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

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*
[2] *RMIT University, 3000 Melbourne, Australia*
[3] *Universidad Politécnica de Valencia (UPV), 46022 Valencia, Spain*

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 extended lab overview describes the tasks, the dataset, the participants' approaches, the evaluation methodology, and the results obtained.

Keywords

sexism detection, sexism categorization, learning with disagreement, data bias, EXIST

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.

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

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

[†] These authors contributed equally.

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. All forms of sexism, despite their differences, contribute to the perpetuation of gender inequality and restrict the full potential of women.

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. [1]. The Internet not only perpetuates these gender differences but also normalizes sexist attitudes, as noted by Burgos [2]. 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.

This issue is particularly worrying considering that a significant portion of Internet users, particularly teenagers and social media users, are exposed to this increasing wave of sexism. Previous studies [3] 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. Urgent attention is needed to comprehensively study and engage in a social debate surrounding this issue, leading to concrete actions, particularly in the realm of education. Moreover, social media platforms are not acting efficiently to remove or avoid sexist and hateful content.

Up to date, most work dealing with sexism in online media is focused on detecting misogyny or hatred towards women [4, 5, 6]. However, sexism does not always imply misogyny. Sexism may sound "friendly" as in (1). Sexism may sound "funny", as in sexist jokes or humour (2). Or sexism may sound "offensive" and "hateful", as in (3). However, even the most subtle forms of sexism can be as pernicious as the most violent ones and affect women in many facets of their lives, including domestic and parenting roles, career opportunities, sexual image and life expectations, to name a few.

(1) *Women must be loved and respected, always treat them like a fragile glass.*

(2) *You have to love women... just that... You will never understand them.*

(3) *Humiliate, expose and degrade yourself as the fucking bitch you are if you want a real man to give you attention.*

EXIST 2023 (http://nlp.uned.es/exist2023/) 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 [7, 8]. 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. In 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 consisting in determining 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 aim to undermine women from those that report sexist 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" [9], is the adoption of the "learning with disagreement" (LwD) paradigm [10] for the development of the dataset and for the evaluation of the systems. Our previous work showed that the perception 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 taken into account by the systems and the learning process becomes fairer.

In addition to adopting the LwD paradigm, we have made an effort to control bias in the annotations (see Section 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 (https://gab.com/), 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". Therefore, three tasks are proposed which are described below.

2.1. 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.

(4) **Sexist:** *Woman driving, be careful!.*

(5) **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. 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 follows:

- **Direct** sexist message. The intention was to write a message that is sexist by itself or incites to be sexist, as in:

  (6) *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:

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

- **Judgemental** message. The intention was judgmental, since the tweet describes sexist situations or behaviours with the aim of condemning them.

  (8) *21st century and we are still earning 25% less than men #Idonotrenounce.*

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

  (9) *#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.*

- **Role 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.

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

- **Objectification**. Objectification and physical stereotyping 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.

  (11) *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:

  (12) *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.

  (13) *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 [11]. 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. Moreover, we adopt the "learning with disagreement" paradigm and provide all labels provided by the different annotators to allow systems to learn from conflicting and subjective information.

3.1. Data Sampling

To gather the data, we follow the methodology used in previous EXIST editions [7, 8]. 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 (https://everydaysexism.com/); 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 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** [12]. 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. 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. The workers were 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 (https://www.prolific.co/). 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.

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. Guidelines for Annotators

Firstly, the annotators were informed that no identifying information would be obtained from them, and they were warned that they would be reading potentially violent messages that could be distressing. They were informed that they had the option to discontinue the task if they felt offended.

Secondly, the annotators were informed about the task, which involved reviewing 60 tweets individually. They were specifically instructed to read each tweet attentively in order to provide an answer to the following question:

1. Is the tweet sexist, in any form, or does it describe situations in which such discrimination occurs (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour)?

   a) **YES**, the tweet is sexist itself, describes a sexist situation or criticizes a sexist behavior. Examples of tweets in this category are:

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

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

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

   b) **NO**, the tweet does not prejudice, underestimate or discriminate women. Examples of tweets in this category are:

      (17) *"Where all the white women at?".*

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

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

      IMPORTANT NOTE: Remember that sexism may be expressed in different forms! It may sound friendly, funny, offensive, hateful or even violent. Sexism comprises any form of oppression or prejudice against women because of their sex. This discrimination can be based on different beliefs such as stereotypes, the belief that men are superior to women or irrational hatred of women (misogyny). Note that misogyny is a radical/hate-driven sexism.

   If the answer to the previous question was **YES**, then the two following questions were showed to the annotators and had to be answered:

2. What do you think the intention of the person that wrote the tweet is?

   a) **DIRECT**: the intention is to write a message that is sexist by itself, as in:

      (20) "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".

   b) **REPORTED**: the intention of the author is to report or describe a sexist situation or event suffered by a woman or women in first or third person, as in:

      (21) *"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".*

c) **JUDGEMENTAL:** the intention of the author is to be judgemental, since the tweet describes sexist situations or behaviors with the aim to condemning them.

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

3. What type of sexism does the tweet enclose?

a) Ideological discredit, denying inequality and flipped narrative (**IDEOLOGICAL-INEQUALITY**): this category includes all tweets or gabs that discredit the feminist movement in order to devalue, belittle and defame the struggle of women in all areas. Also included are messages that reject inequality between men and women, or present men as victims of gender-based oppression. Examples of tweets in this category are:

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

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

b) Role stereotyping and dominance (**STEREOTYPING-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. Examples of tweets in this category are:

(25) *"I feel like every time I flirt with a girl they start to imagine all the ways they can utilize me"*.

(26) *"Woman driving, be careful!"*.

c) Objectification and Physical stereotyping (**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 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. Examples of tweets in this category are:

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

(28) *"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."*.

d) Sexual violence (**SEXUAL-VIOLENCE**): this category includes messages where sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made. Examples of tweets in this category are:

(29) *"fuck that cunt, I would with my fist"*.

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

e) Misogyny and violence (**MISOGYNY-NON-SEXUAL-VIOLENCE**): this category includes expressions of hatred and violence towards women. Examples of tweets in this category are:

(31) *"Some woman are so toxic they don't even know they are draining everyone around them in poison. If you lack self awareness you won't even notice how toxic you really are".*

(32) *"Regardless, if you grab a woman by the throat, you're a massive cunt".*

IMPORTANT NOTE: Sexism may affect women in many facets of their lives, including domestic and parenting roles, career opportunities, sexual image and life expectations, to name a few.

3.4. Pre-aggregated Annotations: The Learning with Disagreement Paradigm

As stated in [10], 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. [10] have proposed the Learning With Disagreement paradigm (LwD), which consists mainly in 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 [13, 14, 15], 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.

3.5. EXIST 2023 Dataset statistics

This section provides a detailed description of the EXIST 2023 dataset, which serves as a valuable resource for analyzing and understanding sexism detection in social media. The dataset encompasses a wide range of tweets from various countries and includes annotations provided by a diverse group of annotators. The distribution of data within the EXIST 2023 dataset is presented in Table 1, highlighting the number of tweets, annotators, and countries represented in each subset (Training, Development, and Test).

Table 1
Number of tweets, annotators and countries of annotators per partition in the EXIST 2023 dataset

	Annotators	Countries	Tweets
Training	725	45	6,920
Development	113	23	1,038
Test	227	28	2,076

As shown in Table 1, the EXIST 2023 dataset was created by large and diverse group of 1,065 annotators, who contributed to the annotation process. This diverse pool of annotators ensures a wide range of perspectives and enhances the representativeness of the dataset. The dataset includes annotators from 45 countries in the Training subset, 23 countries in the Development subset, and 28 countries in the Test subset. This broad representation across countries allows for cross-cultural analysis and provides a comprehensive view of the global understanding sexism on social media.

In total, the EXIST 2023 dataset consists of 10,034 annotated tweets. The Training subset contains the majority of the tweets with 6,920 samples, 3660 in Spanish and 3260 in English. The Dev and Test subsets contain 1,038 and 2,076 tweets, respectively, distributed in 549 in Spanish and 489 in English for the Development set and 1098 in Spanish and 978 in English for the Test set.

The final distribution of labels for Task 1 is shown in Figure 1. The label "YES" indicates that the annotator considered the tweet to be sexist, while the label "NO" indicates that the annotator determined that the tweet was non-sexist. The charts in Figure 1 show the percentage of "YES" and "NO" labels in the three subsets. It can be seen that, in the three subsets, the distribution among sexist and not sexist tweet is well-balanced, being the "NO" label slightly more frequent than the "YES" one.

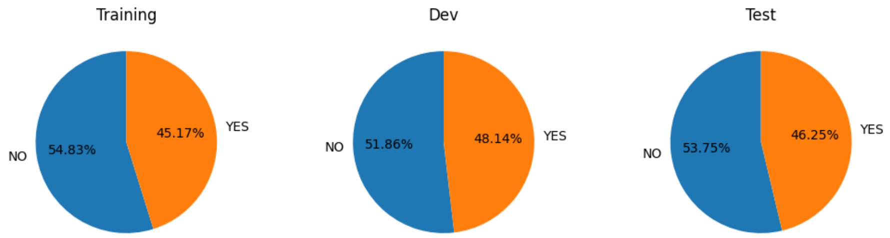

Figure 1: Distribution of labels by subset for Task 1 in the EXIST 2023 dataset

Similarly, the distribution of labels for Task 2 is shown in Figure 2. The *DIRECT* label indicates that the annotator considered the tweet to be inherently sexist in their content. These tweets displayed explicit or overt sexist language or sentiments. On the other hand, the *REPORTED* label indicates that the annotator considered that the tweet was describing specific instances of sexist events or behaviors experienced by women. These tweets shed light on real-life incidents and their impact. Lastly, the *JUDGEMENTAL* label means that the annotator determined that the tweet made judgments about sexist acts or situations. These tweets expressed opinions or evaluations of sexist occurrences.

The charts in Figure 2 show the percentage of "DIRECT" , "REPORTED" and "JUDGEMENTAL" labels in the three subsets. The distribution is homogeneous across the different datasets, being the "DIRECT" label the most frequent one (near 50%), followed by the "JUDGEMENTAL" label (around 28%), and finally the "REPORTED" label (around 25%).

Finally, the distribution of labels for Task 3 is shown in Figure 3. Again, the distribution of labels is quite homogenous across the three subsets, being the "STEREOTYPING-DOMINANCE" the most frequent label, followed by the "IDEOLOGICAL-INEQUALITY", "OBJECTIFICATION",

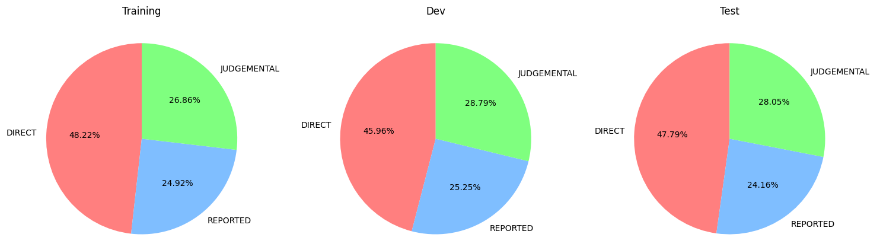

Figure 2: Distribution of labels by subset for Task 2 in the EXIST 2023 dataset

"MISOGYNY-NON-SEXUAL-VIOLENCE", and finally, "SEXUAL-VIOLENCE".

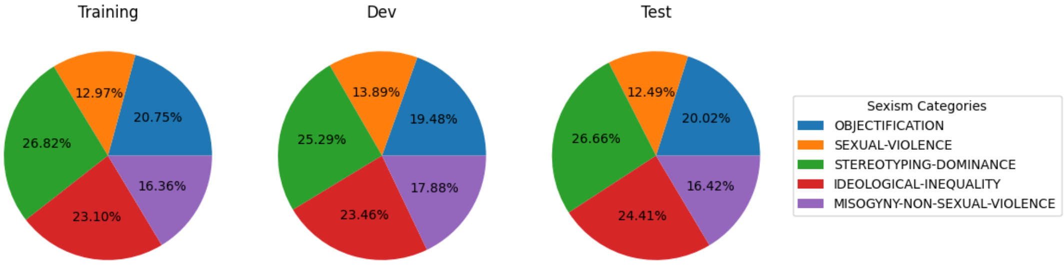

Figure 3: Distribution of labels by subset for Task 3 in the EXIST 2023 dataset

In order to provide a comprehensive analysis of the dataset, we present the statistics for the hard version of the dataset, i.e., that obtained by assigning to each tweet the majority vote among all annotators' labels following the methodology described in Section 6. However, as shown in Table 2, this approach results in a significant loss of information in cases where annotators do not reach a consensus. It is important to note that disagreements among annotators were expected because, to some extent, they are inherent to subjective tasks such as sexism detection.

4. Lab Setup and Participation

The EXIST 2023 Lab was conducted and overseen via the website http://nlp.uned.es/exist2023/. Participants had access to task details, dataset information, as well as the registration and participation process on this website.

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

Table 2
EXIST 2023 dataset hard labels statistics

	Training			Development			Test		
	ES	EN	Total	ES	EN	Total	ES	EN	Total
Task 1									
NO	1634	1733	3367	229	250	479	491	489	980
YES	1560	1137	2697	261	194	455	478	349	827
Task 2									
DIRECT	749	545	1294	117	87	204	244	152	396
REPORTED	265	194	459	40	35	75	78	45	123
JUDGEMENTAL	228	148	376	55	28	83	78	75	153
Task 3									
OBJECTIFICATION	611	492	1103	88	95	183	194	150	344
SEXUAL-VIOLENCE	359	316	675	74	49	123	99	97	196
STEREOTYPING-DOMINANCE	810	613	1423	136	105	241	257	187	444
IDEOLOGICAL-INEQUALITY	632	481	1113	117	95	212	223	161	384
MISOGYNY-NON-SEXUAL-VIOLENCE	552	304	856	90	68	158	166	109	275

participate in any of the three tasks and submit hard and/or soft outputs. Table 3 summarizes the participation in the different tasks and evaluation contexts.

Table 3
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 Section 5 and to the participant papers. 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 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.

5. Overview of Approaches

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

- Team **AIT_FHSTP** [16] utilized embeddings derived from a sentiment analysis and a toxicity analysis model, which were based on XLM_RoBERTa. These embeddings were then used to train a random forest model. Additionally, they explored the application of Principal Component Analysis. They obtained both soft and hard predictions for Tasks 1 and 2, while for Task 3, they only obtained a hard multi-label prediction.
- The **AI-UPV** team [17] utilized BERT-base-multilingual uncased and XLM-RoBERTa-base. They obtained both soft and hard predictions for all three tasks by employing ensembles to combine the probabilities from both models.
- The **Alex_P_UPB** team [18] utilized the MiniLMv2 model [19], a multilingual model based on deep self-attention distillation. The selection of the best parameters was performed manually, as the fine-tuning process involved using all the training data without any splits. They obtained both soft and hard predictions for all three tasks.
- Team **Awakened** employed BiLSTM models along with sentence embeddings built using "pysentimiento/bertweet-hate-speech" for English and "cardiffnlp/twitter-xlm-roberta-base-hate-spanish" for Spanish. They obtained both soft and hard predictions for all three tasks.
- Team **CIC-SDS.KN** [20] participated only in the first task. They fine-tuned a multilingual language model called Bernice specifically for Twitter data. They used the proportion of annotators with a label of YES as the label for training. To obtain hard labels, they applied a rule where the model's prediction had to be greater than 0.5 to predict YES, and NO otherwise.

- Team **CLassifiers** [21] participated in Tasks 1 and 3 using a multilingual approach. They employed XLMRoBERTa for hard classification in Task 3 and the CardiffNLP Twitter XLM-RoBERTa for the hard approach in Task 1. For soft labels, they utilized the CardiffNLP Twitter XLM-RoBERTa model. In Task 1, they utilized data augmentation with EXIST 2021. For Task 3, they employed a cascade model using the model from Task 1.
- Team **CNLP-NITS-PP** [22] utilized the GPT-2 model trained exclusively on the provided training dataset. They generated both soft and hard labels for Task 1, while only hard labels were provided for Tasks 2 and 3.
- Team **DRIM** [23] performed fine-tuning on three BERT models separately for Task 1, Task 2, and Task 3. They then froze the models and concatenated them together, incorporating manual features such as tweet length and the number of hashtags. On top of the concatenated representation, they added three dense layers. For generating labels, they considered the possible distributions, which were finite and had values in multiples of 1/6 (reflecting the number of annotators). They selected the output that was closest to the possible distributions, providing both hard and soft labels for all three tasks.
- Team **I2C-Huelva-1** [24] explored various approaches for Task 1. They utilized an ensemble of models including bert-large-uncased, bert-base-uncased, and distilbert. Additionally, they employed a fine-tuned model based on distilbert_sexism from the Hugging Face library, as well as a fine-tuned model based on bert-large-uncased.
- Team **IIIT-Surat** [25] exclusively participated in Task 1, utilizing a standard Bi-LSTM architecture to provide hard predictions.
- Team **InsightX** [26] solely participated in Task 1, delivering both hard and soft predictions. They employed RoBERTa and BERT models for English and Spanish, along with monolingual transformers such as BERTTweet for English and distill-BERT for Spanish.
- Team **iimasGIL_NLP** [27] participated in Tasks 1 and 3 providing both hard and soft labels. For the binary classification task (Task 1), they evaluated many linguistic patterns combined with bag-of-words and n-gram features as input in classical machine learning algorithms. For the intention and category of sexist messages categorization (Task 3), they fine-tuned transformers models in the English and Spanish datasets.
- Team **IUExist** [28] participated in Task 1, providing both hard and soft predictions. They fine-tuned and optimized various models from huggingface, including XLM-Roberta Base, XLM-Roberta Large, TWHIN Base, and TWHIN Large. Additionally, they utilized extended data from previous EXIST datasets. They trained an XGB Classifier using the predictions generated by the four mentioned models and evaluated its performance on the development dataset. They employed multilingual models and data augmentation techniques with EXIST 2021.
- Team **IU-NLP-JeDi** [29] participated in Task 1, providing both hard and soft predictions. They explored various approaches, including SVM with TF-IDF on bigrams and trigrams, as well as CNN and RNN models.
- Team **IU-Percival** [30] participated in Task 1, providing both hard and soft predictions. They employed a traditional machine learning approach using single-layer perceptrons and utilized 1-, 2-, and 3-grams as features.
- Team **JPM-UNED** [31] participated in Task 1 and Task 2, providing both hard and soft predictions. They utilized data augmentation techniques by translating tweets from one

language to another and trained different models for each cohort based on the gender and age of the annotators. The best-performing models employed were "PlanTL-GOB-ES/roberta-base-bne" for Spanish and "distilbert-base-uncased" for English.

- Team **KUCST** participated in all three tasks, providing only hard labels. They trained a gradient boosting classifier for Task 1 and Task 2, and employed a separate classifier for each label in Task 3. Their feature set included Bag of Words, TF-IDF weighted words, POS tags, n-grams, morphological features, and BERT representations.
- Team **M&S NLP** [32] participated in Task 1 and Task 2, providing both hard and soft labels. In Task 3, they provided only soft labels. To address imbalanced classes, they employed text augmentation techniques. Their approach involved the use of multiple transformers, including BERT, XLM-RoBERTa, and DistilBERT.
- Team **Mario** [33] participated in all three tasks, providing both hard and soft labels. They utilized the "open-source GPT-NeoX" model for English and the "BERTIN-GPT-J-6B" model for Spanish. Their approach involved organizing the LLMs in cascades. One model was fine-tuned with in-domain training data for all three tasks, while the other hate-speech boosted model was sequentially fine-tuned on various hate speech datasets, including an open-sourced hate-speech tweets dataset from the Hugging Face library in the respective target language (English or Spanish). Finally, both models were fine-tuned with task-specific training data.
- Team **MART** participated solely in Task 1, providing hard labels. They employed various pre-processing techniques, such as TF-IDF vectorization, emoji/abbreviation replacement, accent removal, tokenization, stop words removal, and lemmatization. In addition, they utilized a standard BERT model for their approach.
- Team **roh-neil** [34] participated in the three tasks, providing both hard and soft labels. They utilized the "xlm-roberta-large-twitter" model from HuggingFace and employed Optuna for parameter optimization. Additionally, they incorporated data augmentation techniques by leveraging data from previous editions of EXIST.
- Team **shm2023** participated exclusively in Task 1, providing hard labels. They employed the "bert-lapse" embedded model to embed the tweets, and then utilized a random forest algorithm to predict the results based on the embedded text.
- Team **SINAI** [35] participated in the three tasks, providing hard and soft labels. They utilized the mDeBERTa base model and employed data augmentation techniques by replicating instances. Additionally, they incorporated categorical information of the annotators, such as gender and age, during the training process. They experimented with different strategies to incorporate this information, such as concatenating the transformer output and the categorical features, applying a multilayer perceptron on the categorical features, or using attention-based summation of the transformer outputs and categorical features, among others.
- Team **SMS** [36] participated in Tasks 1 and 2, providing hard and soft labels for Task 1, and hard labels for Task 2. They utilized 200-dimensional GloVe Twitter word embeddings for English texts and 300-dimensional FastText word embeddings for Spanish texts, along with an LSTM layer. For the calculation of the soft score, they aimed to replicate human annotator performance by using different models and aggregating the probabilities

predicted by these models. The ensemble of models included Multinomial Naive Bayes, Linear Support Vector Classifier, Multi-Layer Perceptron, XGBoost, LSTM using GloVe embeddings, and LSTM using FastText embeddings. The probabilities predicted by these models were combined to derive the soft score.

- Team **Tlatlamiztli** [37] participated in Task 1, providing both hard and soft labels. Their approach involves preprocessing the data using regex, spaCy, spaCymoji, and wordsegment libraries. They further perform translation to Spanish using the Googletrans library. They then utilize a RoBERTuito model that has been fine-tuned on the EXIST 2021 dataset. This multilingual model allows for the classification of hate speech in both English and Spanish.
- Team **UMU Team** [38] participated in all three tasks. They provided both hard and soft labels for Tasks 1 and 2, and only hard labels for Task 3. Their approach involved using a combination of different Large Language Models (LLMs) with Label Functions (LFs) and training them on multilingual data. They also experimented with the use of sentence embeddings extracted from MarIA and mDeBERTa models.
- Team **UniBo** [39] participated in all three tasks, providing hard labels. They employed a translation approach to convert the Spanish data into English using the Google API. Additionally, they enriched their feature set by incorporating Emotions retrieved using EmoBerta, which were added as features to the RobertaHate embeddings.
- Team **ZaRa-IU-NLP** [40] participated in Task 1 providing only hard labels. They test two different models: XML-RoBERTa and RoBERTa+ BETO.

6. Evaluation Methodology and Metrics

As in SemEval 2021, we have carried out a "**hard evaluation**" and a "**soft evaluation**". In the hard evaluation paradigm, both the system and the gold standard consist of one or more labels assigned to each instance in the dataset. In contrast, the soft evaluation 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.

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 [41]), 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 +∞ and -∞, respectively. To normalize the results, we used the gold standard score as the upper bound and minority class baseline (see Section 4) 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 [41]. 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:

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

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 [41] 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.

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. Note that, 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.

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 [41].

$$\begin{aligned} IC\left(\bigcup_{i=1}^{n}\{\langle c_i, v_i\rangle\}\right) &= IC(\langle c_1, v_1\rangle) + IC\left(\bigcup_{i=2}^{n}\{\langle c_i, v_i\rangle\}\right) \\ &- IC\left(\bigcup_{i=2}^{n}\{\langle \mathrm{lca}(c_1, c_i), min(v_1, v_i)\rangle\}\right) \end{aligned} \tag{33}$$

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

Within that definition, ICM accepts multi-class, multi-label, hierarchical and soft classification problems. Hard labels or large soft values assigned to infrequent classes produce high information quantity. The class specificity affects both the successes and the failures. The more the system and the gold standard differ to each other, the more their labelling conjunction is unlikely, the more the joint information quantity IC(S,G) is high, and the more the resulting score is low.

For example, if the system output and the gold standard are equal then IC(S)=IC(G)=IC(S,G) and the score is proportional to the gold standard information content (ICM=IC(G)). If the system does not label any class (empty system output) then IC(S)=0 and IC(S,G)=IC(G). Therefore, the score is negative and proportional to the gold standard information quantity (ICM=-IC(G)).

7. Results

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

7.1. 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 Section 6, we report three sets of evaluation results.

7.1.1. Soft-Soft Evaluation

Table 4 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 4, 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. Once again, this scores are due to the fact that ICM-Soft highly penalize errors in the minority class, which is really remarkable in a binary classification task. Also, 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 4: 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
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

Continued on next page

Table 4 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
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 4, 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.

7.1.2. Hard-Hard Evaluation

Table 5 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 5: 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
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

Continued on next page

Table 5 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
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	0.6820
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
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

Continued on next page

Table 5 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
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 5, 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).

7.1.3. 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.

The evaluation included 67 systems, 65 outperformed the majority class baseline (all instances labeled as "NO"), and all surpassed the minority class baseline (all instances labeled as "YES"). 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_vote" 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.

Also interesting is the difference found between the first system, with a score of 0.4719 in ICM-Soft, and the oracle most voted baseline, where a large difference can be seen. This difference is, in general, very pronounced in all the results obtained by all participants, specially in those at the end of the table, in this evaluation context. This fact, again, suggests that the errors made by the systems in the minority class are strongly penalised by the ICM-Soft metric, and that these errors are especially identifiable in the hard-soft context.

Table 6: Systems' results for Task 1 in the hard-soft evaluation

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2023_test_gold_soft	0	3.1182	1
EXIST2023_oracle_most_voted	1	1.1977	0.6897
Mario_3	2	0.4719	0.5725
Mario_1	3	0.4507	0.5691
Mario_2	4	0.3634	0.555
roh-neil_1	5	0.3111	0.5465
roh-neil_2	6	0.3111	0.5465
CIC-SDS.KN_2	7	0.2784	0.5412
SINAI_2	8	0.2678	0.5395
SINAI_1	9	0.264	0.5389
CIC-SDS.KN_3	10	0.2606	0.5383
SINAI_3	11	0.2513	0.5368
roh-neil_3	12	0.2497	0.5366
UniBo_1	13	0.2292	0.5333
UniBo_2	14	0.2292	0.5333
IUEXIST_1	15	0.215	0.531
IUEXIST_2	16	0.2109	0.5303
JPM_UNED_3	17	0.2041	0.5292
CIC-SDS.KN_1	18	0.197	0.5281
JPM_UNED_1	19	0.1685	0.5235
CLassifiers_3	20	0.1652	0.5229
UMUTeam_2	21	0.1614	0.5223
AI-UPV_2	22	0.1578	0.5217
UMUTeam_1	23	0.1578	0.5217
CLassifiers_2	24	0.1563	0.5215
Tlatlamiztli_1	25	0.1512	0.5207
AI-UPV_3	26	0.1453	0.5197
AIT_FHSTP_2	27	0.1411	0.519
I2C-Huelva-1_1	28	0.1253	0.5165
I2C-Huelva-1_2	29	0.1253	0.5165
I2C-Huelva-1_3	30	0.1253	0.5165
CLassifiers_1	31	0.1199	0.5156
JPM_UNED_2	32	0.1075	0.5136
AIT_FHSTP_1	33	0.1014	0.5126
iimasGIL_NLP_3	34	0.1	0.5124
AIT_FHSTP_3	35	0.0932	0.5113
MART_1	36	0.0868	0.5103
iimasGIL_NLP_1	37	0.059	0.5058
AI-UPV_1	38	0.0478	0.504

Continued on next page

Table 6 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm
MART_2	39	0.0333	0.5016
iimasGIL_NLP_2	40	-0.0255	0.4921
Alex_P_UPB_1	41	-0.096	0.4807
M&S_NLP_1	42	-0.1419	0.4733
ZaRa-IU-NLP_2	43	-0.1485	0.4723
Awakened_2	44	-0.2577	0.4546
M&S_NLP_2	45	-0.3266	0.4435
IU-Percival_2	46	-0.4435	0.4246
ZaRa-IU-NLP_1	47	-0.4449	0.4244
IU-Percival_1	48	-0.4612	0.4217
InsightX_2	49	-0.4813	0.4185
IU-NLP-JeDi_3	50	-0.5071	0.4143
IU-NLP-JeDi_1	51	-0.5097	0.4139
IU-Percival_3	52	-0.5491	0.4075
SMS_3	53	-0.5638	0.4052
InsightX_1	54	-0.5887	0.4011
SMS_1	55	-0.6017	0.399
SMS_2	56	-0.6017	0.399
IU-NLP-JeDi_2	57	-0.7054	0.3823
UMUTeam_3	58	-0.7329	0.3778
Awakened_3	59	-0.8536	0.3583
IIIT SURAT_1	60	-0.9432	0.3439
CNLP-NITS-PP_1	61	-0.9509	0.3426
CNLP-NITS-PP_2	62	-0.9509	0.3426
Awakened_1	63	-1.1382	0.3124
InsightX_3	64	-1.2069	0.3013
shm2023_2	65	-1.6131	0.2356
KUCST_2	66	-1.8431	0.1985
EXIST2023_test_majority_class	67	-2.3585	0.1152
shm2023_1	68	-2.3774	0.1122
M&S_NLP_3	69	-2.805	0.0431
EXIST2023_test_minority_class	70	-3.0717	0

Regarding the analysis of the top ten systems, the first five positions are exactly the same as in the hard-hard evaluation context, while the next four represent the same systems but in different positions. The only system that has changed in the top ten ranking is SINAI_3, ranked in the 11th position in the hard-hard and ranked 10th in hard-soft. In fact, this behaviour is observed in nearly all the systems in the ranking, suggesting a high correlation between both contexts and metrics, with small variations mainly due to the strong weight given to probabilities in the hard-soft case (classes are assigned with probability 1.0).

7.2. Task 2 Source Intention

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

7.2.1. Soft-Soft evaluation

Table 7 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 7: 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

Continued on next page

Table 7 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
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.

7.2.2. Hard-Hard Evaluation

Table 8 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 8: 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
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

Continued on next page

Table 8 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
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.

7.2.3. Hard-Soft Evaluation

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. This observation gains further significance in this context for two reasons: the increasing number of classes and the absolute probabilities assigned to hard labels (hard labels are interpreted in this context as soft labels with probability 1.0). Also significant is that the difference between the gold standard and participants' systems is even more pronounced than in Task 1. This can be attributed to the multiclass and hierarchical nature of the task, which presents additional complexities.

Finally, 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%.

Table 9: Systems' results for Task 2 in the hard-soft evaluation

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2023_test_gold_soft	0	6.2057	1
EXIST2023_oracle_most_voted	1	-2.3974	0.7803
UMUTeam_2	2	-5.12	0.7108
UMUTeam_3	3	-5.3093	0.706
EXIST2023_test_majority_class	4	-5.446	0.7025
UMUTeam_1	5	-5.5369	0.7001
SMS_2	6	-5.627	0.6978
Mario_2	7	-5.8157	0.693
SMS_3	8	-5.9579	0.6894
KUCST_2	9	-5.9955	0.6884
SMS_1	10	-6.0774	0.6863
AIT_FHSTP_2	11	-6.3494	0.6794
roh-neil_1	12	-6.522	0.675
roh-neil_2	13	-6.522	0.675
AIT_FHSTP_3	14	-6.735	0.6696
AIT_FHSTP_1	15	-6.8143	0.6675
UniBo_2	16	-6.8552	0.6665
AI-UPV_2	17	-7.0678	0.6611
Awakened_2	18	-7.1322	0.6594
UniBo_1	19	-7.2258	0.657
CNLP-NITS-PP_1	20	-7.2467	0.6565
AI-UPV_3	21	-7.2669	0.656
JPM_UNED_2	22	-7.3346	0.6542
JPM_UNED_1	23	-7.5078	0.6498
JPM_UNED_3	24	-7.5205	0.6495
Awakened_1	25	-7.7378	0.6439
AI-UPV_1	26	-7.7615	0.6433
Alex_P_UPB_1	27	-8.8468	0.6156
iimasGIL_NLP_1	28	-10.1498	0.5824
iimasGIL_NLP_2	29	-10.6025	0.5708
iimasGIL_NLP_3	30	-10.7368	0.5674
SINAI_2	31	-10.9983	0.5607
SINAI_1	32	-11.0357	0.5597
Awakened_3	33	-11.9799	0.5356
SINAI_3	34	-15.0466	0.4573
M&S_NLP_1	35	-23.8919	0.2314
EXIST2023_test_minority_class	36	-32.9552	0

In the context of the hard-soft evaluation, the discussion of the top ten systems reveals several interesting findings. Firstly, the best-performing system is based on a monolingual model for Spanish, highlighting the effectiveness of language-specific approaches in contrast with previous evaluations. Interestingly, when comparing the results between the first-ranked system, excluding the majority vote baseline, and the tenth-ranked system, the differences are minimal. The discrepancy is less than 1 point in the ICM-Soft metric and 4 percentage points in the normalized metric, suggesting a similar performance. Moreover, the remaining systems are a mix of monolingual and multilingual models, employing ensemble techniques and data augmentation. Surprisingly, two teams utilizing deep learning (LSTM), and traditional ML (gradient boosting), have achieved top 10 results. A preliminary analysis suggests that these types of techniques tend to prioritize the majority class, which, when correctly classified, contributes to overall higher accuracy.

These conclusions, however, should be considered preliminary, and further analysis within this context is necessary to provide a comprehensive comparison between the hard and soft evaluation contexts.

7.3. 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 Section 2.

7.3.1. Soft-Soft Evaluation

Table 10 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.

Table 10: 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

Continued on next page

Table 10 – continued from previous page

Run	Rank	ICM-Soft	ICM-Soft Norm
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	0.2010
Minority_class	25	−46.1080	0

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.

7.3.2. Hard-Hard Evaluation

In the hard-hard evaluation context for the last task, a total of 30 systems were submitted. As shown in Table 11, 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 11: 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

Continued on next page

Table 11 – continued from previous page

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
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.

7.3.3. Hard-Soft Evaluation

Finally, we provide a brief summary of the main findings for the hard-soft evaluation. 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, and specially significant when the number of classes augmented. Also, the multi-label nature of the task amplifies the significance of penalizing errors in the minority class even further. In the context of sexism categorization, where multiple labels can be present simultaneously, correctly identifying and classifying instances belonging to the minority classes becomes increasingly challenging. Therefore, the importance of addressing the issue of penalization in the context of the minority classes cannot be overstated, as it directly affects the accuracy and fairness of the system’s predictions. Finally, 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%. This difference is even larger than that found in Tasks 1 and 2, once again highlighting the complex nature of the task.

Table 12: Systems' results for Task 3 in the hard-soft evaluation

Run	Rank	ICM-Soft	ICM-Soft Norm
EXIST2023_test_gold_soft	0	9.4686	1
EXIST2023_oracle_most_voted	1	-8.3816	0.6788
EXIST2023_test_majority_class	2	-8.7089	0.6729
Mario_1	3	-9.1398	0.6652
M&S_NLP_1	4	-9.504	0.6586
Mario_2	5	-9.6735	0.6556
Mario_3	6	-10.2297	0.6456
AI-UPV_3	7	-10.9218	0.6331
CNLP-NITS-PP_1	8	-11.3206	0.6259
Alex_P_UPB_1	9	-11.7145	0.6188
AI-UPV_1	10	-11.8571	0.6163
Awakened_2	11	-11.8918	0.6157
SINAI_3	12	-12.1399	0.6112
roh-neil_1	13	-12.195	0.6102
roh-neil_2	14	-12.195	0.6102
AI-UPV_2	15	-12.2553	0.6091
AIT_FHSTP_1	16	-13.6923	0.5833
UniBo_2	17	-14.2263	0.5737
Awakened_3	18	-14.3296	0.5718
iimasGIL_NLP_2	19	-14.3422	0.5716
iimasGIL_NLP_3	20	-14.6869	0.5654
iimasGIL_NLP_1	21	-14.9609	0.5604
Awakened_1	22	-15.2776	0.5547
UniBo_1	23	-15.4834	0.551
KUCST_2	24	-22.9796	0.4162
roh-neil_3	25	-24.3698	0.3911
SINAI_2	26	-27.2984	0.3384
UMUTeam_3	27	-34.5038	0.2088
SINAI_1	28	-34.9858	0.2001
UMUTeam_1	29	-35.0505	0.199
CLassifiers_2	30	-36.0875	0.1803
UMUTeam_2	31	-37.3056	0.1584
CLassifiers_1	32	-37.5046	0.1548
EXIST2023_test_minority_class	33	-46.108	0

In the hard-soft evaluation, the best-performing system is a monolingual GPT-based model. The differences between the first-ranked system, excluding the baselines, and the 10th-ranked system are more noticeable compared to previous tasks, reaching up more than 5 percentage points in terms of the normalized ICM-Soft metric. The remaining systems in the top 10 primarily

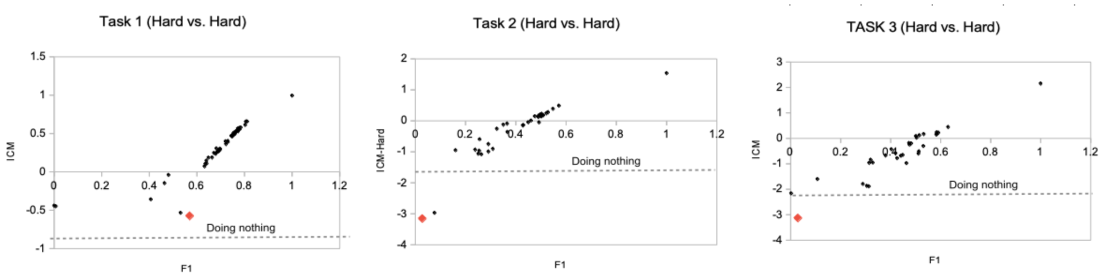

Figure 4: Correspondence between F1 and the ICM metrics across tasks

utilize multilingual systems in combination with ensembles, data augmentation techniques, and domain-specific models. Unlike what was found in Task 2, more traditional machine learning systems have not secured positions within the top 10. This observation may be attributed to the increased complexity of the task, stemming from the larger number of classes and their hierarchical structure.

Furthermore, it is important to note that these results are preliminary, and a more comprehensive comparative analysis between the soft evaluation context and the hard evaluation context should be conducted to draw more robust conclusions from the hard-soft evaluation.

7.4. The Information Contrast Model Analysis

In this campaign, ICM is applied over both hard (discrete labels) and soft outputs and gold standards. Figure 4 shows the correlation between ICM (hard output and hard gold standard) and F1. In these charts, the dashed line represents the score for the empty system output (ICM=-I(G)). According to the graphs, in the hard-hard evaluation, all systems contribute to a certain extent regarding the empty output. Generally, there is a high correlation when ICM is positive, although in Task 3, there are not enough samples to notice it. The reason is that ICM penalises failures in infrequent classes (which have more information quantity) to a greater extent than F1. Particularly in Task 1 and Task 2, an interesting point is that ICM discriminates "returning always the minority class" (red point) from the rest of outputs to a greater extent than F1.

In Figure 5, we compare the ICM scores when considering the hard output (discrete class) versus soft outputs. The gold standard in both cases is soft. As the graphs show, both scores differ substantially. In general, considering hard outputs (vertical axis) often leads to a decrease ICM, especially when considering more than two classes. This means that hard outputs are missing out on a lot of information when they restrict themselves to outputting one or a few classes in hard mode. Once again, the dashed line in the ICM on soft outputs represents "doing nothing" (ICM=-I(G)). Here, it can be observed that as the task becomes more complex, there are more systems that contributes with useful information (Task 3).

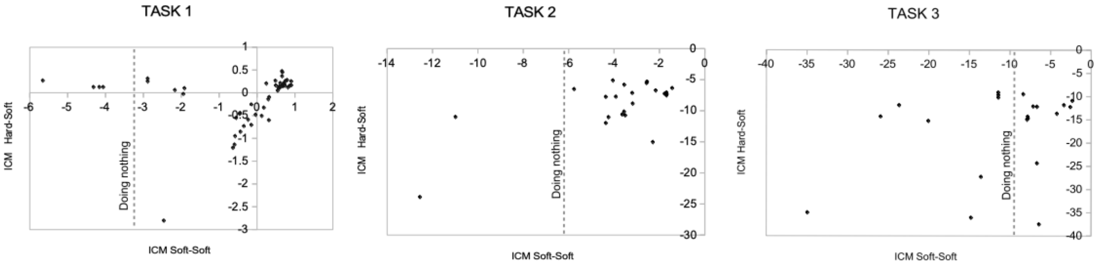

Figure 5: Correspondence of ICM between hard and soft outputs. The figures show that considering hard outputs substantially decreases the ICM scores, suggesting that relevant information is missed when discretizing the system outputs

8. 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 such as hate speech, toxicity, and sentiment analysis. Fortunately, most participants chose to leverage the multiple annotations available and embrace the LwD 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] A. Dhrodia, Social media and the silencing effect: Why misogyny online is a human rights issue, 2017. URL: https://www.newstatesman.com/culture/social-media/2017/11/social-media-and-silencing-effect-why-misogyny-online-human-rights-issue.

[2] A. Burgos, T. Donoso-Vázquez, E. López, E. Cabañó, Y. Martínez, A. Martín, N. Prado-Soto, M. Rubio, A. Velasco-Martínez, R. Vila, Violencias de género 2.0, E-litterae, 2014.

[3] G. Gerbner, L. Gross, M. Morgan, N. Signorielli, Living with television: The dynamics of the cultivation process, Perspectives on media effects 1986 (1986) 17–40.

[4] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2018, pp. 57–64.

[5] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (2020) 102360.

[6] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An expert annotated dataset for the detection of online misogyny, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1336–1350.

[7] 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.

[8] 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.

[9] 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), Association for Computational Linguistics, 2023.

[10] 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.

[11] D. Roselli, J. Matthews, N. Talagala, Managing bias in ai, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 539–544.

[12] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Language and Linguistics Compass 15 (2021) e12432.

[13] V. S. Sheng, F. Provost, P. G. Ipeirotis, 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, KDD ’08, Association for Computing Machinery, New York, NY, USA, 2008, p. 614–622.

[14] F. Rodrigues, F. C. Pereira, 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’18/IAAI’18/EAAI’18, AAAI Press, 2018.

[15] J. Peterson, R. Battleday, T. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 9616–9625.

[16] J. Böck, M. Schütz, N. Q. S. Daria Liakhovets, A. Babic, D. Slijepcevic, M. Zeppelzauer, A. Schindler, 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.

[17] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, 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.

[18] A. Petrescu, Leveraging MiniLMv2 Pipelines for EXIST2023, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[19] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 5776–5788.

[20] J. Angel, S. Aroyehun, A. Gelbukh, Multilingual Sexism Identification using contrastive learning, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[21] G. Radler, B. I. Ersoy, S. Carpentieri, 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.

[22] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, Leveraging GPT-2 for Automated Classification of Online Sexist Content, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[23] J. Erbani, E. Egyed-Zsigmond, D. Nurbakova, P.-E. Portier, 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.

[24] V. P. Pablo Cordón, Jacinto Mata, J. L. Domínguez, 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.

[25] A. Chaudhary, R. Kumar, Sexism Identification In Social Networks, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[26] C. Jhakal, K. Singal, M. Suri, D. Chaudhary, I. Gorton, B. Kumar, Detection of Sexism on Social Media with Multiple Simple Transformers, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[27] A. Sanchez-Urbina, H. Gómez-Adorno, G. Bel-Enguix, V. Rodríguez-Figueroa, A. Monge-

Barrera, 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.

[28] Y. Hatekar, M. Abdo, S. Khanna, S. Kübler, 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.

[29] M. Buzzell, J. Dickinson, N. Singh, S. Kübler, IU-NLP-JeDi: Investigating Sexism Detection in English and Spanish, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[30] E. Gabel, H. Redman, D. Swanson, S. Kübler, IU-Percival: Linear Models for Sexism Detection, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[31] J. Pedrosa-Marín, J. C. de Albornoz, L. Plaza, 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.

[32] H. Mohammadi, A. Giachanou, A. Bagheri, 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.

[33] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via Large Language Models Cascades, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[34] R. Koonireddy, N. Adel, 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.

[35] M. E. Vallecillo-Rodríguez, F. M. P. del Arco, L. A. Ureña-López, M. T. Martín-Valdivia, A. Montejo-Ráez, Integrating Annotator Information in Transformer Fine-tuning for Sexism Detection, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[36] S. Kelkar, S. Ravi, A. K. Madasamy, LSTM-Attention Architecture for Online Bilingual Sexism Detection, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[37] H. Asnani, A. Davis, A. Rajanala, S. Kübler, Tlatlamiztli: Fine-Tuned RoBERTuito for Sexism Detection, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[38] J. A. G. Díaz, R. Pan, R. Valencia-Garcia, 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.

[39] A. Muti, E. Mancini, 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.

[40] R. M. Regulagedda, Z. Leech, S. Kübler, Specialized or Generalized? Sexism Detection for EXIST 2023 at CLEF, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[41] 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.