AI-UPV at EXIST 2023 – Sexism Characterization Using Large Language Models Under The Learning with Disagreements Regime

Notebook for the EXIST Lab at CLEF 2023

Angel Felipe Magnossão de Paula[1,2], Giulia Rizzi[1,3,*], Elisabetta Fersini[3] and Damiano Spina[2]

[1]*Universitat Politècnica de València, Valencia, Spain*

[2]*RMIT University, Melbourne, Australia*

[3]*University of Milano-Bicocca, Milan, Italy*

Abstract

With the increasing influence of social media platforms, it has become crucial to develop automated systems capable of detecting instances of sexism and other disrespectful and hateful behaviors to promote a more inclusive and respectful online environment. Nevertheless, these tasks are considerably challenging considering different hate categories and the author's intentions, especially under the learning with disagreements regime. This paper describes AI-UPV team's participation in the EXIST (sEXism Identification in Social neTworks) Lab at CLEF 2023 [1, 2]. The proposed approach aims at addressing the task of sexism identification and characterization under the learning with disagreements paradigm by training directly from the data with disagreements, without using any aggregated label. Yet, performances considering both soft and hard evaluations are reported. The proposed system uses large language models (i.e., mBERT and XLM-RoBERTa) and ensemble strategies for sexism identification and classification in English and Spanish. In particular, our system is articulated in three different pipelines. The ensemble approach outperformed the individual large language models obtaining the best performances both adopting a soft and a hard label evaluation. This work describes the participation in all the three EXIST tasks, considering a soft evaluation, it obtained fourth place in Task 2 at EXIST and first place in Task 3, with the highest ICM-Soft of -2.32 and a normalized ICM-Soft of 0.79. The source code of our approaches is publicly available at https://github.com/AngelFelipeMP/Sexism-LLM-Learning-With-Disagreement.

Keywords
Sexism Characterization, Learning with Disagreements, Large Language Models, Ensemble

1. Introduction

Sexism in online content represents a significant challenge in maintaining inclusive and fair digital environments. In addition, cultural differences [3], misinformation [4], and hateful behavior

[5] aggravate it. Many women from all around the world have reported abuses, discrimination, and other sexist situations in real life on social media. Social networks are also contributing to the spread of sexism and other disrespectful and hateful behaviors. While social platforms such as Twitter are constantly developing new ways to identify and eliminate hateful content, they face numerous challenges when dealing with the massive amount of data generated by users. In this context, automated tools are adopted not only to assist in detecting and alerting against sexist behaviors, but also in estimating how frequently sexist and abusive phenomena are faced on social media platforms, what forms of sexism are more diffused, and how those harmful messages are spread.

Traditional sexism detection systems often rely on predefined labels and fixed perspectives, which may fail to capture the complexity and subjectivity inherent in sexist statements. Detecting and mitigating sexism remains a challenging task due to the subjective nature of its assessment. Perspectivism, on the other hand, provides a viable approach for improving detection by incorporating multiple opinions and viewpoints.

An important contribution in this field, that aims at tackling the problem of sexism identification under the paradigm of learning with disagreements has been given by the *EXIST 2023: sEXism Identification in Social neTworks* [1, 2].

In this paper, we address all the tasks proposed in the challenge, which cover different granularities and perspectives of sexism classification. In particular, we proposed a Large Language Model (LLM) ensemble-based strategy to predict both soft levels that reflect the set of labels provided by the annotators and hard labels derived from the aggregation of different annotators' perspectives.

The paper is organized as follows. An overview of the state of the art is provided in Section 2 focusing both on Sexism Detection and Perspectivism. Details about the shared task and the related datasets are reported in Section 3 and Section 4 respectively. In Section 5 the proposed approach is detailed focusing on the adopted prediction models, the proposed ensemble approach, and the post-processing operations. In Section 6, the results achieved by the proposed models are reported. Finally, conclusions and future research directions are summarized in Section 7.

2. Related Work

Sexism, defined as prejudice, stereotyping, or discrimination based on gender, is a pervasive issue affecting individuals across various contexts. The advent of digital platforms has amplified the reach and impact of sexist content, necessitating the development of automated systems to detect and counteract sexism. Researchers have proposed several approaches to address this issue, ranging from rule-based methods [6] to more advanced machine-learning techniques to address the problem of sexism detection mostly from a linguistic perspective [7, 8], with only a few attempts to tackle the problem from a visual or multimodal point of view [9, 10].

The majority of state-of-the-art approaches leveraged machine learning techniques, particularly relying on pre-trained LLM.

However, traditional approaches to sexism detection often overlook the complexity and subjectivity inherent in describing sexist behavior.

Sexism can in fact be further categorized in different forms according to the intention of the author, or the type of sexism. In the EXIST 2021 and 2022 challenges [11, 12] five classes of sexism have been detected: "Ideological and inequality", "Stereotyping and dominance", "Objectification", "Sexual violence" and "Misogyny and non-sexual violence". Similarly at SemEval 2023 – Task 10 – Explainable Detection of Online Sexism (EDOS) [13], provided a new taxonomy for the more explainable classification of sexism in three hierarchical levels – binary sexism detection, category of sexism, and fine-grained vector of sexism. As emerged from the leaderboard, top systems use multiple models or ensembles, and many top systems apply further pre-training and multi-task learning [7, 14, 15].

Despite the recent advances in employing attention mechanisms and other deep learning approaches, the detection of sexism [8] and hate speech [16] in general is still considered a major challenge, especially, when dealing with social media text written in low-resource languages, such as Arabic [17].

Multilingual LLM, especially if encapsulated in Ensemble methods, have been proven to be a robust solution to address this task [18, 7]. A considerable contribution in the field is represented by challenges' tasks like AMI IberEval, Evalita and EXIST. AMI IberEval 2018 [19] proposed two shared tasks mainly focused on tackling the problem of misogyny on Twitter, in three different languages, namely English, Italian, and Spanish.

Despite the problem of sexism is mostly addressed from a textual perspective, a growing field of research is tackling the problem from visual or multimodal point of view. Sexist content can in fact be represented also in images or in a multimodal form. From a visual perspective, most of the available investigations relate to offensive, non-compliant, or pornographic content detection [20, 21, 22], while from a multimodal point of view, the main contribution is given by the *Hateful Meme challenge: Detecting Hate Speech in multimodal Memes* [23], where one of the targets of hateful memes were women and by *SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI)*[24] that focused on misogyny recognition in memes.

Finally, many research papers give emphasis to an open issue in sexism detection: the presence of bias that could affect the real performance of the models. In this research area, bias analysis is still in its infancy with only a few contributions addressing the problem considering textual data and sexism identification [25, 26, 27, 9].

While the aforementioned studies have made substantial contributions to detecting sexism, they frequently approach the problem as an objective concept without considering taking into account the complexities and differing perspectives around it. Sexism may, in fact, take several forms depending on cultural, societal, and individual factors.

Recognizing the importance of addressing *perspectivism*, some researchers have investigated methods to include a more nuanced understanding of the tasks in detection systems [28]. However, only a few of them investigated perspectivism in sexism detection. For example, Kalra and Zubiaga [29] explore the errors made by classification models and discusses the difficulty in automatically classifying sexism due to the subjectivity of the labels and the complexity of natural language used in social media.

The attention to perspectivism is also reflected in the challenges with an increasing number of tasks that consider the level of agreement of the annotators in the form of soft labels. For instance, EXIST 2023 [1, 2], addresses sexism identification at different granularities and perspectives under the learning with disagreements paradigm.

3. Task Description

EXIST aims to detect sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviors (EXIST 2021 [11], EXIST 2022 [12]). The main task of the new edition of EXIST is to develop models able to capture sexism in all its forms, while considering the perspective of the learning with disagreements paradigm.

The EXIST Lab at CLEF 2023 [1, 2] is articulated in three different tasks that address sexism identification at different granularities and perspectives:

TASK 1: Sexism Identification. The first task addresses sexism identification as a binary problem, requiring the systems to classify if a tweet contains sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior) or not.

TASK 2: Source Intention. The second task focuses on classification of the message in accordance with the author's intentions (distinguishing between "Direct", "Reported", and "Judgemental"), which sheds light on the part social networks play in the creation and spread of sexist messages.

TASK 3: Sexism Categorization. The third task aims to categorize tweet into different types of sexism (i.e., "Ideological and Inequality", "Stereotyping and Dominance", "Objectification", "Sexual Violence", and "Misogyny and Non-sexual Violence") that reflect different focus of sexism attitudes ranging from domestic and parenting roles, career opportunities to sexual image.

The challenge also faces the sexism identification from the perspective of the learning with disagreements paradigm introducing two evaluations approaches for each task:

- **Soft Evaluation**. This is the evaluation under the learning with disagreements paradigm. Systems performances are evaluated through a soft-soft evaluation that compares the probabilities assigned by the system with the probabilities assigned by the set of human annotators. In this case, the evaluation measure is ICM-Soft [1, 2], while additional evaluation measures such as Cross Entropy are also reported.
- **Hard Evaluation**. Systems performances are also evaluated on the hard label derived from the different annotators' labels, through a probabilistic threshold computed for each task. The adopted evaluation measure is the original ICM [30]. F1 Score is also reported.

Additionally, for systems that provide a hard output a hard-soft evaluation is evaluated, comparing the categories assigned by the system with the probabilities assigned to each category in the ground truth with ICM-Soft.

4. Dataset

The EXIST 2023 dataset incorporates multiple types of sexist expressions, including descriptive or reported assertions where the sexist message is a description of sexist behavior.

In particular, the dataset is composed of more than 10,000 tweets both in English and Spanish,

divided into a test set (2,076 tweets), a development set (1,038 tweets), and a training set (6,920 tweets).

For each sample, the following attributes are provided in a JSON format:

- **id_EXIST**: a unique identifier for the tweet.
- **lang**: the language of the text ("en" or "es").
- **tweet**: the text of the tweet.
- **number_annotators**: the number of persons that have annotated the tweet.
- **annotators**: a unique identifier for each of the annotators.
- **gender_annotators**: the gender of the different annotators ("F" or "M", for female and male respectively).
- **age_annotators**: the age group of the different annotators (grouped in "18–22", "23–45", or "46+").
- **labels_task1**: a set of labels (one for each of the annotators) that indicate if the tweet contains sexist expressions or refers to sexist behaviors or not ("YES" or "NO").
- **labels_task2**: a set of labels (one for each of the annotators) recording the intention of the person who wrote the tweet ("DIRECT", "REPORTED", "JUDGEMENTAL", "–", and "UNKNOWN").
- **labels_task3**: a set of arrays of labels (one array for each of the annotators) indicating the type or types of sexism that are found in the tweet ("IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE", "MISOGYNY-NON-SEXUAL-VIOLENCE", "–", and "UNKNOWN").
- **split**: subset within the dataset the tweet belongs to ("TRAIN", "DEV", "TEST" + "EN"/"ES").

It is important to note that since hard labels for Tasks 2 and 3 are assigned only if the tweet has been labeled as sexist (label "YES" for Task 1), the label "–" is assigned to not sexist tweets in Tasks 2 and 3, while the label "UNKNOWN" is assigned to those tweets where the annotators did not provide a label. The test set includes only the following attributes: "id_EXIST", "lang", "tweet", and "split".

5. Proposed Approach

In this Section, the proposed approach to address all the tasks is described.

Inspired by the state-of-the-art results, we propose an ensemble method to combine two transformer-based models: namely mBERT and XLM-RoBERTa. The proposed approach focuses on soft label predictions, while the hard labels are directly derived by selecting the most probable ones.

BERT Base Multilingual. BERT Base Multilingual [31], also called "mBERT", is a widely used language model that has made substantial progress in the field of natural language processing. mBERT performs well in tasks like text classification, named entity identification, and question answering thanks to its bidirectional context. Additionally, mBERT has been pre-trained on a

large corpus of text from various languages, enabling it to capture language-specific patterns and nuances.

mBERT has been used by researchers to successfully identify and categorize hate speech across a variety of areas and languages [32, 33]. Furthermore, mBERT is an effective tool for identifying hate speech and creating safer online settings because of its contextual awareness and capacity to identify subtle linguistic patterns.

XLM-RoBERTa. Cross-lingual Language Model-RoBERTa [34], also called "XLM-RoBERTa", is a state-of-the-art language model that combines the benefits of cross-lingual pretraining and RoBERTa. The usage of parallel data from various languages during pre-training enables it to effectively transfer information between languages and therefore to comprehend and produce text in a variety of languages. XLM-RoBERTa was adopted on a variety of natural language processing tasks, including text categorization, named entity identification, and machine translation. XLM-RoBERTa is widely adopted by researchers in multilingual situations because of its robustness in managing multilingual data and its proficiency in cross-lingual transfer. Moreover, researchers have used XLM-RoBERTa to identify hate speech in multilingual settings, a task for which the model's capacity to handle several languages is crucial. It has been demonstrated that XLM-RoBERTa can detect hate speech in a variety of languages, including English, Spanish, and Arabic [35, 36, 37]. It can generalize well and identify hate speech in many language situations thanks to its cross-linguistic transfer abilities.

The proposed approach is articulated in the following steps that have been repeated for each EXIST task:

- **Hyperparameters Selection.** The aim of the first phase is to identify the most suitable number of fine-tuning epochs for our two LLMs. The two selected models have been implemented using the hugging face framework [38] and fine-tuned in the EXIST training dataset with 1 up to 10 epochs. Each model has then been evaluated on the EXIST validation dataset.
- **Model Training.** The training phase has the main goal of learning directly with soft labels. The parameter identified through the previous phase has been used to fine-tune the final models, on the whole dataset obtained via the union of the training and validation dataset.
- **Model Predictions.** The predictions related to the soft labels, both from the fine-tuned single models and from their aggregation given by an ensemble model, have been computed. First, we obtain the predicted probabilities for each model separately on the EXIST test set. Then, we combined the probabilities given by the mBERT and XLM-RoBERTa models by estimating the mean of their predicted probabilities.
- **Prediction Adjustment.** To ensure that the predicted probabilities are compliant with respect to the number of annotators, we performed the following operation: given the probability distribution of a model, we selected the most similar distribution according to cosine similarity with the feasible distributions. This operation allowed adjusting the obtained discrete probability distribution to the nearest ones that match the number of annotators, while ensuring a stochastic distribution for Tasks 1 and 2. This ensures that

the final prediction is consistent with the number of annotators and avoids any potential errors that may arise from the usage of raw predictions.

The hard labels can be derived directly from the predicted soft labels by selecting the most probable label. The proposed framework has been instantiated according to the following three different approaches, depicted in Figure 1:
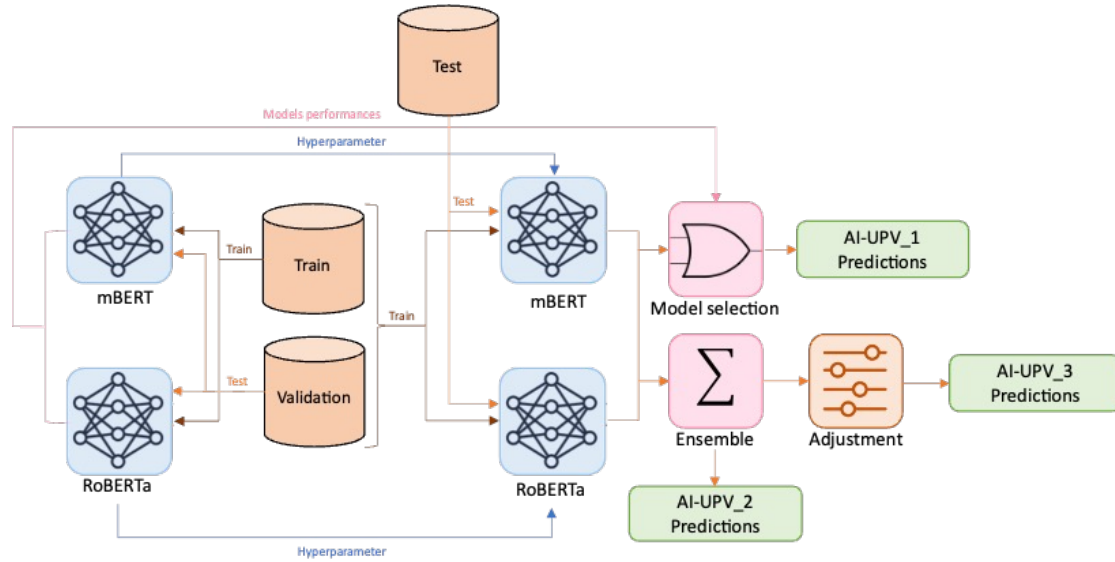

Figure 1: Schematic representation of the proposed pipeline.

AI-UPV_1: Model Selection. The first submission refers to the predictions returned by the best-performing model. In particular, both mBERT and XLM-RoBERTa are trained on EXIST training data and evaluated on the EXIST validation data in order to identify, not only the most suitable number of fine-tuning epochs (as described above) but also to identify the best performing model for each of the EXIST tasks. The predictions obtained from the selected model (i.e., BERT or XLM-RoBERTa) trained on the whole EXIST dataset (obtained by the union of the training and validation datasets) are reported as AI-UPV_1.

AI-UPV_2: Ensemble Model. The second submission refers to the predictions returned by a naive ensemble model. In particular, the final label is determined through a mean aggregation function that considers both fine-tuned models (i.e., mBERT and XLM-RoBERTa).

AI-UPV_3: Adjusted Ensemble Model. The third submission refers to the adjusted ensemble model predictions. Once the hyperparameters have been selected and the models have been trained, the ensemble computes the average probabilities considering the predictions given by mBERT and XLM-RoBERTa to then adjust such estimations according to the number of evaluators.

6. Results and Discussion

In this section, the results obtained through our participation in the challenge are reported and compared with the baselines provided by the organizers. Regarding the **Baselines**, three main approaches have been considered:

- **Majority Class**: non-informative baseline that classifies all instances according to the majority class.
- **Minority Class**: non-informative baseline that classifies all instances using the minority class.
- **Gold**: since the ICM measure is unbounded, a baseline based on an oracle that perfectly predicts the ground truth is considered to provide the best possible reference.

We report in the following tables the results obtained by the proposed approaches and the baselines, distinguishing between the tasks and the languages considered in the challenge. Table 1 shows the results for EXIST Task 1 on all (English and Spanish) instances; the best-performing model (i.e., AI-UPV_2) achieved seventh place in the global ranking considering a soft evaluation.

Table 1
Results of the proposed approaches for Task 1 on all test instances (English + Spanish).

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	3.12	1	0.55	0.99	1	1
EXIST2023_majority_class	−2.36	0.12	4.61	−0.44	0.08	0
EXIST2023_minority_class	−3.07	0	5.36	−0.57	0	0.57
AI-UPV_1 (Model Selection)	0.54	0.58	1.55	0.47	0.67	0.74
AI-UPV_2 (Ensemble)	**0.73**	**0.61**	**1.36**	**0.51**	**0.69**	**0.76**
AI-UPV_3 (Adj. Ensemble)	0.68	**0.61**	1.64	**0.51**	**0.69**	**0.76**

Analogously Tables 2 and 3 present the results for Task 1 on Spanish and English instances, respectively. In this case, we can highlight that the proposed approaches perform in an analogous way in both languages, considering that no strong performance variations between them can be observed.

Table 2
Results of the proposed approaches for Task 1 on ES (Spanish) instances.

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	3.12	1	0.52	1.00	1	1
EXIST2023_majority_class	−2.54	0.01	4.96	−0.49	0.01	0
EXIST2023_minority_class	−2.57	0	5.01	−0.51	0	0.61
AI-UPV_1 (Model Selection)	0.60	0.56	1.52	0.46	0.64	0.76
AI-UPV_2 (Ensemble)	**0.74**	**0.58**	**1.33**	0.49	0.66	0.77
AI-UPV_3 (Adj. Ensemble)	0.70	**0.58**	1.61	**0.51**	**0.67**	**0.78**

Considering the differences of the ICM-Soft – which is the official evaluation measure – it is possible to conclude that: (i) all the proposed models outperformed the proposed baselines, both

Table 3
Results of the proposed approaches for Task 1 on EN (English) instances.

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	3.11	1	0.58	0.98	1	1
EXIST2023_majority_class	−2.20	0.23	4.22	−0.40	0.16	0
EXIST2023_minority_class	−3.82	0	5.75	−0.66	0	0.53
AI-UPV_1 (Model Selection)	0.42	0.61	1.60	0.47	0.69	0.72
AI-UPV_2 (Ensemble)	**0.67**	**0.65**	**1.40**	**0.52**	**0.72**	**0.74**
AI-UPV_3 (Adj. Ensemble)	0.59	0.64	1.67	0.50	0.71	0.73

considering all samples and the language-dependent splits, (ii) the inclusion of an ensemble approach introduces a significant improvement in performances, and (iii) the adjustment operation slightly penalizes the ensemble performances.

We report in Tables 4, 5 and 6, the results related to the Exist Task 2, distinguishing between the results on the overall dataset and single languages. Regarding the overall dataset, the proposed AI-UPV_2 achieved the fourth place in the global ranking considering the soft evaluation.

Table 4
Results of the proposed approaches for Task 2 on all test instances (English + Spanish).

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	6.21	1	0.91	1.54	1	1
EXIST2023_majority_class	−5.45	0.70	4.62	−0.95	0.47	0.16
EXIST2023_minority_class	−32.96	0	8.85	−3.15	0	0.03
AI-UPV_1 (Model Selection)	−4.36	0.73	3.02	0.12	0.70	0.49
AI-UPV_2 (Ensemble)	**−1.68**	**0.80**	**2.17**	**0.20**	**0.71**	**0.50**
AI-UPV_3 (Adj. Ensemble)	−1.77	**0.80**	2.54	0.19	**0.71**	**0.50**

Table 5
Results of the proposed approaches for Task 2 on ES (Spanish) instances

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	6.24	1	0.89	1.60	1	1
EXIST2023_majority_class	−5.67	0.66	4.97	−1.04	0.42	0.15
EXIST2023_minority_class	−28.71	0	8.76	−2.94	0	0.03
AI-UPV_1 (Model Selection)	−4.01	0.71	3.04	0.13	0.68	**0.51**
AI-UPV_2 (Ensemble)	**−1.61**	**0.78**	**2.16**	**0.18**	**0.69**	**0.51**
AI-UPV_3 (Adj. Ensemble)	−1.66	0.77	2.53	0.16	0.68	**0.51**

It is important to note that, although Task 2 is more challenging than task 1, the proposed approach performs in a very competitive way when comparing the obtained ICM-Soft-Norm and ICM-Hard Norm with the corresponding values obtained in Task 1 (See Table 1). This suggests that by deepening the labeling taxonomy, the model is prone to make suitable predictions at lower granularity levels. This means that the intention of sexism can be considered as simpler task than distinguishing what is sexist from what is not. These results are also confirmed by

Table 6

Results of the proposed approaches for Task 2 on EN (English) instances.

	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	6.12	1	0.94	1.44	1	1
EXIST2023_majority_class	−5.20	0.75	4.23	−0.85	0.53	0.17
EXIST2023_minority_class	−39.49	0	8.96	−3.47	0	0.02
AI-UPV_1 (Model Selection)	−5.09	0.75	3.00	0.09	0.72	0.45
AI-UPV_2 (Ensemble)	**−1.91**	**0.82**	**2.18**	**0.19**	**0.75**	**0.48**
AI-UPV_3 (Adj. Ensemble)	−2.06	0.82	2.56	0.19	0.74	0.47

the performance on single languages, where no substantial differences are observed between English and Spanish (Tables 6 and 5 respectively).

Table 7

Results of the proposed approaches for Task 3 on all test instances (English + Spanish).

	ICM-Soft	ICM-Soft Norm	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	9.47	1	2.15	1	1
EXIST2023_majority_class	−8.71	0.67	−15.98	0.29	0.11
EXIST2023_minority_class	−46.11	0	−31.30	0	0.03
AI-UPV_1 (Model Selection)	−3.34	0.77	**−0.19**	**0.56**	0.47
AI-UPV_2 (Ensemble)	−2.56	0.78	−0.25	0.54	**0.48**
AI-UPV_3 (Adj. Ensemble)	**−2.32**	**0.79**	−0.58	0.48	0.42

Table 7 finally shows the results for the EXIST Task 3 on the entire dataset, where the main goal was to distinguish between different types of sexism. In this case, the best-performing model, i.e. AI-UPV_3, ranked first in the global ranking considering the soft evaluation, achieving an ICM-Soft Norm value equal to 0.79. Differently from what has been observed for the previous tasks, the prediction adjustment significantly improves the ensemble performances related to the soft evaluation. In this case, the task is even more complex than the previous ones because it is based on a fine-grained labeling schema that also depends on the previous levels of annotation. By comparing the current ICM-Soft value achieved by the AU-UPV_3 approach (0.79) with the corresponding one reported in Table 4 (0.80), the capabilities of the model at the lower level of the hierarchy are still maintained. This confirms the ability of the model at addressing fine-grained predictions. Analogously results can be grasped by looking at Tables 8 and 9 that focus on Spanish and English instances, respectively.

7. Conclusion

In this work, we proposed a large language models-based ensemble strategy to address the task of sexism identification under the paradigm of learning with disagreements.
The achieved results highlight how the adoption of ensembles can significantly improve the results obtained by the individual models. We developed a simple average ensemble strategy, however, the investigation of more complex ensemble strategies with the proposed system

Table 8
Results of the proposed approaches for Task 3 on ES (Spanish) instances.

	ICM-Soft	ICM-Soft Norm	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_gold	9.61	1	2.24	1	1
EXIST2023_majority_class	−9.03	0.66	−1.73	0.29	0.10
EXIST2023_minority_class	−45.43	0	−3.32	0	0.03
AI-UPV_1 (Model Selection)	−3.65	0.76	**−0.23**	**0.56**	**0.48**
AI-UPV_2 (Ensemble)	−2.74	**0.78**	−0.31	0.54	**0.48**
AI-UPV_3 (Adj. Ensemble)	**−2.58**	**0.78**	−0.65	0.48	0.42

Table 9
Results of the proposed approaches for Task 3 on EN (English) instances.

Run	ICM-Soft	ICM-Soft Norm	ICM-Hard	ICM-Hard Norm	F1
EXIST2023_test_gold_soft	9.13	1	2.04	1	1
EXIST2023_test_majority_class	−8.21	0.69	−1.46	0.30	0.11
EXIST2023_test_minority_class	−46.95	0	−2.93	0	0.03
AI-UPV_1 (Model Selection)	−3.02	0.78	**−0.15**	**0.56**	**0.46**
AI-UPV_2 (Ensemble)	−2.39	0.79	−0.20	0.55	**0.46**
AI-UPV_3 (Adj. Ensemble)	**−2.05**	**0.80**	−0.50	0.49	0.41

could be conducted in future works. Additionally, the proposed ensemble model can easily be extended to include also other Large Language Models that have shown promising performance on similar tasks (e.g., LlaMA [39] or ELECTRA [40]).

Another possible improvement of the proposed approach refers to the inclusion of additional features. While the proposed approach only considers the text within the tweet, sentiment information and lexical characteristics (e.g., the usage of uppercase or emoji) have been shown to be important clues for hate-related tasks. Moreover, despite previous works showing the importance of author profiling and demonstrating the utility of exploiting annotators' characteristics in disagreement detection, this information for the test dataset has not yet been released – making those strategies unfeasible for the participation in the challenge – but is a line of work that we plan to investigate in the near future.

Acknowledgments

Department Research Project.

References

[1] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Thessaloniki, Greece, 2023.

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

[3] I. Baris Schlicht, A. F. Magnossão de Paula, P. Rosso, UPV at CheckThat! 2021: Mitigating Cultural Differences for Identifying Multilingual Check-worthy Claims, in: Proceedings of The 12th Conference and Labs of the Evaluation Forum (CLEF), volume 2936, 2021, pp. 465–475.

[4] I. Baris Schlicht, A. F. Magnossão de Paula, P. Rosso, UPV at TREC Health Misinformation Track 2021 Ranking with SBERT and Quality Estimators, in: Proceedings of The Thirtieth Text REtrieval Conference (TREC), 2021.

[5] I. Baris Schlicht, A. F. Magnossão de Paula, Unified and Multilingual Author Profiling for Detecting Haters, in: Proceedings of The 12th Conference and Labs of the Evaluation Forum (CLEF), volume 2936, 2021, pp. 1837–1845.

[6] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 573–584.

[7] A. F. Magnossão de Paula, R. F. da Silva, I. B. Schlicht, Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN), volume 2943, 2021, pp. 356–373.

[8] A. F. Magnossão de Paula, R. F. da Silva, Detection and Classification of Sexism on Social Media Using Multiple Languages, Transformers, and Ensemble Models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the XXXVIII International Conference of the Spanish Society for Natural Language Processing (SEPLN), volume 3202, 2022.

[9] G. Campisi, S. Corchs, E. Fersini, F. Gasparini, M. Mantovani, Automatic detection of sexist content in memes, Image 46 (2018) 53–9.

[10] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist meme on the web: A study on textual and visual cues, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2019, pp. 226–231.

[11] 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.

[12] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[13] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, arXiv preprint arXiv:2303.04222 (2023).

[14] A. Rydelek, D. Dementieva, G. Groh, AdamR at SemEval-2023 Task 10: Solving the Class Imbalance Problem in Sexism Detection with Ensemble Learning, arXiv preprint arXiv:2305.08636 (2023).

[15] S. Rallabandi, S. Singhal, P. Seth, SSS at SemEval-2023 Task 10: Explainable Detection of Online Sexism using Majority Voted Fine-Tuned Transformers, arXiv preprint arXiv:2304.03518 (2023).

[16] A. F. Magnossão de Paula, I. Baris Schlicht, AI-UPV at IberLEF-2021 DETOXIS task: Toxicity Detection in Immigration-Related Web News Comments Using Transformers and Statistical Models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN), 2021, pp. 547–566.

[17] A. F. Magnossão de Paula, P. Rosso, I. Bensalem, W. Zaghouani, UPV at the Arabic Hate Speech 2022 Shared Task: Offensive Language and Hate Speech Detection Using Transformers and Ensemble Models, in: Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, 2022, pp. 181–185.

[18] A. F. Magnossão de Paula, I. Bensalem, P. Rosso, W. Zaghouani, Transformers and Ensemble Methods: A Solution for Hate Speech Detection in Arabic Languages, in: Proceedings of the 1st Natural Language Processing (NLP) Challenge at Centre De Recherche Sur L'information Scientifique et Technique (CERIST), 2023.

[19] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, IberEval @ SEPLN 2150 (2018) 214–228.

[20] S. Gandhi, S. Kokkula, A. Chaudhuri, A. Magnani, T. Stanley, B. Ahmadi, V. Kandaswamy, O. Ovenc, S. Mannor, Scalable Detection of Offensive and Non-compliant Content / Logo in Product Images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2247–2256.

[21] A. Tabone, K. Camilleri, A. Bonnici, S. Cristina, R. Farrugia, M. Borg, Pornographic Content Classification Using Deep-Learning, in: 21st ACM Symposium on Document Engineering, 2021, pp. 1–10.

[22] S. L. Hor, H. A. Karim, M. H. L. Abdullah, N. AlDahoul, S. Mansor, M. F. A. Fauzi, J. See, A. S. B. Wazir, An Evaluation of State-of-the-Art Object Detectors for Pornography Detection, in: IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2021, pp. 191–196.

[23] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, Advances in Neural Information Processing Systems 33 (2020) 2611–2624.

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

[25] J. Alba-Cepero, M. Cabezas-Puerto, V. F. López-Batista, Á. M. Moreno-Montero, Bias Analysis on Twitter, in: New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET 2022 Collection, Springer, 2022, pp. 131–142.

[26] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: the Problem of Biased Datasets, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 602–608. doi:10.18653/v1/N19-1060.

[27] X. Zhou, Challenges in Automated Debiasing for Toxic Language Detection, University of Washington, 2021.

[28] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from Disagreement: A Survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:10.1613/jair.1.12752.

[29] A. Kalra, A. Zubiaga, Sexism Identification in Tweets and Gabs using Deep Neural Networks, arXiv preprint arXiv:2111.03612 (2021).

[30] E. Amigó, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.

[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[32] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepčević, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic Sexism Detection with Multilingual Transformer Models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN), volume 2943, 2021. URL: https://ceur-ws.org/Vol-2943/exist_paper1.pdf.

[33] K. Bengoetxea, A. Aguirregoitia, Multiaztertest@Exist-Iberlef2022: Sexism Identification in Social Networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the XXXVIII International Conference of the Spanish Society for Natural Language Processing (SEPLN), volume 3202, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper8.pdf.

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[35] R. Ali, U. Farooq, U. Arshad, W. Shahzad, M. O. Beg, Hate Speech Detection on Twitter Using Transfer Learning, Computer Speech & Language 74 (2022) 101365.

[36] M. R. Karim, S. Kanti Dey, T. Islam, M. Shajalal1, B. R. Chakravarthi, Multimodal Hate Speech Detection from Bengali Memes and Texts, in: International conference on Speech & Language Technology for Low-resource Languages (SPELLL), 2022, pp. 1–15.

[37] P. K. Roy, S. Bhawal, C. N. Subalalitha, Hate Speech and Offensive Language Detection in Dravidian Languages Using Deep Ensemble Framework, Computer Speech & Language 75 (2022) 101386.

[38] H. Face, Transformers: State-of-the-art Natural Language Processing, https://huggingface.co/, 2022. Accessed on May 31, 2023.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971.

[40] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, 2020. arXiv:2003.10555.