0.5286	59
JorgeyQuique_2	JorgeyQuique	−0.2706	0.4372	0.5348	60
ArPa Project_1	ArPa Project	−0.2791	0.4352	0.5231	61
ArPa Project_2	ArPa Project	−0.2791	0.4352	0.5231	62
CogniCIC_3	CogniCIC	−0.2855	0.4337	0.5242	63
TransformerHotspur_2	TransformerHotspur	−0.3348	0.4223	0.5266	64
Cyberpuffs_3	Cyberpuffs	−0.3499	0.4188	0.5205	65
Lirili-Larila_3	Lirili-Larila	−0.3748	0.4130	0.4521	66
Vectorious_1	Vectorious	−0.3762	0.4127	0.5093	67
SalaPlanes_1	SalaPlanes	−0.3877	0.4100	0.5144	68
alBERTAberg_1	alBERTAberg	−0.4086	0.4051	0.4613	69
CogniCIC_1	CogniCIC	−0.4200	0.4025	0.4385	70
Lirili-Larila_2	Lirili-Larila	−0.4615	0.3928	0.4566	71
Lirili-Larila_1	Lirili-Larila	−0.4710	0.3906	0.4610	72
rubenmanu_1	rubenmanu	−0.4720	0.3904	0.4643	73

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	M F1	Rank
NYCU-NLP_1	NYCU-NLP	−0.5419	0.3742	0.3210	74
A-squared_2	A-squared	−0.5760	0.3663	0.4094	75
SalaPlanes_2	SalaPlanes	−0.6089	0.3586	0.4482	76
Mouctar Diakhaby_1	Mouctar Diakhaby	−0.6744	0.3434	0.4834	77
SergioAngel_1	SergioAngel	−0.7656	0.3222	0.4382	78
IratxeCarla_1	IratxeCarla	−0.8960	0.2919	0.4373	79
Tweetbusters_1	Tweetbusters	−0.9419	0.2813	0.4773	80
E.T._1	E.T.	−0.9692	0.2749	0.3026	81
NYCU-NLP_2	NYCU-NLP	−0.9882	0.2705	0.2779	82
Lim-go-home_1	Lim-go-home	−1.0386	0.2588	0.4005	83
GuerraMendez_2	GuerraMendez	−1.0877	0.2474	0.4369	84
BRAINSTORMERS_1	BRAINSTORMERS	−1.0982	0.2450	0.3908	85
Vallecas_1	Vallecas	−1.1051	0.2434	0.4112	86
BRAINSTORMERS_3	BRAINSTORMERS	−1.1055	0.2433	0.3902	87
GuerraMendez_1	GuerraMendez	−1.1061	0.2432	0.4375	88
A-squared_1	A-squared	−1.1291	0.2378	0.2945	89
Joses_1	Joses	−1.1463	0.2338	0.3455	90
IratxeCarla_2	IratxeCarla	−1.1664	0.2292	0.3763	91
TransformerHotspur_1	TransformerHotspur	−1.1718	0.2279	0.4445	92
BRAINSTORMERS_2	BRAINSTORMERS	−1.1864	0.2245	0.3994	93
Cyberpuffs_2	Cyberpuffs	−1.1872	0.2243	0.4080	94
NYCU-NLP_3	NYCU-NLP	−1.1945	0.2226	0.2418	95
Data-force_1	Data-force	−1.2165	0.2175	0.2236	96
JOW_1	JOW	−1.2690	0.2053	0.3254	97
LNP-SP_1	LNP-SP	−1.2874	0.2011	0.3684	98
MakeTwitterGreatAgain_1	MakeTwitterGreatAgain	−1.3093	0.1960	0.4218	99
Cyberpuffs_1	Cyberpuffs	−1.3161	0.1944	0.3691	100
CogniCIC_2	CogniCIC	−1.3830	0.1789	0.4468	101
Vallecas_2	Vallecas	−1.3927	0.1766	0.3804	102
COMFOR_1	COMFOR	−1.3994	0.1751	0.3077	103
AlbaandRita_2	AlbaandRita	−1.4531	0.1626	0.4021	104
ecorbi-upv_3	ecorbi-upv	−1.4621	0.1605	0.3319	105
BERTinators_1	BERTinators	−1.5034	0.1509	0.4045	106
DuthThace_1	DuthThace	−1.5980	0.1289	0.3897	107
EXIST2025-test_majority-class	–	−1.5984	0.1289	0.1069	108
exist@CeDRI_1	exist@CeDRI	−1.5984	0.1289	0.1069	109
AlbaandRita_1	AlbaandRita	−1.7115	0.1026	0.3762	110
DoubleA_1	DoubleA	−1.7424	0.0954	0.3726	111
SQUADRAMP_1	SQUADRAMP	−1.7461	0.0946	0.2946	112
JENLCB_1	JENLCB	−1.7726	0.0884	0.3641	113
Güeypingüino_1	Güeypingüino	−1.7869	0.0851	0.3688	114
JosepYSergio_1	JosepYSergio	−1.8373	0.0734	0.3586	115
Nogroupnocry_1	Nogroupnocry	−1.8454	0.0715	0.2557	116
EXISTsValenciaWork_2	EXISTsValenciaWork	−1.8729	0.0651	0.1613	117
Güeypingüino_2	Güeypingüino	−1.9106	0.0564	0.3529	118
Cachapas_1	Cachapas	−1.9257	0.0529	0.3465	119
moniclaudia_1	moniclaudia	−1.9281	0.0523	0.3685	120
PabloyFede_1	PabloyFede	−2.0854	0.0158	0.3203	121
Sofia_1	Sofia	−2.2868	0.0000	0.3010	122
CS-GO_3	CS-GO	−2.4413	0.0000	0.2514	123
ecorbi-upv_2	ecorbi-upv	−2.4551	0.0000	0.1816	124
NeuralNomads_1	NeuralNomads	−2.5139	0.0000	0.3208	125
PorTod@s_1	PorTod@s	−2.5143	0.0000	0.2832	126
ecorbi-upv_1	ecorbi-upv	−2.5920	0.0000	0.1788	127
EXIST2025-test_minority-class	–	−3.1295	0.0000	0.0288	128

(continued on next page)

(continued from previous page)

System	Team	ICM-H	ICM-H Nr	M F1	Rank
BeatrizRuiz_2	BeatrizRuiz	−3.1920	0.0000	0.2208	129
BeatrizRuiz_1	BeatrizRuiz	−3.8352	0.0000	0.2414	130
TheMagicToken_1	TheMagicToken	−4.3696	0.0000	0.2072	131
BeatrizRuiz_3	BeatrizRuiz	−4.9143	0.0000	0.2295	132

6.4. Task 2.1: Sexism Identification in Memes

6.4.1. Soft Evaluation

Table 12 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 baseline (*EXIST2025-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.

Table 12
Leaderboard for EXIST 2025 Task 2.1 (sexism detection in memes), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	3.1107	1.0000	0.5852	0
TrankilTwice_1	TrankilTwice	**0.0683**	**0.5110**	1.0096	1
TrankilTwice_3	TrankilTwice	0.0526	0.5084	0.9798	2
TrankilTwice_2	TrankilTwice	0.0129	0.5021	1.2963	3
surrey-mm-group_1	surrey-mm-group	−0.7061	0.3865	0.9364	4
I2C-UHU-Altair_1	I2C-UHU-Altair	−0.8558	0.3624	0.9469	5
UMUTeam_2	UMUTeam	−0.9623	0.3453	1.0554	6
UMUTeam_1	UMUTeam	−1.2113	0.3053	1.3238	7
EXIST2025-test_majority-class	–	−2.3568	0.1212	4.4015	8
Nogroupnocry_1	Nogroupnocry	−2.7060	0.0650	**0.6782**	9
EXIST2025-test_minority-class	–	−3.5089	0.0000	5.5672	10

6.4.2. Hard Evaluation

Table 13 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 *EXIST2025-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. All submissions surpassed the *EXIST2025-test_minority-class* baseline. 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.

Table 13
Leaderboard for EXIST 2025 Task 2.1 (sexism detection in memes), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, F1Y = F1 YES.

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
EXIST2025-test_gold	–	0.9832	1.0000	1.0000	0
CogniCIC_1	CogniCIC	**0.3691**	**0.6877**	**0.7810**	1
GrootWatch_3	GrootWatch	0.3589	0.6825	0.7740	2
ArcosGPT_1	ArcosGPT	0.3200	0.6627	0.7571	3
GrootWatch_2	GrootWatch	0.1898	0.5965	0.7253	4
TrankilTwice_2	TrankilTwice	0.1667	0.5848	0.7508	5
TrankilTwice_1	TrankilTwice	0.1332	0.5678	0.7304	6
TrankilTwice_3	TrankilTwice	0.1239	0.5630	0.7349	7
GrootWatch_1	GrootWatch	0.0062	0.5032	0.6898	8
I2C-UHU-Altair_2	I2C-UHU-Altair	−0.0134	0.4932	0.7125	9
I2C-UHU-Altair_1	I2C-UHU-Altair	−0.1035	0.4474	0.6987	10
NaturalThinker_1	NaturalThinker	−0.1303	0.4337	0.6837	11
surrey-mm-group_1	surrey-mm-group	−0.1575	0.4199	0.7056	12
UMUTeam_2	UMUTeam	−0.2957	0.3496	0.6420	13
UMUTeam_1	UMUTeam	−0.3043	0.3452	0.6064	14
CLTL_3	CLTL	−0.3529	0.3206	0.5113	15
CLTL_1	CLTL	−0.3809	0.3063	0.4891	16
EXIST2025-test_majority-class	–	−0.4038	0.2947	0.6821	17
CLTL_2	CLTL	−0.4096	0.2917	0.5029	18
Nogroupnocry_1	Nogroupnocry	−0.5604	0.2150	0.4783	19
EXIST2025-test_minority-class	–	−0.6468	0.1711	0.0000	20

6.5. Task 2.2: Source Intention in Memes

6.5.1. Soft Evaluation

Table 14 presents the results for the classification of memes according to the intention of the author, with the outputs provided as the 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 majority baseline. 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.

Table 14
Leaderboard for EXIST 2025 Task 2.2 (author intention detection in memes), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	4.7018	1.0000	0.9325	0
UMUTeam_1	UMUTeam	**-1.6327**	**0.3264**	1.7316	1
I2C-UHU-Altair_1	I2C-UHU-Altair	−2.0736	0.2795	1.5556	2
surrey-mm-group_1	surrey-mm-group	−2.4423	0.2403	2.0468	3
UMUTeam_2	UMUTeam	−2.4994	0.2342	1.8697	4
Nogroupnocry_1	Nogroupnocry	−4.1395	0.0598	**0.3164**	5
EXIST2025-test_majority-class	–	−5.0745	0.0000	5.5565	6
EXIST2025-test_minority-class	–	−18.9382	0.0000	8.0245	7

6.5.2. Hard Evaluation

Table 15 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 *EXIST2025-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 (near 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 15
Leaderboard for EXIST 2025 Task 2.2 (author intention detection in memes), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-test_gold	–	1.4383	1.0000	1.0000	0
CogniCIC_1	CogniCIC	**0.2254**	**0.5784**	**0.5634**	1
GrootWatch_3	GrootWatch	0.1868	0.5649	0.5513	2
ArcosGPT_1	ArcosGPT	0.0597	0.5208	0.5109	3
GrootWatch_2	GrootWatch	−0.0588	0.4796	0.4917	4
GrootWatch_1	GrootWatch	−0.3055	0.3938	0.4738	5
NaturalThinker_1	NaturalThinker	−0.5429	0.3113	0.3762	6
I2C-UHU-Altair_1	I2C-UHU-Altair	−0.6519	0.2734	0.2685	7
UMUTeam_2	UMUTeam	−0.7265	0.2474	0.3641	8
CLTL_2	CLTL	−0.7566	0.2370	0.3309	9
surrey-mm-group_1	surrey-mm-group	−0.7621	0.2351	0.3314	10
CLTL_3	CLTL	−0.7629	0.2348	0.3395	11
UMUTeam_1	UMUTeam	−0.7730	0.2313	0.3613	12
CLTL_1	CLTL	−0.7732	0.2312	0.3299	13
EXIST2025-test_majority-class	–	−1.0445	0.1369	0.1839	14
I2C-UHU-Altair_2	I2C-UHU-Altair	−1.2641	0.0606	0.2599	15
Nogroupnocry_1	Nogroupnocry	−1.3924	0.0160	0.1045	16
EXIST2025-test_minority-class	–	−2.0637	0.0000	0.0697	17

6.6. Task 2.3: Sexism Categorization in Memes

6.6.1. Soft Evaluation

Table 16 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as class probabilities. This task received 6 submissions from 4 different teams. Among these, 5 systems outperformed the majority-class baseline, while all of then outperformed the minority-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.

Table 16
Leaderboard for EXIST 2025 Task 2.3 (sexism categorization in memes), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm.

System	Team	ICM-S	ICM-S Nr	Rank
EXIST2025-test_gold	–	9.4343	1.0000	0
UMUTeam_1	UMUTeam	**-4.7791**	**0.2467**	1
UMUTeam_2	UMUTeam	−4.8825	0.2412	2
I2C-UHU-Altair_1	I2C-UHU-Altair	−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
EXIST2025-test_majority-class	–	−9.8173	0.0000	6

(continued on next page)

(continued from previous page)

System	Team	ICM-S	ICM-S Nr	Rank
Nogroupnocry_1	Nogroupnocry	−17.5040	0.0000	7
EXIST2025-test_minority-class	–	−50.0353	0.0000	8

6.6.2. Hard Evaluation

Finally, Table 17 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 from different teams was 59.5%, indicating competitive but not saturated performance across top ranks. All systems clearly outperformed the `EXIST2025-test_minority-class` baseline.

Table 17
Leaderboard for EXIST 2025 Task 2.3 (sexism categorization in memes), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-test_gold	–	2.4100	1.0000	1.0000	0
CogniCIC_1	CogniCIC	**0.0244**	**0.5051**	**0.5763**	1
GrootWatch_3	GrootWatch	−0.0798	0.4834	0.5472	2
GrootWatch_2	GrootWatch	−0.3550	0.4263	0.5119	3
ArcosGPT_1	ArcosGPT	−0.4187	0.4131	0.5501	4
GrootWatch_1	GrootWatch	−0.5812	0.3794	0.4921	5
I2C-UHU-Altair_1	I2C-UHU-Altair	−0.9958	0.2934	0.4223	6
I2C-UHU-Altair_2	I2C-UHU-Altair	−1.1838	0.2544	0.3786	7
CLTL_2	CLTL	−1.4243	0.2045	0.3143	8
CLTL_3	CLTL	−1.5325	0.1820	0.2851	9
UMUTeam_1	UMUTeam	−1.5624	0.1758	0.3582	10
CLTL_1	CLTL	−1.6077	0.1664	0.2573	11
UMUTeam_2	UMUTeam	−1.8869	0.1085	0.3153	12
NaturalThinker_1	NaturalThinker	−2.0376	0.0773	0.1599	13
EXIST2025-test_majority-class	–	−2.0711	0.0703	0.0919	14
surrey-mm-group_1	surrey-mm-group	−2.9992	0.0000	0.3135	15
EXIST2025-test_minority-class	–	−3.3135	0.0000	0.0318	16

6.7. Task 3.1: Sexism Identification in Videos

6.7.1. Soft Evaluation

Table 18 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 (*EXIST2025-test_majority-class*). 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.

Table 18
Leaderboard for EXIST 2025 Task 3.1 (sexism detection in videos), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	2.8488	1.0000	0.1962	0
LaVellaPremium_2	LaVellaPremium	**0.3362**	**0.5590**	1.5731	1
LaVellaPremium_1	LaVellaPremium	**0.3362**	**0.5590**	1.5731	2
MIARFID ducks_2	MIARFID ducks	0.2968	0.5521	1.7725	3
MIARFID ducks_1	MIARFID ducks	0.2956	0.5519	1.3931	4
YesWeEXIST_1	YesWeEXIST	0.2759	0.5484	1.9034	5
MIARFID ducks_3	MIARFID ducks	0.2257	0.5396	1.3827	6
profLayton_1	profLayton	0.1779	0.5312	0.9870	7
profLayton_3	profLayton	−0.1064	0.4813	1.1656	8
profLayton_2	profLayton	−0.1444	0.4747	1.0069	9
EmbeddingGuards_1	EmbeddingGuards	−0.2429	0.4574	0.8413	10
DanKyuPre_1	DanKyuPre	−0.2846	0.4500	0.9422	11
Raulet_1	Raulet	−0.2856	0.4499	0.9153	12
Raulet_2	Raulet	−0.2952	0.4482	0.9026	13
Raulet_3	Raulet	−0.4437	0.4221	0.8735	14
ECA-SIMM-UVa_1	ECA-SIMM-UVa	−0.4517	0.4207	**0.8378**	15
ScalaR_2	ScalaR	−0.4586	0.4195	0.9184	16
ECA-SIMM-UVa_2	ECA-SIMM-UVa	−0.4774	0.4162	0.8387	17
ECA-SIMM-UVa_3	ECA-SIMM-UVa	−0.5234	0.4081	0.8475	18
The Gamblers_1	The Gamblers	−0.5329	0.4065	3.7152	19
ScalaR_1	ScalaR	−0.5541	0.4027	0.9106	20
jdsanroj_1	jdsanroj	−0.5652	0.4008	0.8740	21
biasedmodels_1	biasedmodels	−0.7662	0.3655	2.4142	22
The Gamblers_2	The Gamblers	−0.9665	0.3304	0.9432	23
The Gamblers_3	The Gamblers	−1.0384	0.3177	1.0117	24
jdsanroj_2	jdsanroj	−1.1982	0.2897	0.9136	25
EXIST2025-test_majority-class	–	−1.2877	0.2740	4.4285	26
LaVellaPremium_3	LaVellaPremium	−1.4688	0.2422	0.9951	27
DanKyuPre_2	DanKyuPre	−1.4834	0.2396	0.9926	28
EXISTencialCrisis_2	EXISTencialCrisis	−1.5375	0.2302	1.0307	29
UMUTeam_1	UMUTeam	−1.9857	0.1515	3.4790	30
EXIST2025-test_minority-class	–	−2.0051	0.1481	5.5402	31
EXISTencialCrisis_1	EXISTencialCrisis	−2.4801	0.0647	1.6832	32
Nogroupnocry_1	Nogroupnocry	−2.6323	0.0380	1.2770	33
DS@GT EXIST_1	DS@GT EXIST	−2.7570	0.0161	1.5763	34
DS@GT EXIST_2	DS@GT EXIST	−3.1497	0.0000	1.9086	35
DS@GT EXIST_3	DS@GT EXIST	−3.2477	0.0000	2.1939	36

6.7.2. Hard Evaluation

Finally, Table 19 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 (*EXIST2025-test_majority-class*), showing strong global performance. The top five teams, as can be observed from Table 19, were closely matched, with only a 4.0% difference between the best and lowest performer among the top five.

Table 19
Leaderboard for EXIST 2025 Task 3.1 (sexism detection in videos), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, F1Y = F1 YES.

System	Team	ICM-H	ICM-H Nr	F1Y	Rank
EXIST2025-test_gold	–	0.9907	1.0000	1.0000	0
ECA-SIMM-UVa_3	ECA-SIMM-UVa	**0.1984**	**0.6001**	0.6935	1
CogniCIC_1	CogniCIC	0.1940	0.5979	0.6835	2
ECA-SIMM-UVa_2	ECA-SIMM-UVa	0.1827	0.5922	0.6833	3
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
ECA-SIMM-UVa_1	ECA-SIMM-UVa	0.1445	0.5730	0.6643	7
profLayton_1	profLayton	0.1395	0.5704	0.7008	8
EXISTencialCrisis_2	EXISTencialCrisis	0.1391	0.5702	0.6780	9
Raulet_1	Raulet	0.1287	0.5650	0.6800	10
DanKyuPre_1	DanKyuPre	0.1128	0.5569	0.6597	11
jdsanroj_2	jdsanroj	0.1121	0.5566	0.6755	12
EXISTencialCrisis_1	EXISTencialCrisis	0.1090	0.5550	0.6857	13
Raulet_2	Raulet	0.1067	0.5539	0.6632	14
DanKyuPre_2	DanKyuPre	0.1059	0.5534	0.6700	15
AIDONTTOKSEXISM_1	AIDONTTOKSEXISM	0.0926	0.5467	0.6656	16
EXISTencialCrisis_3	EXISTencialCrisis	0.0760	0.5384	0.6931	17
profLayton_2	profLayton	0.0757	0.5382	0.6781	18
Raulet_3	Raulet	0.0675	0.5341	0.6318	19
KeTEAM_3	KeTEAM	0.0671	0.5339	0.6410	20
jdsanroj_1	jdsanroj	0.0487	0.5246	0.6440	21
YesWeEXIST_1	YesWeEXIST	0.0448	0.5226	0.6340	22
LaVellaPremium_3	LaVellaPremium	0.0337	0.5170	0.6527	23
profLayton_3	profLayton	0.0325	0.5164	0.6667	24
KeTEAM_1	KeTEAM	0.0270	0.5137	0.6784	25
KeTEAM_2	KeTEAM	0.0261	0.5132	0.6528	26
ScalaR_1	ScalaR	−0.0094	0.4953	0.6502	27
SantiMG_2	SantiMG	−0.0192	0.4903	0.6281	28
ScalaR_2	ScalaR	−0.0258	0.4870	0.6299	29
SantiMG_1	SantiMG	−0.1022	0.4484	0.5828	30
I2C-UHU-Sirius_1	I2C-UHU-Sirius	−0.1278	0.4355	0.5768	31
The Gamblers_1	The Gamblers	−0.1331	0.4328	0.5197	32
biasedmodels_1	biasedmodels	−0.1533	0.4227	0.4272	33
The Gamblers_2	The Gamblers	−0.1837	0.4073	0.5374	34
The Gamblers_3	The Gamblers	−0.2197	0.3891	0.5723	35
DS@GT EXIST_3	DS@GT EXIST	−0.2282	0.3848	0.6108	36
DS@GT EXIST_2	DS@GT EXIST	−0.2456	0.3761	0.5782	37
DS@GT EXIST_1	DS@GT EXIST	−0.3933	0.3015	0.4782	38
I2C-UHU-Sirius_2	I2C-UHU-Sirius	−0.4212	0.2874	0.1379	39
EXIST2025-test_majority-class	–	−0.4244	0.2858	0.0000	40
Nogroupnocry_1	Nogroupnocry	−0.5014	0.2469	0.4010	41
EXIST2025-test_minority-class	–	−0.6036	0.1954	0.6117	42
UMUTeam_1	UMUTeam	−0.6926	0.1504	0.4549	43

6.8. Task 3.2: Source Intention in Videos

6.8.1. Soft Evaluation

Table 20 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 (*EXIST2025-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.

Table 20
Leaderboard for EXIST 2025 Task 3.2 (author intention detection in videos), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm, CE = Cross Entropy.

System	Team	ICM-S	ICM-S Nr	CE	Rank
EXIST2025-test_gold	–	4.6948	1.0000	0.2550	0
MIARFID ducks_2	MIARFID ducks	**-1.1940**	**0.3728**	1.7731	1
MIARFID ducks_3	MIARFID ducks	−1.3319	0.3581	1.6850	2
EXISTencialCrisis_1	EXISTencialCrisis	−1.3535	0.3558	3.0998	3
profLayton_1	profLayton	−1.3821	0.3528	1.4974	4
MIARFID ducks_1	MIARFID ducks	−1.5329	0.3367	1.7514	5
YesWeEXIST_1	YesWeEXIST	−1.6151	0.3280	1.5712	6
LaVellaPremium_1	LaVellaPremium	−1.6159	0.3279	1.3258	7
profLayton_2	profLayton	−1.6519	0.3241	1.4987	8
EXISTencialCrisis_2	EXISTencialCrisis	−1.7235	0.3164	3.1922	9
LaVellaPremium_3	LaVellaPremium	−1.8572	0.3022	1.2861	10
DanKyuPre_2	DanKyuPre	−2.0099	0.2859	1.2937	11
LaVellaPremium_2	LaVellaPremium	−2.0166	0.2852	1.4651	12
DanKyuPre_1	DanKyuPre	−2.0681	0.2797	1.2959	13
profLayton_3	profLayton	−2.0966	0.2767	1.5257	14
AIDontTokSexism_1	AIDontTokSexism	−2.1499	0.2710	2.0872	15
jdsanroj_2	jdsanroj	−2.3405	0.2507	1.1432	16
biasedmodels_1	biasedmodels	−2.6193	0.2210	2.9887	17
Raulet_2	Raulet	−2.6593	0.2168	1.4018	18
ScalaR_1	ScalaR	−2.6853	0.2140	1.2383	19
The Gamblers_1	The Gamblers	−2.7037	0.2120	4.3420	20
Raulet_3	Raulet	−2.7801	0.2039	1.3898	21
ScalaR_2	ScalaR	−2.8495	0.1965	1.2694	22
EmbeddingGuards_1	EmbeddingGuards	−2.8659	0.1948	1.2380	23
jdsanroj_1	jdsanroj	−2.9608	0.1847	1.2072	24
The Gamblers_2	The Gamblers	−2.9968	0.1808	1.3496	25
UMUTeam_1	UMUTeam	−3.0703	0.1730	3.6690	26
EXIST2025-test_majority-class	–	−3.1337	0.1663	4.4354	27
The Gamblers_3	The Gamblers	−3.3874	0.1392	1.4123	28
Raulet_1	Raulet	−3.6180	0.1147	1.4756	29
Nogroupnocry_1	Nogroupnocry	−4.5051	0.0202	**0.3763**	30
EXIST2025-test_minority-class	–	−15.4368	0.0000	8.8286	31

6.8.2. Hard Evaluation

Table 21 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 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 21
Leaderboard for EXIST 2025 Task 3.2 (author intention detection in videos), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-test_gold	–	1.3244	1.0000	1.0000	0
CogniCIC_1	CogniCIC	**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
EXISTencialCrisis_2	EXISTencialCrisis	−0.0953	0.4640	0.5609	6
EXISTencialCrisis_3	EXISTencialCrisis	−0.0990	0.4626	0.5658	7
EXISTencialCrisis_1	EXISTencialCrisis	−0.1209	0.4544	0.5511	8
YesWeEXIST_1	YesWeEXIST	−0.1317	0.4503	0.5495	9
LaVellaPremium_3	LaVellaPremium	−0.1679	0.4366	0.5436	10
LaVellaPremium_2	LaVellaPremium	−0.1690	0.4362	0.5557	11
ScalaR_1	ScalaR	−0.1778	0.4329	0.5444	12
AIDontTokSexism_2	AIDontTokSexism	−0.1831	0.4309	0.5346	13
AIDontTokSexism_1	AIDontTokSexism	−0.1933	0.4270	0.4198	14
Raulet_1	Raulet	−0.2068	0.4219	0.5253	15
DanKyuPre_2	DanKyuPre	−0.2097	0.4208	0.5606	16
YesWeEXIST_2	YesWeEXIST	−0.2303	0.4131	0.5260	17
ScalaR_2	ScalaR	−0.2464	0.4070	0.5255	18
Raulet_2	Raulet	−0.2477	0.4065	0.5190	19
jdsanroj_1	jdsanroj	−0.2680	0.3988	0.5001	20
KeTEAM_1	KeTEAM	−0.2846	0.3926	0.5230	21
DanKyuPre_1	DanKyuPre	−0.2925	0.3896	0.5362	22
YesWeEXIST_3	YesWeEXIST	−0.3180	0.3800	0.4982	23
Raulet_3	Raulet	−0.3374	0.3726	0.4987	24
KeTEAM_2	KeTEAM	−0.3555	0.3658	0.4963	25
profLayton_2	profLayton	−0.3596	0.3643	0.4835	26
profLayton_3	profLayton	−0.3676	0.3612	0.4685	27
KeTEAM_3	KeTEAM	−0.3910	0.3524	0.4902	28
The Gamblers_1	The Gamblers	−0.4489	0.3305	0.4619	29
biasedmodels_1	biasedmodels	−0.4597	0.3265	0.3864	30
The Gamblers_2	The Gamblers	−0.5351	0.2980	0.4501	31
The Gamblers_3	The Gamblers	−0.6355	0.2601	0.4177	32
SantiMG_2	SantiMG	−0.7200	0.2282	0.4353	33
EXIST2025-test_majority-class	–	−0.7537	0.2155	0.2375	34
SantiMG_1	SantiMG	−1.0128	0.1177	0.3668	35
UMUTeam_1	UMUTeam	−1.1856	0.0524	0.2944	36
Nogroupnocry_1	Nogroupnocry	−1.2902	0.0129	0.0860	37
EXIST2025-test_minority-class	–	−2.4749	0.0000	0.0586	38

6.9. Task 3.3: Sexism Categorization in Videos

6.9.1. Soft Evaluation

Table 22 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.

Table 22
Leaderboard for EXIST 2025 Task 3.3 (sexism categorization in videos), for the soft evaluation. Metrics: ICM-S = ICM Soft, ICM-S Nr = ICM Soft Norm.

System	Team	ICM-S	ICM-S Nr	Rank
EXIST2025-test_gold	–	8.3833	1.0000	0
EXISTencialCrisis_1	EXISTencialCrisis	**-5.7131**	**0.1593**	1
AIDONTTOKSEXISM_1	AIDONTTOKSEXISM	−6.0447	0.1395	2
AIDONTTOKSEXISM_ESEN1	AIDONTTOKSEXISM	−6.0447	0.1395	3
LaVellaPremium_1	LaVellaPremium	−6.2730	0.1259	4
LaVellaPremium_2	LaVellaPremium	−6.3367	0.1221	5
biasedmodels_1	biasedmodels	−6.5149	0.1114	6
AIDONTTOKSEXISM_3	AIDONTTOKSEXISM	−6.5633	0.1085	7
AIDONTTOKSEXISM_ES1	AIDONTTOKSEXISM	−6.5633	0.1085	8
EXIST2025-test_majority-class	–	−6.8222	0.0931	9
profLayton_2	profLayton	−6.9313	0.0866	10
LaVellaPremium_3	LaVellaPremium	−6.9919	0.0830	11
jdsanroj_2	jdsanroj	−7.1798	0.0718	12
AIDONTTOKSEXISM_EN1	AIDONTTOKSEXISM	−7.2655	0.0667	13
AIDONTTOKSEXISM_2	AIDONTTOKSEXISM	−7.2655	0.0667	14
profLayton_1	profLayton	−7.2743	0.0661	15
YesWeEXIST_1	YesWeEXIST	−7.5679	0.0486	16
ScalaR_1	ScalaR	−7.8219	0.0335	17
profLayton_3	profLayton	−7.9543	0.0256	18
EmbeddingGuards_1	EmbeddingGuards	−8.0243	0.0214	19
jdsanroj_1	jdsanroj	−8.0721	0.0186	20
ScalaR_2	ScalaR	−8.1615	0.0132	21
DanKyuPre_1	DanKyuPre	−8.1977	0.0111	22
Nogroupnocry_1	Nogroupnocry	−8.2025	0.0108	23
DanKyuPre_2	DanKyuPre	−8.4404	0.0000	24
The Gamblers_2	The Gamblers	−9.0347	0.0000	25
UMUTeam_1	UMUTeam	−9.0825	0.0000	26
The Gamblers_1	The Gamblers	−9.1817	0.0000	27
EXIST2025-test_minority-class	–	−11.6668	0.0000	28
MIARFID ducks_3	MIARFID ducks	−13.1053	0.0000	29
Raulet_3	Raulet	−13.1988	0.0000	30
Raulet_2	Raulet	−14.9062	0.0000	31
MIARFID ducks_1	MIARFID ducks	−14.9274	0.0000	32
MIARFID ducks_2	MIARFID ducks	−18.3113	0.0000	33
Raulet_1	Raulet	−23.1900	0.0000	34

6.9.2. Hard Evaluation

Finally, Table 23 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 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.

Table 23
Leaderboard for EXIST 2025 Task 3.3 (sexism categorization in videos), for the hard evaluation. Metrics: ICM-H = ICM Hard, ICM-H Nr = ICM Hard Norm, M F1 = Macro F1.

System	Team	ICM-H	ICM-H Nr	M F1	Rank
EXIST2025-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
KeTEAM_3	KeTEAM	−0.4057	0.3687	0.2622	4
ScalaR_1	ScalaR	−0.4102	0.3673	0.2533	5
ScalaR_2	ScalaR	−0.4315	0.3604	0.2478	6
jdsanroj_1	jdsanroj	−0.4373	0.3585	0.2516	7
profLayton_1	profLayton	−0.4780	0.3453	0.3345	8
EXISTencialCrisis_1	EXISTencialCrisis	−0.5322	0.3278	0.3518	9
AIDONTTOKSEXISM_EN1	AIDONTTOKSEXISM	−0.5501	0.3220	0.2489	10
AIDONTTOKSEXISM_2	AIDONTTOKSEXISM	−0.5501	0.3220	0.2489	11
EXISTencialCrisis_2	EXISTencialCrisis	−0.5571	0.3197	0.3632	12
LaVellaPremium_1	LaVellaPremium	−0.5843	0.3110	**0.3925**	13
AIDONTTOKSEXISM_1	AIDONTTOKSEXISM	−0.5863	0.3103	0.2871	14
AIDONTTOKSEXISM_ESEN1	AIDONTTOKSEXISM	−0.5863	0.3103	0.2871	15
EmbeddingGuards_1	EmbeddingGuards	−0.5874	0.3099	0.3470	16
CogniCIC_1	CogniCIC	−0.6151	0.3010	0.3519	17
profLayton_2	profLayton	−0.6174	0.3002	0.2499	18
LaVellaPremium_3	LaVellaPremium	−0.6517	0.2891	0.3173	19
I2C-UHU-Sirius_1	I2C-UHU-Sirius	−0.6644	0.2850	0.2436	20
LaVellaPremium_2	LaVellaPremium	−0.6818	0.2794	0.3441	21
biasedmodels_1	biasedmodels	−0.7097	0.2704	0.1928	22
KeTEAM_2	KeTEAM	−0.7558	0.2554	0.3172	23
jdsanroj_2	jdsanroj	−0.7681	0.2515	0.1812	24
AIDONTTOKSEXISM_3	AIDONTTOKSEXISM	−0.7903	0.2443	0.2533	25
AIDONTTOKSEXISM_ES1	AIDONTTOKSEXISM	−0.7903	0.2443	0.2533	26
YesWeEXIST_3	YesWeEXIST	−0.7995	0.2413	0.3097	27
DanKyuPre_2	DanKyuPre	−0.8725	0.2177	0.2919	28
SantiMG_2	SantiMG	−0.8970	0.2098	0.1924	29
The Gamblers_1	The Gamblers	−0.9215	0.2018	0.2036	30
EXIST2025-test_majority-class	–	−0.9530	0.1916	0.1188	31
DanKyuPre_1	DanKyuPre	−1.0039	0.1752	0.2638	32
The Gamblers_2	The Gamblers	−1.0071	0.1741	0.1971	33
YesWeEXIST_2	YesWeEXIST	−1.0275	0.1675	0.3241	34
SantiMG_1	SantiMG	−1.2490	0.0959	0.1717	35
Nogroupnocry_1	Nogroupnocry	−1.3309	0.0694	0.1350	36
Raulet_3	Raulet	−1.4911	0.0175	0.2962	37
Raulet_1	Raulet	−1.7781	0.0000	0.3037	38
Raulet_2	Raulet	−1.8602	0.0000	0.2776	39
UMUTeam_1	UMUTeam	−2.7332	0.0000	0.2095	40
EXIST2025-test_minority-class	–	−6.7467	0.0000	0.0025	41

6.10. Cross-task Performance Analysis

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.

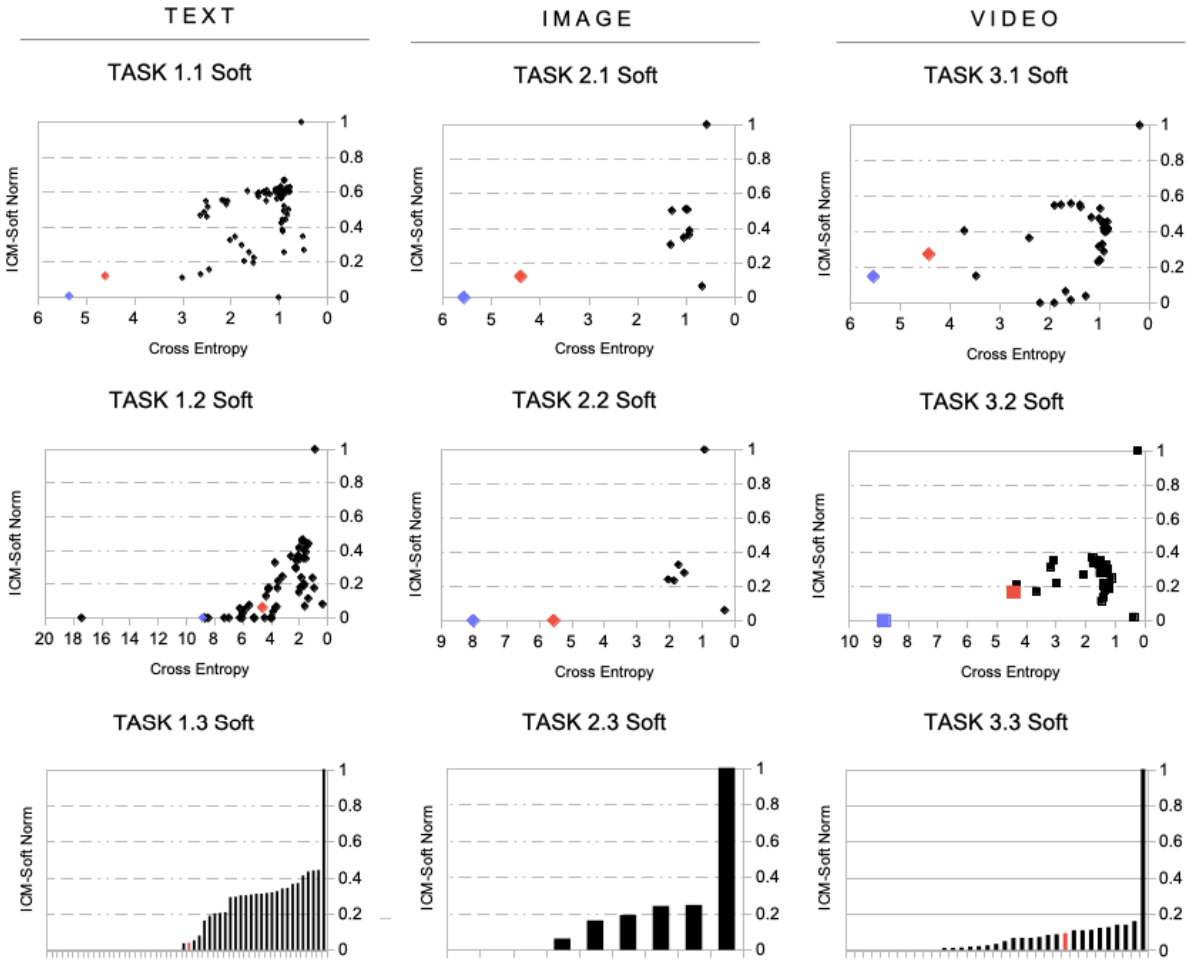

Figure 4: ICM-Soft and Cross Entropy results across tasks. The baseline approaches consisting of labeling everything as the majority class or as the minority class are marked in blue and red, respectively. The system results are marked in black.

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. While on the horizontal axis (cross-entropy), clusters of outputs with high ICM similarity are located far from the baseline in terms of cross-entropy, all the graphs show ICM ranges with high cross-entropy values spanning from the maximum down to the baseline. This may be due, 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 **intention 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 than 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 surpass the zero threshold in normalized ICM. This suggests that some 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.

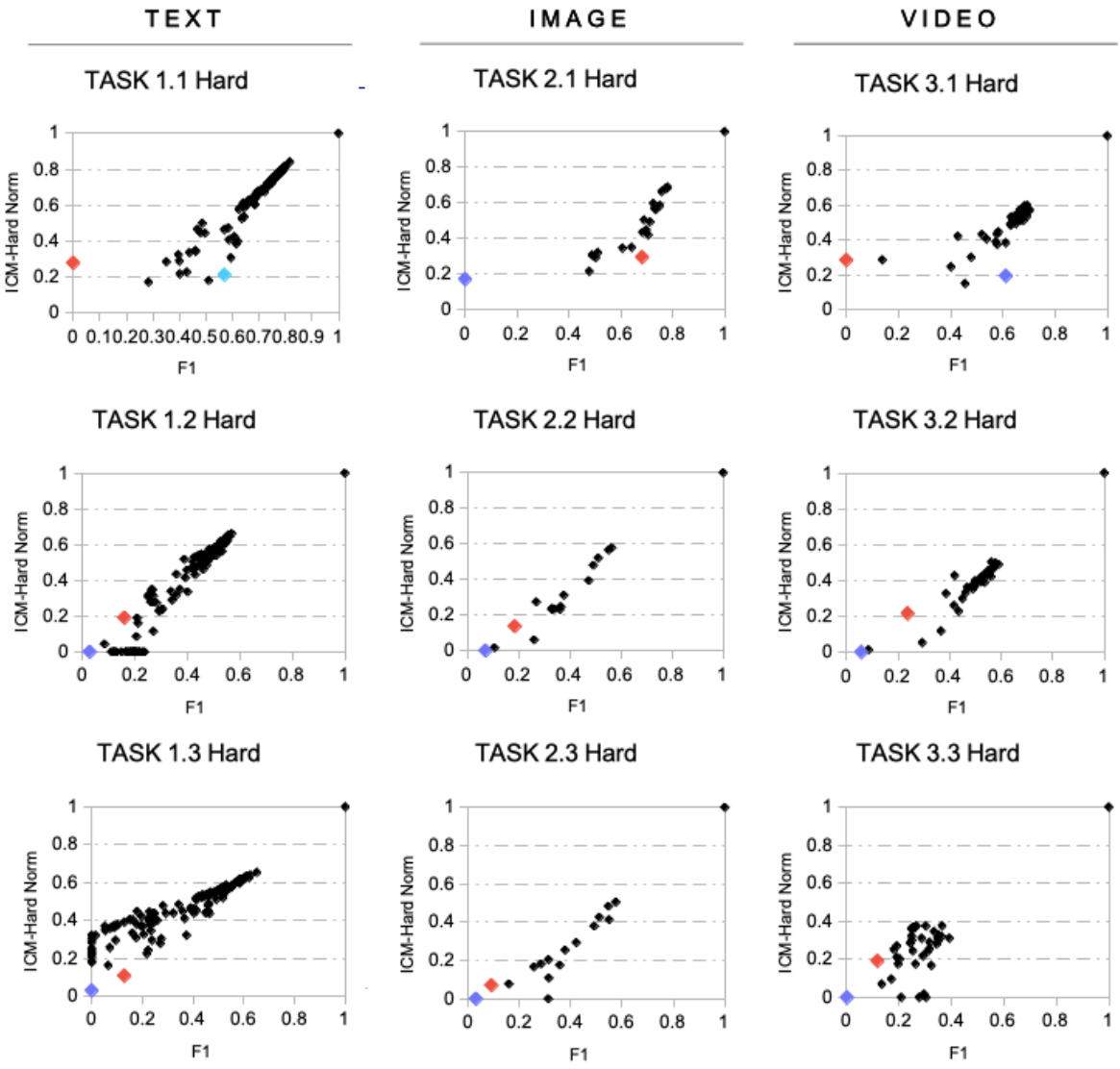

Figure 5: ICM-Hard and F1 results across tasks. The baseline approaches consisting of labeling everything as the majority class or as the minority class are marked in blue and red, respectively. The system results are marked in black.

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. Discussion

The following discussion analyzes system performance across the full range of tasks proposed in EXIST 2025, which include the detection, intent classification, and fine-grained categorization of sexist content. For the first time in the series, these tasks have been applied not only to textual data (tweets), but also to memes and short-form videos (TikToks), enabling a broad multimodal evaluation. The section is structured into three parts, each focusing on one of the core challenges: sexism detection, source intention, and categorization, allowing us to examine how the nature of the input content (text, image, or video) affects model effectiveness.

7.1. System Performance Across Text, Memes, and Video in Sexism Detection

As it can be observed in Table 24, which summarizes the best results for the subtasks 1.1, 2.1 and 3.1 (sexism detection in tweets, memes and TikTok videos, respectively), the **tweets (text)** dataset yielded the highest detection performance, while **memes** and especially **videos** proved more challenging. In the Soft-Soft evaluation (probabilistic outputs), the top system on tweets achieved an ICM-Soft Norm of $\sim$0.67, notably higher than the top systems on memes (0.511) and videos (0.559), as shown in Table 24. In the Hard-Hard evaluation (binary outputs), tweet data again saw the best results with the top F1 (positive class) $\sim$0.817 and a normalized ICM-Hard of $\sim$0.84. Memes were intermediate (top F1 $\sim$0.781, Norm $\sim$0.688), and videos the lowest (top F1 $\sim$0.694, Norm $\sim$0.600). These gaps suggest that the data source significantly influences system performance. Models detect sexism in raw text more effectively than in images or videos, likely due to the noise and information loss introduced when dealing with multimedia content.

Table 24
Top system results for each data source and evaluation context (Soft-Soft: ICM-Soft Norm, Hard-Hard: ICM-Hard Norm, F1 for the YES class).

Source	Soft-Soft	Hard-Hard	ICM-Soft Norm	ICM-Hard Norm	F1 (YES)
Tweets	TeamX	TeamY	0.670	0.840	0.817
Memes	CogniCIC	GrootWatch	0.511	0.688	0.781
Videos	CogniCIC	VideoTeam	0.559	0.600	0.694

Even state-of-the-art multimodal systems face difficulties with blurry or stylized text and background clutter in memes, which can explain the reduced accuracy on the meme and video datasets. The lower results on Subtask 3.1 (videos) align with the expectation that multimodal sexism detection is a novel and challenging problem, less studied than text-based sexism and complicated by needing to interpret visual or audio context. Overall, tweet-based models outperformed those on OCR-derived text, underlining how a clean text signal (tweets) is easier for current NLP systems to handle compared to extracted text from images or videos.

The lower performance observed in memes and videos is not solely attributable to the multimodal nature of these formats. Beyond the technical challenges of processing visual and audio data, these media often rely on implicit cultural references, sarcasm, irony, and contextual humor that are difficult to interpret automatically. Memes, in particular, tend to condense layered meanings into very short texts superimposed on images, often requiring familiarity with platform-specific discourse, internet slang, or ongoing social debates. Similarly, TikTok videos frequently reference adolescent trends, in-group codes,

and popular audio tracks, which may be opaque to both annotators and systems unless they share that sociocultural context. These aspects introduce a level of pragmatic and cultural ambiguity that goes beyond the limitations of current vision or language models, and point to the need for systems that can integrate both multimodal understanding and world knowledge to interpret such content effectively.

7.2. System Performance Across Text, Memes, and Video in Sexism Source Intention

This task required systems to predict the intention behind online sexist content, with a hierarchical multiclass setup. The classification pipeline first determines whether the content is sexist, and then predicts the fine-grained intention: *DIRECT*, *REPORTED* (tweets only), or *JUDGEMENTAL*. Table 25 presents the top systems and their evaluation metrics for each modality and context.

Table 25
Top system results for each data source and evaluation context in Task 2. Soft-Soft reports ICM-Soft Norm, Hard-Hard reports ICM-Hard Norm and Macro F1.

Source	Soft-Soft	Hard-Hard	ICM-Soft Norm	ICM-Hard Norm	Macro F1
Tweets	GrootWatch_1	Mario_1	0.4647	0.6623	0.5692
Memes	UMUTeam_1	CogniCIC_1	0.3264	0.5784	0.5634
Videos	MIARFID ducks_2	CogniCIC_1	0.3728	0.5018	0.5623

As observed in Table 25, tweet-based systems once again outperform meme and video systems, especially in the Soft-Soft (probabilistic) evaluation. However, absolute values of all metrics are lower than in binary sexism detection, reflecting the increased difficulty of intention identification, particularly in noisy or OCR-extracted content. Notably, the performance gap between modalities is less pronounced in Macro F1 than in ICM-Soft, suggesting that top systems are better at predicting the main class, but struggle with fine calibration to the true distribution of annotator votes.

The gap between tweet, meme, and video results is partially explained by the challenges posed by multimodal and OCR-derived content, as in Tasks 1.1, 2.1 and 3.1. Additionally, the removal of the REPORTED class from memes and videos (a design choice based on data inspection) means that systems face a simpler but less nuanced label space in those domains. This may contribute to the relatively high Macro F1 in memes and videos, as models need only differentiate between fewer classes.

Moreover, the higher prevalence of the DIRECT class in memes aligns with the nature of meme content, which often features explicit or humorous sexist material. Systems tuned to this distribution may perform well in memes but generalize poorly to tweets, where REPORTED and JUDGEMENTAL are more common and context-dependent.

7.3. System Performance Across Text, Memes, and Video in Sexism Categorization

Tasks 1.3, 2.3 and 3.3 addressed the multilabel, multiclass, and hierarchical classification of online sexist content, where systems must not only detect sexist content, but also assign one or more fine-grained categories indicating the *facet of womanhood* under attack. The categories include *IDEOLOGICAL AND INEQUALITY*, *STEREOTYPING AND DOMINANCE*, *OBJECTIFICATION*, *SEXUAL VIOLENCE*, and *MISOGYNY AND NON-SEXUAL VIOLENCE*.

Table 26 presents the top system performances for each data source and context. The overall pattern mirrors previous tasks: tweet-based systems consistently outperform those on memes and videos, especially in the probabilistic (Soft-Soft) context. However, absolute metrics are lower than for binary or intention-based sexism detection, reflecting the increased complexity of the multilabel, hierarchical setup and the annotation ambiguity intrinsic to these subtle categories.

In all modalities, ICM-Soft Norm scores are considerably lower than in previous tasks, indicating that systems struggle to accurately capture the distribution of annotator opinions and to model multilabel uncertainty. Notably, even the best systems on tweets barely exceed 0.41 in ICM-Soft Norm, with further drops for memes and videos.

Table 26
Top system results for each data source and evaluation context in Task 3. Soft-Soft: ICM-Soft Norm, Hard-Hard: ICM-Hard Norm and Macro F1.

Source	Soft-Soft	Hard-Hard	ICM-Soft Norm	ICM-Hard Norm	Macro F1
Tweets	GrootWatch_3	Mario_3	0.4112	0.5791	0.4468
Memes	CogniCIC_3	CogniCIC_3	0.3158	0.4942	0.4212
Videos	GrootWatch_3	CogniCIC_3	0.3220	0.4208	0.3896

7.4. Performance Trends on Tweet-based Tasks (2023–2025)

To better understand the progress in sexism detection over time, we compared the best-performing systems across the three tweet-based tasks (Tasks 1.1, 1.2, and 1.3) in the last three editions of EXIST. The results, shown in Table 27, include both ICM-Soft scores and their normalized counterparts (when available).

Table 27
Best system ICM-Soft scores on tweet-based tasks across EXIST editions (2023–2025). The normalized scores are provided in parenthesis when available.

Year	Task 1.1	Task 1.2	Task 1.3
2023	0.90	-1.34	-2.32
2024	1.09 (0.68)	-0.25 (0.48)	-1.18 (0.44)
2025	1.06 (0.67)	-0.43 (0.46)	-1.10 (0.44)

The data suggests a clear performance improvement from 2023 to 2024, likely reflecting the broader adoption of large language models and increasingly refined prompt engineering and fine-tuning strategies. This gain is particularly visible in the source intention and category classification tasks (1.2 and 1.3), which traditionally require more nuanced modeling.

Interestingly, 2025 shows no clear progress over 2024, despite a significant increase in the number of participants and submitted runs. In fact, the best normalized scores for Tasks 1.2 and 1.3 in 2025 are slightly lower than the previous year. This raises the important question: are we reaching a performance ceiling on these tasks when using the same dataset? One possible explanation is saturation — as systems converge toward similar architectures and training data, gains become increasingly marginal. Moreover, when using the same test data over multiple editions, top systems may begin to approach the upper bounds of what can be achieved without new annotation rounds or more diverse evaluation settings.

These findings highlight the importance of refreshing datasets, increasing task complexity, or shifting focus to novel and underexplored modalities to maintain scientific progress and distinguish truly innovative approaches.

8. Conclusions

The objective of the EXIST challenge is to foster research on the automatic detection and modeling of sexism in online environments, with a particular emphasis on social networks. The 2025 edition of the lab, organized as part of CLEF, attracted 114 participant teams and received a total of 873 system runs. Participants explored a wide range of approaches, including vision transformer models, data augmentation via automatic translation and duplication, the use of data from previous EXIST editions, multilingual and Twitter-specific language models, as well as transfer learning from related domains such as hate speech, toxicity, and sentiment analysis.

The tasks in EXIST 2025 addressed the problem of sexism detection and classification across three types of content—text (tweets), images (memes), and video (TikToks)—demonstrating the comprehensive and

multimodal scope of the challenge. This multimodal design reflects the complexity of real-world social media platforms, where sexist messages may be conveyed through language, visuals, or a combination of both.

While many participating systems followed the conventional strategy of producing hard-label outputs, a substantial number took advantage of the multi-annotator nature of the dataset to submit soft-label predictions. This shift indicates a growing interest within the research community in building models that can handle subjectivity, disagreement, and nuanced interpretations of harmful content.

Acknowledgments

This work has been financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Digital Transformation and by the UNED University. However, the points of view and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be considered responsible for them. 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 (DE200100064 and CE200100005).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to check grammar and spelling.

References

[1] NewStatesman, Social media and the silencing effect: Why misogyny online is a human rights issue, https://bit.ly/3n3ox68, n.d. Last accessed 18 Oct 2023.

[2] J. L. Gil Bermejo, C. Martos Sánchez, O. Vázquez Aguado, E. B. García-Navarro, Adolescents, ambivalent sexism and social networks, a conditioning factor in the healthcare of women, Healthcare (Basel) 9 (2021) 721. doi:10.3390/healthcare9060721.

[3] G. Morales Rodríguez, J. Lopez-Figueroa, The portrayal of women in media, Journal of Student Research 13 (2024).

[4] S. E. Davis, Objectification, sexualization, and misrepresentation: Social media and the college experience, Social Media + Society 4 (2018).

[5] J. Harriger, J. Thompson, M. Tiggemann, TikTok, TikTok, the time is now: Future directions in social media and body image, ody Image B (2023) 222–226.

[6] 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.

[7] 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.

[8] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 497, CEUR Working Notes, 2023, pp. 813–854.

[9] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International

Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347.
[10] M. Billig, Humour and hatred: the racist jokes of the Ku Klux Klan, Discourse & Society 12 (2014) 267–289.
[11] A. Mendiburo-Seguel, T. E. Ford, The Effect of Disparagement Humor on the Acceptability of Prejudice., Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues (2019) No Pagination Specified–No Pagination Specified. doi:10.1007/s12144-019-00354-2.
[12] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter, in: Procesamiento del Lenguaje Natural (SEPLN), 71, 2023, pp. 383–395.
[13] B. Chulvi, L. Fontanella, R. Labadie, R. P, 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.
[14] G. Hodson, J. Rush, C. C. MacInnis, 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 (2010) 660–682. doi:10.1037/a0019627.
[15] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark Dataset of Memes with Text Transcriptions for Automatic Detection of Multi-modal Misogynistic Content, Data in Brief 44 (2022) 108526.
[16] 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.
[17] 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.
[18] G. B. Amor, N. Medimagh, S. Ben Chaabene, O. Trigui, ANLP-Uniso at Exist! 2025: Sexism Identification and Characterization in Tweets, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[19] C. Christodoulou, NLPDame at EXIST: Sexism Categorization in Tweets via Multi-Head Multi-Task Models, LLM & RAG Voting Synergy, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[20] E. Corbí Verdú, M. Franco Salvador, ECORBI-UPV at EXIST 2025: Large Language Models and Embedding Strategies for Sexism Detection in Tweets, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[21] Q. Chen, L. Kong, Y. Chen, Modeling Annotator Subjectivity for Sexism Detection on Social Media, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[22] D. Chen, H. Qi, Beyond Binary: 7-Class Sexism Identification via ModernBERT and SCL, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[23] A. Alajmi, G. Pergola, Leveraging Model Confidence and Diversity: A Multi-Stage Framework for Sexism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[24] P. Pastells, M. Vázquez, M. Farrús, M. Taulé, CLiC at EXIST 2025: Combining Fine-tuning and Prompting with Learning with Disagreement for Sexism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[25] M.-D. Cotelin, C.-O. Truică, E.-S. Apostol, NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 –

Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[26] S. Khan, A. Jhumka, G. Pergola, Multilingual Sexism Detection through Domain Adaptation and Label-Augmented Translation, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[27] A. Ibrahim, R. Lopes, Sexism Detection in Multilingual Tweets, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[28] A. Petrescu, E.-S. Apostol, C.-O. Truică, Awakened at EXIST2025: Adaptive Mixture of Transformers, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[29] M. Hurtado, A. Tarrasó, Dandys-de-BERTganim at EXSIST 2025: a Multi-task Learning Architecture for Sexism Identification, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[30] G. Arampatzis, A. Arampatzis, DuthThrace at EXIST 2025: Multilingual Sexism Detection with Soft Labels and Transformers, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[31] K. Villarreal Haro, G. Segura Gómez, J. Tavarez Rodríguez, F. Sánchez Vega, A. Pastor López Monroy, Leveraging Reasoning of Auto-Revealed Insights via Knowledge Injection and Evolutionary Prompting for Sexism Analysis, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[32] L. Dominguez-Sol Sastre, I. Segura Bedmar, L. Dominguez-Sol, L. Dominguez-Sol, Sexism Identification in Social Networks using LLMs, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[33] M. S. Khursheed, S. R. Hasan Abidi, S. Faisal Sikandar, S. Zahra, F. Alvi, A. Samad, Sexism Identification in Tweets Using Ensembles & Augmentation: A Multilingual Approach, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[34] F. Fritzsche, J. Felser, M. Spranger, COMFOR at EXIST 2025: Support Vector Machines vs. Large Language Models in Sexism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[35] K. Villarreal Haro, F. Sanchez-Vega, A. Pastor López Monroy, Knowledge Expansion Guided by Justification for Improved Sexism Categorization, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[36] L. Tian, J. R. Trippas, M.-A. Rizoiu, Mario at EXIST 2025: A Simple Gateway to Effective Multilingual Sexism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[37] R. Labadie-Tamayo, A. J. Böck, D. Slijepčević, X. Chen, A. Babic, M. Zeppelzauer, FHSTP@EXIST 2025 Benchmark: Sexism Detection with Transparent Speech Concept Bottleneck Models, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[38] J. C. Prajogo, L.-H. Lee, H.-I. Lin, NYCU-NLP at EXIST 2025: An Empirical Study of Annotator-Aware Two-Stage Pipeline for Sexism Detection in Tweets, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[39] P. Italiani, F. Maqbool, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, TrankilTwice at EXIST2025: Detecting Sexism in Memes under Multi-Lingual Settings, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[40] D. K. Jha, M. K. Mandal, A. K. M, Sexism Identification using Annotator Ranking in Memes: A

Multimodal Approach Using Transformers, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[41] I. Arcos, Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[42] A. Britez, I. Markov, CLTL at EXIST 2025: Identifying Sexist Memes Using an Ensemble of Shallow and Transformer Models, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[43] M. Guerrero-García, F. Carrillo García, J. Mata, V. Pachón-Álvarez, I2C-UHU-Altair at EXIST2025: Multimodal Sexism Detection and Classification Using Advanced Vision-Language Models BLIP2 and Qwen, Large Language Models, and Learning with Disagreement Frameworks, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[44] N. Nowakowski, L. Calogiuri, E. Egyed-Zsigmond, D. Nurbakova, J. Erbani, S. Calabretto, Automatic Sexism Detection on Social Networks: Classification of Tweets and Memes, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[45] D. Fernández, E. Amigó, V. Cardeñoso, ECA-SIMM-UVa at EXIST 2025: A Segmentation Oriented Approach to Sexism Detection in TikTok Videos Based on a "One Is Enough" Paradigm, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[46] M. Ali, L. Yendapalli, B. Tawfik, M. Winzenried, Tackling Sexism in Multimodal Social Media: Exploring Hybrid Generative-Transformer models, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[47] R. Pan, T. Bernal Beltrán, J. A. García Díaz, R. Valencia-Garcia, UMUTeam at EXIST 2025: Multimodal Transformer Architectures and Soft-Label Learning for Sexism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[48] T. Alcantara, O. Garcia-Vazquez, H. Calvo, J. E. Valdez-Rodríguez, CogniCIC at EXIST 2025: Identifying Sexist Content in Text and Visual Media using Transformers and Generative AI Models, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[49] T. Chopard, D. Rawlings, Generalisable BERT-Based Cross-Media Sexism Classification, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[50] B. Ruiz, Comparative Analysis of Transformer-Based Models for Sexism Detection in Text, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.