EXIST 2025 Participation Report
Mateusz Król, Teamname: Güey Pingüino

1. Introduction

The goal of this project is to develop and evaluate transformer-based models to automatically detect and categorize sexist content in social media posts, as part of the EXIST 2025 challenge.

I addressed these tasks using English and Spanish datasets and evaluated performance using standard classification metrics and the ICM score.

2. Data Preprocessing

At the beginning, in order to provide better efficiency of the models I introduced some preprocessing steps of the dataset. It included:

- Removing **URLs**, **mentions**, **hashtags**
- Removing **emojis** and **special characters**
- Converting text to **lowercase**
- Ensuring **language consistency** (per task)
- Tokenization using language-specific BERT/Roberta tokenizers

This ensures clean, uniform inputs for training robust models.

3. Learning Models

I experimented with **three different transformer-based models**, training each using the **HuggingFace Transformers library** with early stopping and evaluation on a development set. These architectures operate by processing input text through multiple layers of self-attention and feed-forward networks. Each transformer block includes a multi-head self-attention layer followed by a position-wise feed-forward layer, both of which are followed by layer normalization and residual connections to enhance stability and learning efficiency. The final representation of the input is then passed through a classification head, in this case a simple linear layer followed by a softmax or sigmoid activation depending on the task setup to generate the final prediction. The models were additionally fine-tuned using training data from the EXIST 2025 dataset. I also confronted the fine-tuned models with the LoRA tuning methods to optimize the learning process.

The table below presents the description and explanation of different hyperparameters used in the training process for different models.

Tab 1. Parameters overview:

Parameter	Purpose	Effect on Fine-Tuning
Learning_rate	Step size in weight updates	Controls how fast/slow the model learns
Batch_size	Batch size per device	Balances speed, memory, and gradient stability
Weight_decay	Regularization to avoid overfitting	Encourages simpler models
Eval_strategy	When to evaluate during training	Evaluates performance at each epoch
Early_stopping	Stop early if no improvement	Prevents overfitting and saves compute

Rank	The dimensionality of the low-rank matrices	Not used by transformers directly, used in LoRA
Num_train_epoch	Number of epochs	Controls training duration
Max_grad_norm	Gradient clipping	Prevents gradient explosions
LR_scheduler _type	Controls how the learning rate changes during training.	Helps the model converge more effectively.
Lora Alpha	Scales the output of the low-rank adapters.	Helps balance between original weights and LoRA updates during fine-tuning.
Lora Dropout	Adds dropout to the LoRA layers during training.	Improves regularization, reduces overfitting.

4. Results

a) Task 1

This task involves **binary classification** of tweets to determine whether a given tweet contains **sexist content**. The label set is:

- **YES** → sexist content is present (explicit, reported, or critical),
- **NO** → no sexist content.

The tables below present the hyperparameters used for the learning process and evaluation metrics obtained for task 1 classification.

Tab. 2. Configuration parameters for Task 1

Parameter	Value (Model 1)	Value (Model 2)	Value (Model 3)
Learning_rate	2e-5	5e-6	2e-5
batch_size	32	16	32
Weight_decay	0.05	0.01	0.05
Eval_strategy	“epoch”	“epoch”	“epoch”
Early_stopping	2	2	2
Rank	16	16	8
Num_train_epochs	10 (ended after 5th)	10	10 (ended after 6th)
Max_grad_norm	1.0	0.5	1.0

Tab 3. Task 1 results

#	Lang	Model Name	F1	Accuracy	Precision	Recall	Notes
1	EN	cardiffnlp/twitter -roberta-base	0.824	0.844	0.814	0.835	Robust general English Twitter model
2	EN	cardiffnlp/twitter -roberta-base	0.839	**0.860**	0.844	0.835	More stable, higher accuracy
3	SP	dccuchile/bert-base-spanish-wwm-cased	**0.855**	0.849	**0.876**	0.835	Tailored for Spanish, strongest F1

Observations

- **Model 2** performed best on English tweets, striking a strong balance between **F1** and **accuracy**, indicating the benefit of using **lower learning rates** and **smaller batch sizes** for stable training.

- **Model 3 (Spanish BERT)** achieved the **highest F1 score overall (0.855)**, showing the value of language-specific models for detecting nuanced, culture-dependent expressions of sexism.

- All models used early stopping and evaluation strategies to prevent overfitting and ensure robust generalization.

Conclusion

- **Language-specific fine-tuning** (e.g., Spanish BERT for Spanish tweets) substantially boosts performance in multilingual sexism detection.

- Careful tuning of training parameters (like learning rate, batch size) improves performance even within the same model architecture.

- Our models show strong potential to support automated moderation or analysis systems aimed at identifying and understanding online sexism.

b) Task 2

This task involves **multiclass classification** of tweets to determine the type of sexism content. We need to specify if the tweet belongs to the group of texts like: Reported, Direct or Judgemental.

Table 4. Configuration parameters for Task 2

Parameter	Value (Model 4)	Value (Model 5)	Value (Model 6)
Learning_rate	1e-5	5e-6	5e-6
batch_size	4	8	4
Weight_decay	0.01	0.01	0.01
Eval_strategy	“epoch”	“epoch”	“epoch”
Early_stopping	4	4	4
Rank	8	4	4
Num_train_epochs	10	10	10 (ended after 9th)
Max_grad_norm	1.0	1.0	1.0
lr_scheduler_type	“cosine”	“cosine”	“cosine”

Tab 5. Task 2 results

#	Lang	Model Name	F1 Score	Acc.	Prec.	Rec.	Notes
4	EN	cardiffnlp/twitter-roberta-base	**0.635**	**0.712**	**0.654**	**0.624**	Smaller batch improved generalization (best F1)
5	EN	Bert-base-uncased	0.558	0.671	0.567	0.557	More overfit, better loss but lower F1
6	SP	dccuchile/bert-base-spanish-wwm-cased	0.566	0.637	**0.578**	0.558	Lower F1, possibly due to smaller dataset or less pretraining on relevant tasks

Key Takeaways

1. **Batch size matters**: Reducing the batch size (from 8 to 4) significantly improved F1 from 0.54 → 0.64. Smaller batches can act as a regularizer.
2. **Cosine LR scheduling**: May have contributed to better convergence stability across all models.
3. **Precision vs. Recall Tradeoff**: All models show reasonably balanced precision and recall, with better recall in the best-performing setup.
4. **Spanish model struggles**: Despite using a native Spanish BERT, performance was slightly lower. Could be due to:
    - Dataset size
    - Label imbalance
    - Semantic variation in Spanish-language tweets

c) Task 3

This task detects the presence of one or more of five fine-grained sexism types ('IDEOLOGICAL-INEQUALITY', 'OBJECTIFICATION', 'STEREOTYPING-DOMINANCE', 'MISOGYNY-NON-SEXUAL-VIOLENCE', 'SEXUAL-VIOLENCE'). Evaluation emphasizes macro-averaged metrics and a hierarchical metric called ICM (Integrated Contextual Measure).

Table 6. Configuration parameters for Task 3

Parameter	Value (Model 7)	Value (Model 8)	Value (Model 9)
Learning_rate	5e-6	1e-6	5e-6
batch_size	16	8	16
Weight_decay	0.05	0.01	0.05
Eval_strategy	“epoch”	“epoch”	“epoch”
Early_stopping	4	4	4
Rank	16	8	16
Num_train_epochs	8	10	8
Lora_alpha	16	8	16
Lora_dropout	0.05	0.05	0.05

Table 7. Task 3 results

#	Lang	Model Name	F1 Macro	ICM	Prec. macro	Rec. macro	Notes
7	EN	cardiffnlp/twitter-roberta-base	**0.681**	**-0.03**	**0.704**	**0.624**	Strong macro-F1, but ICM < 0 indicates structural mismatch
8	EN	cardiffnlp/twitter-roberta-base	0.661	0.02	0.700	0.664	Solid F1 Macro, model's predictions **mostly respect the label structure** and are **logically consistent**
9	SP	dccuchile/bert-base-spanish-wwm-cased	0.709	0.09	**0.689**	0.748	Best balance, higher recall & ICM

Insights

- **model9 (Spanish)** performs **best overall**, particularly in terms of:
    - **Hierarchical ICM score**, which rewards correct predictions with consideration for category relationships.
    - **Recall**: Captures more true positive sexist behaviors.
    - Slightly better **F1** and **loss**, showing better generalization.

- **model7 (English)**, although strong in **precision**, shows **negative ICM**, indicating **structural mismatches** (i.e., predicting incorrect or incompatible combinations of labels). This matters in hierarchical tasks like Task 3.

- **model8 (English)** used a **smaller batch size and lower LR**, which may lead to underfitting or slower convergence.

6. Overall Conclusions

1. **English Twitter-RoBERTa excels in binary and multi-class tasks**, suggesting strong contextual understanding and domain adaptation to social media text.

2. **Spanish BERT performs best in multi-label classification**, especially in ICM, highlighting its strength in nuanced, hierarchical classification with structural constraints.

3. **ICM metric is crucial in Task 3** — some high F1 models still failed due to logically inconsistent label predictions.

4. **Class imbalance** and label sparsity are evident issues, particularly in Task 2 and 3.