I2C-UHU at EXIST 2024: Transformer-Based Detection of Sexism and Source Intention in Memes Using a Learning with Disagreement Approach

Alvaro Carrillo-Casado*, Javier Román-Pásaro, Jacinto Mata-Vázquez and Victoria Pachón-Álvarez

I2C Research Group, University of Huelva, Spain

Abstract

In this paper, the I2C-UHU Group addresses the Exist-2024 challenges of Sexism Identification and Source Intention in Memes. We developed an ensemble of classifiers based on Transformer technology and adopted a Learning with Disagreement (LeWiDi) approach to analyze data from multiple annotators' perspectives. Techniques for constructing datasets and optimizing hyperparameters were explored, enhancing model performance through varied combinations. The optimal models were refined by weighting according to prediction accuracy. Our submissions for Task 4 achieved ranks of 4th with ICM-Hard and ICM-Soft scores of 0.5668 and 0.4476, respectively. For Task 5, we secured 2nd and 10th places with ICM-Hard and ICM-Soft scores of 0.4119 and 0.2023, respectively.

Keywords

Transformers, Ensemble of classifiers, Learning with Disagreement, Memes, Hyperparameter, Sexism

1. Introduction

Recent years have seen a marked increase in the prevalence of memes on social media, a distinct type of imagery characterized by humorous textual content. This study investigates how such memes can be used to entertain and disseminate sexist content. This type of humour is often utilized to harm others, for instance through sexism. However, natural language processing (NLP) is an effective tool for understanding and analysing such content.

This paper presents our research on developing a system to detect sexism and the creator's intention in memes, using natural language processing techniques as part of the tasks Sexism Identification in Memes and Source Intention in Memes of EXIST 2024 [1]. For this purpose, models based on Transformers [2] were developed, different types of dataset constructions were performed [3], followed by utilizing the Learning with Disagreement (LeWiDi) [4] approach to build models based on the various perspectives of the annotators, and finally, they were assembled to improve the performance of the models.

In Section 2, we delineate prior research efforts, while Section 3 provides a detailed exposition of Tasks 4 and 5 within the EXIST 2024 framework. Subsequently, Sections 4 and 5, expound

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France

*Corresponding author.

alvaro.carrillo121@alu.uhu.es (A. Carrillo-Casado); javier.roman780@alu.uhu.es (J. Román-Pásaro); mata@uhu.es (J. Mata-Vázquez); vpachon@dti.uhu.es (V. Pachón-Álvarez)

upon the methodology employed and the resultant findings. Finally, Section 6 encapsulates the study's conclusions and outlines prospective avenues for future research endeavors.

2. Related Works

As previously indicated, one of the foundational elements employed in this study is the Learning with Disagreement (LeWiDi) approach. When information from multiple annotators was available during the classifier's creation, the decision generally favoured the majority's opinion. Nonetheless, this method could overlook valuable insights that might enhance the models' effectiveness.

In [5], participation by the AIT_FHSTP team in the EXIST2021 benchmark was noted, concentrating on the automated detection of sexism across social networks using machine learning techniques. This effort was approached as both a binary classification problem and a more detailed task that categorized various forms of sexist content. Two multilingual Transformer models were utilized for their analysis: one based on Multilingual BERT and the other on XLM-R. These models underwent adaptation through unsupervised pre-training and were subsequently fine-tuned with additional data to optimize performance.

Furthermore, in [6], irony is analyzed based on the principles of data perspectivism. It was observed how data, varying by origin, age, and gender, were managed. The performance derived from the standard test set was compared with that from a perspective-based test set. The latter detected the positive class more accurately, demonstrating the effectiveness of incorporating diverse annotator viewpoints.

The detection of sexism in memes presents unique challenges due to the multimodal nature of memes, which combine text and images to convey messages. Techniques such as image-text alignment, sentiment analysis, and context understanding are crucial for accurately identifying sexist content in memes. Recent advancements in computer vision and natural language processing (NLP) have enabled more sophisticated analysis of such multimodal content.

To provide a broader overview of existing techniques, we review additional notable studies in the field. The work by [7] introduced a novel approach that combines convolutional neural networks (CNNs) for image analysis with transformer-based models for text analysis to detect hate speech and offensive content on social media platforms. Their approach leverages the synergy between visual and textual cues in memes to enhance detection accuracy.

Another relevant study by [8] employed a hybrid model that integrates both supervised and unsupervised learning techniques to improve the detection accuracy of subtle forms of hate speech, including sexist remarks, in online discussions. Their approach demonstrates the effectiveness of combining linguistic and behavioral signals to detect nuanced forms of offensive content.

Overall, the integration of various machine learning techniques, including deep learning models, ensemble methods, and data augmentation strategies, has significantly advanced the field of sexism detection and meme analysis. This study builds on these foundational works, incorporating the LeWiDi approach to further enhance the robustness and effectiveness of our models.

3. Tasks and Dataset Description

The objective of *Task 4: Sexism Identification in Memes* is to determine which memes are sexist, while *Task 5: Source Intention in Memes* involves categorizing memes based on the author's intention to understand the role of social media in disseminating sexist messages. The dataset labels are "DIRECT," "JUDGEMENTAL," "-," and "UNKNOWN." For this study, the classification is focused on distinguishing between "DIRECT," where the intention is to spread a sexist message, and "JUDGEMENTAL," where the intention is to condemn a sexist situation or behavior. Both tasks are binary classification tasks.

The features of each meme are:

- `id_EXIST` : a unique identifier for the meme.
- `lang` : languages of the meme ("en" or "es").
- `text` : text automatically extracted from the meme.
- `meme` : name of the file that contains the meme.
- `path_memes` : path to the file that contains the meme.
- `number_annotators` : number of persons that have annotated the meme.
- `annotators` : a unique identifier for each of the annotators.
- `gender_annotators` : gender of the different annotators. Possible values are: "F" and "M", for female and male respectively.
- `age_annotators` : age group of the different annotators. Possible values are: 18-22, 23-45 and 46+.
- `ethnicity_annotators` : self-reported ethnicity of the different annotators. Possible values are: "Black or African America", "Hispano or Latino", "White or Caucasian", "Multiracial", "Asian", "Asian Indian" and "Middle Eastern".
- `study_level_annotators` : self-reported level of study achieved by the different annotators. Possible values are: "Less than high school diploma", "High school degree or equivalent", "Bachelor's degree", "Master's degree" and "Doctorate".
- `country_annotators` : self-reported country where the different annotators live in.
- `labels_task4` : a set of labels (one for each of the annotators) that indicate if the meme contains sexist expressions or refers to sexist behaviours or not. Possible values are: "YES" and "NO".
- `labels_task5` : a set of labels (one for each of the annotators) recording the intention of the person who created the meme. Possible labels are: "DIRECT", "JUDGEMENTAL", "", and "UNKNOWN".
- `split` : subset within the dataset the meme belongs to ("TRAIN-MEME", "TRAIN- MEME" + "EN"/"ES").

The organizers provided only a training dataset; therefore, an 80%-20% split was performed for training and testing purposes. Furthermore, the training dataset was subdivided into 85% for training and 15% for validation. To establish an initial baseline, a single label was assigned using hard voting [9] among the labels proposed by the six annotators. Given the even number of annotators, ties were resolved by randomly selecting a label. Table 1 displays the class distribution for Task 4 following the voting process.

Table 1
Class distribution for Task 4

Class	Total	YES	NO
Train	2749	1810	939
Valid	486	245	241
Test	809	476	333

For Task 5, since only two labels ("DIRECT" and "JUDGEMENTAL") need to be detected, a hard voting strategy was also used to generate the hard label among the annotators. The values "-" and "UNKNOWN" were discarded in the voting process. Table 2 shows the class distribution for Task 5 after the voting process.

Table 2
Class distribution for Task 5

Class	Total	DIRECT	JUDGEMENTAL
Train	2498	1668	830
Valid	440	293	147
Test	721	482	239

4. Methodology and Experiments

In this section, we delineate the methodologies employed in our investigation. Despite the availability of visual content in the provided meme datasets, our analytical approach was exclusively focused on the textual data extracted from these memes. This decision was driven by our aim to develop and refine text-based classifiers capable of effectively discerning sexism and source intentions within the content.

It's worth noting that the decision to use only text stemmed from several considerations. Firstly, we observed a significant overlap in the visual content between both classes of memes. Images in both categories often bore striking resemblances, making it challenging to distinguish between them purely based on visual cues. Additionally, within the dataset labeled as containing sexist content, there were instances where seemingly neutral or innocuous images appeared, further complicating the visual classification process. Therefore, to maintain clarity and focus in our analysis, we opted to rely exclusively on textual data extracted from these memes. This approach allowed us to develop and refine text-based classifiers specifically designed to discern nuances of sexism and underlying intentions embedded within the meme content.

One of the primary innovations of this study lies in the utilization of three distinct training datasets for experimentation. Given that the data encompass two languages, English and Spanish, we employed two translation techniques to generate supplementary training datasets. For task resolution, we leveraged language models founded on Transformer architectures. Specifically, our approach entailed the utilization of two multilingual models: BERT [10] and RoBERTa [11]. The fine-tuning process of these models was meticulously optimized through

a comprehensive search for optimal hyperparameter values, as elaborated in Section 4.3. The models chosen for inclusion in the study were:

- `bert-base-multilingual-uncased` [10]: This model is the multilingual version of BERT.
- `xlm-roberta-base` [12]: This model is the multilingual version of RoBERTa.

In addition to using a single hard label, we have explored and trained the models from the perspective of the annotators using various strategies, which will be described in the following sections.

To compare the results, a baseline was constructed using the two selected models with default hyperparameters: a batch size of 32, a learning rate of 3e-5, a maximum sequence length of 128, and a weight decay of 0.01. Tables 3 and 4 show the F1 score achieved by the models.

Table 3
Baselines for Task 4

Model	F1 Score
BERT	0.6395
XLM-RoBERTa	0.6626

Table 4
Baselines for Task 5

Model	F1 Score
BERT	0.5481
XLM-RoBERTa	0.5520

4.1. Data Pre-processing

Data preprocessing in this study involved an initial comprehensive processing of textual content from memes. This processing included converting all text to lowercase, and removing links, usernames, and hashtag symbols ('#'). Subsequent empirical evaluations demonstrated that additional preprocessing steps did not yield significant improvements in test outcomes. Consequently, the final preprocessing strategy was refined to include only the conversion of text to lowercase.

4.2. Dataset Construction

The dataset, as illustrated in Tables 1 and 2 comprises a constrained quantity of instances. To address this constraint, various strategies were employed to increase the amount of data, similar to those used in data augmentation. We leveraged the fact that the data provided by the organization are in both English and Spanish by translating each instance into the opposite language, thereby creating a new dataset with double the data.

The other technique employed was back-translation [13], where each instance was translated into a different language (in this case, German) and then translated back into the original language. We leveraged the accuracy of ChatGPT [14] for this process. These augmented datasets were then combined with the original dataset to create three datasets for experimentation:

- `Original` : The training dataset provided by the organization.
- `Simple` : Original plus simple translation extension.
- `Back` : Original plus back-translation extension.

4.3. Hyperparameter Search

Hyperparameter search [15] is one of the most important steps for model fine-tuning. Various combinations of hyperparameters were evaluated, and the number of instances was reduced to shorten experimentation time. The Optuna library [16] in Python was used, which allows us to establish the hyperparameter space to find the best ones according to a specified metric.

Table 5
Hyperparameters space

Hyperparameter	Values
Batch Size	[8, 16, 32]
Learning Rate	[1e-05, 3e-05, 5e-05]
Weight Decay	[0.01, 0.1]

Table 5 shows the hyperparameter space, and Tables 6 and 7, show the best hyperparameters for each task.

Table 6
Best hyperparameters for Task 4

Hyperparameter	BERT	XLM-RoBERTa
Batch Size	16	16
Learning Rate	3e-05	1e-05
Weight Decay	0.1	0.01

Table 7
Best hyperparameters for Task 5

Hyperparameter	BERT	XLM-RoBERTa
Batch Size	32	32
Learning Rate	3e-05	1e-05
Weight Decay	0.1	0.01

4.4. Model Perspectives

Models training based on annotators' perspectives were employed, motivated by the abundance of features available within the dataset. This approach allows using only a specific perspective or combining as many as desired, although it is computationally more expensive. In our case, the eight perspectives with the most number of examples were chosen and trained with the three datasets mentioned above to create a final model by combining all the best perspectives.

Furthermore, to enhance the reliability and validity of our annotations, we intend to implement a sophisticated "learning with disagreement" approach. This method involves clustering annotators into groups based on specific characteristics stipulated by the organization, such as expertise in linguistics, cultural sensitivity, or familiarity with meme contexts. By grouping annotators in this manner, we aim to minimize biases and inconsistencies that may arise during the annotation process, thereby ensuring the quality and accuracy of our dataset. This structured approach not only enhances the robustness of our analysis but also reflects current best practices in managing subjective content classification tasks, where nuanced interpretations and contextual understanding play pivotal roles.

For each perspective, the data were balanced by means a undersampling technique. The selected perspectives are: gender("M", "F"), age("23-45", "18-22", "46+"), studies("Bachelor's degree", "High school degree or equivalent"), and ethnicity ("White or Caucasian").

Table 8
F1-Score results for the perspectives for Task 4

Model	Dataset	M	F	23-45	18-22	46+	Bachelor's	High school	White
BERT	Original	0.6580	**0.6205**	0.5920	0.6360	**0.6231**	0.6131	**0.6270**	0.6611
BERT	Simple	0.6213	0.6191	0.6255	0.6417	0.6223	0.6170	0.6150	**0.6679**
BERT	Back	**0.6613**	0.5699	**0.6710**	**0.6507**	0.5967	**0.6349**	0.6228	0.6393
XLM-RoBERTa	Original	0.6251	0.6264	0.6700	**0.6493**	0.6270	**0.6559**	**0.6553**	0.6471
XLM-RoBERTa	Simple	0.6445	**0.6457**	**0.6744**	0.6469	**0.6425**	0.6375	0.6387	**0.6768**
XLM-RoBERTa	Back	**0.6468**	0.6305	0.6637	0.6398	0.6299	0.6431	0.6501	0.6688

Table 9
F1-Score results for the perspectives for Task 5

Model	Dataset	M	F	23-45	18-22	46+	Bachelor's	High school	White
BERT	Original	**0.5375**	**0.5805**	**0.5609**	0.5132	0.5500	0.5357	**0.5402**	**0.5616**
BERT	Simple	0.5148	0.5170	0.5324	0.4991	0.4923	0.5033	0.5222	0.5204
BERT	Back	0.5225	0.5159	0.5314	**0.5460**	**0.5649**	**0.5516**	0.5095	0.5567
XLM-RoBERTa	Original	0.5322	0.5275	0.5474	0.5311	0.4906	0.5317	0.5372	**0.5738**
XLM-RoBERTa	Simple	**0.5472**	0.4938	**0.5572**	0.5517	0.5589	**0.5502**	0.5390	0.5635
XLM-RoBERTa	Back	0.5299	**0.5336**	0.5449	**0.5591**	**0.5760**	0.5432	**0.5479**	0.5072

In Tables 8 and 9, the selected models for each perspective are highlighted. Given our approach of treating the models separately, we choose the best model for each perspective based

on the dataset employed for its training. For example, the Model 1 is composed of perspective "M" with the training dataset "Back", "F" with "Original", "23-45" with "Back", "18-22" with "Back", "46+" with "Original", "Bacherlor's" with "Back", "High school" with "Original" and "White" with "Simple". The architecture of our ensemble models is structured as follows:

- `Model 1 and Model 4`: More efficient BERT models from each perspective for Task 4 and Task 5 respectively.
- `Model 2 and Model 5`: More efficient XLM-RoBERTa models from each perspective for Task 4 and Task 5 respectively.
- `Model 3 and Model 6`: More efficient BERT/XLM-RoBERTa models from each perspective for Task 4 and Task 5 respectively.

4.5. Ensemble Approach

This section describes our ensemble approach to obtain a single prediction based on the predictions obtained individually from each perspective. This strategy involves assigning a weight to each individual prediction through a joint weight search process to obtain overall F1.

Table 10
Weigth values space

Weigths
{0.5, 0.75, 1, 1.25, 1.5, 1.75}

In Table 10, the possible weight values assigned to the predictions of each perspective are displayed.

Table 11
Final combination for Task 4

Number	Model	M	F	23-45	18-22	46+	Bachelor's	High school	White	Overall F1 score
1	**Model 1**	**1.25**	**0.75**	**1.25**	**0.75**	**0.5**	**0.5**	**1.5**	**1.5**	**0.7294**
2	**Model 1**	**1.25**	**0.75**	**1.25**	**0.75**	**0.5**	**0.5**	**1.25**	**1.75**	**0.7288**
3	Model 2	1.75	0.5	1.5	0.75	0.75	0.5	0.5	1.75	0.7052
4	Model 2	1.75	0.5	1.25	0.75	0.75	0.5	0.75	1.75	0.7029
5	**Model 3**	**0.75**	**0.5**	**1.75**	**1.75**	**0.5**	**1**	**1**	**0.75**	**0.7224**
6	Model 3	1.5	1	0.75	1.75	0.5	0.5	1.5	0.5	0.7187

Table 12
Final combination for Task 5

Number	Model	M	F	23-45	18-22	46+	Bachelor's	High school	White	Overall F1 score
7	**Model 4**	**1.25**	**0.5**	**1.75**	**0.5**	**1**	**1**	**1.5**	**0.5**	**0.6352**
8	**Model 4**	**1**	**0.5**	**1.75**	**0.5**	**1**	**1**	**1.75**	**0.5**	**0.6329**
9	Model 5	0.5	1	1.75	0.75	1.75	0.75	0.5	1	0.6061
10	Model 5	0.5	1	1.5	1	1.5	0.75	0.5	1.25	0.6049
11	Model 6	1.5	1.5	1.5	0.75	1.25	0.5	0.5	0.5	0.6147
12	Model 6	1.75	1.25	1.5	0.5	1	0.5	1	0.5	0.6114

As observed in Tables 11 and 12, the approach based on training models using annotators' perspectives and the weight-based ensemble significantly improve the results over the baselines shown in Tables 3 and 4, respectively. The three best models for Task 4 found Table in 11 are:

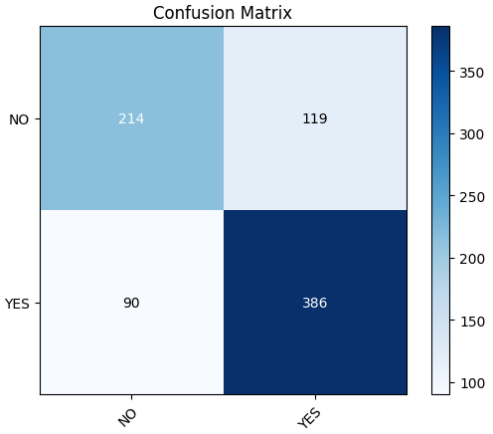
- `Run 1 (I2C-Huelva_1)`: Model 1 with balanced weigths number 1.
- `Run 2 (I2C-Huelva_2)`: Model 1 with balanced weigths number 2.
- `Run 3 (I2C-Huelva_3)`: Model 3 with balanced weigths number 5.

For Task 5, a run from Task 4 was chosen and its result was evaluated with the following models in Table 12:

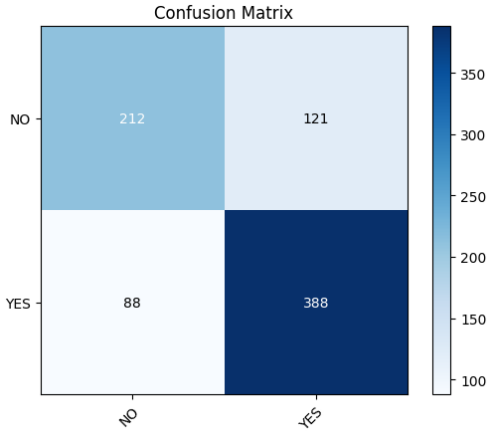
- `Run 4 (I2C-Huelva_1)` : Run 1 with Model 4 and balanced weigths number 7.
- `Run 5 (I2C-Huelva_2)` : Run 3 with Model 4 and balanced weigths number 7.
- `Run 6 (I2C-Huelva_3)` : Run 2 with Model 4 and balanced weigths number 8.

4.6. Error Analysis

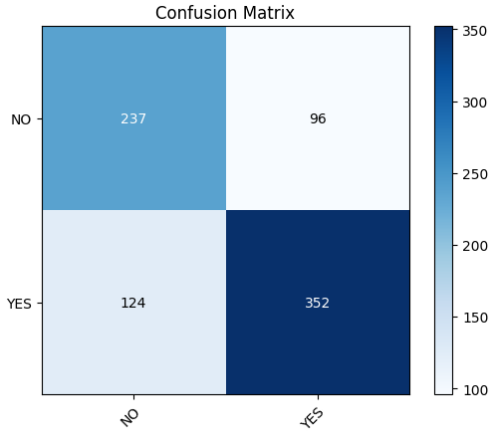
In this section, the errors of the models will be examined through the analysis of their confusion matrices. This approach will allow a detailed understanding of the models' performance, identifying both their successes and failures in classifying the samples. This critical evaluation will provide valuable information for improving the accuracy and reliability of the models, thereby contributing to the advancement of the field of study.



(a) Confusion matrix for Run 1



(b) Confusion matrix for Run 2



(c) Confusion matrix for Run 3

Figure 1: Confusion matrices for Task 4

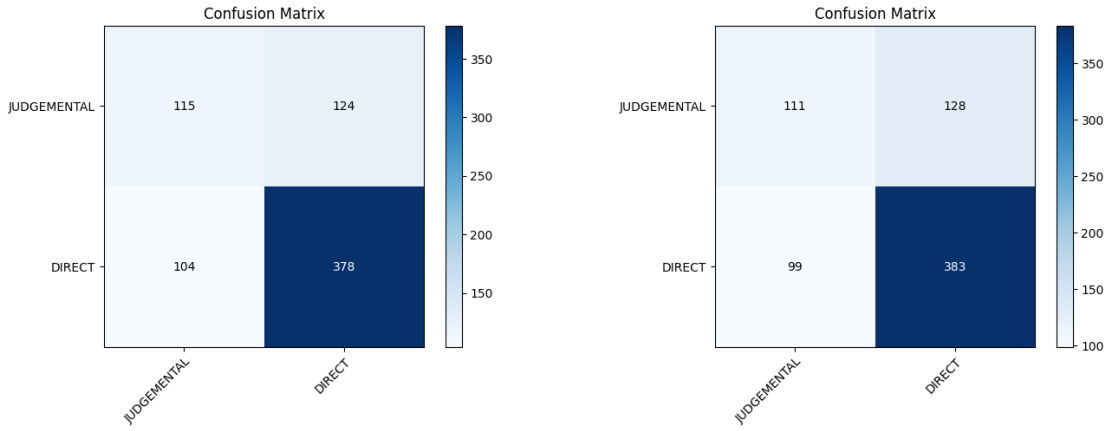
For Task 4, all figures show similar overall performance patterns, with a notably high proportion of correct predictions (TP and TN) compared to incorrect ones (FP and FN). This suggests a consistent ability of the models to accurately classify samples from both positive and negative classes. However, differences between the models reveal distinct trends. Figure 1c exhibits a slightly higher number of true positives (TP) compared to Figures 1a and 1b, indicating a potentially better capability of the mixed BERT/XLM-RoBERTa model to identify positive class samples. Conversely, Figures 1a and 1b demonstrate similar trends in false positives (FP) and false negatives (FN), while Figure 1c shows a slightly higher proportion of false negatives (FN). These discrepancies could stem from variations in model architectures (solely BERT vs. mixed BERT/XLM-RoBERTa) and the specific characteristics of the dataset and training processes. Combining these observations may inform future research on model selection and optimization for specific classification tasks.

Table 13
Examples labeled for Task 4 (Original text highlighted and translation)

Text	Test	Run 1	Run 2	Run 3
SP: metro q estilo de vida alexa recomienda a una madre asesinar a sus hijos: amazon pide disculpas por ”error en la configuración” el asistente inteligente ofreció una respuesta polémica cuando una mujer le preguntó sobre ”cómo evitar que los niños rían” alexa le pusieron ese nombre por no llamarla skynet más en cuantarazon.com EN: metro q lifestyle alexa recommends a mother to kill her children: amazon apologizes for ”configuration error” the smart assistant offered a controversial response when a woman asked her about ”how to stop children from laughing” alexa was given that name for not calling her skynet more on cuantarazon.com	0	1	1	1
SP: ME DIJO QUE ME FUERA A FREGAR memegenerator.es EN: TOLD ME TO GO SCRUB memegenerator.es	1	0	0	0

Table 13 illustrates the difficulty in classifying certain texts accurately. In the first example, the text addresses controversial topics and specific entities (e.g., Amazon, Alexa), which can lead to misclassification due to lack of context. In the second example, it demonstrates the variety of topics and the presence of humorous elements that can complicate the task of automated classification.

For Task 5, Run 4 and 5 are identical, whereas Run 6 is based on the same model but with different weights used for prediction. Therefore, we will only compare the confusion matrices of Runs 4 and 6.



(a) Confusion matrix for Run 4

(b) Confusion matrix for Run 6

Figure 2: Confusion matrices for Task 5

Both Figures 2a and 2b are based on the same BERT architecture. The differences in error distribution and the final model's value system suggest that the models have been trained slightly

differently. However, they share fundamental similarities due to their common foundation in BERT and their identical matrix structure.

Table 14
Examples labeled for Task 5 (Original text highlighted and translation)

Text	Test	Run 4	Run 6
SP: La mecánica es solo para hombres, toma mi bolso, sé más que tú de motores, putito. EN: Mechanics is only for men, take my bag, I know more about engines than you, little faggot.	0	1	1
SP: pero los hombres no tienen los mismos derechos que las mujeres, como el derecho a compartir su opinión sobre el aborto. **EN**: yet men don't have the same rights as women like the right to share their opinion on abortion.	1	0	0

In the first example of Table 14 the misclassification of this text could be attributed to the lack of consideration for cultural, social, and linguistic context, as well as the incapacity of an automated algorithm to capture nuances in tone and communicative intent. In the second example, the text discusses gender rights with a specific focus on the disparity in opinions on abortion, which introduces sensitive and context-dependent themes. These cases demonstrate the challenge of classifying texts with similar vocabulary but different contexts or fragmented and disjointed content.

5. Results

This section presents the results obtained from the competition, detailing the performance of our top submissions across various tasks. The metrics to be evaluated for the competition are:

- `Hard-Hard`: The 'hard' labels are derived from the annotators' labels using probabilistic thresholds specific to each task.
  - `Task 4`: The class annotated by more than 3 annotators is selected.
  - `Task 5`: The class annotated by more than 2 annotators is selected.

  Items without a majority class are removed from the evaluation. The official metric is the original ICM, and F1 (the harmonic mean of precision and recall) is also used for comparison.
- `Soft-Soft`: Compares the probabilities assigned by the system with those assigned by the human annotators. As in the previous case, ICM-soft will be used as the official evaluation metric.

Our final models returned a percentage corresponding to the Soft-Soft measure. For the Hard-Hard measure, it was filtered if that percentage was greater than 50%. Tables 15 to 18 show the official results obtained by the submitted runs.

Table 15
Ranking of participants for Task 4 Hard-Hard

Rank	Run	ICM-Hard	ICM-Hard Norm	F1_YES
1	RoJiNG-CL_3	0.3182	0.6618	0.7642
2	RoJiNG-CL_2	0.2272	0.6155	0.7437
3	RoJiNG-CL_1	0.1863	0.5947	0.7274
4	**I2C-Huelva_2**	**0.1313**	**0.5668**	**0.7241**
5	**I2C-Huelva_1**	**0.1166**	**0.5593**	**0.7154**
-	-	-	-	-
9	**I2C-Huelva_3**	**0.0987**	**0.5502**	**0.6933**
-	-	-	-	-
53	melialo-vcassan_1	-0.8109	0.0876	0.5316

Table 16
Ranking of participants for Task 4 Soft-Soft

Rank	Run	ICM-Soft	ICM-Soft Norm	Cross Entropy
1	Victor-UNED_1	-0.2925	0.4530	1.1028
2	Victor-UNED_2	-0.3135	0.4496	1.2834
3	Elias&Sergio_1	-0.3225	0.4482	0.9903
4	**I2C-Huelva_3**	**-0.3263**	**0.4476**	**1.5189**
5	**I2C-Huelva_1**	**-0.3390**	**0.4455**	**1.4096**
6	**I2C-Huelva_2**	**-0.3446**	**0.4446**	**1.4112**
-	-	-	-	-
37	CNLP-NITS-PP_1	-2.6987	0.0662	1.3445

Table 17
Ranking of participants for Task 5 Hard-Hard

Rank	Run	ICM-Hard	ICM-Hard Norm	Macro F1
1	Victor-UNED_1	-0.2397	0.4167	0.3873
2	**I2C-Huelva_2**	**-0.2535**	**0.4119**	**0.4761**
3	Victor-UNED_2	-0.2668	0.4073	0.3850
4	**I2C-Huelva_3**	**-0.2772**	**0.4036**	**0.4714**
5	**I2C-Huelva_1**	**-0.2880**	**0.3999**	**0.4714**
-	-	-	-	-
22	epistemologos_1	-8.7012	0.0000	0.0557

Table 18
Ranking of participants for Task 5 Soft-Soft

Rank	Run	ICM-Soft	ICM-Soft Norm	Cross Entropy
1	Victor-UNED_2	-1.2453	0.3676	1.6235
-	-	-	-	-
8	melialo-vcassan_3	-2.0653	0.2804	1.5295
9	melialo-vcassan_1	-2.6821	0.2148	1.6291
10	**I2C-Huelva_3**	**-2.7996**	**0.2023**	**3.9604**
11	**I2C-Huelva_2**	**-2.7997**	**0.2023**	**3.9857**
12	**I2C-Huelva_1**	**-2.8007**	**0.2022**	**3.9735**
-	-	-	-	-
17	Penta-ML_2	-5.9832	0.0000	5.4845

6. Conclusions and Future Works

In this study, the identification of sexism and source intentions in memes was explored, and the findings were presented at the EXIST 2024 competition. Various methodologies were evaluated to develop the most effective classifiers, employing both conventional models based on hard voting and innovative models utilizing the Learning with Disagreement (LeWiDi) approach. It was found that the latter approach, which incorporates perspectives from diverse annotators, exhibited superior performance compared to the traditional models. Consequently, notable rankings were achieved: fourth place was secured in both the Hard-Hard and Soft-Soft measures for Task 4, and second and tenth places were obtained for Task 5, respectively.

Looking forward, the methodologies applied in this research are planned to be refined, and the focus is intended to be expanded to include image analysis. This enhancement aims to develop a more comprehensive model that integrates visual elements with textual analysis, thereby advancing the capability to detect sexist content in memes.

Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]"*, PID2021-123983OB-I00, funded by *MCIN/AEI/10.13039/501100011033/* and by *"ERDF/EU"*.

References

[1] 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, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[3] Y. Li, X. Li, Y. Yang, R. Dong, A diverse data augmentation strategy for low-resource neural machine translation, Information 11 (2020) 255.
[4] 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.
[5] 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, arXiv preprint arXiv:2106.04908 (2021).
[6] S. Frenda, A. Pedrani, V. Basile, S. M. Lo, A. T. Cignarella, R. Panizzon, C. Sánchez-Marco, B. Scarlini, V. Patti, C. Bosco, et al., Epic: Multi-perspective annotation of a corpus of irony, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13844–13857.
[7] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[8] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.
[9] D. M. Tax, M. Van Breukelen, R. P. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying?, Pattern recognition 33 (2000) 1475–1485.
[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[12] 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).
[13] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.
[14] Y. Gao, R. Wang, F. Hou, How to design translation prompts for chatgpt: An empirical study, arXiv e-prints (2023) arXiv–2304.
[15] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing 415 (2020) 295–316.
[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.