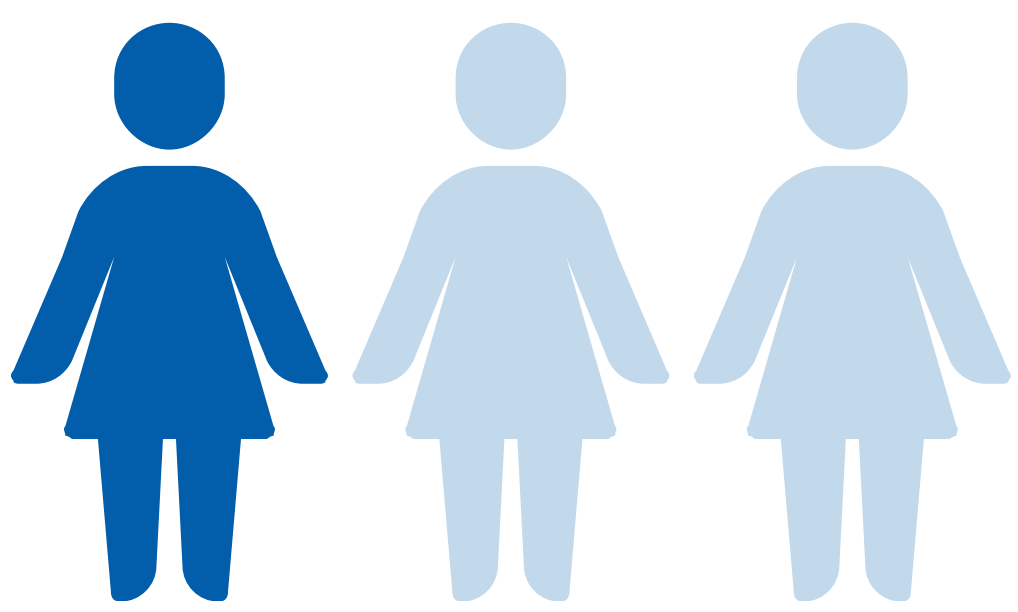


Detecting Sexist Language in Spanish Social Media Using NLP Models



Developed with UNICC, this project applies natural language processing to detect sexist content in Spanish-language social media. We fine-tuned transformer models (BETO, Josefina, XLM-T) on a unified dataset, with extensive preprocessing and label standardization. The resulting binary and multiclass classifiers offer scalable, context-aware moderation tools that support ethical AI use and align with UN SDG 5.2.

Authors: Sebastián Felipe Zambrano Julio, Yihang Li, Juan Martin Echeverri, Gizela Susan Thomas, Pablo Camacho Fernández, Alejandro Felipe Pérez Vargas



Over 1 in 3 Women Experience Sexual Harassment Online

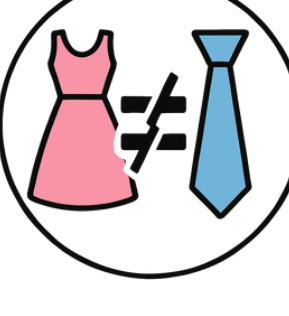
Introduction

Sexist abuse is rife on X [Twitter] and other networks, yet automated moderation still prioritises English content. To close this gap for more than 500 million Spanish speakers, our team, in partnership with UNICC and building on work from the Universitat Politècnica de València, consolidated three public tweet datasets (EXIST 2021, HatEval 2019, ManRosSexism) into the largest cleaned Spanish sexism corpus to date. We standardised noisy labels into five clear categories, translated all English tweets into Spanish, and benchmarked leading Spanish-specific and multilingual transformers, laying the groundwork for a transparent, scalable detection pipeline.

Objective

Deliver a unified Spanish sexism corpus with five refined categories and an open, reproducible transformer pipeline that platforms, NGOs, and the UN can use to flag abusive content.

Sexism Categories Used for Classification

- **Sexual Insults & Objectification**
Explicit sexual remarks, body shaming, or comments reducing women to physical attributes.
- **General Hostile Language**
Aggressive, profane, or demeaning content without specific gendered references.
- **Victim Blaming & Justification**
Posts that downplay abuse, blame victims, or excuse sexist behavior as justified.
- **Gendered Stereotypes & Insults**
Language reinforcing traditional gender roles or mocking feminist beliefs and identity.
- **Threats & Physical Harm**
Tweets that suggest violence, encourage harm, or issue direct threats toward women.

Data Sources			
Feature	EXIST (2021)	HatEval (2019)	ManRosSexism
Languages	English, Spanish	English, Spanish	Spanish
Size	11,325 tweets	~15,000 tweets	3,077 tweets
Label Type	5-class multiclass	Binary + subtasks	Binary + 5 subtypes

Data Cleaning & Combining - We merged 3 datasets removed duplicates, filtered off-topic content, & standardized formats for consistent schema and labeling.

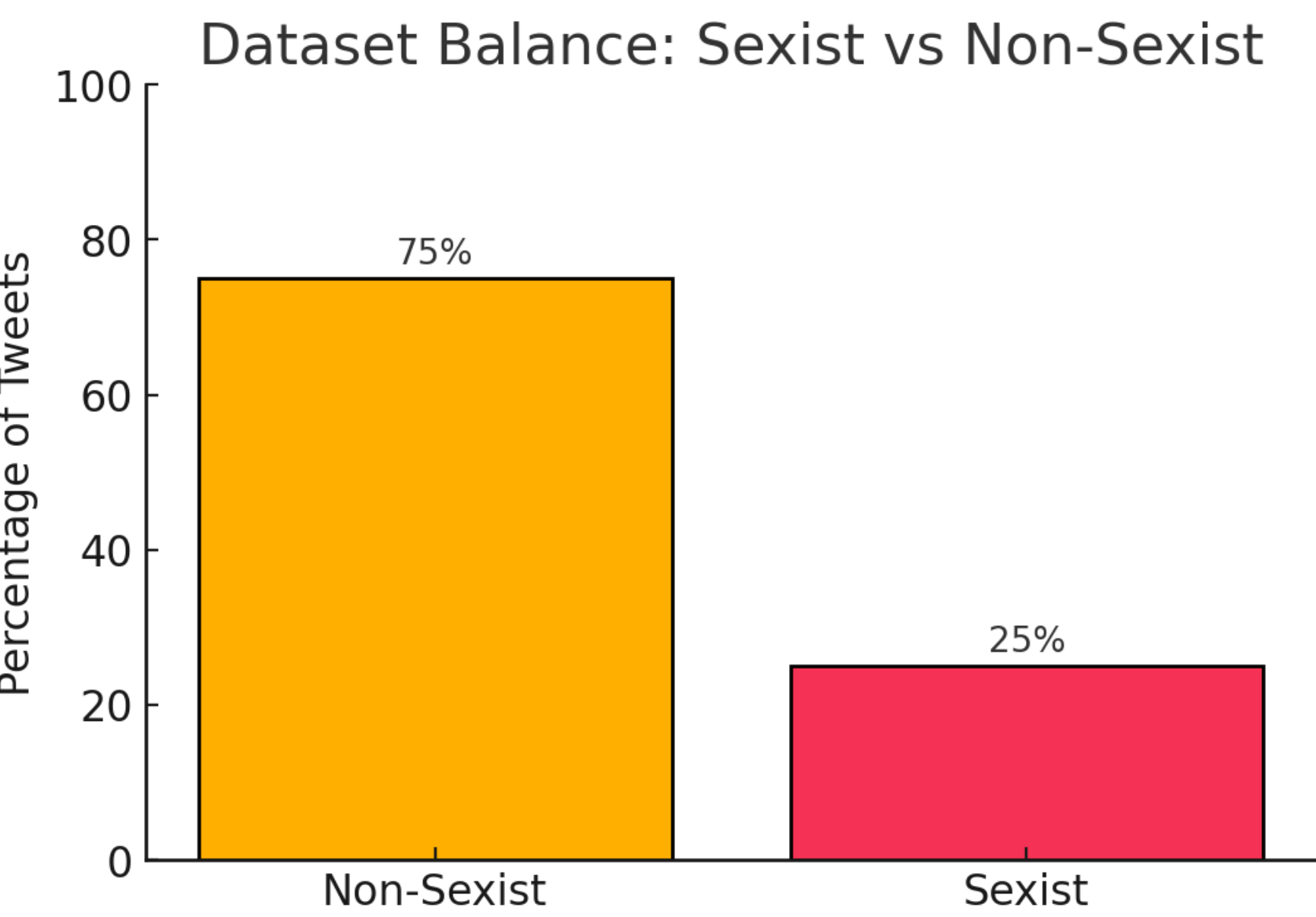
Translating the Data - All tweets were translated into both English & Spanish using the Google Translate feature in Excel.

Refining Data Labels - We consolidated 16 overlapping subcategories into 5 clear sexism types to reduce noise & improve classification performance.

“The catgirl could probably be argued to be a bit of a bimbo, but she’s got some of the smallest tits in the game, so I guess she’s safe.”

“The View. Empty head emotional woman’s daily bitchcraft lesson.”

“It was rape because she regretted it afterwards, duh. There’s no way she could have sex with a nerd. Sheesh.”



Benchmarking & Model Comparison Analysis

We benchmarked against Josefina, a RoBERTa-based transformer fine-tuned by UPV and UNICC on Spanish tweets and pre-trained with National Library of Spain text, which gave us a strong domain baseline . Beyond Josefina, we reviewed **fourteen transformer candidates** spanning multilingual backbones (mBERT, XLM RoBERTa, mDeBERTa, XLM T) and Spanish-specific models such as BETO, MarIA, RoBERTuito and BERTIN, plus lightweight LLMs for prompt experiments . After quick pilots in Google Colab we fine-tuned the most promising models, discarding others when accuracy lagged or GPU limits blocked training, a narrowing that kept our comparison fair and resource-aware

LLM Integration Strategy

We tested local and cloud large language models for classifying sexist content. Locally we ran LLaMA 3 and Mistral with Ollama, and in the cloud we used OpenAI GPT-4o-mini in zero shot mode. The setup was scalable and flexible, but local inference suffered high latency and all models lost precision without fine tuning. This proof of concept shows the value of adding domain specific tuning in future work.

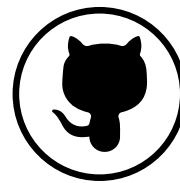
Modeling Approach

Tokenization

Model Loading

Model Fine-Tuning

Hyperparameter Optimization

 <https://github.com/pablo60217/capstoneprojectunicc>

Limitations

Mismatch between Latin American data and Spain-trained Spanish models, limited computation that forced training on only 50 percent of the corpus and curtailed tuning or scaling, plus no formal gender or race-studies background on the team, all skewed label interpretation.

Future Work

Future work includes adapting the model for policy-compliant moderation tools, building a public-facing web application, expanding coverage to more Spanish dialects and regions, and integrating large language models to enhance performance.

Ethical Considerations

This study analyses public tweets that contain explicit sexist language and was conducted under a responsible-AI pledge to respect academic integrity and sensitive social issues

Binary Results

Model	Accuracy	Macro F1	Precision	Recall
BETO	0,83	0,78	0,82	0,83
Josefina	0,81	0,8	0,75	0,87
XLM-T	0,83	0,77	0,82	0,83
BERTIN	0,75	0,63	0,74	0,75
GPT-4o-mini	0,71	0,63	0,49	0,86

Multi Class Results

Model	Accuracy	Macro F1	Precision	Recall
XLM-T	0,82	0,51	0,82	0,83
BETO	0,79	0,37	0,6	0,57
BERTIN	0,77	0,17	0,15	0,2

