Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization

Laura Plaza[1,2(✉)], Jorge Carrillo-de-Albornoz[1,2], Roser Morante[1], Enrique Amigó[1], Julio Gonzalo[1], Damiano Spina[2], and Paolo Rosso[3]

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
{lplaza,jcalbornoz,rmorant,enrique,julio}@lsi.uned.es
[2] RMIT University, 3000 Melbourne, Australia
damiano.spina@rmit.edu.au
[3] Universitat Politècnica de València (UPV), 46022 Valencia, Spain
prosso@dsic.upv.es

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 since 2021. The third edition of EXIST, hosted at the CLEF 2023 conference, consists of three tasks, two of which are the continuation of EXIST 2022 (*sexism identification* and *sexism categorization*), and a third and novel one is on *source intention identification*. For this edition, new test and training data are provided and the "learning with disagreement" paradigm is adopted to address disagreements in the labelling process and promote the development of equitable systems that are able to learn from different perspectives on the sexism phenomena. 28 teams participated in the three EXIST 2023 tasks, submitting 232 runs. This lab overview describes the tasks, dataset, evaluation methodology, approaches and results.

Keywords: Sexism Detection · Sexism Categorization · Data Bias · Learning with Disagreement

1 Introduction

Sexism is defined by the Oxford English Dictionary as "prejudice, stereotyping, or discrimination, typically against women, on the basis of sex". This phenomenon remains prevalent even in contemporary, developed societies, and among the younger generations who have grown up in democratic societies. Sexism continues to pose significant challenges for women in various areas of their lives, such as work, family life, and personal growth, acting as a barrier that impedes their progress.

A. Arampatzis et al. (Eds.): CLEF 2023, LNCS 14163, pp. 316–342, 2023.
https://doi.org/10.1007/978-3-031-42448-9_23

Sexism manifests in many different forms and can be categorized according to different dimensions. Regarding the facet of the women that is attacked, we can find attitudes such as sexual objectification, stereotyping or patriarchy. Depending on the level of society at which discrimination occurs, we can find institutional sexism, interpersonal sexism and individual sexism. Depending on the expression and underlying motivation, sexism can be categorized into two main types: hostile sexism and benevolent sexism. While hostile sexism involves openly negative and antagonistic attitudes towards women, benevolent sexism appears positive but patronizing, involving protective attitudes towards women while still reinforcing restrictive gender roles.

The persistence of gender inequality and discrimination against women in society is now being replicated and amplified in the online realm, as highlighted by Azmina et al. [8]. The Internet not only perpetuates these gender differences but also normalizes sexist attitudes, as noted by Burgos [4]. Specially concerning is the spread of sexist messages through social networks. Social networks have allowed women to rise their voices to report abuses, discrimination and sexist experiences, but the anonymity that they provide has also facilitated the transmission of hateful behaviours against women. With the increase of social media use by children and adolescents, detecting and fighting against online sexism becomes a priority. Previous studies [12] have shown that media content influences how social realities are perceived, so that the exposure to sexist and even misogynous content may contribute to develop sexist attitudes or to perceive them as natural or acceptable. Moreover, social media platforms are not acting efficiently to remove or avoid sexist and hateful content.

EXIST 2023[1] at CLEF is the third edition of the EXIST (sEXism Identification in Social neTworks) challenge that aims at combating sexism on social media. In 2021 and 2022, the EXIST shared tasks were proposed at the IberLEF forum [30,31]. These editions 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 that involve implicit sexist behaviours. The 2021 and 2022 EXIST editions more than 50 teams participated from research institutions and companies from all around the world. While the two previous editions only focused on classifying sexist messages according to the facet of the women that was being undermined, the 2023 edition tackles an additional task that aims to determine the intention of the author of a sexist message. Since social networks are usually used to report and criticize sexist situations, it is important to distinguish the messages that are sexist by themselves from those that report experiences with the aim of raising awareness against sexism.

Additionally, the main novelty of the 2023 EXIST edition, and what makes it different to other recent initiatives such as the SemEval-2023 Shared Task 10: "Explainable Detection of Online Sexism" [17], is the adoption of the "Learning with Disagreement" (LwD) paradigm [36] for the development of the dataset and for the evaluation of the systems. Our previous work showed that the perception

[1] http://nlp.uned.es/exist2023/. Accessed 14 June 2023.

of sexism is strongly dependent on the demographic and cultural background of the subject, so that when identifying sexist attitudes and expressions, and even when classifying them in different sexist categories, people sometimes disagree. In the LwD paradigm, instead of relying on a single "correct" label for each example, the model is trained to handle and learn from conflicting or diverse annotations. In this way, that different annotators' perspectives, biases, or interpretations may be taken into account by the systems and the learning process becomes more fair.

In addition to adopting the LwD paradigm, we have made an effort to control bias in the annotations (see Sect. 3) and have developed new evaluation metrics that take into account the disagreement. This will allow us to assess whether including the different views and sensibilities of the annotators contributes to the development of more accurate and equitable NLP systems.

In the following sections, we provide comprehensive information about the tasks, the dataset, the evaluation methodology, the results and the different approaches of the systems that participated in the EXIST 2023 Lab. The competition features three distinct tasks: (i) sexism identification, (ii) source intention classification, and (iii) sexism categorization. A total of 103 teams from 29 different countries registered to participate. Ultimately, we received 232 results from 28 teams, and 24 of them completed the process by submitting the working notes. 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 in social networks.

2 Tasks

The two first editions of EXIST focused on detecting sexist messages in two social networks, Twitter and Gab,[2] as well as on categorizing these messages according to the type of sexist behaviour they enclose. For the 2023 edition, we focus on Twitter only and we address an additional task, namely "source intention classification". The three tasks are described below.

2.1 Task 1: Sexism Identification

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

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

[2] https://gab.com/. Accessed 14 June 2023.

(2) **Not sexist:** *Just saw a woman wearing a mask outside spank her very tightly leashed dog and I gotta say I love learning absolutely everything about a stranger in a single instant.*

2.2 Task 2: Source Intention

This task aims to categorize the message according to the intention of the author. We propose a ternary classification of tweets: (i) direct sexist message, (ii) reported sexist message, and (iii) judgemental message. 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. The three categories are defined as:

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

2.3 Task 3: Sexism Categorization

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. Automatically detecting which of these facets of women are being more frequently attacked in social networks will facilitate the development of policies to fight against sexism. According to this, each sexist tweet must be assigned one or more of the following categories:

- **Ideological and inequality**. This category includes messages that discredit the feminist movement in order to devalue, belittle and defame the struggle of women in any aspect of their lives. It also includes messages that reject inequality between men and women, or present men as victims of gender-based oppression.
  (6) *#Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.*

- **Stereotyping and dominance**. This category includes messages that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks. It also includes any claim that implies that men are somehow superior to women.

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

- **Objectification**. This category 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, for example, ideas that suggest that women should maintain a standard and ideal of beauty or attacks on a woman's physique.

  (8) *No offense but I've never seen an attractive African American hooker. Not a single one.*

- **Sexual violence**. This category includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made.

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

- **Misogyny and non sexual violence**. This category includes expressions of hatred and violence towards women.

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

3 Dataset: Managing Bias in Data and Providing Pre-aggregated Annotations

An important problem in NLP that has gained attention in the recent years in parallel to the growing protagonism of large language models is bias, both in the data that are used to train and test systems, and in the way algorithms learn, mainly due to the bias in the data [32]. In EXIST 2023 we tackle one aspect of this problem, i.e., the data bias that may be introduced both during the data selection and during the labeling process.

3.1 Data Sampling

To gather the data, we follow the methodology used in previous EXIST editions [30,31]. We first collected different popular expressions and terms, both in English and Spanish, commonly used to underestimate the role of women in our society. These expressions have been extracted from different sources: (a) previous works in the area; (b) Twitter accounts (journalist, teenagers, etc.) or hashtags used to report sexist situations; (c) expressions extracted from The Every Day Sexism Project;[3] and d) a compendium of feminist dictionaries. These expressions were later used as seeds to retrieve Twitter data. To mitigate the **seed bias**, we have also gathered other common hashtags and expressions less

[3] https://everydaysexism.com/. Accessed 14 June 2023.

frequently used in sexist contexts to ensure a balanced distribution between sexist/not sexist expressions. This first set of seeds contains more than 400 expressions.

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

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

3.2 Labeling Process

Before starting the annotation process we considered possible sources of **label bias** [14]. Label bias may be introduced by socio-demographic differences of the persons that participate in the annotation process.

The labeling of the data was carried out by crowd-workers,[4] selected according to their different demographic characteristics in order to minimize the label bias. We consider gender (male/female) and age (18–22 y.o./23–45 y.o./+46 y.o). Each tweet was annotated by 6 crowdsourcing annotators selected through Prolific.[5] The Prolific crowdsourcing platform was specifically selected because of the features it provides to define participant criteria in the recruiting process – in our case, gender, age, and fluency in the different languages.

Different quality control mechanisms were employed, including small/medium size batches to ensure that the data is labeled by a significant/diverse amount of annotators, control of the time employed to perform the task, outlier analysis, and the use of attention mechanisms and ground truth data. We communicated frequently with the workers to solve doubts and to correct errors was kept.

[4] No personally identifiable information about the crowd-workers was collected. Workers were informed that the tweets could contain offensive information and were allowed to withdraw voluntarily at any time. Full consent was obtained.

[5] https://www.prolific.co/. Accessed 14 June 2023.

The annotators were provided with annotation guidelines that included a detailed description of the different tasks along with numerous examples, both positive and negative, for all different categories/labels. The guidelines were developed by two experts in gender issues.

3.3 Pre-aggregated Annotations: The Learning with Disagreement Paradigm

As stated by Uma et al. [36], 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. To deal with this, Uma et al. [36] have proposed the Learning With Disagreement (LwD) paradigm, which consists mainly of letting systems learn from datasets with information about the annotations from all annotators, in an attempt to gather the diversity of views. In the case of sexism identification, this is particularly relevant, since the perception of a situation as sexist or not can be subjective and may depend on the gender, age and cultural background of the person who is judging it. Following methods proposed for training directly from the data with disagreements, instead of using an aggregated label [24,29,34], the EXIST 2023 dataset provides multiple annotations per example. The LwD paradigm may also help to mitigate bias and produce equitable NLP systems. The selection of annotators for the development of the EXIST 2023 dataset took into account the heterogeneity necessary to avoid gender and age biases.

4 Evaluation Methodology and Metrics

As in SemEval 2021, we have carried out a **"soft evaluation"** and a **"hard evaluation"**. The soft evaluation corresponds to the LwD paradigm and is intended to measure the ability of the model to capture disagreements, by considering the probability distribution of labels in the output as a soft label and comparing it with the probability distribution of the annotations. The hard evaluation is the most standard evaluation paradigm and assumes that a single label is provided by the systems for every instance in the dataset.

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

- Task 1 (sexism identification): binary classification, monolabel.
- Task 2 (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.
- Task 3 (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 and 2, which are monolabel problems, the sum of 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 2023, for each of the tasks, three types of evaluation have been performed:

1. **Soft-soft evaluation**. For systems that provide probabilities for each category, we provide a soft-soft evaluation that compares the probabilities assigned by the system with the probabilities assigned by the set of human annotators. The probabilities of the classes for each instance are calculated according to the distribution of labels and the number of annotators for that instance. We use a modification of the original ICM metric (Information Contrast Measure [1]), ICM-Soft (see details below), as the official evaluation metric in this variant and we also provide results for the normalized version of ICM-Soft (ICM-Soft Norm). It is important to note that ICM is a measure that quantifies information, and its upper and lower bounds are $+\infty$ and $-\infty$, respectively. To normalize the results, we used the gold standard score as the upper bound and minority class baseline as the lower bound. We also provide results for Cross Entropy.
2. **Hard-hard evaluation**. For systems that provide a hard, conventional output, we provide a hard-hard evaluation. To derive the hard labels in the ground truth from the different annotators' labels, we use a probabilistic threshold computed for each task. As a result, for Task 1, the class annotated by more than 3 annotators is selected; for Task 2, the class annotated by more than 2 annotators is selected; and for Task 3 (multilabel), the classes annotated by more than 1 annotator are selected. The instances for which there is

no majority class (i.e., no class receives more probability than the threshold) are removed from this evaluation scheme. The official metric for this task is the original ICM, as defined by Amigó and Delgado [1]. We also report a normalized version of ICM (ICM Norm) and F1. In Task 1, we use F1 for the positive class. In Tasks 2 and 3, we use the average of F1 for all classes. 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.

3. **Hard-soft evaluation**. For systems that provide a hard output, we will also provide a hard-soft evaluation comparing the categories assigned by the system with the probabilities assigned to each category in the ground truth. As in the previous case, we use ICM-Soft as the official evaluation metric in this variant. In this evaluation, the hard outputs are transformed into soft outputs by assigning a probability of 1.0 to the selected class and 0.0 to the other classes.

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

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

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

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

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

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

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

Note that the information content of assigning a category c with an agreement v grows inversely with the probability of finding an instance that receives

category c with agreement equal or larger than v. To this end, we compute the mean and deviation of the agreement levels for each class across instances, and applying the cumulative probability over the inferred normal distribution.[6]

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

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

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

5 Overview of Approaches

Although 103 teams from 29 different countries registered for participation, the number of participants who finally submitted results were 28, submitting 232 runs. Teams were allowed to participate in any of the three tasks and submit hard and/or soft outputs. Table 1 summarizes the participation in the different tasks and evaluation contexts.

Table 1. Runs submitted per task and evaluation scenario.

	Task 1		Task 2		Task 3	
	Hard	Soft	Hard	Soft	Hard	Soft
# runs	67	54	33	25	30	23
# teams	28		15		14	

The evaluation campaign started on February 13, 2023 with the release of the training set. The development set was released on March 27, and the test set was made available on April 10. The participant teams were provided with the official evaluation script. Runs had to be submitted by May 15. Each team could submit up to three runs per task, that may contain soft and/or hard outputs.

For a comprehensive description of the systems submitted by the participants, please refer to extended overview [26] and the participants' working notes [2,3,5–7,9–11,13,15,16,18–23,25,27,28,33,35,37,38]. Here we summarize the main approaches.

Approximately 90% of the systems submitted utilized large language models, both monolingual and multilingual. Some teams employed ensembles of multiple language models to enhance the overall performance. Only two teams utilized

[6] In the case of zero variance, we must consider that the probability for values equals or below the mean is 1 (zero IC) and the probability for values above the mean must be smoothed.

deep learning architectures such as BiLSTM, CNN, and RNN, while two others opted for traditional machine learning methods, including perceptrons, Naive Bayes, and SVM.

Data augmentation techniques were used by several teams, involving the translation of tweets, the utilization of data from similar tasks (such as previous EXIST editions), and the duplication of instances within the training set. Additionally, Twitter-specific models and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis were also utilized.

Most participants made use of the soft labels and applied the LwD paradigm, rather than opting for a traditional approach and providing only hard labels as outputs.

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

- **Majority class:** non-informative baseline that classifies all instances as the majority class.
- **Minority class:** non-informative baseline that classifies all instances as the minority class.
- **Oracle most voted:** hard approach that selects the most voted label following the same procedure as the one used to generate the gold hard. Note that this baseline is only employed in the hard-soft evaluation.

6 Results

In the next subsections, we report the results of the participants and the baseline systems for each task.

6.1 Task 1 Sexism Identification

We first report and analyze the results for Task 1, which focuses on sexism identification. This task involves a binary classification. As discussed in Sect. 4, we report three sets of evaluation results.

Soft-Soft Evaluation. Table 2 presents the results for this evaluation context. 54 runs were submitted. 46 runs outperformed the non-informative majority class baseline (all instances are labeled as "NO"), while 51 runs surpassed the non-informative minority class baseline (all instances are labeled as "YES").

Looking at the results in Table 2, we observe a notable variation in the performance of the runs, ranging from an ICM-Soft score of 0.903 (equivalent to a 64% ICM-Soft Norm) to −5.6659 (lower than the empirically determined lower bound). However, it is worth mentioning that the best run only reached a 64% ICM-Soft Norm. This suggests that there is still room for improvement when it comes to capturing the appropriate distribution that represents real data.

These findings highlight the complexity of modeling the distribution of disagreements in a subjective task such as sexism identification.

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

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold_soft	0	3.1182	1	0.5472
SINAI_3	1	0.9030	0.6421	0.7960
CLassifiers_3	2	0.9027	0.6421	0.9754
CLassifiers_2	3	0.8698	0.6368	0.9823
CLassifiers_1	4	0.8172	0.6283	0.9672
CIC-SDS.KN_2	5	0.7960	0.6248	0.7770
CIC-SDS.KN_3	6	0.7555	0.6183	0.7620
AI-UPV_2	7	0.7343	0.6149	1.3607
CIC-SDS.KN_1	8	0.7200	0.6126	0.7846
IUEXIST_1	9	0.7115	0.6112	1.1537
Tlatlamiztli_1	10	0.6879	0.6074	1.0538
UMUTeam_1	11	0.6818	0.6064	0.8707
JPM_UNED_1	12	0.6779	0.6058	0.8023
AI-UPV_3	13	0.6772	0.6056	1.6400
Mario_1	14	0.6696	0.6044	1.9247
Mario_2	15	0.6629	0.6033	1.8536
Mario_3	16	0.6603	0.6029	1.9503
IUEXIST_2	17	0.6141	0.5955	1.8418
JPM_UNED_2	18	0.5972	0.5927	0.8852
AIT_FHSTP_3	19	0.5955	0.5924	0.9392
AIT_FHSTP_1	20	0.5648	0.5875	1.1491
AI-UPV_1	21	0.5448	0.5843	1.5543
DRIM_1	22	0.5433	0.5840	0.8932
UMUTeam_2	23	0.4969	0.5765	0.8100
SINAI_1	24	0.4863	0.5748	1.5759
Alex_P_UPB_1	25	0.3214	0.5482	1.1709
SMS_1	26	0.3142	0.5470	1.6650
SMS_2	27	0.3142	0.5470	1.6650
M&S_NLP_1	28	0.2990	0.5445	0.8496
JPM_UNED_3	29	0.2467	0.5361	2.2342
M&S_NLP_2	30	0.1802	0.5254	0.9724
IU-NLP-JeDi_3	31	0.1244	0.5163	1.0878
InsightX_2	32	−0.0357	0.4905	0.9122
Awakened_2	33	−0.1496	0.4721	0.8706
IU-NLP-JeDi_2	34	−0.1499	0.4720	0.9097
InsightX_1	35	−0.2351	0.4583	0.9145

(*continued*)

Table 2. (*continued*)

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
UMUTeam_3	36	−0.3460	0.4403	0.9318
Awakened_3	37	−0.4369	0.4257	0.9282
IU-Percival_2	38	−0.4435	0.4246	3.1682
IU-Percival_1	39	−0.4612	0.4217	3.1777
IU-Percival_3	40	−0.5491	0.4075	3.2560
CNLP-NITS-PP_2	41	−0.5775	0.4029	2.0155
Awakened_1	42	−0.5931	0.4004	0.9625
InsightX_3	43	−0.6356	0.3936	0.9402
iimasGIL_NLP_3	44	−1.9161	0.1867	1.8619
iimasGIL_NLP_2	45	−1.9438	0.1822	1.8151
iimasGIL_NLP_1	46	−2.1678	0.1460	2.2546
Majority_class	47	−2.3585	0.1152	4.6115
M&S_NLP_3	48	−2.4596	0.0989	3.1670
roh-neil_2	49	−2.8848	0.0302	1.5472
roh-neil_1	50	−2.8851	0.0301	7.3091
roh-neil_3	51	−2.8851	0.0301	6.7608
Minority_class	52	−3.0717	0	5.3572
I2C-Huelva-1_3	53	−4.0609	−0.1598	2.5776
I2C-Huelva-1_1	54	−4.1626	−0.1762	2.4439
I2C-Huelva-1_2	55	−4.3175	−0.2013	2.9025
SINAI_2	56	−5.6559	−0.4175	7.3080

We next analyze the performance of the top ten systems. As shown in Table 2, the top five systems achieve similar results in terms of ICM-Soft, with a difference of less than 2% percentage points in terms of ICM-Soft Norm. The difference among the top ten systems was also less than 4% percentage points.

Among the top ten systems, the top nine utilized multilingual learning models, while the tenth system used a monolingual Spanish model. The variations between the systems primarily stemmed from the use of data augmentation techniques, such as in the fourth and tenth ranked systems, as well as the incorporation of domain-specific Twitter models and ensembles.

Hard-Hard Evaluation. Table 3 presents the results for the Hard-hard evaluation. In this scenario, the annotations from the six annotators are combined into a single label using the majority vote, resulting in the loss of information about the different perspectives provided by each annotator. Out of the 67 systems submitted for this task, 65 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 for the ICM-Hard metric, ranging from 0.6575 (78.50% ICM-Hard Norm) to -0.5335 (2.59%). However, the impact of this variation is not as prominent when considering the F1 score, indicating that ICM-Hard penalizes systems that predominantly suggest only one class for all instances more severely. For example, systems like "shm2023_1" labeled 2063 out of 2076 instances as "NO", while "M&S_NLP_3" labeled 1594 out of 2076 instances as "YES", thus getting very poor results in the ICM metrics.

Furthermore, the comparison between the ICM-Hard and F1 scores, as reflected in the ranking, shows a similar distribution, particularly at the top of the table. A strong correlation between the two metrics has been observed. However, in contrast to the Soft-soft evaluation, the behavior of the best systems in the Hard-hard evaluation aligns more closely with the gold standard. This highlights the higher complexity of the soft scenario and the inherent differences between the soft and hard evaluation contexts.

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

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold_hard	0	0.9948	1	1
Mario_3	1	0.6575	0.7850	0.8109
Mario_1	2	0.6540	0.7828	0.8058
Mario_2	3	0.6120	0.7560	0.8029
roh-neil_1	4	0.5795	0.7353	0.7840
roh-neil_2	5	0.5795	0.7353	0.7840
CIC-SDS.KN_2	6	0.5715	0.7302	0.7775
CIC-SDS.KN_3	7	0.5647	0.7259	0.7721
SINAI_1	8	0.5584	0.7219	0.7804
SINAI_2	9	0.5543	0.7192	0.7719
roh-neil_3	10	0.5474	0.7149	0.7754
SINAI_3	11	0.5440	0.7127	0.7715
CLassifiers_2	12	0.5390	0.7095	0.7702
UniBo_1	13	0.5381	0.7089	0.7716
UniBo_2	14	0.5381	0.7089	0.7716
IUEXIST_2	15	0.5341	0.7064	0.7717
IUEXIST_1	16	0.5313	0.7046	0.7734
CIC-SDS.KN_1	17	0.5303	0.7040	0.7677
CLassifiers_3	18	0.5282	0.7026	0.7642

(*continued*)

Table 3. (*continued*)

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
JPM_UNED_3	19	0.5223	0.6989	0.7623
AI-UPV_3	20	0.5119	0.6922	0.7574
CLassifiers_1	21	0.5113	0.6918	0.7615
AI-UPV_2	22	0.5106	0.6914	0.7589
AIT_FHSTP_2	23	0.5086	0.6901	0.7571
UMUTeam_2	24	0.5083	0.6899	0.7604
I2C-Huelva-1_1	25	0.5075	0.6894	0.7611
I2C-Huelva-1_2	26	0.5075	0.6894	0.7611
I2C-Huelva-1_3	27	0.5075	0.6894	0.7611
JPM_UNED_1	28	0.5057	0.6883	0.7560
UMUTeam_1	29	0.5053	0.6880	0.7611
iimasGIL_NLP_3	30	0.5024	0.6862	0.7568
Tlatlamiztli_1	31	0.5013	0.6855	0.7535
MART_1	32	0.4937	0.6806	0.7587
JPM_UNED_2	33	0.4863	0.6759	0.7533
AIT_FHSTP_1	34	0.4850	0.6751	0.7550
AIT_FHSTP_3	35	0.4832	0.6739	0.7544
iimasGIL_NLP_1	36	0.4751	0.6688	0.7484
AI-UPV_1	37	0.4700	0.6655	0.7445
MART_2	38	0.4672	0.6637	0.7490
iimasGIL_NLP_2	39	0.4626	0.6608	0.7477
Alex_P_UPB_1	40	0.4021	0.6222	0.7302
M&S_NLP_1	41	0.3975	0.6193	0.7202
ZaRa-IU-NLP_2	42	0.3914	0.6154	0.7305
Awakened_2	43	0.3623	0.5969	0.7222
M&S_NLP_2	44	0.3057	0.5608	0.682
IU-Percival_1	45	0.3024	0.5587	0.6971
IU-Percival_2	46	0.2964	0.5549	0.6981
ZaRa-IU-NLP_1	47	0.2842	0.5471	0.6955
IU-NLP-JeDi_3	48	0.2753	0.5414	0.6909
InsightX_2	49	0.2689	0.5373	0.6883
IU-NLP-JeDi_1	50	0.2676	0.5365	0.6872
IU-Percival_3	51	0.2675	0.5365	0.6907
SMS_3	52	0.2469	0.5233	0.6717
InsightX_1	53	0.2458	0.5226	0.6809
SMS_1	54	0.2369	0.5170	0.6787

(*continued*)

Table 3. (*continued*)

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
SMS_2	55	0.2369	0.5170	0.6787
UMUTeam_3	56	0.1882	0.4859	0.6639
IU-NLP-JeDi_2	57	0.1851	0.4839	0.6485
Awakened_3	58	0.1457	0.4588	0.6400
CNLP-NITS-PP_1	59	0.1093	0.4356	0.6409
CNLP-NITS-PP_2	60	0.1093	0.4356	0.6409
IIIT SURAT_1	61	0.1042	0.4324	0.6355
Awakened_1	62	0.0723	0.4120	0.6322
InsightX_3	63	−0.0403	0.3403	0.4804
shm2023_2	64	−0.1470	0.2723	0.4638
KUCST_2	65	−0.3578	0.1379	0.4062
Majority_class	66	−0.4413	0.0847	0
shm2023_1	67	−0.4473	0.0809	0.0071
M&S_NLP_3	68	−0.5335	0.0259	0.5312
Minority_class	69	−0.5742	0	0.5698

As shown in Table 3, the performance of the top ten runs is more remarkable in the hard context evaluation, with a difference of 7% percentage points in terms of ICM-Hard Norm. This difference is primarily due to the good performance of the top three systems ("Mario" team) which outperform the 4th ranked system by 5%. The "Mario" team utilizes a cascade model consisting of two fine-tuned monolingual GPT-based models trained on EXIST 2023 data and on data from other related hate speech tasks. Interestingly, the efficiency of the "Mario" team's runs in the soft context is considerably lower, as they are ranked 14th, 15th, and 16th. Among the other top ten teams, all of them employ multilingual approaches, with some using Twitter-specific models and none utilizing data augmentation techniques. It is also worth noting the performance of the "CIS-SDK.KN" team, which achieves similar rankings in both evaluations (5th and 6th in the soft context, and 6th and 7th in hard context).

Hard-Soft Evaluation. The Hard-soft evaluation context aims to assess the impact of information loss when using the majority vote strategy. Although the results of the Soft-soft and hard-soft evaluations are not strictly comparable, they provide insights into the performance of hard systems compared to soft systems in a real-world context.

Due to length restrictions, a detailed discussion of the hard-soft evaluation is provided in the extended version of the overview [26]. Here, we highlight the main findings. The evaluation included 67 systems, 65 outperformed the majority class baseline, and all surpassed the minority class baseline. Notably, the loss of information in the hard-soft context was more significant than expected when compared to the Soft-soft context. This had a clear impact on the behaviour of the systems, as evidenced by the first-ranked system achieving 64.21% in terms of ICM-Soft in the Soft-soft evaluation context compared to the 57.25% achieved in the Hard-soft context. This phenomenon is also reflected in the performance of the "EXIST2023_oracle_most_voted" baseline, which considers the most voted label by the annotators, and achieves a score of 1.1977 in the hard-soft context and a score of 3.1182 in the Soft-soft context.

6.2 Task 2 Source Intention

The second task is a hierarchical multiclass classification problem where systems must determine if a tweet is sexist or not, and categorize the sexist tweets according to the source intention in: "JUDGEMENTAL", "REPORTED", and "DIRECT".

Soft-Soft Evaluation. Table 4 presents the results for the Soft-soft evaluation of Task 2. The table shows that 25 runs were submitted. Among them, 21 runs achieved better results compared to the majority class baseline (where all instances are labeled as "NO"). Furthermore, all of the submitted runs outperformed the minority class baseline (where all instances are labeled as "REPORTED"). The differences between the scores achieved by the gold standard and the worst-performing system, the non-informative minority class baseline, are higher compared to Task 1. This can be attributed to the hierarchical and multiclass nature of Task 2, which allows for a wider range of values in the quantification of information by the ICM-Soft metric. 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. Our initial analysis suggests that ICM-Soft is better at capturing the hierarchical nature of the task, as evidenced by a preliminary analysis of runs "JPM_UNED_2" and "SINAI_3". This observation is further supported by the fact that "JPM_UNED_2" utilizes a cascade model to determine whether a tweet is sexist or not sexist as a first step. However, further analysis is required to confirm this observation.

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

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold_soft	0	6.2057	1	0.9128
DRIM_1	1	−1.3443	0.8072	1.7833
AIT_FHSTP_2	2	−1.4350	0.8049	1.6486
JPM_UNED_2	3	−1.6750	0.7988	2.5549
AI-UPV_2	4	−1.6836	0.7985	2.1697
JPM_UNED_3	5	−1.6888	0.7984	2.5561
AI-UPV_3	6	−1.7691	0.7964	2.5424
AIT_FHSTP_3	7	−2.1619	0.7863	2.0897
SINAI_3	8	−2.2900	0.7831	1.6753
UMUTeam_3	9	−2.5405	0.7767	2.2271
UMUTeam_1	10	−2.5674	0.7760	1.8102
Alex_P_UPB_1	11	−3.1765	0.7604	3.2050
Awakened_2	12	−3.1954	0.7599	1.6668
iimasGIL_NLP_3	13	−3.5072	0.7520	1.8860
Mario_2	14	−3.5509	0.7509	3.0061
iimasGIL_NLP_1	15	−3.5570	0.7507	1.9067
iimasGIL_NLP_2	16	−3.6387	0.7486	1.9149
Awakened_1	17	−3.9152	0.7416	1.7741
UMUTeam_2	18	−4.0482	0.7382	4.0452
SINAI_1	19	−4.2437	0.7332	2.3710
Awakened_3	20	−4.3598	0.7302	1.8594
AI-UPV_1	21	−4.3632	0.7301	3.0172
Majority_class	22	−5.4460	0.7025	4.6233
roh-neil_1	23	−5.7590	0.6945	3.7519
roh-neil_2	24	−5.7592	0.6945	3.2828
SINAI_2	25	−10.9851	0.5610	4.6237
M&S_NLP_1	26	−12.5531	0.5210	7.5639
Minority_class	27	−32.9552	0	8.8517

The top-performing run achieved a score of −1.3443 in terms of ICM-Soft, which is quite similar to the scores obtained by the other top nine systems. The tenth system achieved a score of −2.5674, resulting in a difference of only 3.1% in terms of ICM-Soft normalized. The first system ("DRIM_1") proposed an ensemble model consisting of three monolingual models, each trained on one of the three EXIST tasks. These models were fine-tuned using the ICM-Soft measure and optimized with manual features and annotator distribution information to calculate similarities.

Regarding the other nine systems, most of them utilized multilingual models, with the exception of "JPM_UNED". The systems employed a combination of techniques such as data augmentation, transfer learning, and ensembles, as in the previous task. One notable difference among the top ten systems is the use of socio-demographic knowledge by the "JPM_UNED" team, where six models were trained for each population's cohort to calculate the final distribution of probabilities.

Finally, upon comparing the performance of the top ten systems in Task 1 and Task 2 in the Soft-soft context, we find that the "AI-UPV_2" and "SINAI_3" teams consistently excel in accurately identifying sexism content and its properties in both tasks.

Hard-Hard Evaluation. Table 5 presents the results of the Hard-hard evaluation for Task 2, where 33 systems were assessed against the hard gold standard. 28 runs outperform the majority class baseline (all instances labeled as "NO"), while all systems demonstrate superior performance compared to the minority class baseline (all instances labeled as "REPORTED"). Similar to the Soft-soft evaluation context, the discrepancies between the gold standard and the worst-performing system (minority class baseline) are higher in Task 2 than in Task 1. The correlation between ICM-Hard and F1 is generally high, with slight variations observed among the top-ranked systems, and higher variability towards the end of the table. This discrepancy can be attributed to the inability of F1 to capture the hierarchical nature of the task, unlike ICM-Hard, which penalizes misclassifications between different levels of the hierarchy more severely.

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

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold_hard	0	1.5378	1	1
Mario_2	1	0.4887	0.7764	0.5715
roh-neil_1	2	0.3883	0.7550	0.5480
roh-neil_2	3	0.3883	0.7550	0.5480
UniBo_2	4	0.2786	0.7316	0.5283
UniBo_1	5	0.2439	0.7243	0.5217
AIT_FHSTP_1	6	0.2229	0.7198	0.5029
AI-UPV_2	7	0.1951	0.7139	0.4962
AI-UPV_3	8	0.1870	0.7121	0.4993
JPM_UNED_2	9	0.1862	0.7120	0.5054
JPM_UNED_3	10	0.1806	0.7108	0.5092
JPM_UNED_1	11	0.1673	0.7079	0.5032
AIT_FHSTP_3	12	0.1662	0.7077	0.4911

(*continued*)

Table 5. (*continued*)

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
AIT_FHSTP_2	13	0.1475	0.7037	0.4759
UMUTeam_1	14	0.1409	0.7023	0.5013
AI-UPV_1	15	0.1217	0.6982	0.4897
Awakened_2	16	0.0088	0.6741	0.4606
UMUTeam_2	17	−0.0453	0.6626	0.4495
SINAI_2	18	−0.0496	0.6617	0.4924
SMS_1	19	−0.0892	0.6533	0.3654
SMS_3	20	−0.1226	0.6461	0.3504
UMUTeam_3	21	−0.1349	0.6435	0.4300
Alex_P_UPB_1	22	−0.1481	0.6407	0.4278
SMS_2	23	−0.2571	0.6175	0.3246
CNLP-NITS-PP_1	24	−0.3601	0.5955	0.3663
SINAI_1	25	−0.5959	0.5453	0.2562
Awakened_1	26	−0.7515	0.5121	0.2910
Awakened_3	27	−0.9048	0.4794	0.3087
KUCST_2	28	−0.9333	0.4734	0.2383
Majority_class	29	−0.9504	0.4697	0.1603
SINAI_3	30	−0.9646	0.4667	0.2544
iimasGIL_NLP_3	31	−0.9925	0.4608	0.2910
iimasGIL_NLP_1	32	−1.0631	0.4457	0.2505
iimasGIL_NLP_2	33	−1.0778	0.4426	0.2629
M&S_NLP_1	34	−2.9687	0.0396	0.0765
Minority_class	35	−3.1545	0	0.0280

Regarding the comparison between the top ten runs in terms of the ICM-Hard, the first system ("Mario_2") achieves a score of 0.4887 and the tenth system ("JPM_UNED_3") achieves a score of 0.1806 (only a 6.6% point difference). Unlike the previous task and context evaluation, where multilingual models were predominant, this task comprises an equal presence of multilingual and monolingual approaches among the top systems. Specifically, the first-ranked system is a monolingual GPT-based model that utilizes transfer learning and a cascade approach to identify and filter sexist tweets. The other top systems employ similar techniques, including transfer learning, data augmentation, and Twitter domain-specific models. Once again, the "JPM_UNED" team stands out as the only top-performing system that leverages the demographic features of the annotators, training six models, one for each cohort. It is interesting to note the comparison with Task 1 in the Hard-hard context, where the systems from the "roh-neil" and "Mario" teams are also among the top 5 in the ranking.

Hard-Soft Evaluation. The Hard-soft evaluation for Task 2 is discussed in the extended version of the overview [26]. However, we provide here a brief summary of the main findings. In this evaluation, 35 hard approaches were evaluated against the soft gold standard. Only 2 systems outperformed the majority class baseline, while all systems showed significant improvement over the minority class baseline. This behaviour is attributed to the fact that ICM-Soft penalizes errors in the minority classes to a greater extent, as they provide more informative signals than the majority classes. When comparing the results with the Soft-soft evaluation, similar trends were observed as in Task 1. The system that performed the best in the Soft-soft evaluation achieved an ICM-Soft Norm score of 80.72%, while the top-performing system in the hard-soft evaluation scored 71.08%.

6.3 Task 3 Sexism Categorization

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

Soft-Soft Evaluation. Table 6 displays the results of the Soft-soft evaluation for Task 3. A total of 23 runs were submitted, with 14 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"). This highlights the complexity of categorizing sexism in social networks. The comparison among the system scores and the gold standard further emphasizes this difficulty, with a notable difference of more than 10 points in ICM-Soft scores: 9.4686 for the gold standard and −2.3183 for the best system ("AI-UPV_3"). Additionally, the differences between the best and the worst systems are significantly higher than in any of the previous tasks. However, despite the complexity, the results of the ICM-Soft Norm metric indicate that the systems are still able to correctly capture relevant information concerning the different types of sexism.

The best system in this task utilizes an ensemble approach with two multilingual models that have been optimized with the number of annotators to calculate probability distributions, achieving an ICM-Soft score of −2.3183. The differences between the top ten systems in this task are higher compared to previous tasks, with a nearly 9% difference in terms of ICM-Soft Norm between the 1st and 10th system. In terms of the techniques employed by the top ten systems, the prevalent architecture is multilingual, often combined with data augmentation techniques, domain-specific models for Twitter, and ensembles. It is noteworthy that the "AI-UPV_2" and "SINAI_3" teams consistently perform well in all three Soft-soft evaluation tasks, securing a position in the top ten of the ranking.

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

Run	Rank	ICM-Soft	ICM-Soft Norm
Gold_soft	0	9.4686	1
AI-UPV_3	1	−2.3183	0.7879
AI-UPV_2	2	−2.5616	0.7835
AI-UPV_1	3	−3.3437	0.7695
DRIM_1	4	−3.6842	0.7633
Alex_P_UPB_1	5	−4.2139	0.7538
CLassifiers_1	6	−6.4072	0.7143
roh-neil_1	7	−6.6622	0.7098
roh-neil_2	8	−6.6622	0.7098
roh-neil_3	9	−6.6622	0.7098
SINAI_3	10	−7.1306	0.7013
iimasGIL_NLP_3	11	−7.7704	0.6898
iimasGIL_NLP_2	12	−7.8073	0.6892
iimasGIL_NLP_1	13	−7.8867	0.6877
M&S_NLP_1	14	−8.3574	0.6793
Majority_class	15	−8.7089	0.6729
Mario_1	16	−11.4241	0.6241
Mario_2	17	−11.4241	0.6241
Mario_3	18	−11.4241	0.6241
SINAI_2	19	−13.5493	0.5858
CLassifiers_2	20	−14.7828	0.5636
Awakened_1	21	−20.0399	0.4690
Awakened_2	22	−23.6389	0.4043
Awakened_3	23	−25.9233	0.3632
SINAI_1	24	−34.9362	0.201
Minority_class	25	−46.108	0

Hard-Hard Evaluation. In the Hard-hard evaluation context for the last task, a total of 30 systems were submitted. As shown in Table 7, 26 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"). Similar to Task 2, the discrepancies between the gold standard and the worst-performing system (minority class baseline) are higher in Task 3 due to its more complex nature of being hierarchical, multiclass, and multilabel.

The variation in results among different runs follows a similar distribution to that observed in Task 2, except for the last four systems which obtained substantially lower results due to that a high number of not sexist instances

have been incorrectly assigned to different sexist subclasses, resulting in a strong penalization by the ICM-Hard metric that considers the class hierarchy.

The correlation between the F1 and the ICM-Hard measure is not as strong as in Task 2. This can be attributed to the fact that the ICM-Hard measure takes into account the hierarchy and penalizes errors between hierarchy levels more severely.

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

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

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_test_gold_hard	0	2.1533	1	1
roh-neil_1	1	0.4433	0.6763	0.6296
roh-neil_2	2	0.4433	0.6763	0.6296
AIT_FHSTP_1	3	0.2366	0.6372	0.5842
UniBo_2	4	0.2263	0.6352	0.5909
UniBo_1	5	0.1776	0.6260	0.5850
Mario_3	6	0.1700	0.6246	0.5323
SINAI_2	7	0.1472	0.6203	0.5822
Mario_2	8	0.1228	0.6156	0.5145
Mario_1	9	0.0896	0.6094	0.5011
SINAI_3	10	0.0249	0.5971	0.5033
AI-UPV_1	11	−0.1862	0.5571	0.4732
Alex_P_UPB_1	12	−0.1948	0.5555	0.4817
AI-UPV_2	13	−0.2516	0.5448	0.4757
SINAI_1	14	−0.3020	0.5352	0.5306
Awakened_2	15	−0.4276	0.5115	0.4027
UMUTeam_3	16	−0.5121	0.4955	0.5130
AI-UPV_3	17	−0.5788	0.4828	0.4195
UMUTeam_1	18	−0.5963	0.4795	0.5108
iimasGIL_NLP_3	19	−0.6510	0.4692	0.4482
Awakened_3	20	−0.6731	0.4650	0.3794
iimasGIL_NLP_1	21	−0.6859	0.4626	0.4406
iimasGIL_NLP_2	22	−0.7786	0.4450	0.4255
CNLP-NITS-PP_1	23	−0.8412	0.4332	0.3199
Awakened_1	24	−0.9507	0.4124	0.3283
roh-neil_3	25	−0.9626	0.4102	0.3139
UMUTeam_2	26	−0.9727	0.4083	0.4630
EXIST2023_test_majority_class	27	−1.5984	0.2898	0.1069
KUCST_2	28	−1.7934	0.2529	0.2889
CLassifiers_2	29	−1.8664	0.2391	0.3047
CLassifiers_1	30	−1.8852	0.2355	0.3126
M&S_NLP_1	31	−2.1587	0.1838	0.0017
EXIST2023_test_minority_class	32	−3.1295	0	0.0288

As in the Soft-soft evaluation of Task 3, the differences among the top ten systems are higher than in other tasks, up to 8% points between the 1st ("roh-neil_1") and 10th ("SINAI_3") systems. In this scenario, the best system proposed a multilingual domain specific model trained on Twitter data, and uses parameter optimization using the Optuna framework. Roughly an equal number of systems utilized monolingual and multilingual approaches. The majority of these systems employed a combination of techniques, including data augmentation, domain-specific models, and transfer learning. Interestingly, the only team that exploit the demographic knowledge is the "SINAI" team. Upon analyzing the performance of the top ten hard approaches across all tasks, it is evident that the "Mario" and "roh-neil" teams consistently achieve favorable results, ranking within the top 10.

Hard-Soft Evaluation. Finally, for completeness we provide a brief summary of the main findings for the hard-soft evaluation, while a detailed description can be found in the extended overview paper [26]. In this evaluation, a total of 30 systems were evaluated, and none of them outperformed the majority class baseline, while all of them improved upon the minority class baseline. As mentioned in the hard-soft evaluation for Task 2, the ICM-Soft measure heavily penalizes errors in the minority classes, while errors in the majority class are not as heavily penalized. This behaviour is particularly significant in the hard-soft evaluation due to the automatic assignment of higher probabilities to the hard labels. When comparing the results with the Soft-soft evaluation, we observe that the best performance system in the Soft-soft evaluation achieved an ICM-Soft Norm score of 78.79%, while the top-performing system in the hard-soft evaluation scored 66.52%.

7 Conclusions

The objective of the EXIST challenge is to encourage research on the automated detection and modeling of sexism in online environments, with a specific focus on social networks. The EXIST Lab held in 2023 as part of CLEF garnered significant interest, attracting over 100 registered teams for participation. We received a total of 232 submissions. Participants adopted a wide range of approaches including data augmentation through automatic translation, data duplication and utilization of past EXIST editions' data, multilingual language models, Twitter-specific language models, and transfer learning techniques from domains like hate speech, toxicity, and sentiment analysis. Fortunately, most participants chose to leverage the multiple annotations available and embrace the learning with disagreements paradigm, rather than opting for a traditional approach and providing only hard labels as outputs.

In addition to obtaining new insights into the detection and categorization of sexist messages in social networks, the Lab has also contributed to raise awareness about the importance of addressing annotator disagreements and label bias by selecting annotators from diverse population groups. This is particularly true

for tasks where subjectivity and moral considerations come into play, and interpretations may vary across cultures and over time. Moreover, the EXIST lab has provided a valuable dataset that can be utilized for future experimentation and benchmarking purposes, further contributing to the advancement of research in this field.

For future editions of EXIST, we plan to extend our study to other languages, communication channels and media, as well as studying sexism in particular scenarios and population groups.

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

References

1. Amigó, E., Delgado, A.: Evaluating extreme hierarchical multi-label classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. vol. 1: Long Papers, pp. 5809–5819. ACL, Dublin, Ireland (2022)
2. Angel, J., Aroyehun, S., Gelbukh, A.: Multilingual Sexism Identification using contrastive learning. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
3. Asnani, H., Davis, A., Rajanala, A., Kübler, S.: Tlatlamiztli: fine-tuned RoBERTuito for sexism detection. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
4. Burgos, A., et al.: Violencias de género 2.0. (2014)
5. Buzzell, M., Dickinson, J., Singh, N., Kübler, S.: IU-NLP-JeDi: investigating sexism detection in English and Spanish. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
6. Böck, J., et al.: AIT_FHSTP at EXIST 2023 benchmark: sexism detection by transfer learning, sentiment and toxicity embeddings and hand-crafted features. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
7. Chaudhary, A., Kumar, R.: Sexism identification in social networks. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
8. Dhrodia, A.: Social media and the silencing effect: why misogyny online is a human rights issue (2017). https://www.newstatesman.com/culture/social-media/2017/11/social-media-and-silencing-effect-why-misogyny-online-human-rights-issue
9. Díaz, J.A.G., Pan, R., Valencia-Garcia, R.: UMUTeam at EXIST 2023: sexism identification and categorisation fine-tuning multilingual large language models. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
10. Erbani, J., Egyed-Zsigmond, E., Nurbakova, D., Portier, P.E.: When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media with pretrained transformers in a soft label

context. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
11. Gabel, E., Redman, H., Swanson, D., Kübler, S.: IU-Percival: linear models for sexism detection. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
12. Gerbner, G., Gross, L., Morgan, M., Signorielli, N.: Living with television: the dynamics of the cultivation process. Perspect. Media Effects **1986**, 17–40 (1986)
13. Hatekar, Y., Abdo, M., Khanna, S., Kübler, S.: IUEXIST: multilingual pre-trained language models for sexism detection on twitter in EXIST2023. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
14. Hovy, D., Prabhumoye, S.: Five sources of bias in natural language processing. Lang. Linguist. Compass **15**(8), e12432 (2021)
15. Jhakal, C., Singal, K., Suri, M., Chaudhary, D., Gorton, I., Kumar, B.: Detection of sexism on social media with multiple simple transformers. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
16. Kelkar, S., Ravi, S., Madasamy, A.K.: LSTM-attention architecture for online bilingual sexism detection. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
17. Kirk, H.R., Yin, W., Vidgen, B., Röttger, P.: SemEval-2023 task 10: explainable detection of online sexism. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval) (2023)
18. Koonireddy, R., Adel, N.: ROH_NEIL@EXIST2023: detecting sexism in tweets using multilingual language models. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
19. Mohammadi, H., Giachanou, A., Bagheri, A.: Towards robust online sexism detection: a multi-model approach with BERT, XLM-RoBERTa, and DistilBERT for EXIST 2023 Tasks. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
20. Muti, A., Mancini, E.: Enriching hate-tuned transformer-based embeddings with emotions for the categorization of sexism. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
21. Cordón, P., Jacinto Mata, V.P., Domínguez, J.L.: I2C at CLEF-2023 EXIST task: leveraging ensembling language models to detect multilingual sexism in social media. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
22. de Paula, A.F.M., Rizzi, G., Fersini, E., Spina, D.: AI-UPV at EXIST 2023 - sexism characterization using large language models under the learning with disagreement regime. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
23. Pedrosa-Marín, J., de Albornoz, J.C., Plaza, L.: Combining large language models with socio-demographic information for improving sexism detection in social media. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
24. Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9616–9625. IEEE Computer Society, Los Alamitos, CA, USA (2019)
25. Petrescu, A.: Leveraging MiniLMv2 pipelines for EXIST2023. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)

26. Plaza, L., et al.: 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)
27. Radler, G., Ersoy, B.I., Carpentieri, S.: CLassifiers at EXIST 2023: detecting sexism in spanish and english tweets with XLM-T. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
28. Regulagedda, R.M., Leech, Z., Kübler, S.: Specialized or generalized? Sexism detection for EXIST 2023 at CLEF. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
29. Rodrigues, F., Pereira, F.C.: Deep learning from crowds. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI 2018/IAAI 2018/EAAI 2018, AAAI Press (2018)
30. Rodríguez-Sánchez, F., et al.: Overview of EXIST 2021: sexism identification in social networks. Procesamiento del Lenguaje Nat. **67**, 195–207 (2021)
31. Rodríguez-Sánchez, F., et al.: Overview of EXIST 2022: sexism identification in social networks. Procesamiento del Lenguaje Nat. **69**, 229–240 (2022)
32. Roselli, D., Matthews, J., Talagala, N.: Managing bias in AI. In: Companion Proceedings of The 2019 World Wide Web Conference, pp. 539–544. WWW 2019, Association for Computing Machinery, New York, NY, USA (2019)
33. Sanchez-Urbina, A., Gómez-Adorno, H., Bel-Enguix, G., Rodríguez-Figueroa, V., Monge-Barrera, A.: iimasGIL_NLP@Exist2023: unveiling sexism on twitter with fine-tuned transformers. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
34. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 614–622. KDD 2008, Association for Computing Machinery, New York, NY, USA (2008)
35. Tian, L., Huang, N., Zhang, X.: Efficient multilingual sexism detection via large language models cascades. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
36. Uma, A., et al.: SemEval-2021 task 12: learning with disagreements. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 338–347. Association for Computational Linguistics, Online (2021)
37. Vallecillo-Rodríguez, M.E., del Arco, F.M.P., Ureña-López, L.A., Martín-Valdivia, M.T., Montejo-Ráez, A.: Integrating annotator information in transformer fine-tuning for sexism detection. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)
38. Vetagiri, A., Adhikary, P.K., Pakray, P., Das, A.: Leveraging GPT-2 for automated classification of online sexist content. In: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023)