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

Notebook for the EXIST Lab at CLEF 2024

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

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

Abstract

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, continuing from 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. To address disagreements in the labeling process, the "learning with disagreement" paradigm is adopted. This approach promotes the development of equitable systems capable of learning from different perspectives on the sexism phenomenon. The 2024 edition of EXIST has surpassed the success of previous editions, with the participation of 57 teams submitting 412 runs. This extended lab overview describes the tasks, dataset, evaluation methodology, participant approaches, and results. Additionally, it highlights the advancements made in understanding and tackling online sexism through more diverse data sources and innovative methodologies. Finally, it briefly introduces future intended work for next editions of EXIST.

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 [1, 2], 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 [3] took place as a CLEF Lab and added a third task consisting in determining the intention of the author of sexist messages with the aim of understanding the purpose behind people posting sexist messages on social networks. Additionally, the main novelty of the 2023 edition was the adoption of the "Learning with Disagreements" (LwD) paradigm [4] 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 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

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

lplaza@lsi.uned.es (L. Plaza); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); victor.ruiz@lsi.uned.es (V. Ruiz); amaeolm@inf.upv.es (A. Maeso); berta.chulvi@upv.es (B. Chulvi); prosso@dsic.upv.es (P. Rosso); enrique@lsi.uned.es (E. Amigó); julio@lsi.uned.es (J. Gonzalo); r.morant@lsi.uned.es (R. Morante); damiano.spina@rmit.edu.au (D. Spina)

individual. Adopting this paradigm was a distinguishing feature in comparison to the SemEval-2023 Shared Task 10: "Explainable Detection of Online Sexism" [5].

EXIST 2024, 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 conceal hate speech by combining stylistic devices of humour [6], since people tolerate humorously communicated prejudices better than explicit irrespectful remarks [7]. Thus, memes contribute to spreading derogatory humour and to strengthen preexisting prejudices and maintaining hierarchies between social groups [8]. As Gasparini et al. indicate [9], 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" [10] focused on 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 shared task on "Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes" [11] 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 languages 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 teams 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. Interestingly, a significant number of teams leveraged the diverse labels representing various demographic groups and provided soft labels as the outputs of their systems. Their results showcase the effectiveness and advantages of employing the LwD paradigm in our specific domain: sexism detection and categorization in social networks.

2. Tasks

The 2024 edition of EXIST feature six 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 is a binary classification task 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. The following statements show examples of sexist and not sexist messages, respectively.

- **Sexist**. The tweet is sexist or describes or criticizes a sexist situation.

  (1) *Woman driving, be careful!.*

  (2) *It's less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely.*

(3) *I'm sorry but women cannot drive, call me sexist or whatever but it is true.*

(4) *You look like a whore in those pants" - My brother of 13 when he saw me in a leather pant*

- **Not sexist**. The tweet is not sexist, nor it describes or criticizes a sexist situation.

(5) *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.*

(6) *Where all the white women at?.*

(7) *The shocking video of a woman at the wheel who miraculously escapes an assassination attempt.*

(8) *Don't my arguments convince you? Let's try to debate. Do you use "feminazi"? You stay alone.*

2.2. Task 2: Source Intention in Tweets

This task aims to categorize the message according to the intention of the author. This distinction will allow us to differentiate sexism that is actually taking place online from sexism which is being suffered by women in other situations but that is being reported in social networks with the aim of complaining and fighting against sexism. 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:

(9) *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.*

(10) *Women shouldn't code...perhaps be influencer/creator instead...it's their natural strength.*

- **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:

(11) *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.*

(12) *Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.*

- **Judgemental** message. The intention is to condemn sexist situations or behaviours, as in:

(13) *As usual, the woman was the one quitting her job for the family's welfare...*

(14) *21st century and we are still earning 25% less than men #Idonotrenounce.*

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. Some examples of this type of sexism discourse can be found in the following tweets:

(15) *#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.*

(16) *Think the whole equality thing is getting out of hand. We are different, that's how were made!*

(17) *OK. Whew, that's good. Get a real degree and forget this poison of victimhood known as feminism.*

- **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. Some examples of this type of sexism discourse can be found in the following tweets:

  (18) *Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.*

  (19) *I feel like everytime I flirt with a girl they start to imagine all the ways they can utilize me.*

  (20) *Woman driving, be careful!*

- **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. Some examples of this type of sexism discourse can be found in the following tweets:

  (21) *No offense but I've never seen an attractive african american hooker. Not a single one.*

  (22) *Most of the times I just want women for sex just because everytime I try to make a relationship out of it they always break my heart.*

  (23) *Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring.*

- **Sexual violence**. It includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made. Some examples of this type of sexism discourse can be found in the following tweets:

  (24) *I wanna touch your tits..you can't imagine what I can do on your body.*

  (25) *Fuck that cunt, I would with my fist.*

  (26) *Now I understand why woman that get no action are miserable. Semen retention almost drove me insane.*

- **Misogyny and non sexual violence**. It includes expressions of hatred and violence towards women. Some examples of this type of sexism discourse can be found in the following tweets:

  (27) *Domestic abuse is never okay... Unless your wife is a bitch.*

  (28) *Some woman are so toxic they don't even know they are draining everyone around them in poison.*

  (29) *I hate misogyny more than I hate women.*

2.4. Task 4: Sexism Identification in Memes

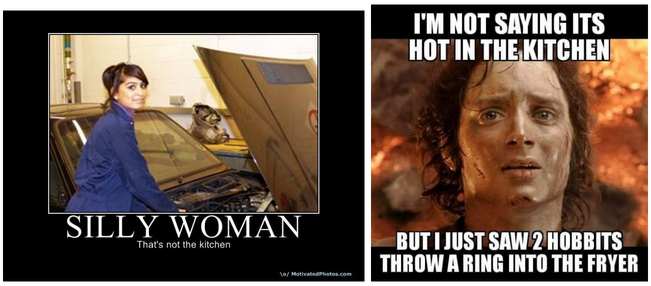
As in Task 1, this is a binary classification consisting on deciding whether or not a meme is sexist. Figure 1 shows examples of sexist and non sexist memes, respectively.

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, we barely found examples of memes within the "reported" category, so this category was not considered. As a result, in this task systems should only classify memes in two classes: "direct" or "judgemental", as shown in Figure 2.

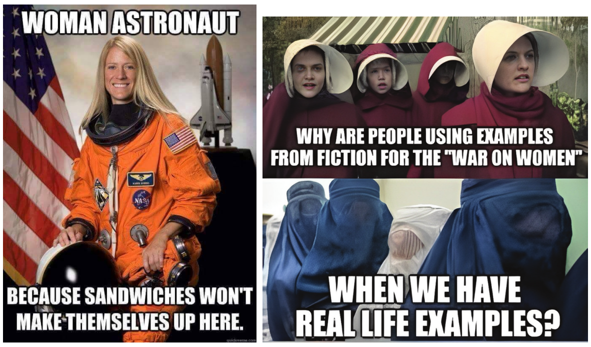
2.6. Task 6: Sexism Categorization in Memes

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.


SILLY WOMAN
That's not the kitchen
\o/ MotivatedPhotos.com
I'M NOT SAYING ITS HOT IN THE KITCHEN
BUT I JUST SAW 2 HOBBITS THROW A RING INTO THE FRYER

(a) Sexist meme (b) Non sexist meme

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


WOMAN ASTRONAUT
BECAUSE SANDWICHES WON'T MAKE THEMSELVES UP HERE.
WHY ARE PEOPLE USING EXAMPLES FROM FICTION FOR THE "WAR ON WOMEN"
WHEN WE HAVE REAL LIFE EXAMPLES?

(a) Direct (b) Judgemental

Figure 2: Examples of direct and judgemental memes.

3. Dataset

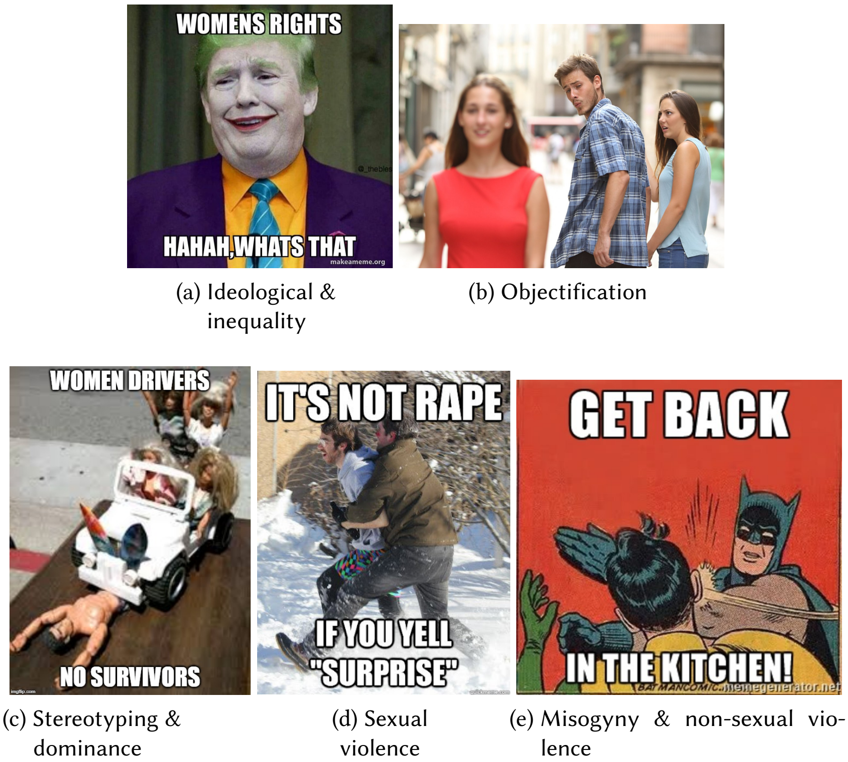
The EXIST 2024 dataset comprises two types of data: the tweets from the EXIST 2023 dataset and a completely new dataset of memes. Plaza et al. [3] provide a detailed description of the EXIST 2023 tweet dataset. Here, we briefly describe the process followed to curate the meme dataset.

Since we adopt the LwD paradigm, the dataset is released with with the labels assigned by the different annotators. This allows systems to learn from conflicting and subjective information. This paradigm not only proved to improve the systems' accuracy, robustness and generalizability, but it also helped to mitigate bias.

3.1. Data Sampling

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.



(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.

3.2. Dealing with Label Bias

We have considered some sources of "label bias" that 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 particular, we consider two sociodemographic parameters: gender (MALE/FEMALE) and age (18–22/23–45/+46 y.o.). Each meme was annotated by six annotators selected through the Prolific crowdsourcing platform (https://www.prolific.com/). No personally identifiable information about the crowd workers was collected. Crowd workers were informed that the tweets could contain offensive information and were allowed to withdraw voluntarily at any time. Full consent was obtained.

Also, as a new feature of the 2023 and 2024 datasets, we have reported three additional demographic characteristics of annotators: level of education, ethnicity, and country of residence.

3.3. Learning with Disagreement

The assumption that natural language expressions have a single and clearly identifiable interpretation in a given context is a convenient idealization, but far from reality, especially in highly subjective task as sexism identification. The learning with disagreements paradigm aims to deal with this by letting systems learn from datasets where no gold annotations are provided but information about the annotations from all annotators, in an attempt to gather the diversity of views. In accordance with methods proposed for training directly from the data with disagreements instead of using an aggregated label, the dataset is provided with the annotations of the six different strata of annotators per instance.

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

4. Lab Setup and Participation

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 Section 6 and to the participants' papers.

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 participants 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.

A wide range of approaches and strategies were used by the participants. Nearly all participant systems utilized large language models, both monolingual and multilingual. Most employed LLMs include BERT, DistilBERT, MarIA, MDEBERTA, RoBERTa, DeBERTa, Llama, and GPT-4. For processing memes, popular vision models were employed: 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 four teams opted for traditional machine learning methods, including SVM, Random Forest, and XGBoost, among others. As in EXIST 2023 [12], 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 followed the traditional approach of providing only hard labels as outputs. Textual tasks received greater engagement, although participation is also high in the tasks on memes. The binary classification tasks had more participants, 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 two different baseline runs:

- **EXIST2024 majority**, a non-informative baseline that classifies all instances as the majority class.
- **EXIST2024 minority**, a non-informative baseline that classifies all instances as the minority class.
- The evaluation results of the gold standard (**EXIST2024 gold**) are also provided in order to set the upper bound for the ICM metrics.

5. Evaluation Methodology and Metrics

As in EXIST 2023, we have carried out a **"soft evaluation"** and a **"hard evaluation"**. The soft evaluation relates 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 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 LwD paradigm can be considered in both sides of the evaluation process:

- **The ground truth.** In a “hard” 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” 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 2024, 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 [13]), 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 [13]. 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:

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

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 [13] 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:

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

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 LwD 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 [13] for assignment sets and avoid the redundant information of parent categories.

$$\begin{aligned} IC\left(\bigcup_{i=1}^{n}\{\langle c_i, v_i\rangle\}\right) &= IC(\langle c_1, v_1\rangle) + IC\left(\bigcup_{i=2}^{n}\{\langle c_i, v_i\rangle\}\right) \\ &- IC\left(\bigcup_{i=2}^{n}\{\langle \mathrm{lca}(c_1, c_i), min(v_1, v_i)\rangle\}\right) \end{aligned} \tag{30}$$

where lca(a, b) is the lowest common ancestor of categories a and b.

6. Overview of approaches EXIST 2024

In this section, we provide a description of the approaches adopted by the participants. More detailed information of each work is provided in the participants' working notes.

Team **FraunhoferSIT** [14] utilized a stacking ensemble of machine learning models. For predicting hard labels, they used an ensemble of classification models: Multinomial Naive Bayes, Stochastic Gradient Descent, Decision Tree, k-Nearest Neighbors, Logistic Regression and Extra Trees. For predicting soft labels, they used an ensemble of regression models: Random Forest, Gradient Boosting, Stochastic Gradient Descent and AdaBoost. They employed two augmentation methods at word level: synonym replacement using WordNet for English tweets and contextual augmentation with three transformer models: BERTIN, ALBERT Base Spanish, and RoBERTuito. They participated in the first three tasks.

Team **frms** [15] participated in Tasks 1, 2 and 3, for both hard and soft evaluations. For all three tasks, their first run used BERT multilingual, and their second run used XLM-RoBERTa model. In the third task, they created an ensemble of BERT and XLM-RoBERTa combining the predictions from both models. For Task 1, RoBERTa model performed the best in the hard evaluation, but, for Task 2, BERT performed better for both hard and soft evaluations. For Task 3, the ensemble obtained the best results in the hard evaluation and tied with RoBERTa in the soft evaluation.

Team **RMIT-IR** [16] proposed different approaches to Tasks 1-3 and Task 4. For Tasks 1–3 (on tweets), they studied the effectiveness of zero-shot In-Context Learning (ICL) with off-the-shelf pre-trained Large Language Models (LLMs). Their approaches for meme classification (Task 4) utilize CLIP (Contrastive Language-Image Pre-training) to experiment with multi-modal embeddings and zero-shot sexism identification models. They participated on hard and soft evaluations, obtaining soft labels by generating six answers and calculating the proportions. The annotator's genders and/or study levels were included in some runs for the first three tasks. For Task 4, they used three systems TI-CLIP (feedforward network), TIMV-CLIP (Transformer encoder), and Prompt-CLIP (zero-shot). TIMV-CLIP (Transformer encoder) stands out by its performance, especially in the memes in English dataset, where it achieves the best result in soft evaluation with a ICM-Soft Norm score of 0.4998.

Team **dap-upv** [17] proposed hard labels in Tasks 4 and 6. For Task 4, they utilized a two-stage approach by fine-tuning the Contrastive Language-Image Pre-training (CLIP) model followed by a classifier. A different classifier was tested per run, where the best one was Light Gradient Boosting Machine (LightGBM) reaching a 0.72 in F1. For Task 6, they fine-tuned both a RoBERTa model for text, a Google Vision Transformer (ViT) for images, and they concatenated them training a LightGBM classifier on the concatenated embeddings. The ensemble notably improved results, obtaining 0.49 in F1.

Team **Aditya** [18] only participated in Task 1 submitting hard labels. They preprocessed tweets by removing any emoji, URLs or mentions. They used XLM-RoBERTa fine-tuned on the dataset, and their three runs differ on the number of epochs of training and whether they used development in training too. Their best model, trained with 12 epochs, ranked 14th with F1 of 0.7691.

Team **BAZI** [19] fine-tuned various transformer models, some monolingual in English and Spanish and some multilingual, and they chose XLM-RoBERTa as the best performing of them. With this method, they provided hard labels and soft labels. Hard labels were the direct outputs from the model, and soft labels were obtained by adding a softmax function to the last layer. Their approach using XLM-RoBERTa achieved 4th place in the soft evaluation for Task 1, and 2nd place for Task 2. Also, they employed few-shot learning with GPT-3.5 with three examples in English and three in Spanish from the training set, providing hard labels with this method.

Team **mc-mistral_2** [20] submitted hard labels for Task 1. Their approach leveraged a Mistral 7B model along with a few-shot learning strategy and prompt engineering to address the task in the hard labelling setup. They translated cases in Spanish to English with an online and real-time use of Google Translator from the *deep_translator* library. Then they randomly selected 10 samples from the provided labelled training set and formatted the samples of test set as: Tweet1 // NO, Tweet2 // YES. In the global ranking, they achieved a F1 score of 0.51.

Team **CNLP-NITS-PP** [21] presented two systems to cover all tasks in EXIST 2024. The model for textual modalities is a Convolutional Neural Network - Bidirectional Long Short-Term Memory (CNN-BiLSTM) and it is used in Tasks 1, 2 and 3, and to characterize textual data in Tasks 4, 5 and 6. Texts are tokenized into words utilizing GloVe, they are lowercased, and common stopwords are deleted. For memes, a combination of Residual Network 50 (ResNet-50), used to analyze images, and text-based analysis is utilized. Images are resized to 224x224 pixels, and pixel values are normalized to the interval [0, 1]. To enhance model robustness, data augmentation techniques such as random rotation, flipping, and colour jitter are applied. Hyperparameter tuning is conducted via grid search and k-fold cross-validation. A single run was submitted for every task in both hard and soft labels. Results stand out in Task 5, where they reached the 5th position in the ranking.

Team **Awakened** [22] studied the most appropriate way of assembling transformer models by comparing performances of different models and assigning different weights in the ensemble to the best model. They used both English models and multilingual models, and both general language models (like DistilBERT of xlm-RoBERTa), and domain-specific models (like twitter-xlm-roberta-base-sentiment or roberta-hate-speech-dynabench-r4). They tested three ways of assembling models: assigning half of the weight to the most dominant model, assigning 75% of the weight, and assigning all the weight to the most dominant model. Results showed that, in most of the tasks, the best model was the one in which the most dominant was assigned a 75% of the weight.

The **NYCU-NLP** team [23] employed extensive data preprocessing techniques, which included removing irrelevant elements, standardizing text formats, back-translation using the Google Translator API, and implementing the AEDA method for text augmentation. Additionally, they adapted the Round to Closest Value approach to handle non-continuous annotation values. The system relies on two transformer-based language models: DeBERTa-v3 and xlm-RoBERTa. The team integrated annotator information such as gender, age, and ethnicity, creating a unified vector representation for each tweet. They further incorporated Hard Parameter Sharing to optimize shared layers across tasks, improving generalization and computational efficiency. Notably, their model achieved outstanding performance in the EXIST 2024 challenge, securing first place in Tasks 1, 2, and 3 in the soft evaluation setting. In the hard setting, their system ranked first in Task 1, second in Task 2, and third in Task 3.

Team **shm2024** [24] participated with hard labels to Tasks 1 and 2. They implemented three different models. The first system was a MultiLayer Perceptron (MLP) classifier with Language Agnostic BERT Sentence Embeddings (LaBSE). The second one was a eXtreme Gradient Boosting (XGBoost) Classifier, and the third approach was an ensemble of CNN models. The first model was the best performing on the test set with a ICM-Hard Norm value of 0.6623 and F1_YES value of 0.7044 for Task 1 and, for Task 2, 0.2115 for ICM-Hard Norm and 0.1200 for Macro F1.

UMUTeam [25] submitted soft predictions to Tasks 1 and 2, and hard labels to the rest of the tasks. For textual modalities, they created an ensemble of two Spanish LLMs, BETO and MarIA; and two multilingual LLMs, deBERTa v3 and XLMTwitter. They also extracted the linguistic features (LFs) using the UMUTextStats tool and performed hyperparameter tuning on 10 models. They tried two ways of ensembling: Knowledge Integration (KI), and Ensemble Learning (EL). Their three runs were KI, EL and LFs. Their results were better at Spanish than English. Their best results were obtained in Task 2, where KI reached the 8th position. For multi modalities, they used a CLIP model to extract images and text. Their algorithm was formed by an Image Encoder (CLIP image encoder), a Text Encoder (CLIP text encoder), Diagonal multiplication and a Classification head. Their results were near the baselines.

The **3 Musketeers** [26] team applied traditional classification machine learning algorithms such as Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM), and BERT transformer model to the EXIST Task 1. They followed a pipeline of preprocessing, including lowercasing, removing punctuation, emoticons, links, mentions and stopwords. They lemmatized words and used TF-IDF vectorization. They used GridSearch to optimize hyperparameters. Their best model was an SVM that got a F1 score of 0.6299.

Team **TextMiner** [27] submitted hard labels to Task 1. They preprocessed the text to standardize it and text embeddings were created using TF-IDF. They performed feature selection of word n-grams and character n-grams. They explored a diverse range of classifiers: Random Forest Classifier, Extra Trees Classifier, LightGBM Classifier, AdaBoost Classifier, Bernoulli Naive Bayes, Support Vector Classifier. They experimented with different combinations of hyperparameters, and created three ensembles of models: one with the 10 best models, another one with the 50 best models, and the last one with the 100 best models. The best model was the one with the top 50 models, coming 39th place.

The team **Victor-UNED** [28] participated in every task of EXIST 2024. Their system employed a concatenation of models based on the predictions of the level of agreement of the instances. They trained various transformer based models on the textual training dataset to generate predictions to

Tasks 1 and 4. Models with more information shown in their learning phase were used to determine soft labels of instances with lower level of agreement. Models were trained with soft labels, and hard labels were obtained from the resulting soft labels. For Tasks 2 & 5, and Tasks 3 & 6, results from previous tasks were incorporated, taking into account the hierarchical nature of the challenge. First run was mDeBERTa-v3 trained on this year's dataset, second run was the concatenated models to determine low agreement cases, and third run took into account an annotator ensemble to distinguish low agreement cases. Their approach achieved top rankings in Tasks 4 and 5, and was one of the most consistent models among all tasks.

Team **PINK** [29] participated in Task 4 on sexism detection in memes. They proposed a unified, multimodal Transformer-based architecture capable of dealing with multiple languages, namely English and Spanish. Their architecture extracts high-level features using large-scale, pre-trained models that are kept frozen during training. These features are then normalized and projected into the same dimensional space. They are then conditioned based on the language of the sample and its modality before being processed by a Transformer encoder backbone. The final classification is predicted through average pooling and a linear projection. With this architecture, they created three types of systems: Single Models, Majority Voting Ensembles (MVE) and Average Probability Ensembles (APE). Each of these systems was used to obtain a set of predictions, both hard and soft labels. Their approach reached the 10th and 20th places in the final ranking for soft- and hard-label evaluations, respectively.

Team **RoJiNG-CL** [30] investigated using large language models (LLMs), specifically GPT-4, to extract textual descriptions from images in Task 4. They obtained these descriptions from GPT-4's zero-shot prompting, with textual prompts alongside memes. They integrated these descriptions with related texts to fine-tune both monolingual and multilingual models, enhancing their ability to identify sexist content in memes with hard labels. They experimented with several transformer-based models and their hyperparameter's optimization with Optuna. The first run used BERT fine-tuned in English data, and BETO fine-tuned in Spanish data. The second run used mDeBERTa, and the third run used GPT-4 output results. Their submissions secured the top three positions on the hard-hard evaluation leaderboard, encompassing both English and Spanish instances. The GPT-4 based predictions emerged as the most effective, delivering top results in a zero-shot setting.

Team **NICA** [31] participated in tasks from 1 to 5. For textual models, they used various multilingual transformer models to detect sexism in English and Spanish tweets. Their runs worked with xlm-RoBERTa-Large-Twitter, multilingual BERT, and multilingual DistilBERT. First, they preprocessed texts, focusing on eliminating tags and URLs in the tweets. Their experiments showed that BERT outperformed other models, however xlm-RoBERTa-Large-Twitter showed notable better results in the test set. For tasks 4 and 5, they employed the CLIP model, which leverages both image and corresponding text data to identify sexist elements. CLIP performance yielded promising results, reaching the 9th position in the soft evaluation in Task 4, and the 4th position in both hard and soft evaluations in Task 5.

Team **Mind** [32] presented an approach for detecting sexism in memes in Task 4. In their approach, they used ResNet50 as image encoder and m-BERT to create the text embeddings, fine-tuned on EXIST 2024 dataset. Once they had the encodings, they applied a projection layer for dimensionality reduction and feature transformation on input vectors. In order to combine the image and text features and get concatenated data, they used the feature interaction matrix (FIM). Then, they trained a contrastive learning-based model on these embeddings. They computed the cosine similarity between each test sample and all the training samples. They used the K-Nearest Neighbors (KNN) algorithm to select the 10 training embeddings with the highest cosine similarity to each sample. The performance of the model achieved ICM scores of 0.2778 for English, 0.2152 for Spanish, and 0.2465 for the combined dataset.

Team **DiTana-PV** [33] focused on hard evaluation of Tasks 4 and 6. Their objectives were to evaluate the effect of machine translation on model performance and explore data augmentation techniques. They automatically translated Spanish data to English and leveraged a trained version of the BERTweet model, BERTweet-large-sexism-detector, fine-tuned in the dataset of SemEval-2023 Task 10 EDOS. Then they used data augmentation techniques to increase the amount of data and reduce the imbalance. They used BERT contextual embeddings for paraphrasing the words in the original text. One separate model was trained for each language. For Task 4, runs included models that added a weighted-loss function and a weighted-loss function plus data augmentation. The one that performed the best was the model with weighted-loss function and no data augmentation. For Task 6, their models tried to predict 5 labels or 6 labels at a time. Results showed that the model predicting 5 labels performed better.

Team **Atresa-I2C-UHU** [34] participated in Tasks 4 and 5 submitting both hard and soft labels. They focused on working with perspectives and Learning with Disagreement. They trained the multilingual versions of BERT and RoBERTa with different hyperparameters to analyze their effect in each perspective with enough values (gender, age, level of studies, Bachelor's High school White). They chose the versions of BERT and RoBERTA that worked best for every perspective, and combined them. Data was preprocessed and translated to generate supplementary training datasets. Final runs were combinations of BERT models or combinations of BERT and Roberta models to deal with hard and soft evaluations in each task. For Task 4, they ranked 4th, with ICM-Hard and ICM-Soft scores of 0.5668 and 0.4476, respectively. For Task 5, they secured 2nd and 10th places with ICM-Hard and ICM-Soft scores of 0.4119 and 0.2023, respectively.

Team **CAU&ITU_2** [35] investigated a broad range of models, including traditional machine learning methods, such as ensemble models and probability-based model like Random Forest and XGboost, and deep learning architectures models with the use of multiligual BERT. In order to obtain vector representations of texts, they implemented BOW and TF-IDF and the machine learning methods were trained with both alternatives. Hyperparameter fine-tuning was performed with RandomizedSearchCV. These methods were explored for binary classification in Task 1, and multi-class classification in Task 2. Alternatively, multilingual BERT performed better than the rest of the models in Tasks 1 and 2. For Task 3, only multilingual BERT was evaluated. Every experiment was evaluated with hard labels.

Participation of team **I2C-UHU_2** [36] in Tasks 1 and 2 aimed to employ Learning with Disagreement techniques and explore annotators' perspectives to obtain more robust models. Firstly, they preprocessed data with cleaning and normalization techniques. Then they applied data augmentation strategies and hyperparameter optimization with Optuna. They explored different transformer-based models. For Task 1, the first run used XLM RoBERTa Base to predict all instances, and the second run separated DeBERTa v3 Base for the English dataset and RoBERTa Base BNE for the Spanish dataset. For Task 2, they submitted three runs: the first one used XLM RoBERTa Base trained with the whole dataset, the second one implemented an ensemble of XLM RoBERTa Base for every annotator group, and the third one was an ensemble of XLM RoBERTa Base for every age group. In Task 1, they secured the 10th position in the hard evaluation, and the 13th position in the soft evaluation. In Task 2, they achieved the 11th position for the hard evaluation, and the 17th position for the soft evaluation.

Team **maven** [37] submitted hard labels to textual Tasks 1, 2 and 3. They focused on creating a stacking classifier composed by an ensemble of four LLMs. They built the ensemble calculating the highest Complementary Error Correction. The stacking classifier was based on LightGBM, whose parameters were fine-tuned with Optuna, and fed with the output scores of the four LLMs. They created two datasets from the original EXIST dataset by translating all data from English to Spanish, and from Spanish to English. They performed preprocessing: lowercasing and eliminating mentions, hashtags, links, numerals, etc. From the four LLMs, two of them were trained with English data (DistilBERT + RoBERTa), and the other two with Spanish data (somosnlp-hackathon-2022/twitter-sexismo-finetuned-robertuito-exist2021 and annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal). For Tasks 2 and 3, their system was based on BERT. In Task 1, the best model obtained a F1-score of 0.7359, 0.4563 for Task 2, and 0.4491 for Task 3.

Team **CIMAT-CS-NLP** [38] participated in Task 1, with hard and soft labels, and in Task 2 with hard labels. The proposed methods for both tasks are based on unifying the knowledge of two different systems: zero-shot classification with LLMs through prompting, and supervised fine-tuning of multilingual transformer models. The zero-shot classification was performed with the Gemini API (gemini-1.0-pro). Four kinds of results were taken into account, depending on the type of the prompt engineering processed. On the other hand, three types of transformer models were fine-tuned on the dataset, namely XLM-RoBERTa, mBERT and Twitter-XLM-Roberta. With seven types of results (one per prompt and model), hard labels were obtained by three methods: creation of new input for fine-tuning, proportion of votes and Best LLM response, or best fine-tuned model. Soft labels were obtained by proportion of votes. The best system achieved the 3rd place in the hard evaluation for all tweets with a F1 (positive class) of 0.7899. The highest ranked model for soft labels was in 5th place, and the two best results obtained for Task 2 were ranked 7th and 8th.

Team **Medusa** [39] adressed Task 3, sexism categorization in tweets, with hard and soft labels. They aimed to study the performance of two architectural archetype: Classifier Chain and Binary Relevance. The binary relevance architecture assumes that each label is independent of the others and can therefore be treated separately. In the classifier chain architectures, classifiers are chained together so that predictions from individual labels become features for other classifiers. These architectures constitute the multi-label classifier head situated at the top of the pretrained models from the XML-RoBERTa family. They trained various models and selected the Best BR, Best Chain, and "Best ICM-Soft", that is, the models with less BCE loss on the validation set. Their models achieved 4th, 5th, and 6th positions in the soft evaluation ranking.

Team **Penta ML** [40] participated with soft labels in Tasks 4 and 5, and with hard labels in Tasks 4, 5 and 6. They presented a multimodal architecture with five different components: (i) A Pretrained Vision-Language (ViLT) Model, which employs BERT as the text processor and ViT (Vision Transformer) as the image processor; (ii) Semantics from Pooled Representations; (iii) Attention Enhanced Context Vector for each Modality; (iv) Modality Fusion and (v) Classification Head, based on Cross Entropy loss. Then, they experimented with CLIP and VILT as their baseline models. They concatenated representations from images and texts and passed them to an MLP for classification. Since ViLT outperformed CLIP, it was used as the model in their approach. Results showed that ViLT alone obtained the best results in ICM metrics, however, their approach achieved better performance in Macro F1 in Task 6.

Team **CIMAT-GTO** [41] only participated in Task 1 with hard labels. They explored the reasoning capabilities of Llama 3 in a two-step process. In the first stage, they generated "reasoning" texts using a LLM that aims to understand the tweets' nature. These rationales are added to the tweets and, in the second stage, they are processed further with a pre-trained XLM-RoBERTa model trained on multilingual tweets. They differentiated between various types of reasoning: positive and negative reasoning and comparative reasoning. They included answers from the LLM (Llama 3) to questions about sexism identification. They processed these answers with a multi-layer FFN and then concatenated with the text. The models corresponded to a RoBERTa model with negative reasoning, an ensemble of models with negative and comparative reasoning and an ensemble of models with negative, comparative and answering reasoning. The latter was the best performing model, reaching the 4th position in the hard evaluation ranking for Task 1.

Team **EquityExplorers** [42] submitted two runs of hard labels to Task 1. These two runs corresponded to the results of their two approaches: the Dual-Transformer Fusion Network (DTFN) and the Multimodel Fusion Ensemble (MFE). The DTFN is based on the fusion of two Transformer models, RoBERTa-Large and DeBERTa-V3-Large. This ensemble model leverages the distinctive characteristics of each constituent model to enhance text classification. MFE is a more complex approach based on the ensemble of LLMs (Mistral-7b, RoBERTa-Large and DeBERTa-V3-Large), and the DTFN using a majority voting mechanism. Evaluation showed that these methodologies significantly outperform existing models, with MFE and DTFN ranking 1st and 2nd, respectively, in the English segment, and 4th and 13th in the combined English and Spanish segments of the official leaderboard.

Umera Wajeed Pasha [43] participated in Task 4 on sexism identification in memes. The study starts by importing and visualizing a meme dataset, then pre-processing the images using techniques including cropping, scaling, and normalization to get them ready for model training. A pre-trained model called CLIP is used to extract features, and the dataset is split into training and validation sets for memes in both Spanish and English. The collected features are used to train and assess a variety of machine learning models, such as Logistic Regression, SVM, XGBoost, Decision Trees, Random Forest, Neural Network, AdaBoost, and SGD. The Random Forest model performed the best out of all of them.

Team **VerbaNex AI** [44] proposed a method to deal with Task 1 in the hard setting. They implemented a profiling approach based on demographic factors: gender, education level, and age. This allowed to categorize the profiles into four groups based on their likelihood of labelling messages as sexist or not sexist. Then they performed feature extraction and trained four distinct systems based on the grouped profiles and their responses. To address class imbalance, they used K-Fold Stratified Shuffle-Split. They incorporated the Twitter-roBERTa-base model specifically fine-tuned for sentiment analysis. This method was evaluated using the testing profiles, achieving a F1 score of 0.745. In the evaluation phase, their approach yielded a F1 score of 0.63.

Team **MMICI** [45] participated in every task of the EXIST challenge. For textual tasks, they used transformer models ("cardiffnlp/twitter-roberta-base-sentiment" for English and "pysentimiento/robertuito-base-uncased" for Spanish). They created two types of ensembles. The first one used a majority vote from the outputs of six different models, one for each annotator. The second ensemble used a majority vote from the outputs of five different models, focusing on gender and age. For multimodal tasks, they utilized CLIP embeddings using a Vision Transformer (ViT) model and two types of classifiers: FNNs and Factorization Machines. In runs 1 and 2, demographic information was represented using one-hot encoding, whereas, in run 3, a descriptive text was created for the annotator features, from which embeddings were extracted. For runs 2 and 3, the classifier used a FNN, and in run 3 they proposed a Factorization Machine model. Their best performances include a 10th place in Task 1, a 15th place in Task 2, and a 13th place in Task 3 for Spanish tweets. For memes, they achieved a 3rd place in Task 4 for English.

Team **Penta-nlp** [46] participated in Tasks 1 to 3. They explored multiple approaches: Machine Learning (ML), Deep Learning (DL) and Transformer-based Pretrained Models. The ML models included Support Vector Machine, Random Forest, XGBoost (Tasks 1 & 2) and Logistic Regression (Task 3). The DL models leverage both LSTM and LSTM + Attention models. Transformer models explored XLM-RoBERTa, mBERT and BETO. They conducted experiments using various preprocessing methods: removing usernames, URLs, punctuation and emojis. It was shown that models performed best without URLs for Tasks 1 and 3. A single run with hard labels per task was submitted to each task, reaching the 29th position in Task 1, and the 9th position in Task 2.

The **ABCD** [47] team participated in Tasks 1, 2 and 3 with both hard and soft labels. In their approaches, they employed both LLMs like Llama 2 and T5 and smaller models like XLM_RoBERTa. They divided the datasets into six subsets corresponding to each annotator group. Subsamples are preprocessed, and prompt engineering is applied to LLMs. Then, smaller transformer models are fine-tuned on each subset and predictions are collected for each model. To incorporate the hierarchical structure of Tasks 2 and 3, they only made predictions for subsamples classified as sexist in Task 1. Their best performance model achieved 2nd in Task 1, 1st in Task 2, and 1st in Task 3 for the hard evaluation.

7. Results

In the next subsections, we report the results of the participants and the baseline systems for each task. Disaggregated results for each language, English and Spanish, may be found at the EXIST 2024 website (http://nlp.uned.es/exist2024/).

7.1. Task 1: Sexism Identification in Tweets

We first report and analyze the results for Task 1, which focuses on sexism identification in tweets. This task involves a binary classification. As discussed in Section 5, we report two sets of evaluation results (hard and soft).

7.1.1. Soft Evaluation

Table 2 presents the results for the soft-soft evaluation for 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.

Table 2: Results of 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 [23]	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 [47]	4	0.9291	0.6490	1.2637
CIMAT-CS-NLP_3 [38]	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 [19]	9	0.8179	0.6311	0.9750
Awakened_2 [22]	10	0.7196	0.6154	0.8106
Victor-UNED_1 [28]	11	0.6952	0.6115	1.0691
Awakened_3	12	0.6909	0.6108	0.8542
I2C-UHU_2 [36]	13	0.6871	0.6102	0.9184
Victor-UNED_2	14	0.6797	0.6090	0.9818
UMUTEAM_1 [25]	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 [45]	21	0.4589	0.5736	2.0316
clac_1	22	0.1431	0.5230	2.9543
RMIT-IR_1 [16]	23	−0.0011	0.4998	2.7892
FraunhoferSIT_1 [14]	24	−0.0658	0.4895	0.8801
CNLP-NITS-PP_1 [21]	25	−0.2086	0.4666	1.0390
RMIT-IR_3	26	−0.3016	0.4516	2.8235
Atresa-I2C-UHU_1 [34]	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 [31]	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

7.1.2. 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. 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.

Table 3: Results of 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 [42]	4	0.5883	0.7957	0.7775
CIMAT-GTO_3 [41]	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 [18]	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 [45]	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 [15]	35	0.4398	0.7211	0.7462

Continued on next page

Table 3 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
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 [27]	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 [35]	45	0.3460	0.6739	0.7024
DLRG_1	46	0.3446	0.6732	0.7085
TextMiner_1	47	0.3412	0.6715	0.7048
shm2024_3 [24]	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 [20]	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 [44]	64	−0.4048	0.2965	0.4588
The 3 Musketeers_1 [26]	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

7.2. Task 2: Source Intention in Tweets

In this section, we report and analyze the results for Task 2, which focuses on determing the intention of the author when posting a sexist tweet. This task is a multi-class, mono-label classification. We report two sets of evaluation results (hard and soft).

7.2.1. 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 than 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 range from the 0.4795 of the best system (“nycu-nlp_2”) to 0.0000 of “fmrs_2”, indicating significant variability in the effectiveness of the submitted models. It is worth mentioning that the best system outperforms the second-best by more than 8 percentage points. Overall, performance is considerably lower compared to Task 1. This can be attributed to the hierarchical and multi-class 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 differences can still be observed due to the fact that cross entropy does not have into account the specificity of the different classes.

Table 4: Results of 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

7.2.2. 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 as "REPORTED"). Similar to the soft-soft evaluation, discrepancies between the best and the worst-performing systems are more pronounced in Task 2 than in Task 1. The top-ranking system, "ABCD Team_1," achieved the highest ICM-Hard normalized score (0.6320). The top 5 best systems range between 0.5937 and 0.6320. The lower end of the table includes five systems which score 0 in the ICM-Hard norm metric.

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.

Table 5: Results of Task 2 in the hard-hard evaluation.

Run	Rank	ICM-Hard	ICM-Hard 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
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 [37]	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

Continued on next page

Table 5 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
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

7.3. Task 3: Sexism Categorization in Tweets

The third task is a hierarchical multi-class and multi-label 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.

7.3.1. 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" 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: Results of Task 3 in the soft-soft evaluation.

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2024 gold	0	9.4686	1.0000
NYCU-NLP_1	1	−1.1762	0.4379
NYCU-NLP_2	2	−1.2169	0.4357
NYCU-NLP_3	3	−1.4555	0.4231
Medusa_1 [39]	4	−2.2055	0.3835
Medusa_2	5	−2.4010	0.3732

Continued on next page

Table 6 – continued from previous page

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

7.3.2. 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 is less than in Task 2. Finally, comparing the performance of the three different textual tasks in the hard-hard evaluation, the efficiency of the systems in this task, in terms of ICM-Hard Norm, is lower than in previous tasks. This further highlights the complexity of categorizing sexism.

Table 7: Results of 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

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
ABCD Team_1	1	0.3713	0.5862	0.6004
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 [46]	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

7.4. Task 4: Sexism Identification in Memes

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

7.4.1. Soft Evaluation

Table 8 presents the results for the classification of 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. However, the variability between the best and worst-performing systems is reduced compared to that of the tasks described above. 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 might often be conveyed through the imagery, even when the accompanying text seems to be neutral.

Table 8: Results of 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 [29]	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 [40]	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
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

7.4.2. Hard Evaluation

Table 9 presents the results for the hard-hard evaluation of Task 4. 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: Results of 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 [30]	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 [33]	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
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

Continued on next page

Table 9 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
Umera Wajeed Pasha_1 [43]	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 [32]	42	−0.4986	0.2465	0.5674
ALC-UPV-JD-2_1	43	−0.5446	0.2231	0.4878
dap-upv_1 [17]	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

7.5. Task 5: Source Intention in Memes

In this section, we report and analyze the results for Task 5, which focuses on determing the intention of the author when posting a sexist meme. This task is a multi-class, mono-label classification. We report two sets of evaluation results (hard and soft).

7.5.1. Soft Evaluation

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.

Table 10: Results of 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

Continued on next page

Table 10 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
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

7.5.2. Hard Evaluation

Table 11 presents the results for the hard-hard evaluation of Task 5. 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, but are quite homogeneous among the top 5 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: Results of 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

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
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

7.6. Task 6: Sexism Categorization in Memes

The sixth task is a hierarchical multi-class and multi-label classification problem, where systems must determine if a meme is sexist or not, and if so, categorize it according to the five categories of sexism defined in Section 2.

7.6.1. 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: Results of 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

Continued on next page

Table 12 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm
MMICI_2	20	−19.3246	0.0000
MMICI_3	21	−45.0237	0.0000
EXIST2024 minority	22	−50.0353	0.0000

7.6.2. 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.

Table 13: Results of 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

8. Discussion

After the study of the 34 submitted working notes, we have discovered patterns of system performance that are of the greatest interest to the task. Even though in this analysis we only considered systems whose working notes were available, most of the best performing models were considered. We will focus on the conclusions from the study of the top ten systems for the six tasks in hard and soft evaluations, and the classification in English and Spanish for each one. We can divide two types of approaches submitted: textual systems for Tasks 1, 2 and 3, and systems for Tasks 4, 5 and 6.

8.1. Tasks 1, 2 and 3: Sexism Detection and Classification in Tweets

In the textual tasks, most of the top-ranking models were encoding-based transformer, fine-tuned on the EXIST dataset with an additional component. This additional component could be:

- A meticulous data preprocessing step. For example, that is the main contribution of the *NYCU-NLP* team: they removed irrelevant elements of the text, and they increased the size of the dataset by applying data augmentation techniques such as AEDA, and by translating from English to Spanish and vice versa.
- Ensembles of encoding-based transformer models. A huge variety of models were employed: most of them include BERT-like systems like BERT, RoBERTa, DeBERTa, multilingual versions of them trained in Spanish datasets, and some were fine-tuned in tweets, hate-speech or sentiment analysis datasets. Systems following this architecture excel particularly in the soft evaluation. This success appears to be linked to how teams trained their models, as encoding-based models are easily fine-tuned with soft labels. Depending on the models used to obtain results for the ensemble, different types of voting methods are considered:
  - If the results differ highly between models, systems seem to perform better when a higher weight is given to the best model and the influence of the rest of the models is reduced. The *Awakened* team considered different models in their ensembles, but their best performing runs gave higher weight to their best model in the ensemble.
  - If the differences between the performance of models are small, a proportion of votes can take into account aspects detected by every model. This method obtained good results for *CIMAT-CS-NLP* in the first task, and *Medusa* in the third task.
- Sharing results with LLMs. This method proved to be successful in the hard evaluation, where teams that submitted labels obtained by both LLMs and encoding-based transformer models reached the best results. Due to the increasing amount of LLMs published and their constant improvement, there is a whole range of models that the community can try. In EXIST tasks, there were attempts testing Llama-2, Llama-3, Gemini, Mistral and GPT-4. Most of them are used to obtained labels using zero-shot or few-shots because of their computational cost of fine-tuning. Thus, approaches rely on the use of prompt engineering to make the model understand the task. Because of these differences in models and ways of formulating prompts, systems cannot be directly compared; however, we can mention as some of the best performing ones: the system of *CIMAT-CS-NLP*, that uses an ensemble of 4 zero-shot answers of Gemini; the team *ABCD*, which using prompt engineering created one Llama-2 answer per annotator; the team *EquityExplorers*, who used Mistral results in their ensemble, and the team *CIMAT-GTO*, which studied types of reasoning with Llama-3.

In general, we can draw some general conclusions about systems for textual tasks. Models that use encoding-based transformers performed better in the soft evaluation. Even systems which use one model or one model trained in different ways, like *BAZI* and *Victor-UNED*, achieved top-10 ranking results in the soft evaluation because they are trained with soft labels. On the other hand, LLMs performance stands out in the hard evaluation, but drops in the soft evaluation. This can be clearly seen with *ABCD*, where runs obtained by Llama-2 got better results than encoding models in the hard evaluation for the three tasks, whereas encoding models outperformed them in the soft evaluation. In general, demographic information was not included in models, but those teams who have included it have obtained improvements in their systems in Task 3.

8.2. Tasks 4, 5 and 6: Sexism Detection and Classification in Memes

Tasks 4, 5 and 6 are multimodal, that is, teams could use both images and text to detect and categorize sexism in the instances. Although most of the teams have faced these tasks as multimodal, the best performances correspond to models which only use text to analyze memes. Top ranking positions for most of the tasks, especially Tasks 4 and 5, were obtained by textual models. However, images remain important. *RoJiNG-CL* used GPT-4 to create a textual description of the image and analyzed it, outperforming the other approaches. Models that were used to analyze the text were mostly encoding-based transformers: combinations of BERT, DeBERTa and RoBERTa with different weights or different fine-tunings. Systems that only considered text where similar to those presented in the previous section.

Next in the ranking are multimodal approaches. The influence of textual analysis implies that even in multimodal approaches the ones that focus on textual treatment with a ViT module using BERT or similar outperform models than do not include it. This phenomenon is shown in *RMIT-IR* runs, whose best run is the one that used mBERT and obtained several positions ahead of their next run which did not include it. Most of the systems utilized CLIP to analyze images. However, the use of CLIP alone led to poor results. The concatenation of text and images stands as the necessary way of dealing with these tasks, but new ways of obtaining representations of images are needed to outperform text-only systems.

9. 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 60 participant teams, and received more than 400 runs. Participants adopted a wide range of approaches, including vision transformer models, data augmentation through automatic translation, data duplication, utilization of data from past EXIST editions, multilingual language models, Twitter-specific language models, and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis. While many systems opted for the traditional approach of providing only hard labels as outputs, a significant number of systems leveraged the multiple annotations available in the dataset, and provided soft outputs, proving that there is an increasing interest by the research community in developing systems able to deal with disagreements and with different perspectives.

Concerning the results, in the textual tasks (Tasks 1, 2 and 3), top-performing models were typically encoding-based Transformers fine-tuned on the EXIST dataset with an additional component, such as meticulous data preprocessing, data augmentation or the use of model ensembles. In multimodal tasks (Tasks 4, 5, and 6), where both images and text could be used to detect and categorize sexism, top performances were achieved by models focusing solely on text.

For future editions of EXIST, we plan to expand our study in order 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 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).

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval), 2023.

[6] M. Billig, Humour and hatred: the racist jokes of the Ku Klux Klan, Discourse & Society 12 (2014) 267–289.

[7] 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.

[8] 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.

[9] 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.

[10] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.

[11] B. R. Chakravarthi, S. Rajiakodi, R. Ponnusamy, K. Pannerselvam, A. K. Madasamy, R. Rajalakshmi, H. LekshmiAmmal, A. Kizhakkeparambil, S. S Kumar, B. Sivagnanam, C. Rajkumar, Overview of Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes, in: Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, 2024, pp. 139–144.

[12] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[13] 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.

[14] S. Fan, R. A. Frick, M. Steinebach, FraunhoferSIT@EXIST2024: Leveraging Stacking Ensemble Learning for Sexism Detection, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[15] M. Usmani, R. Siddiqui, S. Rizwan, F. Khan, F. Alvi, A. Samad, Sexism Identification in Tweets using BERT and XLM – Roberta, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[16] T. Smith, R. Nie, J. Trippas, D. Spina, RMIT-IR at EXIST Lab at CLEF 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[17] M. Obrador Reina, A. García Cucó, LightGMB for Sexism Identification in Memes, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[18] A. Shah, A. Gokhale, Team Aditya at EXIST 2024 — Detecting Sexism In Multilingual Tweets Using Contrastive Learning Approach, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[19] A. Azadi, B. Ansari, S. Zamani, Bilingual Sexism Classification: Fine-Tuned XLM-RoBERTa and GPT-3.5 Few-Shot Learning, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[20] M. Siino, I. Tinnirello, Prompt Engineering for Identifying Sexism using GPT Mistral 7B, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[21] A. Vetagiri, P. Mogha, P. Pakray, Cracking Down on Digital Misogyny with MULTILATE a MULTImodaL hATE Detection System, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[22] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[23] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, NYCU-NLP at EXIST 2024 – Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[24] G. Shimi, J. Mahibha, D. Thenmozhi, Automatic Classification of Gender Stereotypes in Social Media Post, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[25] R. Pan, J. A. García Díaz, T. Bernal Beltrán, R. Valencia-Garcia, UMUTeam at EXIST 2024: Multimodal Identification and Categorization of Sexism by Feature Integration, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[26] M. Sreekumar, S. K, T. Durairaj, S. Gopalakrishnan, K. Swaminathan, Sexism Identification in Tweets using Traditional Machine Learning Approaches, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[27] R. Keinan, 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: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[28] V. Ruiz, J. Carrillo-de-Albornoz, L. Plaza, Concatenated Transformer Models based on Levels of AgreementsfFor Sexism Detection, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[29] G. Rizzi, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, PINK at EXIST2024: A Cross-Lingual and Multi-Modal Transformer Approach for Sexism Detection in Memes, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[30] J. Ma, R. Li, RoJiNG-CL at EXIST 2024: Sexism Identification in Memes by Integrating Prompting and Fine-Tuning, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[31] A. Naebzadeh, M. Nobakhtian, S. Eetemadi, NICA at EXIST CLEF Tasks 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[32] F. Maqbool, E. Fersini, A Contrastive Learning based Approach to Detect Sexism in Memes, in:

Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[33] A. Menárguez Box, D. Torres Bertomeu, DiTana-PV at sEXism Identification in Social neTworks (EXIST) Tasks 4 and 6: The Effect of Translation in Sexism Identification, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[34] Álvaro Carrillo-Casado, J. Román-Pásaro, J. Mata-Vázquez, V. Pachón-Álvarez, I2C-UHU at EXIST 2024: Transformer-Based Detection of Sexism and Source Intention in Memes Using a Learning with Disagreement Approach, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[35] N. Maqbool, Sexism Identification in Social Networks: Advances in Automated Detection – A Report on the Exist Task at CLEF, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[36] M. Guerrero-García, M. Cerrejón-Naranjo, J. Mata-Vázquez, V. Pachón-Álvarez, I2C-UHU at EXIST2024: Learning from Divergence and Perspectivism for Sexism Identification and Source Intent Classification, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[37] A. Shanbhag, S. Jadhav, A. Date, S. Joshi, S. Sonawane, The Wisdom of Weighing: Stacking Ensembles for a More Balanced Sexism Detector, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[38] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better Together: LLM and Neural Classification Transformers to Detect Sexism, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[39] G. Aru, N. Emmolo, A. Piras, S. Marzeddu, J. Raffi, L. C. Passaro, RoBEXedda: Enhancing Sexism Detection in Tweets for the EXIST 2024 Challenge, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[40] D. D. Barua, M. S. U. R. Sourove, F. Haider, F. T. Shifat, M. F. Ishmam, M. Fahim, F. A. Bhuiyan, Penta ML at EXIST 2024: Tagging Sexism in Online Multimodal Content With Attention-enhanced Modal Context, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[41] K. Villarreal-Haro, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Stacked Reflective Reasoning in Large Neural Language Models, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[42] S. Khan, G. Pergola, A. Jhumka, Multilingual Sexism Identification via Fusion of Large Language Models, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[43] U. W. Pasha, Multilingual Sexism Detection in Memes, A CLIP-Enhanced Machine Learning Approach, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[44] E. Martinez, J. Cuadrado, J. C. Martinez-Santos, E. Puertas, VerbaNex AI at CLEF EXIST 2024: Detection of Online Sexism using Transformer Models and Profiling Techniques, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[45] M. P. Jimenez-Martinez, J. M. Raygoza-Romero, C. E. Sánchez-Torres, I. H. Lopez-Nava, M. Montes-y Gómez, Enhancing Sexism Detection in Tweets with Annotator-Integrated Ensemble Methods and Multimodal Embeddings for Memes, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[46] F. T. Shifat, F. Haider, M. S. U. R. Sourove, D. D. Barua, M. F. Ishmam, M. Fahim, F. A. Bhuiyan, Penta-nlp at EXIST 2024 Task 1–3: Sexism Identification, Source Intention, Sexism Categorization In Tweets, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[47] L. M. Quan, D. V. Thin, Sexism Identification in Social Networks with Generation-based Approach, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.