Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes

Laura Plaza[1], Jorge Carrillo-de-Albornoz[1], Enrique Amigó[1], Julio Gonzalo[1], Roser Morante[1], Paolo Rosso[2,3], Damiano Spina[4], Berta Chulvi[2], Alba Maeso[2], Víctor Ruiz[1]

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

Abstract. In recent years, the rapid increase in the dissemination of offensive and discriminatory material aimed at women through social media platforms has emerged as a significant concern. This trend has had adverse effects on women's well-being and their ability to freely express themselves. The EXIST campaign has been promoting research in online sexism detection and categorization in English and Spanish since 2021. The fourth edition of EXIST, hosted at the CLEF 2024 conference, consists of three groups of tasks, which are a continuation of EXIST 2023: *sexism identification*, *source intention identification*, and *sexism categorization*. However, while EXIST 2023 focused on processing tweets, the novelty of this edition is that the three tasks are also applied to memes, resulting in a total of six tasks. The "learning with disagreement" paradigm is adopted to address disagreements in the labelling process and promote the development of equitable systems that are able to learn from different perspectives on the sexism phenomena. The 2024 edition of EXIST has exceeded the success of previous editions, with the participation of 57 teams submitting 412 runs. This lab overview describes the tasks, dataset, evaluation methodology, participant approaches and results.

Keywords: sexism identification · sexism categorization · learning with disagreement · memes · data bias

1 Introduction

EXIST (sEXism Identification in Social neTworks) is a series of scientific events and shared tasks on sexism identification in social networks. The editions of

2021 and 2022 [37, 38], celebrated under the umbrella of the IBERLEF forum, were the first in proposing tasks focusing on identifying and classifying online sexism in a broad sense, from explicit and/or hostile to other subtle or even benevolent expressions. The 2023 edition [34] took place as a CLEF Lab and added a third task consisting in determining the intention of the author of sexist messages, with the goal of distinguishing sexist messages from those that report experiences with the aim of raising awareness against sexism. Additionaly, the main novelty of the 2023 edition was the adoption of the "Learning with Disagreement" (LwD) paradigm [46] for the development of the dataset and for the evaluation of the systems. In the LwD paradigm, models are trained to handle and learn from conflicting or diverse annotations so that different annotators' perspectives, biases, or interpretations are taken taken into account. This approach fits the findings of our previous work that showed that the perception of sexism is strongly dependent on the demographic and cultural background of the individual. Adopting this paradigm was a distinguishing feature in comparison to the SemEval-2023 Shared Task 10: "Explainable Detection of Online Sexism" [21].

EXIST 2024,[5] organised also as a CLEF Lab, aims to continue contributing datasets and tasks that help developing applications to combat sexism on-line, as a form of hate on-line. This edition embraces also the LwD paradigm and, as novelty, incorporates three new tasks that center around memes. Memes are images that are spread rapidly by social networks and Internet users. While by nature memes are humorous, there is a growing tendency to use them for harmful purposes, as an strategy to hide hate speech behind by combining stylistic devices of humour [4], since people tolerate humorously communicated prejudices better than explicit irrespectful remarks [12]. Thus, memes contribute to spreading derogatory humour and to strengthen preexisting prejudices and maintaining hierarchies between social groups [18]. As Gasparini et al. indicate [13], misogyny and sexism against women are widespread attitudes within the social media communities, reinforcing age-old patriarchal establishments of baseless name-calling, objectifying their appearances, and stereotyping gender roles. By including sexist memes in the EXIST 2024 dataset, we aim to encompass a broader spectrum of sexist manifestations in social networks and to contribute to the development of automated multimodal tools capable of detecting harmful content targeting women.

Meme detection has also been the focus of other competitions. The SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification [11] focused on the the detection of misogynous memes on the web in English and proposed two tasks: recognising whether a meme is misogynous or not and recognising types of misogyny in memes. The Overview of Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes [7] consisted in classifying misogynistic content and troll memes, focusing specifically on memes in Tamil and Malayalam languages. The originality of EXIST lies in that the lan-

[5] http://nlp.uned.es/exist2024/. Accessed 28 May 2024.

guages addressed are English and Spanish, it introduces also the task on source intention recognition and it adopts the LwD paradigm.

In the following sections, we provide comprehensive information about the tasks, the dataset, the evaluation methodology, the results and the different approaches of the systems that participated in the EXIST 2024 Lab. The competition features six distinct tasks: sexism identification, source intention classification, and sexism categorization both in tweets and in memes. A total of 148 teams from 32 different countries registered to participate. Ultimately, we received 412 results from 57 teams.

2 Tasks

The 2024 edition of EXIST feature 6 tasks, which are described below. The languages addressed are English and Spanish and the datasets are collections of tweets and memes. For the tasks on memes, all the partitions of the dataset are new, whereas for the tasks on tweets we employ the EXIST 2023 dataset.

2.1 Task 1: Sexism Identification in Tweets

This task is a binary classification where systems must decide whether or not a given tweet expresses ideas related to sexism in any of the three forms: it is sexist itself, it describes a sexist situation in which discrimination towards women occurs, or criticizes a sexist behaviour.

(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 2: Source Intention in Tweets

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

- **Direct** sexist message. The intention is to write a message that is sexist by itself or incites to be sexist, as in:
  (3) *A woman needs love, to fill the fridge, if a man can give this to her in return for her services (housework, cooking, etc), 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 3: Sexism Categorization in Tweets

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. 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 express false ideas about women that suggest they 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 4: Sexism Identification in Memes

As in Task 1, this is a binary classification consisting on deciding whether or not a meme is sexist. Figure 1 shows examples of sexist and non sexist memes.

2.5 Task 5: Source Intention in Memes

As in Task 2, this task aims to categorize the meme according to the intention of the author. Due to the characteristics of the memes, the REPORTED label is virtually null, so 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

Fig. 1: Examples of sexist and non-sexist memes

(a) Direct (b) Judgemental

Fig. 2: Examples of direct and judgemental memes

2.6 Task 6: Sexism Categorization in Memes

This task aims to classify sexist memes according to the categorization provided for Task 3: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence. Figure 3 shows one meme of each category.

3 Dataset

The EXIST 2024 dataset comprises two types of data: the tweets from the EXIST 2023 dataset and a completely new dataset of memes. Here, we briefly describe the process followed to curate the meme dataset. More details about the dataset, including bias considerations, the annotation process, quality experiments, and inter-annotator agreement, can be found in the extended overview [33]. In contrast, a detailed description of the tweets dataset can be found in [34].

Since we adopt the LwD paradigm, we provide all labels assigned by the different annotators to allow systems to learn from conflicting and subjective

(a) Ideological & inequality

(b) Objectification

(c) Stereotyping & dominance

(d) Sexual violence

(e) Misogyny & non-sexual violence

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

information. This paradigm not only proved to improve the systems' accuracy, robutness and generalizability, but also helped to mitigate bias.

3.1 Data Sampling

We first curated a lexicon of terms and expressions leading to sexist memes, derived from expressions proven representative in identifying sexism in previous EXIST editions. The set of seeds encompasses diverse topics, incorporating terms with varying degrees of use in both sexist and non-sexist contexts. The final set 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 removal of noise like textless images, text-only images, ads, and duplicates from the dataset. The final set of memes 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. Furthermore, the final data set is the result of obtaining the most equitable distribution of memes per seed. To avoid introducing selection bias, we randomly

selected memes, adhering to the appropriate distribution per seed. As a result, we have more than 2,000 memes per language for the training set and more than 500 memes per language for the test set.

3.2 Labeling the Date with Disagreement

As in the previous edition, we have considered some sources of "label bias". Label bias may be introduced by the socio-demographic differences of the persons that participate in the annotation process, but also when more than one possible correct label exists or when the decision on the label is highly subjective. In order to mitigate label bias, we consider two different social and demographic parameters: gender (MALE/FEMALE) and age (18-22 y.o./23-45 y.o./+46 y.o). Each meme was annotated by 6 annotators selected through the Prolific app (https://www.prolific.com/), following the guidelines developed by two experts in gender issues.

As new feature in the datasets, both 2023 and 2024, we will include three additional demographic characteristic of each anotator: level of education, ethnicity and country of residence.

4 Evaluation Methodology and Metrics

As in EXIST 2023, we have carried out a **"soft evaluation"** and a **"hard evaluation"**. The soft evaluation corresponds to the LwD 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 most standard evaluation paradigm and assumes that a single label is provided by the systems for every instance in the dataset.

1. **Soft-soft evaluation**. For systems that provide probabilities for each category, we provide 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 [1]), 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). Explicar como se normaliza
2. **Hard-hard evaluation**. For systems that provide a hard, conventional output, we provide 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 Task 1, the class annotated by more than 3 annotators is selected; for Task 2, the class annotated by more than 2 annotators is selected; and for Task 3 (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 Amigó and Delgado [1]. 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 system 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 [1] 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.

As there is not, to the best of our knowledge, any current metric that fits hierarchical multilabel classification problems in a learning with disagreement 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:

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

Note that the information content of assigning a category c with an agreement v grows inversely with the probability of finding an instance that receives category c with agreement equal or larger than v. To this end, 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.[6]

[6] 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.

Table 1: Runs submitted and teams participating on each EXIST 2024 task.

	Tweets			Memes		
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
# Runs	106	77	63	87	36	43
# Teams	46	38	27	41	18	22

The system output and the gold standards are sets of assignments. Therefore, in order to estimate their information content, we apply a recursive function similar to the one described by Amigó and Delgado [1].

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

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

5 Overview of Approaches

In this section, we provide a concise overview of the approaches presented at EXIST 2024. For a comprehensive description of the systems, please refer to the participant papers in the working notes [33].

Although 148 teams from 32 different countries registered for participation, the number of participants who finally submitted results were 57, submitting 412 runs. Teams were allowed to participate in any of the six tasks and submit hard and/or soft outputs. Table 1 summarizes the participation in the different tasks and evaluation contexts.

The evaluation campaign started on March 4, 2024 with the release of the training set. The test set was made available on April 15. The participant teams were provided with the official evaluation script. Runs had to be submitted by May 10. Each team could submit up to three runs per task, that may contain soft and/or hard outputs.

A wide range of approaches and strategies were used by the participants. For a comprehensive description of the systems submitted, please refer to extended overview [33].

Here we summarize the techniques and tools employed:

Nearly all participants the systems submitted utilized large language models, both monolingual and multilingual. Most employed LLMs include like BERT, DistilBERT, MarIA, MDEBERTA, RoBERTa, DeBERTa, LLAMA, and GPT-4.

For processing memes, more popular vision models employed were CLIP, BEIT and VIT .

Some teams employed ensembles of multiple models to enhance the overall performance. A couple of teams made use of knowledge integration to combine different language models with language features. Data augmentation techniques were used by several teams. Prompt Engineering was also used to adapt pre-trained models to the sexism detection task.

Only two teams utilized deep learning architectures such as BiLSTM and CNN, while other team opted for traditional machine learning methods, including SVM, Random Forest, XGBoost, among others.

As in EXIST 2023 [32], Twitter-specific models where employed, such as Twitter-RoBERTa and Twitter-XML-RoBERTa.

While 174 systems took advantage of the multiple annotations available and provided soft outputs, 238 employed the traditional approach of providing only hard labels as outputs.

Textual tasks have received greater engagement, although participation is also high in image tasks. There has been higher participation in binary classification tasks, followed by mono-label tasks, and finally, multi-label tasks, which is due to the increasing difficulty of these tasks.

For each of the six tasks, the organization also provided different baseline runs:

- **Majority class:** non-informative baseline that classifies all instances as the majority class.
- **Minority class:** non-informative baseline that classifies all instances as the minority class.

The evaluation metrics for the the gold standard are also provided, in order to set the upper bound for the ICM metric

6 Results

In the next subsections, we report the results of the participants and the baseline systems for each task. We only show the results obtained by the best run submitted by each participant to each task. For more detailed results, please refer to the Lab Working Notes [33].

6.1 Task 1: Sexism Identification in Tweets

We first report and analyze the results for Task 1, which focuses on sexism identification. This task involves a binary classification. As discussed in Section 4, we report two sets of evaluation results (hard and soft).

Soft Evaluation Table 2 presents the results for the soft-soft evaluation of Task 1. A total of 37 runs were submitted. Out of these, 34 runs outperformed the non-informative majority class baseline (where all instances are labeled as "NO"), and all runs surpassed the non-informative minority class baseline (where all instances are labeled as "YES"). We observed a significant discrepancy in performance, with ICM-Soft-norm scores ranging from 0.6755 to 0.0374. However, if we analyze the top 5 systems, we appreciate a difference of less than 5 percentual points. Notably, the best run achieved an ICM-Soft Norm score of 68% for this binary classification task, surpassing the top performance of 64% recorded by the best EXIST 2023 participant. This suggests that new models and approaches are becoming more effective at detecting sexism in social networks. However, it also indicates that there is still room for improvement in capturing an appropriate distribution that represents real data accurately.

Table 2: Systems' results for Task 1 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	3.1182	1.0000	0.5472
NYCU-NLP_1	1	1.0944	0.6755	0.9088
NYCU-NLP_2	2	1.0866	0.6742	0.8826
NYCU-NLP_3	3	1.0810	0.6733	0.9831
ABCD Team_3	4	0.9291	0.6490	1.2637
CIMAT-CS-NLP_3	5	0.9285	0.6489	1.2252
CIMAT-CS-NLP_1	6	0.8468	0.6358	1.2538
ABCD Team_1	7	0.8316	0.6333	1.6727
CIMAT-CS-NLP_2	8	0.8213	0.6317	1.2684
BAZI_1	9	0.8179	0.6311	0.9750
Awakened_2	10	0.7196	0.6154	0.8106
Victor-UNED_1	11	0.6952	0.6115	1.0691
Awakened_3	12	0.6909	0.6108	0.8542
I2C-UHU_2	13	0.6871	0.6102	0.9184
Victor-UNED_2	14	0.6797	0.6090	0.9818
UMUTEAM_1	15	0.6679	0.6071	0.8708
Awakened_1	16	0.6663	0.6068	0.8037
Victor-UNED_3	17	0.6479	0.6039	1.0930
I2C-UHU_1	18	0.5175	0.5830	1.0666
UMUTEAM_2	19	0.5033	0.5807	0.8357
ABCD Team_2	20	0.4594	0.5737	1.2164
MMICI_3	21	0.4589	0.5736	2.0316
clac_1	22	0.1431	0.5230	2.9543
RMIT-IR_1	23	−0.0011	0.4998	2.7892
FraunhoferSIT_1	24	−0.0658	0.4895	0.8801
CNLP-NITS-PP_1	25	−0.2086	0.4666	1.0390

Continued on next page

Table 2 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
RMIT-IR_3	26	−0.3016	0.4516	2.8235
Atresa-I2C-UHU_1	27	−0.3256	0.4478	3.9518
CNLP-NITS-PP_2	28	−0.3286	0.4473	0.9236
MMICI_1	29	−0.3394	0.4456	0.8375
MMICI_2	30	−0.3622	0.4419	0.8382
RMIT-IR_2	31	−0.3941	0.4368	2.9956
UMUITEAM_3	32	−0.4170	0.4331	1.0933
UniLeon-UniBO_1	33	−1.1882	0.3095	1.2449
UniLeon-UniBO_2	34	−1.3306	0.2866	2.0069
UniLeon-UniBO_3	35	−1.3447	0.2844	2.0239
EXIST2024 majority	36	−2.3585	0.1218	4.6115
NICA_3	37	−2.8848	0.0374	1.5286
NICA_2	38	−2.8848	0.0374	1.3862
NICA_1	39	−2.8848	0.0374	1.2301
EXIST2024 minority	40	−3.0717	0.0075	5.3572

Hard Evaluation Table 3 presents the results for the Hard-hard evaluation. In this scenario, the annotations from the six annotators are combined into a single label using the majority vote, resulting in the loss of information about the different perspectives provided by each annotator. Out of the 67 systems submitted for this task, 66 ranked above the majority class baseline (all instances labeled as "NO"). All systems surpassed the minority class baseline (all instances labeled as "YES"). Similar to the Soft-soft evaluation, the results vary considerably. If we focus on the ICM-hard normalized metric, we observe that the best run gets 0.8002 while the worse one gests only 0.2665. If we focus on the top 5 systems ,we observe that they achieve comparable results.

If we compare ICM-Hard norm results with F1, we see that there exists a correlation, specially for the best performance systems.

Table 3: Systems' results for Task 1 in the Hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
EXIST2024 gold	0	0.9948	1.0000	1.0000
NYCU-NLP_1	1	0.5973	0.8002	0.7944
ABCD Team_1	2	0.5957	0.7994	0.7826
CIMAT-CS-NLP_2	3	0.5926	0.7978	0.7899
EquityExplorers_2	4	0.5883	0.7957	0.7775

Continued on next page

Table 3 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
CIMAT-GTO_3	5	0.5848	0.7939	0.7903
CIMAT-GTO_2	6	0.5798	0.7914	0.7887
ABCD Team_3	7	0.5766	0.7898	0.7823
NYCU-NLP_3	8	0.5749	0.7889	0.7813
NYCU-NLP_2	9	0.5619	0.7824	0.7785
I2C-UHU_2	10	0.5557	0.7793	0.7733
BAZI_1	11	0.5490	0.7759	0.7755
CIMAT-CS-NLP_1	12	0.5486	0.7757	0.7746
EquityExplorers_1	13	0.5448	0.7738	0.7615
ADITYA_3	14	0.5418	0.7723	0.7691
CIMAT-GTO_1	15	0.5407	0.7718	0.7694
CIMAT-CS-NLP_3	16	0.5357	0.7692	0.7700
MMICI_3	17	0.5324	0.7676	0.7637
ADITYA_2	18	0.5246	0.7636	0.7669
NICA_1	19	0.5214	0.7621	0.7642
Awakened_3	20	0.5196	0.7611	0.7652
Awakened_2	21	0.5124	0.7575	0.7620
maven_3	22	0.5015	0.7521	0.7596
Awakened_1	23	0.4984	0.7505	0.7582
Victor-UNED_3	24	0.4934	0.7480	0.7602
Victor-UNED_1	25	0.4914	0.7470	0.7542
Victor-UNED_2	26	0.4863	0.7444	0.7535
RMIT-IR_3	27	0.4802	0.7414	0.7548
MMICI_2	28	0.4780	0.7402	0.7460
penta-nlp_1	29	0.4779	0.7402	0.7508
RMIT-IR_1	30	0.4739	0.7382	0.7526
MMICI_1	31	0.4705	0.7365	0.7455
I2C-UHU_1	32	0.4651	0.7338	0.7513
RMIT-IR_2	33	0.4590	0.7307	0.7448
ADITYA_1	34	0.4580	0.7302	0.7447
fmrs_2	35	0.4398	0.7211	0.7462
clac_1	36	0.4380	0.7201	0.7376
NICA_3	37	0.4358	0.7191	0.7429
fmrs_1	38	0.3961	0.6991	0.7194
TextMiner_2	39	0.3926	0.6973	0.7223
ABCD Team_2	40	0.3884	0.6952	0.7292
TextMiner_3	41	0.3876	0.6948	0.7180
maven_1	42	0.3860	0.6940	0.7121
NICA_2	43	0.3750	0.6885	0.7263
BAZI_2	44	0.3472	0.6745	0.7121
CAU&ITU_2	45	0.3460	0.6739	0.7024

Continued on next page

Table 3 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
DLRG_1	46	0.3446	0.6732	0.7085
TextMiner_1	47	0.3412	0.6715	0.7048
shm2024_3	48	0.3230	0.6623	0.7044
maven_2	49	0.3044	0.6530	0.6946
shm2024_1	50	0.2905	0.6460	0.6946
CAU&ITU_1	51	0.2832	0.6423	0.6922
Atresa-I2C-UHU_1	52	0.2782	0.6398	0.6899
FraunhoferSIT_1	53	0.2320	0.6166	0.6823
CNLP-NITS-PP_1	54	0.1977	0.5994	0.6762
CNLP-NITS-PP_2	55	0.1440	0.5724	0.6281
mc-mistral_2	56	0.0614	0.5309	0.5317
mc-mistral_1	57	−0.0094	0.4953	0.4779
UniLeon-UniBO_2	58	−0.1870	0.4060	0.4963
UniLeon-UniBO_3	59	−0.1959	0.4015	0.4906
NIT-Patna-NLP_1	60	−0.2975	0.3505	0.5272
UniLeon-UniBO_1	61	−0.2980	0.3502	0.5972
shm2024_2	62	−0.3410	0.3286	0.4922
DadJokers_1	63	−0.3611	0.3185	0.4365
VerbaNex_1	64	−0.4048	0.2965	0.4588
The 3 Musketeers_1	65	−0.4229	0.2875	0.3371
The 3 Musketeers_2	66	−0.4260	0.2859	0.3719
VerbaNex_2	67	−0.4392	0.2792	0.4560
EXIST2024 majority	68	−0.4413	0.2782	0.0000
The 3Musketeers_3	69	−0.4645	0.2665	0.2999
EXIST2024 minority	70	−0.5742	0.2114	0.5698

6.2 Task 2: Source Intention in Tweets

Soft Evaluation Table 4 presents the results for the Soft-soft evaluation of Task 2. The table shows that 32 runs were submitted. Among them, 25 runs achieved better results compared to the majority class baseline (where all instances are labeled as "NO"). Furthermore, all of the submitted runs outperformed or equaled the minority class baseline (where all instances are labeled as "REPORTED").

The ICM-Soft Norm scores vary considerably, from the 0.4795 of the best system ("nycu-nlp_2") system to 0.0000 of "fmrs_2" and "EXIST2024 minority", indicating significant variability in the effectiveness of the submitted models. Overall, performance is considerably lower compared to Task 1. This can be attributed to the hierarchical and multiclass nature of Task 2. It is also worth noting the correlation between the ICM-Soft and Cross-Entropy measures. The results indicate a strong correlation between the two metrics, but some differ-

ences can still be observed due to the fact that cross entropy does not have into account the specificity of the different classes.

Table 4: Systems' results for Task 2 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	6.2057	1.0000	0.9128
NYCU-NLP_2	1	−0.2543	0.4795	1.8344
NYCU-NLP_1	2	−0.4059	0.4673	1.8549
NYCU-NLP_3	3	−0.5226	0.4579	1.9206
BAZI_1	4	−1.3468	0.3915	1.7812
Victor-UNED_2	5	−1.6440	0.3675	1.7971
Victor-UNED_1	6	−1.6549	0.3667	1.8132
ABCD Team_3	7	−1.8462	0.3513	2.4123
UMUTEAM_1	8	−1.9566	0.3424	1.4726
Awakened_2	9	−2.0091	0.3381	3.0835
ABCD Team_2	10	−2.0149	0.3377	2.3892
Awakened_1	11	−2.0365	0.3359	3.1429
UMUTEAM_2	12	−2.0533	0.3346	1.7890
Awakened_3	13	−2.1502	0.3268	3.0908
fmrs_1	14	−2.1737	0.3249	2.1210
CNLP-NITS-PP_1	15	−2.4732	0.3007	1.6696
Atresa-I2C-UHU_1	16	−2.6802	0.2841	2.1629
I2C-UHU_2	17	−2.6952	0.2828	2.1440
ABCD Team_1	18	−2.9080	0.2657	2.7595
UMUTEAM_3	19	−3.3189	0.2326	4.4490
MMICI_3	20	−3.6350	0.2071	1.7285
FraunhoferSIT_1	21	−4.0856	0.1708	1.7649
I2C-UHU_3	22	−4.2278	0.1594	2.5245
RMIT-IR_1	23	−4.5481	0.1336	3.5776
MMICI_1	24	−4.5753	0.1314	1.6866
MMICI_2	25	−4.6285	0.1271	1.6974
CUET-SSTM_1	26	−5.1320	0.0865	4.8736
EXIST2024 majority	27	−5.4460	0.0612	4.6233
NICA_2	28	−5.7592	0.0360	2.7026
RMIT-IR_3	29	−5.7632	0.0357	3.9903
UniLeon-UniBO_3	30	−5.7633	0.0356	2.1267
UniLeon-UniBO_1	31	−5.9587	0.0199	2.2261
UniLeon-UniBO_2	32	−5.9798	0.0182	2.1542
RMIT-IR_2	33	−6.1535	0.0042	4.0930
fmrs_2	34	−6.9170	0.0000	4.1975
EXIST2024 minority	35	−32.9552	0.0000	8.8517

Hard Evaluation Table 5 presents the Hard-hard evaluation results for Task 2, assessing 43 systems against the hard gold standard. Among these, 37 runs outperform the majority class baseline (where all instances are labeled "NO"), and all systems show equal or better performance compared to the minority class baseline (where all instances are labeled "REPORTED"). Similar to the Soft-soft evaluation, discrepancies between the gold standard and the worst-performing system (minority class baseline) are more pronounced in Task 2 than in Task 1. The correlation between ICM-Hard and F1 is generally strong, with slight variations among the top-ranked systems and greater variability towards the lower end of the table. This variability arises because F1 does not account for the hierarchical nature of the task as effectively as ICM-Hard, which more stringently penalizes misclassifications between different hierarchy levels.

The top-ranking system, "ABCD Team_1," achieved the highest ICM-Hard normalized score (0.6320). The top 5 best systems ranges between 0.5937 and 0.6320. The lower end of the table includes five systems which score 0 in the ICM-Hard norm metric. Generally, a high correlation is observed between all metrics, particularly for the top-performing systems. However, variability increases towards the lower end of the rankings, where discrepancies between ICM-Hard norm and cross entropy become more pronounced.

Table 5: Systems' results for Task 2 in the Hard-hard evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm	Macro F1
EXIST2024 gold	0	1.5378	1.0000	1.0000
ABCD Team_1	1	0.4059	0.6320	0.5677
NYCU-NLP_3	2	0.3522	0.6145	0.5410
NYCU-NLP_1	3	0.3383	0.6100	0.5353
NYCU-NLP_2	4	0.3073	0.5999	0.5273
CUET-SSTM_1	5	0.2883	0.5937	0.5383
ABCD Team_3	6	0.2847	0.5926	0.5289
CIMAT-CS-NLP_2	7	0.2643	0.5859	0.5171
CIMAT-CS-NLP_1	8	0.2346	0.5763	0.5195
penta-nlp_1	9	0.2089	0.5679	0.4856
BAZI_1	10	0.1883	0.5612	0.4843
I2C-UHU_2	11	0.1815	0.5590	0.4980
Awakened_2	12	0.1812	0.5589	0.4826
CIMAT-CS-NLP_3	13	0.1615	0.5525	0.4885
fmrs_1	14	0.1609	0.5523	0.4978
NICA_2	15	0.1506	0.5490	0.4738
Awakened_1	16	0.1487	0.5483	0.4753
Awakened_3	17	0.1306	0.5425	0.4686
RMIT-IR_1	18	0.0855	0.5278	0.4024
Victor-UNED_1	19	0.0851	0.5277	0.3257

Continued on next page

Table 5 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
Victor-UNED_2	20	0.0815	0.5265	0.3256
I2C-UHU_1	21	0.0418	0.5136	0.4708
BAZI_2	22	0.0396	0.5129	0.4278
RMIT-IR_3	23	0.0394	0.5128	0.3856
I2C-UHU_3	24	0.0210	0.5068	0.4663
RMIT-IR_2	25	0.0173	0.5056	0.3926
maven_1	26	−0.0510	0.4834	0.4563
MMICI_1	27	−0.0987	0.4679	0.4548
MMICI_3	28	−0.1076	0.4650	0.4525
DLRG_1	29	−0.1171	0.4619	0.3931
ABCD Team_2	30	−0.1368	0.4555	0.4182
Atresa-I2C-UHU_1	31	−0.1524	0.4504	0.4278
MMICI_2	32	−0.2406	0.4218	0.4383
CNLP-NITS-PP_1	33	−0.2694	0.4124	0.3743
FraunhoferSIT_1	34	−0.4106	0.3665	0.3823
CAU&ITU_1	35	−0.4711	0.3468	0.2998
CAU&ITU_2	36	−0.5024	0.3366	0.3029
shm2024_1	37	−0.8873	0.2115	0.3148
fmrs_2	38	−0.9078	0.2048	0.1899
EXIST2024 majority	39	−0.9504	0.1910	0.1603
NICA_1	40	−0.9504	0.1910	0.1603
UniLeon-UniBO_3	41	−1.2145	0.1051	0.2605
NIT-Patna-NLP_1	42	−1.9410	0.0000	0.1207
shm2024_2	43	−2.0626	0.0000	0.1200
UniLeon-UniBO_1	44	−2.0862	0.0000	0.1736
UniLeon-UniBO_2	45	−2.2986	0.0000	0.1628
EXIST2024 minority	46	−3.1545	0.0000	0.0280

6.3 Task 3: Sexism Categorization in Tweets

The third task is a hierarchical multiclass and multilabel classification problem, where systems must determine if a tweet is sexist or not, and categorize the sexist tweets according to the five categories of sexism defined in Section 2.

Soft Evaluation Table 6 displays the results of the Soft-soft evaluation for Task 3. A total of 30 runs were submitted, with 26 runs surpassing the majority class baseline (all instances labeled as "NO"), and all systems outperforming the minority class baseline (all instances labeled as "SEXUAL-VIOLENCE").

The "NYCU-NLP" **[DS: Remember to replace "..." with "..."]** team has the top three runs, with "NYCU-NLP_1" ranked first (ICM-Soft: −1.1762,

ICM-Soft Norm: 0.4379). The next two runs from the same team, "NYCU-NLP_2" and "NYCU-NLP_3," follow closely, indicating the consistency and robustness of their approach. The fourth and fifth systems, however, show a significantly poorer performance (0.3835 and 0.3732, respectively) The range of ICM-Soft Norm scores (from 0.4379 to 0.0000) underscores a significant variability in system performance. However, despite the complexity of the task, it seems that systems are still able to correctly capture relevant information concerning the different types of sexism.

Table 6: Systems' results for Task 3 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2024 gold	0	9.4686	1.0000
NYCU-NLP_1 [10]	1	−1.1762	0.4379
NYCU-NLP_2	2	−1.2169	0.4357
NYCU-NLP_3	3	−1.4555	0.4231
Medusa_1 [35]	4	−2.2055	0.3835
Medusa_2	5	−2.4010	0.3732
Medusa_3	6	−2.4142	0.3725
ABCD Team_3	7	−3.5160	0.3143
ABCD Team_2	8	−3.5438	0.3129
Awakened_2	9	−4.0748	0.2848
Awakened_3	10	−4.0786	0.2846
Awakened_1	11	−4.1845	0.2790
NICA_2	12	−4.4324	0.2659
ABCD Team_1	13	−4.5913	0.2576
FraunhoferSIT_1	14	−5.1905	0.2259
Victor-UNED_1	15	−5.5936	0.2046
Victor-UNED_2	16	−5.6190	0.2033
CNLP-NITS-PP_1	17	−5.7385	0.1970
CNLP-NITS-PP_2	18	−6.0810	0.1789
RMIT-IR_1 [45]	19	−7.2098	0.1193
MMICI_3	20	−7.6413	0.0965
RMIT-IR_2	21	−7.8944	0.0831
MMICI_1	22	−7.9356	0.0809
MMICI_2	23	−7.9380	0.0808
fmrs_1	24	−8.2508	0.0643
fmrs_2	25	−8.4277	0.0550
fmrs_3	26	−8.4277	0.0550
RMIT-IR_3	27	−8.5680	0.0476
EXIST2024 majority	28	−8.7089	0.0401
UniLeon-UniBO_1	29	−10.3622	0.0000
UniLeon-UniBO_2	30	−10.3622	0.0000

Continued on next page

Table 6 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm
UniLeon-UniBO_3	31	−10.3622	0.0000
Atresa-I2C-UHU_1	32	−10.4052	0.0000
EXIST2024 minority	33	−46.1080	0.0000

[DS: PARTICIPANTS:]Azadi et al. [2], Barua et al. [3], Box and Bertomeu [5], Álvaro Carrillo-Casado et al. [6], Emmolo et al. [8], Fan et al. [9], Fang et al. [10], Gersome et al. [14], Gopalakrishnan et al. [15], Guerrero-García et al. [16], Haro et al. [17], Keinan [19], Khan et al. [20], Ma and Li [22], Maqbool and Fersini [23], Maqbool [24], Martinez et al. [25, 26], Naebzadeh et al. [27], Obrador Reina and García Cucó [28], Pan et al. [29], Pasha [30], Petrescu et al. [31], Quan and Thin [35], Rizzi et al. [36], Rodríguez et al. [39], Ruiz et al. [40], Shah and Gokhale [41], Shanbhag et al. [42], Shifat et al. [43], Siino and Tinnirello [44], Smith et al. [45], Usmani et al. [47], Vetagiri et al. [48]

Hard Evaluation In the Hard-hard evaluation context for the third task, 31 systems were submitted. As shown in Table 7, 28 systems outperformed the majority class baseline (all instances labeled as "NO"), while all systems achieved better results than the minority class baseline (all instances labeled as "SEXUAL-VIOLENCE"). The discrepancy between the best (ABCD Team 1, 0.5862 ICM-Hard norm score) and the worst-performing system (CAU&ITU 1, 0.000 score) is over 0.5 ICM-hard-norm which, surprisingly is less that in the task 2 hard to hard evaluation.

The variation in results among different runs follows a similar distribution to that observed in Task 2, except for the last four systems which obtained substantially lower results due to that a high number of not sexist instances have been incorrectly assigned to different sexist subclasses, resulting in a strong penalization by the ICM-Hard metric that considers the class hierarchy.

Finally, comparing the behaviour of the different tasks in a hard-hard context, the efficiency of the systems in this task, in terms of ICM-Hard Norm, is lower than in previous tasks, further highlighting the complexity of categorizing sexism.

Table 7: Systems' results for Task 3 in the Hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	2.1533	1.0000	1.0000
ABCD Team_1	1	0.3713	0.5862	0.6004

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
ABCD Team_3	2	0.3540	0.5822	0.6042
NYCU-NLP_3	3	0.3069	0.5713	0.6130
NYCU-NLP_1	4	0.2364	0.5549	0.6066
NYCU-NLP_2	5	0.1725	0.5401	0.5933
Awakened_2	6	−0.0042	0.4990	0.4833
Awakened_3	7	−0.0115	0.4973	0.4803
RMIT-IR_3	8	−0.0344	0.4920	0.5049
RMIT-IR_2	9	−0.0394	0.4909	0.4986
RMIT-IR_1	10	−0.0396	0.4908	0.5024
Awakened_1	11	−0.0427	0.4901	0.4743
ABCD Team_2	12	−0.1090	0.4747	0.5286
NICA_2	13	−0.2383	0.4447	0.4564
penta-nlp_1	14	−0.2597	0.4397	0.4379
maven_1	15	−0.2654	0.4384	0.4491
UniLeon-UniBO_1	16	−0.3188	0.4260	0.5032
UniLeon-UniBO_2	17	−0.3188	0.4260	0.5032
UniLeon-UniBO_3	18	−0.3188	0.4260	0.5032
NICA_1	19	−0.3258	0.4243	0.3867
UMUTEAM_1	20	−0.7339	0.3296	0.4942
FraunhoferSIT_1	21	−0.7437	0.3273	0.3724
UMUTEAM_3	22	−0.7901	0.3165	0.4821
MMICI_3	23	−0.8105	0.3118	0.4805
UMUTEAM_2	24	−0.8719	0.2975	0.4738
CNLP-NITS-PP_1	25	−0.9571	0.2778	0.2684
CNLP-NITS-PP_2	26	−0.9684	0.2751	0.2318
MMICI_1	27	−1.4509	0.1631	0.4026
MMICI_2	28	−1.5003	0.1516	0.4017
fmrs_3	29	−1.5952	0.1296	0.1087
EXIST2024 majority	30	−1.5984	0.1289	0.1069
fmrs_2	31	−1.6017	0.1281	0.1069
fmrs_1	32	−1.7482	0.0941	0.1700
CAU&ITU_1	33	−2.3423	0.0000	0.1705
EXIST2024 minority	34	−3.1295	0.0000	0.0288

6.4 Task 4: Sexism Identification in Memes

Soft Evaluation Table 8 presents the results for the task of classifying memes as sexist or not sexist. The performance results are notably low for a binary classification task: Victor-UNED_1, the top-ranked participant, achieved an ICM-Soft Norm score of 0.4530 and a relatively low Cross Entropy of 1.1028.

When comparing these results to those of Task 1 (classifying tweets as sexist or not), we observe a significant drop in performance for image classification (0.4530 versus 0.6755 ICM-Soft Norm). It is important to highlight that most approaches relied solely on the text within the meme for classification, without incorporating image processing. This suggests that sexism in memes is often conveyed through the imagery, even when the accompanying text appears neutral.

Table 8: Systems' results for Task 4 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	3.1107	1.0000	0.5852
Victor-UNED_1	1	−0.2925	0.4530	1.1028
Victor-UNED_2	2	−0.3135	0.4496	1.2834
Elias&Sergio_1	3	−0.3225	0.4482	0.9903
I2C-Huelva_3	4	−0.3263	0.4476	1.5189
I2C-Huelva_1	5	−0.3390	0.4455	1.4096
I2C-Huelva_2	6	−0.3446	0.4446	1.4112
Victor-UNED_3	7	−0.3761	0.4395	1.1562
RMIT-IR_2	8	−0.3780	0.4392	0.9852
NICA_1	9	−0.4360	0.4299	0.9278
PINK_2	10	−0.4396	0.4293	0.9375
PINK_1	11	−0.4537	0.4271	0.9282
ROCurve_3	12	−0.4646	0.4253	0.9609
the gym nerds_2	13	−0.5015	0.4194	0.9201
Elias&Sergio_2	14	−0.5617	0.4097	0.9228
ROCurve_2	15	−0.6097	0.4020	0.9537
MMICI_2	16	−0.6183	0.4006	0.9143
MMICI_1	17	−0.6189	0.4005	0.9151
PINK_3	18	−0.6378	0.3975	0.9318
MMICI_3	19	−0.6410	0.3970	0.9534
ROCurve_1	20	−0.6420	0.3968	0.9431
OppositionalOppotision_1	21	−0.9556	0.3464	3.2025
melialo-vcassan_1	22	−1.0022	0.3389	0.9931
melialo-vcassan_2	23	−1.0239	0.3354	0.9904
RMIT-IR_3	24	−1.0894	0.3249	1.1206
melialo-vcassan_3	25	−1.0957	0.3239	1.0090
the gym nerds_1	26	−1.1035	0.3226	0.9733
CNLP-NITS-PP_2	27	−1.2354	0.3014	1.0918
CHEEXIST_2	28	−1.2710	0.2957	1.1993
RMIT-IR_1	29	−1.2819	0.2940	1.0128
Penta-ML_2	30	−1.2910	0.2925	2.2277
epistemologos_1	31	−1.3486	0.2832	2.9425
Penta-ML_1	32	−1.5664	0.2482	2.4735

Continued on next page

Table 8 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Penta-ML_3	33	−1.7425	0.2199	4.0007
CHEEXIST_3	34	−2.0119	0.1766	0.5017
CHEEXIST_1	35	−2.0388	0.1723	0.5030
EXIST2024 majority	36	−2.3568	0.1212	4.4015
CNLP-NITS-PP_1	37	−2.6987	0.0662	1.3445
EXIST2024 minority	38	−3.5089	0.0000	5.5672

Hard Evaluation Table 9 presents the results for the Hard-hard evaluation. In this scenario, the annotations from the six annotators are combined into a single label using the majority vote, resulting in the loss of information about the different perspectives provided by each annotator. Out of the 50 systems submitted for this task, only 37 ranked above the majority class baseline (all instances labeled as "NO"), while 47 systems surpassed the minority class baseline (all instances labeled as "YES"). Similar to the Soft-soft evaluation, the results vary considerably, from 0.6618 ICM-Hard Norm for the best performing system (RoJiNG-CL_3) to 0.0876 (melialo-vcassan_1).

When comparing ICM-Hard Norm results with F1 scores, we observe little correlation between the two metrics, especially in the lower ranks of the table.

Table 9: Systems' results for Task 4 in the Hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
EXIST2024 gold	0	0.9832	1.0000	1.0000
RoJiNG-CL_3	1	0.3182	0.6618	0.7642
RoJiNG-CL_2	2	0.2272	0.6155	0.7437
RoJiNG-CL_1	3	0.1863	0.5947	0.7274
I2C-Huelva_2	4	0.1313	0.5668	0.7241
I2C-Huelva_1	5	0.1166	0.5593	0.7154
DiTana-PV_2	6	0.1150	0.5585	0.7122
Victor-UNED_2	7	0.1028	0.5523	0.7154
MMICI_2	8	0.1014	0.5515	0.7261
I2C-Huelva_3	9	0.0987	0.5502	0.6933
DiTana-PV_3	10	0.0888	0.5451	0.7082
NICA_1	11	0.0767	0.5390	0.7248
MMICI_1	12	0.0751	0.5382	0.7202
Victor-UNED_1	13	0.0641	0.5326	0.7051
OppositionalOppotision_1	14	0.0494	0.5251	0.7168

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
Elias&Sergio_1	15	0.0433	0.5220	0.6979
Elias&Sergio_2	16	0.0408	0.5208	0.6962
Victor-UNED_3	17	0.0364	0.5185	0.6991
DiTana-PV_1	18	0.0337	0.5171	0.6908
ROCurve_3	19	0.0088	0.5045	0.6834
PINK_1	20	0.0076	0.5039	0.7044
PINK_3	21	−0.0053	0.4973	0.7006
RMIT-IR_2	22	−0.0123	0.4938	0.6726
PINK_2	23	−0.0346	0.4824	0.7102
MMICI_3	24	−0.0361	0.4816	0.6781
ROCurve_2	25	−0.0956	0.4514	0.6654
Miqarn_1	26	−0.1159	0.4411	0.6632
CNLP-NITS-PP_1	27	−0.1234	0.4372	0.6699
Penta-ML_2	28	−0.1308	0.4335	0.6742
Penta-ML_1	29	−0.1745	0.4113	0.6524
epistemologos_1	30	−0.1823	0.4073	0.5503
TokoAI_1	31	−0.1872	0.4048	0.5639
Penta-ML_3	32	−0.2049	0.3958	0.6101
UMUTEAM_1	33	−0.2422	0.3768	0.6963
RMIT-IR_3	34	−0.2601	0.3677	0.6040
ROCurve_1	35	−0.2640	0.3657	0.6318
Umera Wajeed Pasha_1	36	−0.3083	0.3432	0.5956
TargaMarhuenda_1	37	−0.3535	0.3202	0.6487
TargaMarhuenda_2	38	−0.3844	0.3045	0.5568
EXIST2024 majority	39	−0.4038	0.2947	0.6821
CNLP-NITS-PP_2	40	−0.4045	0.2943	0.4937
DLRG_1	41	−0.4206	0.2861	0.6469
MIND_1	42	−0.4986	0.2465	0.5674
ALC-UPV-JD-2_1	43	−0.5446	0.2231	0.4878
dap-upv_1	44	−0.5737	0.2082	0.4188
AI Fusion_1	45	−0.6416	0.1737	0.4651
EXIST2024 minority	46	−0.6468	0.1711	0.0000
RMIT-IR_1	47	−0.6468	0.1711	0.0000
AI Fusion_2	48	−0.6486	0.1702	0.4656
AI Fusion_3	49	−0.6508	0.1691	0.4079
TheATeam_1	50	−0.6644	0.1621	0.4821
melialo-vcassan_2	51	−0.6644	0.1621	0.0281
melialo-vcassan_3	52	−0.6723	0.1581	0.0347
melialo-vcassan_1	53	−0.8109	0.0876	0.5316

6.5 Task 5: Source Intention in Memes

Table 10 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 15 runs were submitted for this task. While all the runs ranked above the minority class baseline (all instances labeled as "JUDGEMENTAL"), only 15 runs surpassed the majority class baseline (all instances labeled as "NO"). The results for this task are notably low, with the best team (Victor-UNED_2) achieving only 0.3676 ICM-Soft Norm. This suggests that identifying whether a meme contains direct sexism or is judgmental is more difficult than identifying the intention behind a sexist tweet.

Furthermore, we do not find a correlation between cross entropy and ICM-Soft Norm, which could be attributed to ???.

Soft Evaluation

Table 10: Systems' results for Task 5 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2024 gold	0	4.7018	1.0000	0.9325
Victor-UNED_2	1	−1.2453	0.3676	1.6235
MMICI_1	2	−1.2660	0.3654	1.4645
MMICI_2	3	−1.3738	0.3539	1.4405
NICA_1	4	−1.5329	0.3370	1.4664
CNLP-NITS-PP_1	5	−1.5907	0.3308	1.5273
melialo-vcassan_2	6	−1.9847	0.2889	1.5211
Victor-UNED_1	7	−2.0053	0.2867	2.0028
melialo-vcassan_3	8	−2.0653	0.2804	1.5295
melialo-vcassan_1	9	−2.6821	0.2148	1.6291
I2C-Huelva_3	10	−2.7996	0.2023	3.9604
I2C-Huelva_2	11	−2.7997	0.2023	3.9857
I2C-Huelva_1	12	−2.8007	0.2022	3.9735
MMICI_3	13	−3.4751	0.1304	3.4504
EXIST2024 majority	14	−5.0745	0.0000	5.5565
Penta-ML_3	15	−5.2668	0.0000	5.1547
Penta-ML_1	16	−5.3096	0.0000	3.2977
Penta-ML_2	17	−5.9832	0.0000	5.4845
EXIST2024 minority	18	−18.9382	0.0000	8.0245

Hard Evaluation Table 11 presents the results for the Hard-hard evaluation. Out of the 19 systems submitted for this task, only 15 ranked above the majority

class baseline (all instances labeled as "NO"), while 18 systems surpassed the minority class baseline (all instances labeled as "JUDGEMENTAL"). The results range from 0.4167 ICM-Hard Norm for the best performing system (Victor-UNED_1) to 0.0000 for the worst performing systems..

When comparing ICM-Hard Norm results with F1 scores, we again observe little correlation between the two metrics, especially in the lower ranks of the table.

Table 11: Systems' results for Task 5 in the Hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	1.4383	1.0000	1.0000
Victor-UNED_1	1	−0.2397	0.4167	0.3873
I2C-Huelva_2	2	−0.2535	0.4119	0.4761
Victor-UNED_2	3	−0.2668	0.4073	0.3850
I2C-Huelva_3	4	−0.2772	0.4036	0.4714
I2C-Huelva_1	5	−0.2880	0.3999	0.4714
NICA_1	6	−0.2881	0.3999	0.3837
MMICI_1	7	−0.3066	0.3934	0.4179
MMICI_3	8	−0.3297	0.3854	0.3814
CNLP-NITS-PP_1	9	−0.3370	0.3829	0.4101
MMICI_2	10	−0.3868	0.3655	0.3770
Penta-ML_3	11	−0.6123	0.2872	0.3841
Penta-ML_1	12	−0.6546	0.2725	0.3856
Penta-ML_2	13	−0.7089	0.2536	0.3841
TokoAI_1	14	−0.7263	0.2475	0.3716
melialo-vcassan_3	15	−0.7758	0.2303	0.3709
melialo-vcassan_2	16	−0.8585	0.2016	0.3500
EXIST2024 majority	17	−1.0445	0.1369	0.1839
UMUTEAM_1	18	−1.1486	0.1007	0.2098
melialo-vcassan_1	19	−1.1971	0.0838	0.2970
DLRG_1	20	−1.4891	0.0000	0.2530
EXIST2024 minority	21	−2.0637	0.0000	0.0697
epistemologos_1	22	−8.7012	0.0000	0.0557

6.6 Task 6: Sexism Categorization in Memes

Soft Evaluation Table 12 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as class probabilities.

Only 19 runs were submitted for this task. While all runs performed better than the minority class baseline (labeling all instances as "MISOGYNY-NON-

SEXUAL-VIOLENCE"), only 11 runs exceeded the majority class baseline (labeling all instances as "NO"). The performance for this task was generally low, with the top team (ROCurve_1) achieving an ICM-Soft Norm score of only 0.2462, which is significantly lower compared to the results for the same task when applied to tweets (task 3).

Table 12: Systems' results for Task 6 in the Soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2024 gold	0	9.4343	1.0000
ROCurve_1	1	−4.7893	0.2462
the gym nerds_2	2	−4.7942	0.2459
ROCurve_2	3	−5.0030	0.2348
ROCurve_3	4	−5.0675	0.2314
Elias&Sergio_1	5	−5.9160	0.1865
Victor-UNED_1	6	−6.4124	0.1602
Victor-UNED_2	7	−6.4777	0.1567
CNLP-NITS-PP_1	8	−6.6782	0.1461
CNLP-NITS-PP_2	9	−7.2381	0.1164
AI Fusion_1	10	−7.6282	0.0957
AI Fusion_2	11	−7.6363	0.0953
AI Fusion_3	12	−7.7043	0.0917
EXIST2024 majority	13	−9.8173	0.0000
dap-upv_1	14	−10.4213	0.0000
Penta-ML_2	15	−11.2593	0.0000
the gym nerds_1	16	−11.2648	0.0000
Penta-ML_1	17	−11.8047	0.0000
Penta-ML_3	18	−13.2556	0.0000
MMICI_1	19	−16.1248	0.0000
MMICI_2	20	−19.3246	0.0000
MMICI_3	21	−45.0237	0.0000
EXIST2024 minority	22	−50.0353	0.0000

Hard Evaluation Finally, Table 13 presents the results for classifying memes based on the aspects of women being attacked, with outputs provided as a single class prediction.

22 runs were submitted for this task. Only 17 runs exceeded the majority class baseline (labeling all instances as "NO"), while 21 runs ranked above the minority class (all instances labeled as "MISOGYNY-NON-SEXUAL-VIOLENCE") The performance for this task was low, with the top team (DiTana-PV_1) achieving an ICM-Soft Norm score of 0.3549. No correlation is found between ICM-Hard Norm and Macro-F1.

Table 13: Systems' results for Task 6 in the Hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2024 gold	0	2.4100	1.0000	1.0000
DiTana-PV_1	1	−0.6996	0.3549	0.4319
DiTana-PV_2	2	−0.8450	0.3247	0.4430
MMICI_1	3	−0.9863	0.2954	0.4342
ROCurve_1	4	−1.0089	0.2907	0.3639
ROCurve_2	5	−1.1075	0.2702	0.3275
ROCurve_3	6	−1.1440	0.2627	0.3085
MMICI_2	7	−1.3446	0.2210	0.4453
Penta-ML_3	8	−1.3631	0.2172	0.3356
DiTana-PV_3	9	−1.3691	0.2160	0.3255
Penta-ML_2	10	−1.4684	0.1954	0.3093
Elias&Sergio_1	11	−1.5276	0.1831	0.4321
Penta-ML_1	12	−1.5499	0.1784	0.3053
Miqarn_1	13	−1.6216	0.1636	0.3211
CNLP-NITS-PP_1	14	−1.7920	0.1282	0.1587
ALC-UPV-JD-2_1	15	−1.8573	0.1147	0.2103
CNLP-NITS-PP_2	16	−1.8813	0.1097	0.1511
dap-upv_1	17	−1.9497	0.0955	0.2227
UMUTEAM_1	18	−1.9511	0.0952	0.3786
EXIST2024 majority	19	−2.0711	0.0703	0.0919
TargaMarhuenda_1	20	−2.0725	0.0700	0.1440
TargaMarhuenda_2	21	−2.2075	0.0420	0.1140
TheATeam_1	22	−2.3159	0.0195	0.1490
EXIST2024 minority	23	−3.3135	0.0000	0.0318
MMICI_3	24	−3.8341	0.0000	0.2347
One-by-zero_1	25	−4.5910	0.0000	0.2304

7 Conclusions

The objective of the EXIST challenge is to encourage research on the automated detection and modeling of sexism in online environments, with a specific focus on social networks. The EXIST 2024 Lab held as part of CLEF attracted nearly 150 registered teams and received a total of 412 submissions. Participants adopted a wide range of approaches, including vision transformer models, data augmentation through automatic translation, data duplication, utilization of past EXIST editions' data, multilingual language models, Twitter-specific language models, and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis.

While many systems opted for the traditional approach of providing only hard labels as outputs, a significant number of systems leveraged the multiple

annotations available and provided soft outputs, proving that there is an increasing interest by the research community in developing systems able to deal with disagreement and with different perspectives.

For future editions of EXIST, we plan to expand our study to include additional communication channels and media formats, such as TikTok videos. By doing so, we aim to address the nuances and unique challenges presented by different formats, enhancing the robustness and applicability of research on automated sexism detection. Additionally, this expansion will allow us to capture a broader spectrum of online interactions and cultural contexts.

Acknowledgments. This work has been financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. It has also been financed by the Spanish Ministry of Science and Innovation (project FairTransNLP (PID2021-124361OB-C31 and PID2021-124361OB-C32)) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe, and by the Australian Research Council (DE200100064 and CE200100005).

References

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

[2] Azadi, A., Ansari, B., Zamani, S.: Bilingual Sexism Classification: Fine-Tuned XLM-RoBERTa and GPT-3.5 Few-Shot Learning. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[3] Barua, D.D., Sourove, M.S.U.R., Haider, F., Shifat, F.T., Ishmam, M.F., Fahim, M., Bhuiyan, F.A.: Penta ML at EXIST 2024: Tagging Sexism in Online Multimodal Content With Attention-enhanced Modal Context. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[4] Billig, M.: Humour and hatred: the racist jokes of the ku klux klan. Discourse & Society 12(3), 267–289 (2014)

[5] Box, A.M., Bertomeu, D.T.: DiTana-PV at sEXism Identification in Social neTworks (EXIST) tasks 4 and 6: The Effect of Translation in Sexism Identification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[6] Álvaro Carrillo-Casado, Román-Pásaro, J., Mata-Vázquez, J., Pachón-Álvarez, V.: I2C-UHU at EXIST 2024: Transformer-Based Detection of Sexism and Source Intention in Memes Using a Learning with Disagreement Approach. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[7] Chakravarthi, B.R., Rajiakodi, S., Ponnusamy, R., Pannerselvam, K., Madasamy, A.K., Rajalakshmi, R., LekshmiAmmal, H., Kizhakkeparambil, A., S Kumar, S., Sivagnanam, B., Rajkumar, C.: Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In: Chakravarthi,

B.R., B, B., Buitelaar, P., Durairaj, T., Kovács, G., García Cumbreras, M.Á. (eds.) Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion. pp. 139–144. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.ltedi-1.13

[8] Emmolo, N., Aru, G., Piras, A., Marzeddu, S., Raffi, J., Passaro, L.C.: RoBEXedda: Enhancing Sexism Detection in Tweets for the EXIST 2024 Challenge. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[9] Fan, S., Frick, R.A., Steinebach, M.: FraunhoferSIT@EXIST2024: Leveraging Stacking Ensemble Learning for Sexism Detection. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[10] Fang, Y.Z., Lee, L.H., Huang, J.D.: NYCU-NLP at EXIST 2024 – Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[11] Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., Lees, A., Sorensen, J.: SemEval-2022 task 5: Multimedia automatic misogyny identification. In: Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., Ratan, S. (eds.) Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). pp. 533–549. Association for Computational Linguistics, Seattle, United States (Jul 2022), https://aclanthology.org/2022.semeval-1.74

[12] andThomas E Ford, A.M.S.: The effect of disparagement humor on the acceptability of prejudice. Current Psychology 42, 16222–16233 (2019)

[13] Gasparini, F., Rizzi, G., Saibene, A., Fersini, E.: Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. Data in Brief 44, 108526 (2022)

[14] Gersome, S., Mahibha, J., Thenmozhi, D.: Automatic Classification of Gender Stereotypes in Social Media Post. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[15] Gopalakrishnan, S., K, S., Sreekumar, M., Thenmozhi, D., Swaminathan, K.: Sexism Identification in Tweets using Traditional Machine Learning Approaches. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[16] Guerrero-García, M., Cerrejón-Naranjo, M., Mata-Vázquez, J., Pachón-Álvarez, V.: I2C-UHU at EXIST2024: Learning from Divergence and Perspectivism for Sexism Identification and Source Intent Classification. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[17] Haro, K.V., Sánchez-Vega, F., Rosales-Pérez, A., López-Monroy, A.P.: Stacked Reflective Reasoning in Large Neural Language Models. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[18] Hodson, G., Rush, J., Macinnis, C.: The effect of disparagement humor on the acceptability of prejudice. Journal of Personality and Social Psychology 99, 660–682 (2010)

[19] Keinan, R.: Sexism Identification in Social Networks using TF-IDF Embeddings, PreProccessing, Feature Selection, Word/Char N-Grams and Various Machine Learning Models In Spanish and English. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[20] Khan, S., Pergola, G., Jhumka, A.: Multilingual Sexism Identification via Fusion of Large Language Models. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[21] Kirk, H.R., Yin, W., Vidgen, B., Röttger, P.: SemEval-2023 Task 10: Explainable Detection of Online Sexism. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval) (2023)
[22] Ma, J., Li, R.: RoJiNG-CL at EXIST 2024: Sexism Identification in Memes by Integrating Prompting and Fine-Tuning. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[23] Maqbool, F., Fersini, E.: A Contrastive Learning based Approach to detect Sexism in Memes. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[24] Maqbool, N.: Sexism Identification in Social Networks: Advances in Automated Detection - A Report on the Exist Task at CLEF. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[25] Martinez, E., Cuadrado, J., Santos, J.C.M., Puertas, E.: VerbaNex AI at CLEF EXIST 2024: Detection of Online Sexism using Transformer Models and Profiling Techniques. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[26] Martinez, M.P.J., Romero, J.M.R., Torres, C.E.S., Nava, I.H.L., Gómez, M.M.Y.: Enhancing Sexism Detection in Tweets with Annotator-Integrated Ensemble Methods and Multimodal Embeddings for Memes. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[27] Naebzadeh, A., Nobakhtian, M., Eetemadi, S.: NICA at EXIST CLEF Tasks 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[28] Obrador Reina, M., García Cucó, A.: LightGMB for sexism identification in memes. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[29] Pan, R., García Díaz, J.A., Bernal Beltrán, T., Valencia-Garcia, R.: UMUTeam at EXIST 2024: Multi-modal Identification and Categorization of Sexism by Feature Integration. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[30] Pasha, U.W.: Multilingual Sexism Detection in Memes, A CLIP-Enhanced Machine Learning Approach. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[31] Petrescu, A., Truică, C.O., Apostol, E.S.: Language-based Mixture of Transformers for EXIST2024. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[32] Plaza, L., Carrillo-de Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview). In: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum (2023)
[33] Plaza, L., Carrillo-de Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., Spina, D.a.: Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview). In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[34] Plaza, L., de Albornoz, J.C., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of exist 2023 – learning with disagreement for sexism identification and characterization (extended overview). In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). vol. 497, pp. 813–854. CEUR Working Notes (2023)
[35] Quan, L.M., Thin, D.V.: Sexism Identification in Social Networks with Generation-based Approach. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[36] Rizzi, G., Gimeno-Gómez, D., Fersini, E., Martínez-Hinarejos, C.D.: PINK at EXIST2024: A Cross-Lingual and Multi-Modal Transformer Approach for Sexism Detection in Memes. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[37] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: Sexism identification in social networks. Procesamiento del Lenguaje Natural 67, 195–207 (2021)
[38] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Spina, D., Gonzalo, J., Rosso, P.: Overview of EXIST 2022: Sexism identification in social networks. Procesamiento del Lenguaje Natural 69, 229–240 (2022)
[39] Rodríguez, J.T., Vega, F.S., Pérez, A.R., Monroy, A.P.L.: Better Together: LLM and Neural Classification Transformers to Detect Sexism. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[40] Ruiz, V., Carrillo-de Albornoz, J., Plaza, L.: Concatenated Transformer Models Based On Levels Of Agreements For Sexism Detection. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[41] Shah, A., Gokhale, A.: Team Aditya at EXIST 2024 – Detecting Sexism In Multilingual Tweets Using Contrastive Learning Approach. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[42] Shanbhag, A., Jadhav, S., Date, A., Joshi, S., Sonawane, S.: The Wisdom of Weighing: Stacking Ensembles for a More Balanced Sexism Detector. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)

[43] Shifat, F.T., Haider, F., Sourove, M.S.U.R., Barua, D.D., Ishmam, M.F., Fahim, M., Bhuiyan, F.A.: Penta-nlp at EXIST 2024 Task 1-3: Sexism Identification, Source Intention, Sexism Categorization In Tweets. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[44] Siino, M., Tinnirello, I.: Prompt Engineering for Identifying Sexism using GPT Mistral 7B. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[45] Smith, T., Nie, R., Trippas, J., Spina, D.: RMIT-IR at EXIST Lab at CLEF 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[46] Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., Poesio, M.: SemEval-2021 task 12: Learning with disagreements. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 338–347. Association for Computational Linguistics, Online (Aug 2021)
[47] Usmani, M., Siddiqui, R., Rizwan, S., Khan, F., Alvi, F., Samad, A.: Sexism Identification in Tweets using BERT and XLM - Roberta. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)
[48] Vetagiri, A., Mogha, P., Pakray, P.: Cracking Down on Digital Misogyny with MULTILATE a MULTImodaL hATE Detection System. In: Faggioli, G., Ferro, N., Galuščáková, P., Seco de Herrera, A.G. (eds.) Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum (2024)