Sexism Detection in Social Media Using Transformers

Mateusz Król, Universitat Politecnica de Valencia

Introduction & Objectives

Binary Classification

Detect if a post contains sexist content or not.

Multi-Class Classification

Identify general types of sexism like stereotyping or objectification.

Multi-Label Classification

Detect multiple specific sexist behaviors simultaneously.

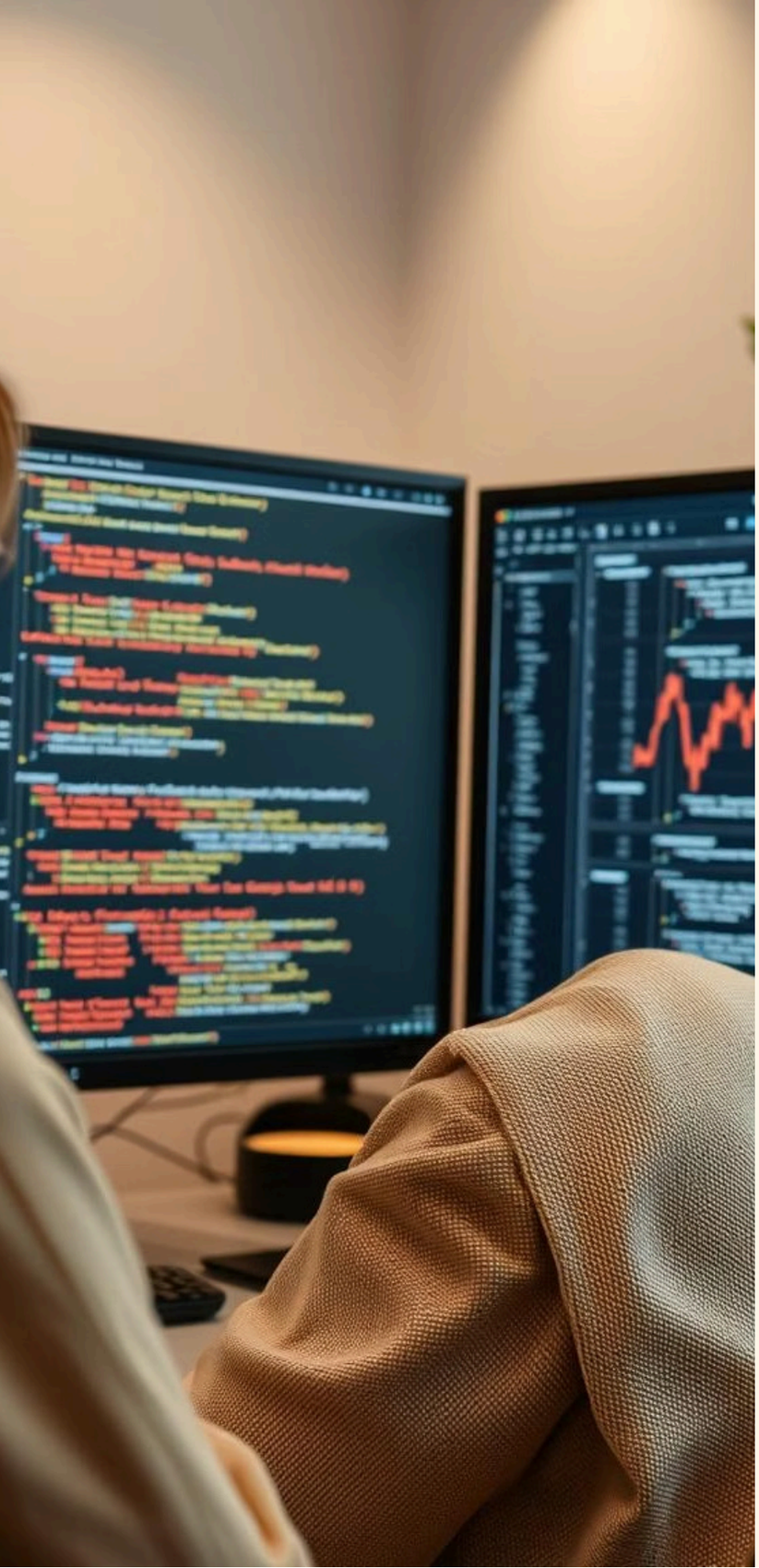
EXIST 2025 Dataset & Preprocessing

Dataset Overview

- Posts from Twitter in English and Spanish
- Labeled for sexism-related content
- Supports binary, multi-class, and multi-label tasks

Preprocessing

- Removed URLs, mentions, hashtags, emojis
- Lowercased text and ensured language consistency
- Tokenized with language-specific BERT/Roberta



Model Setup and Training Parameters

Parameter	Values	Effect
learning_rate	2e-5, 1e-5, 5e-6	Controls training speed
batch_size	8, 16, 32	Balances speed and stability
weight_decay	0.01, 0.05, 0.1	Prevents overfitting
early_stopping_patience	2, 3, 4	Stops training if no improvement
num_train_epochs	10	Training duration
max_grad_norm	1.0, 0.5	Gradient clipping, prevents gradient explosions

I experimented with **three different transformer-based models**, training each using the **HuggingFace Transformers library** with early stopping and evaluation on a development set with LoRA optimization and threshold tuning included.

Task 1: Binary Classification Results

Model Name	F1 Score	Accuracy	Precision	Recall	Notes
cardiffnlp/twitter-roberta-base	**0.824**	0.844	0.814	0.835	Robust, general English Twitter model
cardiffnlp/twitter-roberta-base (lower LR, batch size 16)	**0.839**	**0.860**	0.844	0.835	More stable, higher accuracy
dccuchile/bert-base-spanish-wwm-cased	**0.855**	0.849	**0.876**	0.835	Tailored for Spanish, strongest F1

English Models

Model 2 achieved best F1 (0.839) and accuracy (0.860).

Spanish Model

Model 3 had highest F1 (0.855), showing language-specific strength.

Key Insight

Lower learning rates and batch sizes improved stability and accuracy.

Task 2: Multi-Class Classification Overview

Model Name	F1 Score	Accuracy	Batch-Size	Learning rate	Notes
bert-base-uncased (English)	**0.54**	0.65	16	1e-5	More overfit, better loss but lower F1
cardiffnlp/twitter-roberta-base	**0.64**	**0.71**	8	1e-5	Smaller batch improved generalization (best F1)
dccuchile/bert-base-spanish-wwm-cased	**0.57**	0.64	4	5e-6	Lower F1, struggling possibly due to too small batch_size and too big generalization

English Model

Smaller batch size improved F1 from 0.54 to 0.64.

Balanced precision and recall observed.

Spanish Model

Lower F1 (0.57), possibly due to label imbalance.

Semantic variation affected performance.

Task 3: Multi-Label Classification & Evaluation

Base Model	Batch Size	LR	Epochs	F1 Macro	ICM Score	Notes
cardiffnlp/twitter-roberta-base	8	1e-6	10	0.661	0.024	Solid F1 Macro, model's predictions **mostly respect the label structure** and are **logically consistent**
cardiffnlp/twitter-roberta-base	16	5e-6	8	**0.682**	**-0.038**	Strong macro-F1, but ICM < 0 indicates structural mismatch
dccuchile/bert-base-spanish-wwm-cased	16	5e-6	8	**0.710**	**0.092**	Best balance, higher recall & ICM

1

Spanish Model

Best F1 (0.71) and positive ICM (0.092), strong hierarchical alignment.

2

English Model

Good precision but negative ICM, indicating label structure issues.

3

Recommendations

hierarchy-aware constraints and label-wise threshold tuning might strengthen the training process

Overall Conclusions & Future Improvements

Conclusions

Language-specific fine-tuning boosts performance.

English RoBERTa performs quite good in binary and multi-label tasks.

Spanish BERT leads in multi-label hierarchical detection.

Future Work

Ensemble Models:

Combine predictions from top-performing English and Spanish models using weighted voting or stacking.

LoRA Optimization Tuning:

Adjust rank, alpha, and dropout to find optimal low-rank adaptations per task.

Threshold Tuning (Task 3):

Use validation-based label-wise thresholding instead of fixed 0.5 for better ICM performance.

Thank you for your attention!