UMUTeam at EXIST 2024: Multi-modal Identification and Categorization of Sexism by Feature Integration

Notebook for the EXIST 2024 Lab at CLEF 2024

Ronghao Pan[1], José Antonio García-Díaz[1], Tomás Bernal-Beltrán[1] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

The fourth edition of the EXIST shared task is a multimodal identification and categorization of sexism. This edition will take place as a lab in CLEF 2024. The main innovation in this edition with respect 2023 is the addition of a multimodal task based on sexism identification and categorization with MEMEs. As before, this shared task compromises sexist documents written in English and Spanish. For the textual tasks, we rely on feature integration from some Large Language Models and linguistic features extracted from our custom tool. We rank 15th in Sexism Identification, 8th in Source Intention, and 20th in Sexism Categorization. For the multimodal task, we rely on the CLIP model to extract the embedded text and image, and then combine them by diagonal multiplication to obtain the classification models. We rank 33rd in Sexism Identification, and 18th in both Source Intention and Sexism Categorization.

Keywords

Sexism identification, sexism categorization, source intention detection, multi modal Feature Engineering, natural language processing

1. Introduction

Social networks have become central platforms for social activism and complaints, allowing movements like #MeToo, #8M, and #Time'sUp to spread rapidly. Through these channels, women around the world have shared their real-life experiences of abuse, discrimination, and other forms of sexism. However, Social networks also facilitate the spread of sexist, disrespectful, and hateful behavior. In this context, the development of automated tools is essential. These tools can help detect and warn against sexist behavior and discourse, estimate the frequency of sexist and abusive situations on social media, identify the most common forms of sexism, and analyze how sexism is expressed on these platforms. Traditional systems for detecting sexism frequently depend on predefined labels and fixed perspectives, which can miss the complexity and subjectivity of sexist statements. The challenge in identifying and addressing sexism stems from its inherently subjective assessment. In contrast, perspectivism offers a promising method for enhancing detection by incorporating diverse opinions and viewpoints. An important contribution to this field, which aims to address the problem of sexism identification within the paradigm of learning with disagreement, has been made by the project EXIST 2024: sEXism Identification in Social neTworks [1, 2, 3].

Here we describe our participation in the fourth edition of EXIST [4, 5]. In the first three editions, EXIST focused only on the detection and classification of sexist text messages, and this 2024 edition introduces tasks that focus on images, specifically memes. Memes, generally humorous images, spread rapidly through social networks and the Internet. They can therefore encompass a wider range of sexist manifestations on social networks, especially those disguised as humor. Therefore, this shared

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

*Pan et al.

†These authors contributed equally.

ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); tomas.bernalb@um.es (T. Bernal-Beltrán); valencia@um.es (R. Valencia-García)

0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0003-2457-1791 (R. Valencia-García)

task involves the development of automated multimodal tools capable of detecting sexism in both text and memes. With this addition, the organizers aim to cover a broader spectrum of sexism on social networks, especially that which is disguised as humor.

As a reminder, the tasks in the last edition of EXIST were three challenges, namely, (1) sexism identification, a binary classification task in which participants had to determine whether a tweet was sexist or not; (2) source intention, a multi-classification task focused on determining whether the author's intention was to post a sexist message, to report a sexist situation, or to make a judgment; and (3) sexism categorization, a multi-label classification task focused on identifying sexist characteristics. It should be noted that in EXIST 2024, the organizers are retaining the learning with disagreements paradigm proposed in 2023.

Our research group has experience in the detection hate speech [6] and misogyny [7] through the compilation and evaluation of several corpora. We have focused mainly on Spanish-language data and our work dealing with English-language data is more limited, limited to participating in the previous editions of EXIST [8, 9, 10] and the evaluation of zero and few-shot learning strategies on some existing English datasets focused on hate speech detection [11].

2. Related works

Sexism refers to any abuse or negative sentiment directed at women based on their gender, or their gender in combination with other identity attributes. In particular, sexism is a growing problem on the Internet, with detrimental effects on women and other marginalized groups. It makes online spaces less accessible and less welcoming, perpetuating asymmetries and social injustices.

Automated tools are already widely used to identify and rate sexist content online. Researchers have proposed several approaches to address this problem, ranging from rule-based methods [12] to the use of more complex models based on deep learning and pre-trained language models with Transformer architecture from a linguistic perspective [13, 14], with only a few attempts to address the problem from a visual or multimodal perspective. Another work is described in [15], where the authors study about 6 thousand misogyny memes using a deep learning model that determines which modality plays a more significant role. The dataset included various characteristics such as hate speech, sexism, or cyberbullying, among others. The authors found that all modalities were useful for identifying misogyny, with text playing a significant role.

Sexism can be further categorized into different forms depending on the author's intent or the type of sexism. All editions of EXIST use categories such as "Ideology and Inequality", "Stereotyping and Dominance", "Objectification", "Sexual Violence", and "Misogyny and Non-Sexual Violence". This is similar to SemEval 2023 - Task 10 - Explainable Detection of Online Sexism (EDOS) [16], which defined a taxonomy for the more explainable classification of sexism in three hierarchical levels: binary sexism detection, category of sexism, and fine-grained vector of sexism. In all editions of EXIST, sexism is also considered at three levels: binary sexism detection in tweets, multiclass detection for source intention in tweets, and category of sexism in tweets.

Although the problem of sexism is primarily approached from a textual perspective, a field of research has emerged that approaches the problem from a visual or multimodal perspective, as the multimodal approach has shown superior performance to single modality approaches in detecting hate speech and misogyny, such as [17] and [18]. Sexist content can also appear in images or in multimodal forms. From a visual perspective, most research focuses on detecting offensive, non-conforming, or pornographic content. For this reason, this edition of EXIST includes a new task focused on detecting sexism in memes.

Currently, one of the problems that needs to be addressed in sexism detection is the presence of biases that could affect the actual performance of the models. Studies such as, [19], and [20] have addressed this issue by using textual data to detect sexism. However, many of the contributions of previous sexism detection studies have not considered the complexity and multiple perspectives surrounding sexism, as sexism can manifest itself in multiple ways influenced by cultural, social, and individual factors.

Perspectivism in sexism detection involves considering multiple cultural, social, and individual viewpoints and contexts when analyzing potentially sexist content. It is essential to better understand the different manifestations of sexism, avoid bias, and improve the accuracy of detection models. It promotes equity by recognizing the different ways in which sexism can manifest itself, thus contributing to systems that are fairer and more sensitive to cultural and social diversity. For example, [14] examine the errors made by classification models and discuss the difficulty of automatically classifying sexism due to the subjectivity of labels and the complexity of the natural language used in social networks.

Thus, this edition of EXIST has taken a similar approach to the 2023 edition, adopting the Learning with Disagreement (LeWiDi) paradigm for both dataset development and system evaluation. Unlike traditional methods that rely on a single "correct" label for each example, LeWiDi trains models to process and learn from conflicting or diverse annotations. This allows the system to incorporate different annotator perspectives, biases, and interpretations, resulting in a more equitable learning process.

3. Dataset

The textual challenges on EXIST 2024 followed the same strategies as previous editions [21, 22]; that is, it includes tweets written in Spanish or English crawled with certain keywords that are commonly used to undervalue the role of women. In fact, for the textual dataset the dataset is the same as from 2023.

For the multimodal dataset, the organizers of EXIST 2024 defined a lexicon of 250 terms and phrases that lead to sexist memes, derived from those that have proven effective in identifying sexism in previous editions of EXIST. This set includes 112 terms in English and 138 in Spanish, covering diverse topics and including terms with varying degrees of use in both sexist and non-sexist contexts, all centered around women. These terms were used as search queries on Google Images to obtain the top 100 images per term. Rigorous manual cleaning procedures were applied to define memes and remove noise such as textless images, text-only images, advertisements, and duplicates. This resulted in a final set of over 3,000 memes per language. Given the heterogeneous proportion of memes per term, we discarded the most unbalanced seeds and ensured that each seed had at least five memes. The final dataset was curated to achieve a balanced distribution of memes per seed. To avoid selection bias, memes were randomly selected while maintaining the appropriate distribution per seed. This process resulted in over 2,000 memes per language for the training set and over 500 memes per language for the test set.

For the annotation process, the organizers considered two socio-demographic traits: gender and age range. Each meme was then annotated by six crowd sourced annotators selected through the Prolific app, following guidelines developed by two experts in gender issues. The authors also provided the annotators' education level, ethnicity, and country of residence. The idea is to reduce labeling bias that may arise from cultural differences among annotators.

In addition, following the Learning with Disagreements paradigm removes the assumption that items have a single, unambiguous interpretation in a given context. Therefore, the dataset does not have a single "gold" annotation, but participants can use the full range of annotations from all annotators. This allows us to capture the diversity and develop more robust systems. The organizers release all annotations from all annotators to the participants.

The custom validation split is created using stratification by label in an 80-20 ratio.

First, since we have all the annotations, we decided to train our models for Task 1 (sexism identification) as a regression task rather than a classification task; thus, a text is considered sexist if the regression model outputs a score greater than 2.5. The soft labels are the output of the regression model normalized to a range of 1–10. Note that this approach was also used by our team in the previous edition of EXIST. The hard labels of the textual modality can be found in Table 1.

Second, Table 2 shows the label distribution for Tasks 2 and 5. The labels are: (1) direct, if the intent of the document is to write a message that is itself sexist or incites to be sexist; (2) judgmental, if the intent is to report and share a sexist situation; and (3) reported, if the intent was to judge.

Third, Table 3 shows the label distribution for Tasks 3 and 6. The labels are: (1) ideological and inequality, if the message downplays feminism or, equality between men and women; (2) stereotyping

and dominance, if the message contains stereotypes about social roles; and (3) objectification, if the message contains physical characteristics about beauty standards or hyper-sexualization. As can be seen, we extracted a split from the dataset provided for individual validation (in an 80-20 ratio) using label stratification. For Task 2, there is a significant imbalance between the labels, with direct sexism being the label with the most examples and judgments and reports with a similar number of examples. For Task 3, the dataset has more balance between the characteristics. We have included the number of unknown responses in the textual modality.

Table 1
Datasets statistics for Tasks 1 (left) and 4 (right) concerning sexism identification.

	Task 1			Task 4		
label	train	val	total	train	val	total
non-sexist	2306	1540	3846	2933	734	3667
sexist	2466	1646	4112	302	75	377
total	4772	3186	7958	3235	809	4044

Table 2
Datasets statistics for Tasks 2 (left) and 5 (right) concerning source intention.

	Task 2			Task 5		
label	train	val	total	train	val	total
–	3175	2090	5265	-	-	-
direct	992	665	1657	2609	626	3261
judgmental	296	225	521	652	157	783
reported	309	206	515	-	-	-
total	4772	3186	7958	3261	783	4044

Table 3
Distribution of the datasets of the Tasks 3 (left) and 6 (right) concerning misogyny categorization.

	Task 3			Task 6		
label	train	val	total	train	val	total
ideological inequality	2013	1390	3403	3325	865	4190
misogyny non sexual violence	1526	1053	2579	3890	952	4842
objectification	1857	1173	3030	3662	928	4590
sexual violence	1140	722	1862	1770	474	2244
stereotyping dominance	2253	1475	3728	1639	434	2073
unknown	114	67	181	-	-	-
total	8903	5880	14783	14286	3653	17939

4. Methodology

This section explains our methodology for Tasks 1, 2, and 3 using textual modalities (see Section 4.1), and for Tasks 4, 5, and 6 using multimodalies (see Section 4.2).

4.1. Textual modalities

For the Tasks 1, 2, and 3 we decided to keep our previous strategy of training both languages (Spanish and English) together in order to reduce the carbon footprint of training a model for each language but to reduce the number of models to 4. Since we are evaluating feature integration strategies, we decided to keep two Spanish LLMs, BETO and MarIA; and two multilingual LLMs, deBERTa v3 [23] and XLMTwitter. In this sense, we removed from our pipeline the multilingual BERT, XLM, BERTIN, DistilBETO and ALBETO [24]). We also extracted the linguistic features (LFs) using the UMUTextStats tool [25].

To fine-tune each LLM for each task, we performed hyperparameter tuning on 10 models. This involved testing different learning rates, varying the number of epochs from 1 to 5, experimenting with batch sizes of 8 and 16, and adjusting warm-up steps and weight decay to optimize the learning rate during the initial training phases. The results of this tuning process are detailed in the Table 4.

Table 4
Hyperparameter optimization of LLMs.

LLM	learning rate	epochs	batch size	warmup steps	weight decay
		Task 1			
BETO	3.6e-05	4	16	0	0.044
MARIA	2.9e-05	3	16	250	0.21
MDEBERTA	3.1e-05	4	16	500	0.18
XLMTWITTER	4.7e-05	2	8	0	0.22
		Task 2			
BETO	3.8e-05	3	16	500	0.17
MARIA	3.8e-05	2	8	250	0.031
MDEBERTA	1.4e-05	5	8	500	0.018
XLMTWITTER	4.2e-05	5	8	1000	0.22
		Task 3			
BETO	2.6e-05	4	16	0	0.25
MARIA	4.5e-05	5	16	0	0.054
MDEBERTA	4.5e-05	5	16	500	0.29
XLMTWITTER	2.8e-05	4	8	500	0.17

Next, we extract the classification token for each tweet, LLM, and Task (1, 2, and 3) using Sentence-BERT [26] to extract the `[CLS]` token. Now that each document is represented by a unique fixed-length vector, we can merge it with the LFs using different feature integration strategies. The first strategy, known as Knowledge Integration (KI), is to use the feature vectors to train a new multi-input neural network. The second strategy is based on Ensemble Learning (EL). For this, we train separate simple neural networks for each LLM and for the LFs, and combine the results by averaging probabilities and outputs.

The training of the KI and the individual models are shown in the Table 5. As expected, most architectures are shallow (one or two hidden layers), brick-shaped (same number of neurons in each layer). This is expected because the vectors already capture the meaning of the sentences and are tuned to the output tasks. However, the KI for Tasks 1 and 3 resulted in deep neural networks, but with few neurons in each layer, being a funnel shape in Task 1 and a triangle shape in Task 3. If we compare these results with those obtained in the previous edition, it draws our attention that for Task 3 we obtained the best results in 2023 with deep neural networks and complex shapes.

For the EL, we evaluate different strategies: (1) mode of predictions, (2) averaging of probabilities, and (3) obtaining the highest probability. The only exception is Task 1, where we only averaged the predictions of the model, since we considered it to be a regression task.

Now we report our results with our custom validation split. For Task 1 (see Table 6), as we handled it

as a regression task in order to account for the disagreement between the annotators. Therefore, we report using Explained Variance (EV), Root Mean Squared Logarithmic Error (RMSLE), Pearson R, R Square (R2), Mean Average Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The best results were obtained using the KI strategy.

For Tasks 2 and 3, the results with the custom validation split are shown in the Table 7. The models are scored and ranked using the macro average precision, recall, and F1-Score. Similar to the previous edition, our team achieved their best results with the custom validation split with the with KI for Task 2 with the best recall and F1-Score but with the best precision is obtained with an EL based on the mode. In the case of Task 3, the EL strategies achieve better results with an F1-Score of 53.085% averaging probabilities and, the best recall of 81.242% with the highest probability, but the best precision

Table 5
Results of the hyper-parameter optimization stage using Keras of the LFs (LF), each LLM and the multi-input neural network using Knowledge Integration (KI).

feature set	shape	layers	neurons	dropout	lr	batch size	activation
			Task 1				
LF	brick	2	64	0	0.001	32	linear
BETO	triangle	3	512	0.3	0.001	64	tanh
MARIA	long funnel	3	256	0.1	0.01	64	sigmoid
MDEBERTA	brick	4	16	0.2	0.001	32	selu
XLMTWITTER	brick	2	8	0.3	0.001	32	sigmoid
KI	funnel	8	8	0.3	0.001	64	elu
			Task 2				
LF	brick	2	256	0	0.01	256	linear
BETO	brick	6	128	0	0.01	128	elu
MARIA	triangle	5	512	0	0.01	256	elu
MDEBERTA	brick	2	512	0	0.01	256	linear
XLMTWITTER	brick	5	512	0	0.01	256	selu
KI	brick	3	128	0.2	0.001	512	sigmoid
			Task 3				
LF	brick	2	256	0	0.01	64	linear
BETO	brick	2	128	0	0.001	64	tanh
MARIA	brick	2	512	0	0.01	64	linear
MDEBERTA	brick	2	512	0	0.01	64	linear
XLMTWITTER	brick	2	37	0	0.001	64	sigmoid
KI	triangle	7	16	False	0.01	32	selu

Table 6
Results with the custom validation split for Task 1, reported as a regression task. The results are organized by feature set. The first block is the linguistic features, the second block are the LLMs, and the third and fourth blocks are the Knowledge Integration (KI) strategy and an ensemble learning (EL) respectively.

feature-set	EV	RMSLE	PEARSONR	R2	MAE	MSE	RMSE
LF	0.162	0.414	0.458	0.161	1.573	3.629	1.905
BETO	0.563	0.224	0.752	0.562	1.084	1.892	1.376
MARIA	0.574	0.218	0.757	0.574	1.080	1.843	1.358
MDEBERTA	0.575	0.216	0.758	0.575	1.082	1.837	1.356
XLMTWITTER	0.541	0.235	0.735	0.541	1.130	1.986	1.409
EL	0.584	0.227	0.770	0.584	1.100	1.801	1.342
KI	**0.602**	**0.203**	**0.776**	**0.602**	**1.042**	**1.720**	**1.312**

is achieved by BETO, a Spanish LLM.

Table 7
Results with custom validation for Tasks 2 and 3. The results are organized with the LFs (LF), all LLMs separately, the Knowledge Integration strategy (KI), and the three evaluated ensemble learning strategies (EL). All metrics are macro weighted.

	Task 2			Task 3		
feature-set	precision	recall	F1-Score	precision	recall	F1-Score
LF	38.788	32.025	31.999	36.837	57.090	42.527
BETO	52.360	42.170	43.607	**53.073**	48.725	50.538
MARIA	53.592	44.683	47.010	50.375	55.177	52.343
MDEBERTA	52.402	41.623	42.556	45.748	59.038	50.853
XLMTWITTER	52.636	43.477	45.960	52.535	50.167	50.969
KI	54.288	**52.896**	**53.371**	49.873	54.543	51.760
EL (HIGHEST)	53.906	40.815	42.943	33.973	**81.242**	47.466
EL (MEAN)	59.768	40.618	41.843	51.229	56.362	**53.085**
EL (MODE)	**61.823**	40.601	41.731	50.016	54.478	51.699

4.2. Multi modalities

Figure 1 shows the architecture used to implement the model for Task 4, Task 5 and Task 6. We can see that we have used the CLIP [27] model (specifically, *openai/clip-vit-base-patch32*[1]) to extract both image and text embeddings, and then multiply those embeddings using a diagonal operation to use them as input to a classification neural network to identify whether a meme is sexist or not. Diagonal multiplication refers to a specific mechanism used during the training process to effectively connect and align text and image representations.

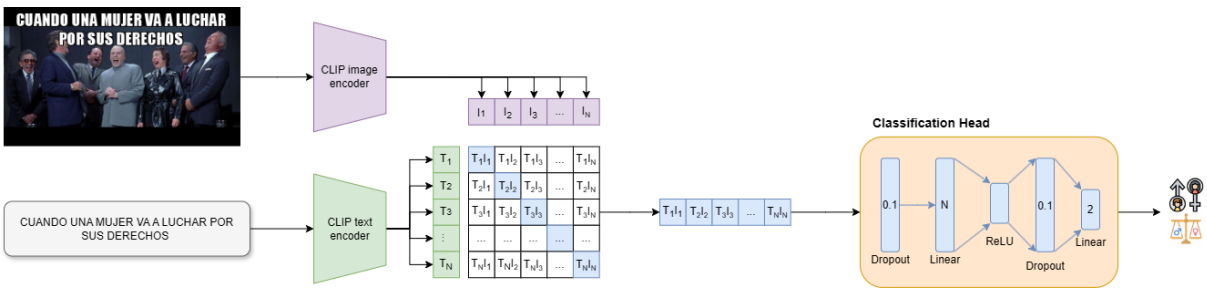

Figure 1: System architecture of Task 4.

As shown in Figure 1, during training with the visual input (meme) and the text input (textual content of the meme), the text embedding vector and the image embedding vector perform diagonal alignment. This means that the dot product between the text embedding vector and its corresponding image embedding vector is high when they are related (i.e., for correct pairs) and low when they are not related (i.e., for incorrect pairs). This is possible because the CLIP model trains and processes the text and images in the same shared feature space. CLIP is a model developed by OpenAI that can efficiently understand and associate images and text. This means that the model learns to correctly associate a descriptive text with its corresponding image and to distinguish it from other unrelated images. Each component of the architecture is described in detail below:

- **Image Encoder (CLIP image encoder).** The meme image is passed through the CLIP image encoder, which extracts a series of embeddings $\{I_1, I_2, I_3, \ldots, I_N\}$ These embeddings represent the

[1] https://huggingface.co/openai/clip-vit-base-patch32

visual properties of the image.

- **Text Encoder (CLIP text encoder).** The text of the meme is fed into the CLIP text encoder, which produces a series of embeddings $\{T_1,T_2,T_3,\ldots,T_N\}$ These embeddings capture the semantic features of the text.
- **Diagonal multiplication**. The image and text embeddings are combined using a diagonal multiplication operation. This involves multiplying each text embedding T_i with the corresponding image embedding I_i to create a new set of combined embedding $\{T_1I_1,T_2I_2,T_3I_3,\ldots,T_NI_N\}$.
- **Classification head**. The combined embeddings are passed through a classification neural network consisting of the following components:
  - **Dropout Layer**. A layer with a dropout rate of 0.1 to prevent overfitting.
  - **Linear Layer**. A linear layer that reduces the dimensionality to N features, where N is the combined embedding length. **Activation Layer (ReLU)**: A ReLU activation layer to introduce nonlinearities.
  - **Dropout Layer**. Another layer with a dropout rate of 0.1 to prevent overfitting.
  - **Final Linear Layer**. A linear layer that reduces the output to 2 neurons, corresponding to the output classes (sexist or non-sexist).

For Tasks 5 and 6 we also assigned the most repeated label of the annotators, and in case of a tie, female annotators have more weight. In addition, we have used the same approach of using the CLIP model to obtain embeddings of text and images and then, through a diagonal multiplication, to obtain a combined embedding that will be the input of a classification neural network.

5. Results

From an evaluation metrics perspective, organizers use two types of evaluation based on the learning with disagreement paradigm [28].

- **Hard-Hard**. This is a comparison between "hard" system outputs (final labels) and "hard" ground truth. The Information Contrast Measure (ICM) metric is used to measure the similarity to the ground truth categories. The F1-Score is also reported, although it is not ideal in this context because it does not take into account the relationships between labels.
- **Soft-Soft**. This is a comparison between "soft" system outputs (probabilities) and "soft" ground truths. In this case, the ICM-soft metric (an extension of ICM) is used as the official metric.

5.1. Textual modality

We sent three runs for Tasks 1, 2, and 3 based on the results on the custom validation split. For Tasks 1, our runs are KI, EL, LFs. Our results for the Task 1 are described in Table 8. We ranked 15th, 19th, and 32nd for the soft-soft scheme for each run. As expected, our results were better in Spanish than English, explained by the usage of two Spanish LLMs, BETO and MarIA. The model based on LFs outperformed the baselines proposed.

The official results for Task 2 are shown in Table 9. The first run is based on KI, while the second and third runs are based on MarIA and multilingual DeBERTA. This is the task in which we obtained our best results, ranking 8th (7th in Spanish and 10th in English) with the KI strategy, totaling 38 submissions among all teams. It is worth noting that MarIA outperformed MDeBERTA in both Spanish and English.

Next, the results for Task 3 are reported in Table 10. In this case, we focus on the hard vs. hard scheme, as we do not calculate probabilities. The first run is based on EL averaging the probabilities, the second run is based on KI, and the third run is based on EL based on mode. We got our best results with the first run, except for the English results, where we got better results with EL based on mode. Since we compete in a Hard vs. Hard scheme, the ranking also includes the Macro F1-Score, which is 49.42% with the first run, 47.38% with the second run, and 48.21% with the third run.

5.2. Multimodal

Table 11 shows the official results for Task 4, which evaluates hard labels only. The UMUTeam consistently ranks around 33-35 in all evaluations, both ALL and language-specific such as English or Spanish. We have exceeded the two baselines suggested by the organizers, ranking 33rd with -0.2422 on ICM-Hard and 0.6963 on F1_YES (F1-Score of sexist labels).

Although the ICM-Hard metrics are negative, indicating space for improvement in similarity to ground truth, we present less negative values compared to the baselines. This means that our predictions are closer to the true labels. If we look only at the F1-Score of the model, our system would be in a better position. However, in terms of similarity to the ground truth (ICM-Hard), we have a lot of room for improvement. This may be due to the fact that our approach does not have a final softmax layer indicating the probability of each label, since we used the model logits directly for the prediction. It is also possible that it is due to a lack of additional adjustments and optimizations. In this case, we only evaluated a 2e-5 learning rate, a training batch size of 16, 20 epochs, and an epoch-based validation strategy.

Table 12 provides the official results for Task 5, evaluating only hard labels. In summary, we have achieved 18th place in all evaluations, slightly behind the baseline majority class (17th place) and ahead

Table 8
Official results for Task 1 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	15	0.6679	12	0.7510	16	0.5200
UMUTeam 2	19	0.50339	18	0.5720	20	0.3745
UMUTeam 3	32	-0.4170	30	-0.3691	29	-0.5207
baseline majority class	36	-2.3585	36	-2.5421	36	-2.1991
baseline minority class	40	-3.0717	37	-2.5742	40	-3.8158

Table 9
Official results for Task 2 with the Soft vs Soft scheme, including the Spanish and English results. Ranking is by runs.

	Soft vs Soft ALL		Soft vs Soft ES		Soft vs Soft EN	
Team	Rank	ICM-Soft	Rank	ICM-Soft	Rank	ICM-Soft
UMUTeam 1	8	-1.9569	7	-1.7667	10	-2.2517
UMUTeam 2	12	-2.0533	10	-1.8027	12	-2.4821
UMUTeam 3	19	-3.3189	19	-3.0432	20	-3.8940
baseline majority class	27	-5.4460	30	-5.6674	25	-5.2028
baseline minority class	35	-32.95527	35	-28.7093	35	-39.4948

Table 10
Official results for Task 3 with the Hard vs Hard scheme, including the Spanish and English results. Ranking is by runs.

	Hard vs Hard ALL		Hard vs Hard ES		Hard vs Hard EN	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam 1	20	-0.733	24	-0.69597	21	-0.8012
UMUTeam 2	24	-0.8719	26	-0.8245	25	-0.9513
UMUTeam 3	22	-0.7901	25	-0.81457	20	-2.4821
baseline majority class	30	-1.5984	30	-1.7269	28	-1.4563
baseline minority class	34	-3.1295	34	-3.3196	33	-2.9279

Table 11
Official results for Task 4, including only hard labels. We report the rank, the ICM-Hard metric and the F1-Score of the YES label between parentheses.

	Hard-Hard ALL		Hard-Hard ES		Hard-Hard ES	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam	33	-0.2422 (69.63)	33	-0.2639 (69.13)	35	-0.2210 (70.17)
baseline majority class	39	-0.4038 (68.21)	45	-0.4001 (67.65)	39	-0.4076 (68.80)
baseline minor class	46	-0.6468	49	-0.6557	43	-0.6381

of the baseline minor class (21st place).

Regarding the ICM-Hard values of our system (-1.1486, -1.0605, -1.2148), they are negative and higher in magnitude compared to the baseline majority class, indicating a lower similarity to the true labels. However, they are significantly better than those of the baseline minor class (-2.0637, -2.0866, -2.0410).

In terms of F1_YES values, our model is superior to both baselines in all scenarios, especially in the second ES column (0.2805), indicating better precision and recall in the positive class.

Table 12
Official results for Task 5, including only hard labels. We report the rank, the ICM-Hard metric and the F1-Score of the YES label between parentheses.

	Hard-Hard ALL		Hard-Hard ES		Hard-Hard ES	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam	18	-1.1486 (20.98)	18	-1.0605 (20.40)	18	-1.2148 (28.05)
baseline majority class	17	-1.0445 (18.39)	17	-1.0504 (18.02)	17	-1.0385 (18.77)
baseline minor class	21	-2.0637 (6.97)	21	-2.0866 (6.77)	21	-2.0410 (7.19)

Table 13 shows the official results for Task 6, which is a multi-label classification problem where only hard labels are evaluated. Our system achieved several ranks depending on the scenario with a total of 36 submissions among all teams: 18th in Hard-Hard ALL, 21st in the first column of Hard-Hard ES, and 12th in the second column of Hard-Hard ES. In comparison, the model performs better than the baselines in most scenarios, especially in the second Hard-Hard ES set.

The ICM-Hard values obtained (-1.9511, -2.3853, -1.5817) are negative, but generally better than the baselines, indicating a higher similarity to the true labels. It also clearly outperforms the baseline minor class in all scenarios and performs better than the baseline majority class in terms of ICM-Hard.

With respect to Macro F1-Score, we obtained significantly higher values than the baselines, especially in the Hard-Hard ES evaluation, indicating a better balance between precision and recall and an overall balanced performance in all classes.

Table 13
Official results for Task 6, including only hard labels. We report the rank, the ICM-Hard metric and the F1-Score of the YES label between parentheses.

	Hard-Hard ALL		Hard-Hard ES		Hard-Hard ES	
Team	Rank	ICM-Hard	Rank	ICM-Hard	Rank	ICM-Hard
UMUTeam	18	-1.9511 (37.86)	21	-2.3853 (34.74)	12	-1.5817 (40.88)
baseline majority class	19	-2.0711 (9.19)	19	-2.1173 (9.01)	17	-2.0015 (9.38)
baseline minor class	23	-3.3135 (3.18)	23	-3.2599 (3.28)	19	-3.3506 (3.07)

6. Conclusions and further work

This document describes UMUTEAM's proposal for EXIST 2024, which focuses on the identification and categorization of SEXISM in Spanish and English. This is a very interesting competition as it deals with the learning with disagreements paradigm, binary and multi-classification tasks as well as a new challenge based on multimodal features. For the textual tasks, we based our proposal on the integration of linguistic features with sentence embeddings extracted for four LLMs, including Spanish and multilingual variants. We achieved our best result, 8th place, in Task 2, the source detection task.

As in previous editions, we are satisfied with our participation, since we achieve competitive results in all tasks, outperforming the proposed baselines. However, in this edition, we only consider the different annotator schemes in the binary classification tasks, since it is considered a regression problem, but we are satisfied that we have been able to participate in all multimodal tasks, incorporating image processing modeling tachniques into our pipeline.

Acknowledgments

This work has been supported by projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe, LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR, "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

[1] F. R.-S. y Jorge Carrillo-de-Albornoz y Laura Plaza y Julio Gonzalo y Paolo Rosso y Miriam Comet y Trinidad Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[2] F. R.-S. y Jorge Carrillo-de-Albornoz y Laura Plaza y Adrián Mendieta-Aragón y Guillermo Marco-Remón y Maryna Makeienko y María Plaza y Julio Gonzalo y Damiano Spina y Paolo Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[3] L. Plaza, J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: Proceedings of ECIR'23, 2023, pp. 593–599. doi:10.1007/978-3-031-28241-6_68.

[4] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes. experimental ir meets multilinguality, multimodality, and interaction., in: Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Springer, 2023, pp. 316–342.

[5] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes (extended overview, in: Working Notes of CLEF 2024- Conference and Labs of the Evaluation Forum, Springer, 2024.

[6] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.
[7] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2021) 506–518. doi:j.future.2020.08.032.
[8] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Umuteam at exist 2021: Sexist language identification based on linguistic features and transformers in spanish and english, in: CEUR Workshop Proceedings, volume 2943, 2021, pp. 512–521.
[9] J. A. García-Díaz, S. M. Jiménez-Zafra, R. Colomo-Palacios, R. Valencia-García, Umuteam at exist 2022: Knowledge integration and ensemble learning for multilingual sexism identification and categorization using linguistic features and transformers, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), volume 3202, 2022.
[10] J. A. García-Díaz, R. Pan, R. Valencia-García, Umuteam at exist 2023: Sexism identification and categorisation fine-tuning multilingual large language models, 2023.
[11] J. A. García-Díaz, R. Pan, R. Valencia-García, Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english, Mathematics 11 (2023) 5004.
[12] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples, in: International Conference on Web and Social Media, 2020. URL: https://api.semanticscholar.org/CorpusID:235303716.
[13] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, ArXiv abs/2111.04551 (2021). URL: https://api.semanticscholar.org/CorpusID:237579065.
[14] A. Kalra, A. Zubiaga, Sexism identification in tweets and gabs using deep neural networks, ArXiv abs/2111.03612 (2021). URL: https://api.semanticscholar.org/CorpusID:243832598.
[15] S. Chen, U. Naseem, I. Razzak, F. Salim, Unveiling misogyny memes: A multimodal analysis of modality effects on identification, in: Companion Proceedings of the ACM on Web Conference 2024, 2024, pp. 1864–1871.
[16] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305. doi:10.18653/v1/2023.semeval-1.305.
[17] A. Chhabra, D. K. Vishwakarma, Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture, Engineering Applications of Artificial Intelligence 126 (2023) 106991. URL: https://www.sciencedirect.com/science/article/pii/S0952197623011752. doi:https://doi.org/10.1016/j.engappai.2023.106991.
[18] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, B. R. Chakravarthi, Mistra: Misogyny detection through text–image fusion and representation analysis, Natural Language Processing Journal 7 (2024) 100073. URL: https://www.sciencedirect.com/science/article/pii/S2949719124000219. doi:https://doi.org/10.1016/j.nlp.2024.100073.
[19] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of abusive language: the problem of biased datasets, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:174799974.
[20] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, N. Smith, Challenges in automated debiasing for toxic language detection, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3143–3155. URL: https://aclanthology.org/2021.eacl-main.274. doi:10.18653/v1/2021.eacl-main.274.
[21] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.
[22] 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.

[23] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, CoRR abs/2111.09543 (2021). URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.

[24] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 4291–4298. URL: https://aclanthology.org/2022.lrec-1.457.

[25] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[26] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/v1/D19-1410.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[28] 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), 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.