Detecting Sexist Language Online

AI Against Sexism: NLP Approaches for Moderating Spanish-Language Social Media

Mentor: Hind Azegrouz

Ethics Statement: We hereby certify that this report represents our original work, developed independently unless otherwise indicated and referenced. We acknowledge the responsible use of AI tools, such as ChatGPT for drafting, editing, and refining written content throughout the project. All contributions made by these tools were guided by our own judgment, and their outputs were critically reviewed and edited to ensure academic integrity and alignment with ethical research practices and respectful engagement with sensitive social issues.

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

Executive Summary

This report, developed in collaboration with the United Nations International Computing Centre (UNICC), delivers a natural language processing solution to detect sexist content in Spanish-language social media. Addressing the growing need for ethical moderation tools, we built and fine-tuned transformer-based models, including BETO, Josefina, and XLM-T, to classify harmful language across both binary and five-category multiclass frameworks. Drawing from three open-source datasets (EXIST, HatEval, and ManRosSexism), we created a unified, multilingual corpus and applied extensive preprocessing and label standardization. Our final models achieved strong precision and recall, enabling scalable, fair, and context-sensitive moderation of online spaces. This work supports UN Sustainable Development Goal 5.2 and sets the foundation for real-world deployment in digital safety applications.

1. Introduction

Social media is a powerful tool that allows people to connect, share opinions, access news, and express creativity. Among these platforms, X (formerly Twitter) stands out for its fast-paced, real-time content, with users sharing thoughts in 280 characters or less. Tweets often reflect and influence broader cultural and political trends, including both progressive movements and more harmful ideologies. Unfortunately, not all online communication is positive. Social media platforms have increasingly become venues for hate speech, cyberbullying, and technology-facilitated violence, particularly against women and marginalized groups.

According to the Oxford English Dictionary, sexism refers to prejudice, stereotyping, or discrimination, typically against women, based on sex. On platforms like Twitter, sexism can manifest in many forms including abuse, misogyny, sexually explicit comments, and violent threats. These expressions are often embedded in cultural contexts and slang, making them difficult to detect with simple filters. To address this challenge, our team partnered with the United Nations International Computing Centre (UNICC) to build a robust language model capable of identifying sexist language in Spanish-language tweets. This work builds upon a 2019 to 2024 initiative by the Universitat Politècnica de València, where researchers developed and fine-tuned transformer-based models, Josefina and Rosita, to detect and classify sexism in Mexican Twitter data.

In this project, we expand and refine that work by:

- Translating and enriching the dataset with English-Spanish tweet pairs
- Consolidating the original 16 label categories into 5 meaningful classes
- Evaluating and comparing performance across state-of-the-art multilingual and Spanish specific transformer models

Our goal is to contribute to a safer and more equitable online environment by enabling more accurate, culturally aware detection of sexist language on social media.

2. Related Work

Advances in natural language processing have significantly improved the detection of online hate speech, including racism, xenophobia, anti-immigrant rhetoric, and sexism. Early approaches often relied on rule-based systems or bag-of-words models, but recent work has shifted toward machine learning techniques, such as logistic regression, support vector machines, and more recently, deep learning models like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. For this project, we build on the work conducted by the Universitat Politècnica de València (UPV), which used transformer-based models, such as BETO and RoBERTa, to detect sexist content in Spanish-language tweets from Latin America. These models were fine-tuned using expert-labeled datasets and adapted to capture the linguistic and cultural nuances of the region. UPV's initiative emphasized robust modeling over dataset creation and aligned with broader goals of digital equity, particularly those outlined in the Sustainable Development Goals (SDGs 5.2 and 5.b), which call for the elimination of violence against women and the promotion of technology for gender equality. Our project advances this work by improving model performance, standardizing classification labels, and expanding dataset coverage, all with the aim of supporting these goals through more effective detection of sexist content in Spanish-language social media.

3. Data

This project builds on prior work by the Universitat Politècnica de València and UNICC, leveraging annotated datasets on online sexism and hate speech. These datasets offer a multilingual corpus of English and Spanish tweets for training and evaluation.

- **Source**: UN-ICC GitHub Repository
- **Format**: CSV files with tweet ID, text, and annotation labels
- **Size**: ~100.000 tweets

3.1 Data Sources

The following table summarizes key characteristics of the core datasets used, which serve as the foundation for our model training and evaluation. A detailed description of each dataset can be found in Appendix A.

Summary of Core Datasets Used for Sexism and Hate Speech Detection

Feature	EXIST (2021)	HatEval (2019)	ManRosSexism_Twitter_MeTwo
Language(s)	English, Spanish	English, Spanish	Spanish
Total Size	11,325	~15,000	3,077
Label Type	Multi-class (5 categories)	Binary + fine-grained subtasks	Binary + 5-category breakdown
Annotation	Manual, expert-informed	Manual with structured guidelines	Manual, aligned with EXIST taxonomy

To provide additional context on the scope and structure of each dataset, we include a brief description of their data source, annotation approach, and relevance to our classification task.

The EXIST 2021 dataset was developed to support fine-grained sexism detection across both English and Spanish texts, collected from Twitter and Gab. It includes over 11,000 samples and features a five-class labeling schema designed to capture diverse expressions of sexism, such as objectification, stereotyping, and violence. The dataset was used in the IberLEF 2021 shared task and is available through the EXIST task site.

The HatEval dataset was released as part of the SemEval 2019 Task 5 on multilingual detection of hate speech against immigrants and women. It includes approximately 15,000 tweets in English and Spanish, all manually annotated using consistent cross-lingual guidelines. The dataset is structured around two subtasks: binary classification of hate speech presence (HATE vs. NOHATE), and fine-grained classification that distinguishes between individual or generic targets as well as levels of aggressiveness.

The ManRosSexism_Twitter_MeToo dataset was developed to capture and classify nuanced expressions of sexism in Spanish-language tweets, particularly those emerging in the wake of the global Me Too movement. Sparked by widespread disclosures of sexual harassment and abuse, the Me Too movement gained momentum across Latin America under hashtags such as #YoTambien and #MeTooEscritoresMexicanos, generating a wave of online discourse about gender-based violence, power dynamics, and societal norms. This dataset consists of 3,077 manually annotated tweets collected using movement-related hashtags and keywords. It applies a two-level annotation schema, first labeling tweets as sexist or non-sexist, and then categorizing sexist content into five subtypes.

3.2 Data Cleaning

3.2.1 Translated Tweet Dataset

The dataset originally contained tweets in both English and Spanish, with approximately 22 percent written in Spanish. However, not all records included a reliable language label. To address this, we used the Langdetect library in Python to automatically infer the language of each tweet. Langdetect applies an n-gram–based approach, comparing character sequences to known language profiles to determine the closest match. Once the language was identified, we translated all tweets to both English and Spanish using the Google Translate function in Google Sheets. This tool, built on transformer-based models, performs translation at the sentence level rather than word by word, allowing for more contextually accurate results. This process ensured consistent bilingual coverage across the dataset and enhanced semantic fidelity for downstream modeling tasks.

3.2.2 Combining Datasets

The raw data was spread across multiple CSV files, each containing different subsets of tweets (original Spanish, translated, synthetic, etc.). To consolidate these into a unified training dataset, we used a Python notebook and referenced sample code in Appendix B.

1. Load and concatenate CSV files
2. Remove duplicate tweets based on text and ID
3. Standardize label formats
4. Drop incomplete or unlabeled entries
5. Filter out off-topic content

To consolidate and standardize multiple public datasets related to sexist and hate speech detection, we developed a preprocessing pipeline in Python, executed in Google Colab. The datasets included EDOS, HatEval, EXIST2021, and ManRoSexism, spanning English and Spanish languages. The main objective was to unify these sources into a single dataframe with consistent schema and a normalized binary label indicating whether each tweet is sexist.

Each dataset was loaded dynamically using a mapping of relative paths stored in a dictionary, and the script inferred the appropriate read method (.csv, .tsv, .txt, .xlsx) based on the file extension. Custom parsing was applied where necessary (e.g., targetResultFile_full2 required ; as delimiter).

The script then added metadata columns:

- source_dataset indicating the origin of each record,
- split_type inferred from filenames or internal dataset structure (e.g., train/test/dev),
- target_column derived from dataset-specific label columns and normalized to binary values (1 = sexist, 0 = non-sexist).

Additionally, the EDOS and HatEval datasets include subcategory annotations, which were extracted and stored in three new columns:

- subcategory_general (e.g., "targeted"),
- subcategory_specific (e.g., "aggressive"),
- subcategory_combined (e.g., 1 | targeted | aggressive).

The script also accounted for ambiguity in labeling: for example, ManRoSexism contains a "DOUBTFUL" label, which was mapped to 0.5 in the target column to preserve semantic ambiguity for future analysis.

After cleaning and normalizing all datasets, a unified data frame was constructed using pd.concat. The output was saved in Pickle format to preserve data types and metadata. The resulting unified dataset includes standardized columns such as text, id, target_column, and source-related annotations, and serves as the foundation for both binary and multiclass classification tasks.

This modular and extensible pipeline ensures reproducibility and facilitates future updates, such as integrating new datasets or expanding label taxonomies.

3.3 Label Mapping and Refinement

The original dataset included 16 subcategories of sexist content, derived from prior academic taxonomies to reflect diverse forms of hostility, objectification, and discrimination. However, many labels had overlapping definitions and inconsistent distinctions. For example, “descriptive attacks” and “emotive attacks” shared similar linguistic features, making them difficult to reliably differentiate during annotation or modeling.

To improve clarity and reduce label noise, we manually reviewed each subcategory using a logic model and consolidated them into 5 broader, semantically distinct classes:

1. Sexual Insults & Objectification
2. Gendered Stereotypes & Insults

3. General Hostile Language
4. Threats & Physical Harm
5. Victim Blaming & Justification

This mapping was guided by qualitative notes (see Appendix C) and grouped tweets based on tone, intent, and target. For example, tweets referencing physical appearance or sexual acts were placed in Category 1, while ideological critiques of feminism or reinforcement of gender roles were grouped under Category 2. Examples of tweets in the new multi class categories can be found in Appendix D.

During review, we also flagged two large subcategories labeled as "untargeted | non aggressive" and "un targeted | aggressive" as problematic. These often contained anti-immigrant or racially motivated speech with little or no reference to gender, suggesting they were misclassified during initial annotation. To preserve the scope and focus of this project on sexist content, we excluded these examples from the final dataset.

Additional Considerations: This process highlighted the challenge of disentangling sexist content from intersectional hate speech. Some tweets combined racism, xenophobia, and sexism, while others lacked sufficient context to determine intent. Removing ambiguous cases improved precision but limited our ability to capture overlapping harms.

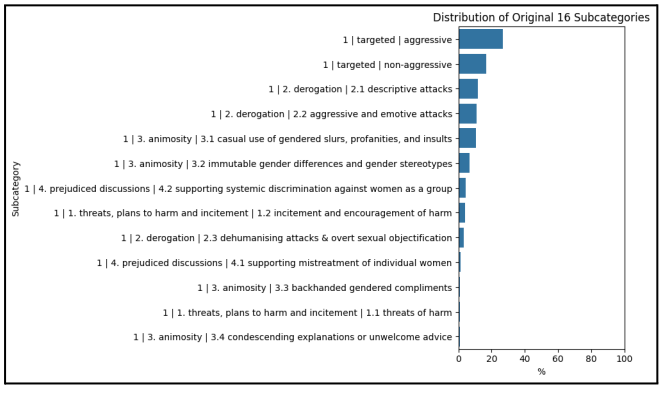
4. Exploratory Data Analysis (EDA)

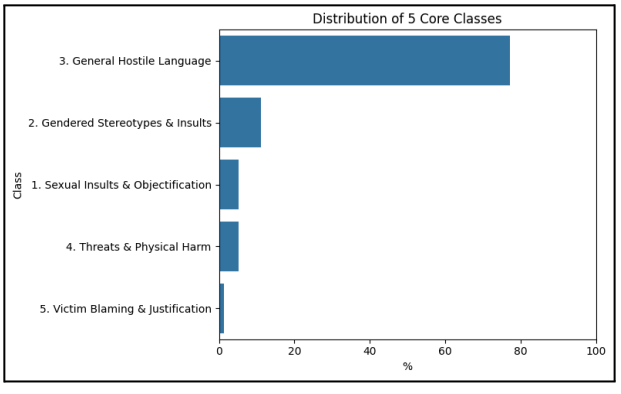
Following the data consolidation and cleaning process outlined in Section 3.2, we performed an exploratory data analysis (EDA) to better understand the structure and content of the tweets. Twitter (now X) is a microblogging platform limited to 280 characters, which often results in abbreviated language, symbols, and colloquialisms. These features can present challenges for natural language processing tasks. Our cleaned dataset includes 44,352 tweets with both binary (sexist vs. non-sexist) and multi-class annotations, as described in Section 3.3, enabling nuanced analysis of sexist content. Each entry includes a unique ID, multilingual text (English and Spanish), and classification labels. Of these, 20,016 tweets are annotated with binary labels, and 8,612 include multi-class labels and 16 original fine-grained subcategories. All English tweets were translated into Spanish to ensure full language coverage for downstream modeling. Additional tables can be found in Appendix E.

4.1 Multiclass and Binary class

Multiclass

To better understand the distribution of sexist content in our dataset, we visualized the normalized class frequencies from both the original 16-category taxonomy and our refined 5-class schema. The 16 subcategories were initially developed to capture the full range of sexist expressions, from objectification to systemic discrimination. However, as noted in



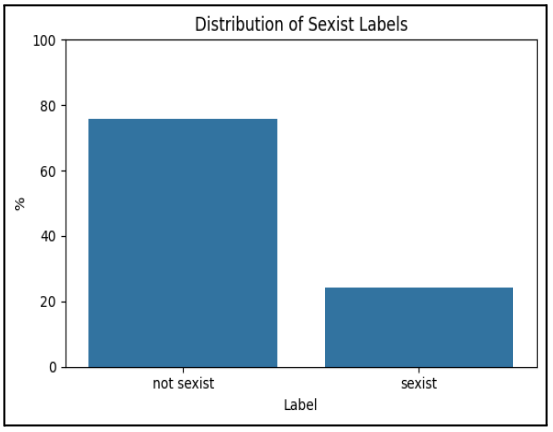


Section 3.3, overlapping themes and labeling inconsistencies limited their usefulness for model training. The first graph below illustrates the class imbalance within this original structure, with "targeted | aggressive" dominating the dataset.

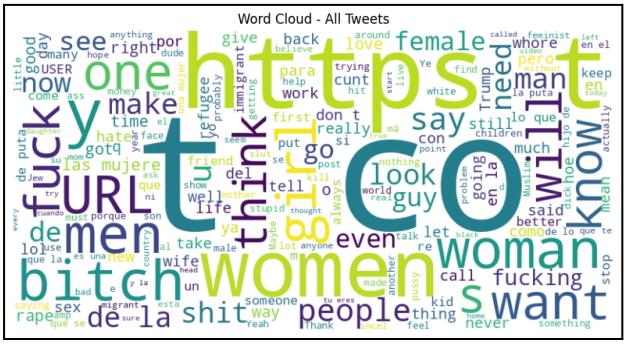
The second graph shows the result of consolidating the subcategories into five core classes, which improved balance and interpretability. These refined categories served as the basis for our multiclass classification task, while the binary classification task relied on a broader sexist vs. non-sexist label.

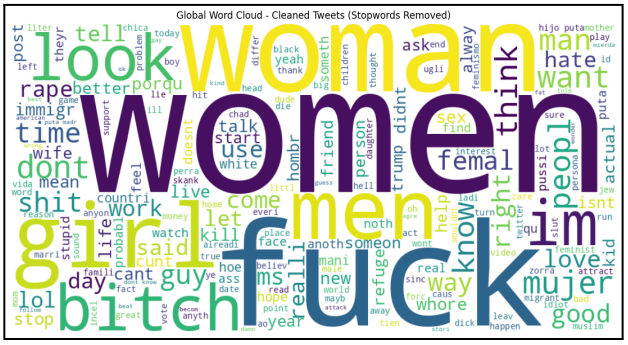
Binary

For the binary classification task, we visualized the distribution of tweets labeled as either *sexist* or *not sexist*. As shown in the figure, the dataset is imbalanced, with approximately 25 percent of tweets labeled as sexist and the remaining 75 percent classified as non-sexist. This imbalance is a key consideration for training and evaluation strategies.



4.2 Stopwords





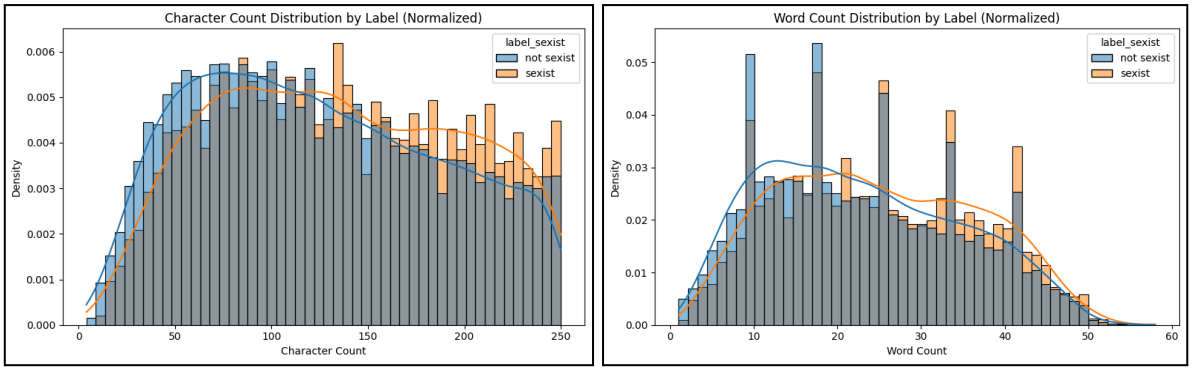
To support our label analysis, we generated two global word clouds from the dataset: one before and one after removing stopwords. Standard stopwords in English and Spanish, along with a custom blacklist of non-informative tokens like “rt,” “lol,” and “https,” were removed to reveal more meaningful content. The refined visualization highlights frequent gendered terms (“women,” “mujer”), explicit language (“bitch,” “rape”), and sentiment-laden verbs (“hate,” “kill”), underscoring the polarized and often aggressive tone of the dataset.

4.3 Word and Character Count Distributions

To examine structural differences in tweet composition, we analyzed normalized distributions by both character and word count across sexist and non-sexist labels. While both groups peak around 60 to 90 characters and 10 to 20 words, their distributions differ in shape and spread. Non-sexist tweets show a tighter, more centralized curve, suggesting more uniform, concise messaging. In contrast, sexist tweets have a broader distribution, with higher relative density in both the short and long ends, indicating variability in tone and format, from brief insults to longer, rant-like expressions.

In the word count plot, sexist tweets sustain higher density beyond 25 words, especially between 30 and 45. This trend may reflect the inclusion of extended justifications, narratives,

or elaborate personal attacks, whereas non-sexist tweets tend to taper off more steeply, suggesting briefer and more standardized phrasing. These patterns highlight the need for models that can recognize both compact hostile language and more verbose, rhetorically complex sexist content.



4.4 Non-Character Elements

Given the informal nature of Twitter, we analyzed emoji usage to explore stylistic and tonal differences between sexist and non-sexist tweets. Emojis appeared infrequently: 2.37% of non-sexist and 1.73% of sexist tweets included at least one. Average counts were similarly low (0.05 vs. 0.035 emojis per tweet), indicating minimal expressive use. Due to this low frequency, emojis were not included as key features in our modeling, though they may be more relevant in other informal or expressive platforms.

5. Model Selection

5.1 Model Benchmarking

Our model draws from the work of the Universitat Politècnica de València (UPV), whose team developed the Josefina model in collaboration with UNICC. Josefina is a fine-tuned version of the PlanTL-GOB-ES/roberta-base-bne model, a RoBERTa variant pre-trained on Spanish data from the National Library of Spain. By reproducing Josefina on our refined dataset, we benchmarked its performance and ensured alignment with prior validated methods. This served as a baseline to assess the gains made by newer models or classification structures.

5.2 Model Landscape Review

To begin our model selection process, we conducted a comprehensive review of 14 transformer-based models for classifying sexist content in Spanish-language tweets referenced in Appendix F. These models were grouped into three major categories based on their design and training focus. Highlights from each model include:

- **General Multilingual Models** (e.g., mBERT, XLM-RoBERTa, mDeBERTa, XLM-T): Trained on multilingual corpora and useful for zero-shot tasks, though they often underperform on Spanish-specific content. XLM-T, optimized for Twitter data, was a notable exception and selected for testing.
- **Spanish-Specific Models** (e.g., BETO, MarIA, RoBERTuito, BERTIN): Trained exclusively on Spanish text, these models capture linguistic nuances more effectively.

BETO and BERTIN were chosen for fine-tuning based on strong general performance.
- **Lightweight LLMs for Prompting** (e.g., GPT-4o-mini): Accessed via API and used for prompt-based classification, GPT-4o-mini enables fast, scalable, and low-code deployment. Its multilingual support and ease of use make it ideal for rapid prototyping.

Although many models were reviewed, only a subset were fine-tuned and evaluated in depth. We prioritized:

- **Language alignment**: BETO and BERTIN were preferred due to their Spanish training data.
- **Task fit**: XLM-T and RoBERTuito were designed for Twitter-style, informal, and noisy data, making them strong candidates for our use case.
- **Performance feasibility**: GPT-4o-mini was accessible via API but showed limited real-world precision without additional tuning.

This strategic narrowing allowed us to focus our computational resources on models with the highest potential for performance and interpretability in the context of Spanish social media moderation.

5.2.1 Local LLMs via Ollama: *LLaMA 3 and Mistral*

We used Ollama, a lightweight tool for running open-weight LLMs locally, to deploy LLaMA 3 and Mistral. These models were queried through a local REST API and integrated into a Python pipeline in VS Code. A sample of the prompt is included in Appendix G.

- Tweets were passed into the model one at a time via structured prompts defining “sexist” and “nonsexist” content.
- The model responded with a single-word label, making parsing simple and outputs clean.
- This method scaled easily to multiclass classification by adjusting the prompt with examples for each of the 5 categories.

Despite the simplicity and control offered by this approach, local inference had drawbacks:

- High latency and memory usage
- No supervised fine-tuning; performance was prompt-dependent

Still, these LLMs showed promise as proof-of-concept tools and offer future potential when paired with embedding extraction or fine-tuned heads. However due to hardware limitations in Google Colab, we relied on dataset subsampling and avoided full fine-tuning for large LLMs. Local LLMs like LLaMA 3 and Mistral, deployed via Ollama, were useful for prompt-based experimentation but suffered from high latency and lacked supervised tuning.

5.2.2 Cloud LLMs via GPT-4o-mini API

To address performance bottlenecks, we tested OpenAI’s GPT-4o-mini via the ChatGPT API for large-scale automated labeling. GPT-4o-mini enabled efficient cloud inference but overpredicted sexist content and lacked domain-specific calibration. These trade-offs limited the performance and generalizability of prompt-only approaches.

- Over 11,000 tweets were labeled using a concise binary classification prompt.
- Cloud-based inference provided rapid, consistent results with minimal setup.
- However, the model overpredicted sexist content, resulting in a high false positive rate. This may stem from lack of domain-specific fine-tuning or prompt calibration.

5.2.3 Selected Models

Ultimately, we selected the following models based on benchmark scores and their fit for detecting sexist content in Spanish tweets. The table below summarizes our rationale.

Model	Reason for Selection	Strengths	Limitations
Josefina	Pretrained and fine-tuned specifically for sexist tweet detection in Spanish	Validated architecture aligned with project goals; tailored to Spanish-language hate speech	Less flexible for adapting to new label schemas or datasets
BETO	Spanish-specific BERT model widely used for general NLP tasks	Strong performance on Spanish text; well-supported in HuggingFace ecosystem	Not optimized for informal or Twitter-style data
XLM-T	Multilingual transformer trained on Twitter data	Optimized for noisy, informal text; handles multiple languages	Higher resource requirements; less transparent than alternatives

6. Modeling Approach

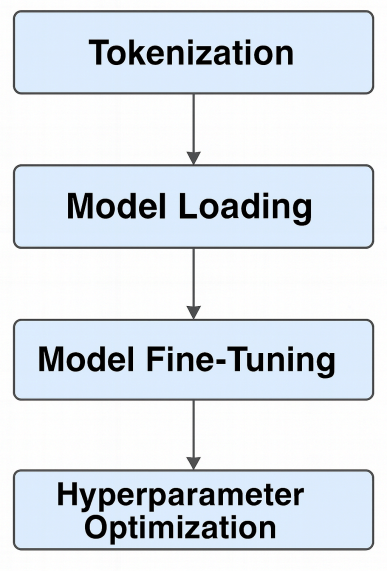
Our modeling strategy was guided by the objective of accurately identifying sexist content in Spanish-language tweets, while also allowing for nuanced categorization where appropriate. The approach was structured around three key areas: defining the classification task (binary vs. multiclass), building a scalable and flexible training pipeline using state-of-the-art NLP tools, and addressing dataset imbalances that could impact model fairness and reliability.

6.1 Binary vs. Multiclass Tasks

Our project addressed two distinct but related classification tasks. The first was a binary classification task in which each tweet was labeled as either sexist (label = 1) or non-sexist (label = 0). This task aimed to establish a foundational detection model capable of flagging any content considered harmful or biased based on gender. The second task was a multiclass classification problem, where sexist tweets were further categorized into one of five refined subcategories, such as sexual insults and objectification (1), gendered stereotypes and insults (2), general hostile language (3), threats and physical harm (4), and victim blaming and justification (5). These categories were derived from the original dataset's annotated labels and reflect more nuanced patterns of sexist language.

6.2 Training Pipeline and Hyperparameter Optimization

To ensure scalable and consistent training, we used a unified pipeline built with Hugging Face Transformers. Models were loaded with AutoModelForSequenceClassification and tokenized using AutoTokenizer for input compatibility. Fine-tuning was managed via the



Trainer API, with adjustable hyperparameters like epochs, batch size, learning rate, and weight decay to optimize performance and reduce overfitting.

To determine the most effective hyperparameter configuration for each model, we integrated a Grid Search optimization step using the Optuna framework. This allowed us to systematically explore the following hyperparameter space:

- learning_rate: [2e-5, 3e-5, 5e-5]
- per_device_train_batch_size: [4, 16, 32]
- num_train_epochs: [3, 4]
- weight_decay: [0.0, 0.01]

Due to GPU constraints in Google Colab, we used a 20% subsample during Grid Search to reduce runtime while maintaining data diversity. Each model underwent 4 trials, and the best configuration (based on validation F1-score) was then used for final training on the full or 50% dataset. This approach ensured consistent, efficient, and reproducible model development within resource limits.

Additionally, we utilized OpenAI GPT 4o mini model, to conduct binary classification run to label tweets without connecting to or training on our internal dataset. This setup was designed to evaluate the model's zero shot inference capability, classifying tweets based solely on prompt instructions without access to prior examples or contextual fine tuning. We processed more than 11,000 tweets using the following standardized parameters for binary text classification:

- Temperature = 0.1, to ensure consistent and deterministic responses
- Max tokens = 50, as the output only required a short label ("sexist" or "nonsexist")

7. Evaluation

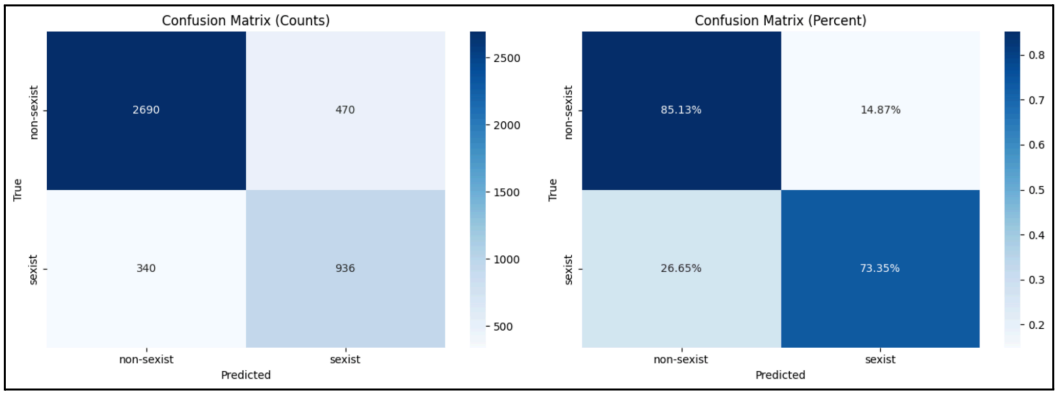
All models were evaluated on both binary and multiclass classification tasks using a consistent training pipeline. Due to computational limitations in Google Colaboratory, each model was trained on a 50% stratified subsample of the full dataset. This allowed us to preserve class distribution while ensuring feasible training times within limited GPU resources. Despite this constraint, the models demonstrated strong performance and stable learning behavior across configurations, enabling fair comparison and robust evaluation. Additional model analysis can be found in Appendix section H-I.

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
BETO - Binary	0.83	0.78	0.82	0.82	0.83
Josefina - Binary	0.81	0.80	0.75	0.75	0.87
XLM-T - Binary	0.83	0.77	0.81	0.82	0.83
roBERTa - Binary	0.63	0.40	0.63	0.78	0.26
BERTIN - Binary	0.75	0.63	0.72	0.74	0.75
ChatGPT - Binary	0.71	0.63	0.63	0.49	0.86

XLM-T - Multi-Class	0.82	0.51	0.82	0.82	0.83
BETO - Multi-Class	0.79	0.37	0.76	0.60	0.57
BERTIN - Multi-Class	0.77	0.17	0.68	0.15	0.20

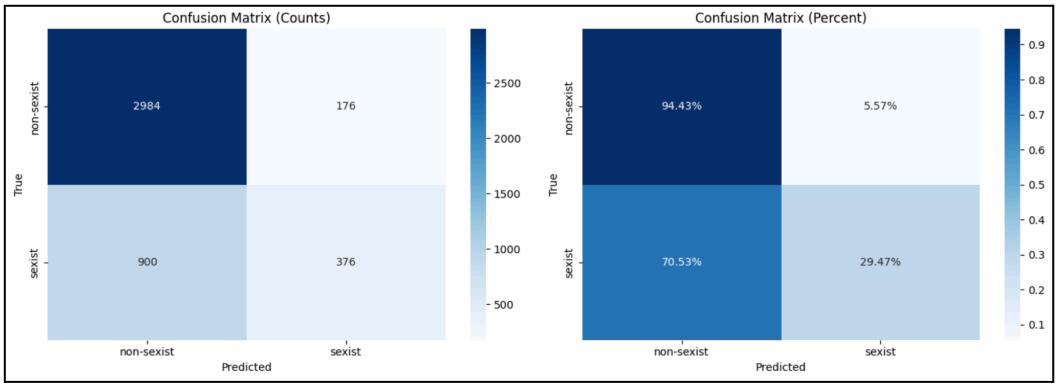
7.1 Binary

7.1.1 BETO - Binary



We fine-tuned the Spanish-language BETO model for binary classification of sexist versus non-sexist tweets, using 50% of the dataset due to computational constraints. A grid search was conducted to optimize hyperparameters, and the final model was trained with a learning rate of 3e-5, batch size of 16, four epochs, and a weight decay of 0.01. The model achieved strong results, with 0.83% accuracy, a macro F1 score of 0.78, and reliable performance across both classes. It showed high precision and recall for non-sexist tweets, and slightly lower performance for sexist content, reflecting class imbalance and linguistic complexity. Confusion matrix analysis confirmed a balanced tradeoff between sensitivity and specificity. While minor overfitting was observed after the second epoch, validation metrics remained stable, indicating good generalization.

7.1.2 BERTIN - Binary



The final BERTIN model was trained with a learning rate of 5e-5, batch size of 16, four epochs, and a weight decay of 0.01. On a validation set of 4,436 tweets, it reached 75.74% accuracy, with a macro F1 score of 0.63 and a weighted F1 of 0.72. The model performed well at detecting non-sexist tweets, correctly identifying 94.43% of them. However, it struggled with sexist tweets, correctly identifying only 29.47%. The confusion matrix shows

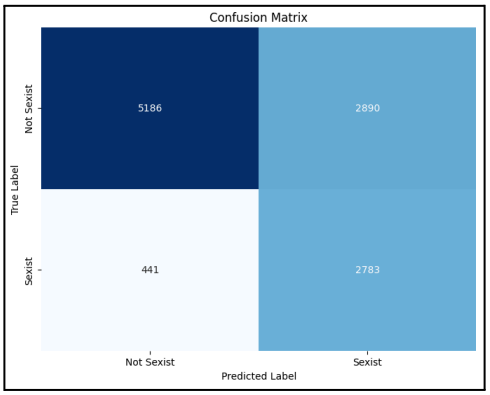
that most sexist tweets were misclassified as non-sexist, revealing a conservative bias that avoids false positives but misses many harmful messages. Still, the model trained steadily, and validation performance improved over time, showing it was learning consistently.

7.1.3 XLM - Binary

The XLM-T model (cardiffnlp/twitter-xlm-roberta-base) was finetuned via grid search conducted to optimize hyperparameters, and the final model was trained with a learning rate of 5e-5, batch size of 16, four epochs, and a weight decay of 0.01. On a validation set of 8,871 tweets, XLM-T achieved strong results with 82% accuracy, a macro F1 score of 0.77, and a weighted F1 of 0.81. It performed particularly well on non-sexist content, while showing moderate results on the sexist class, reflecting the expected challenges due to class imbalance and linguistic nuance. The confusion matrix confirmed high true positives for non-sexist tweets but also revealed that 982 sexist tweets were missed, indicating a conservative prediction bias. Despite this, training was stable and effective, with steadily decreasing loss and no overfitting. Overall, XLM-T proved to be a reliable and efficient model under constrained resources, outperforming BERTIN in recall and offering a strong alternative to BETO.

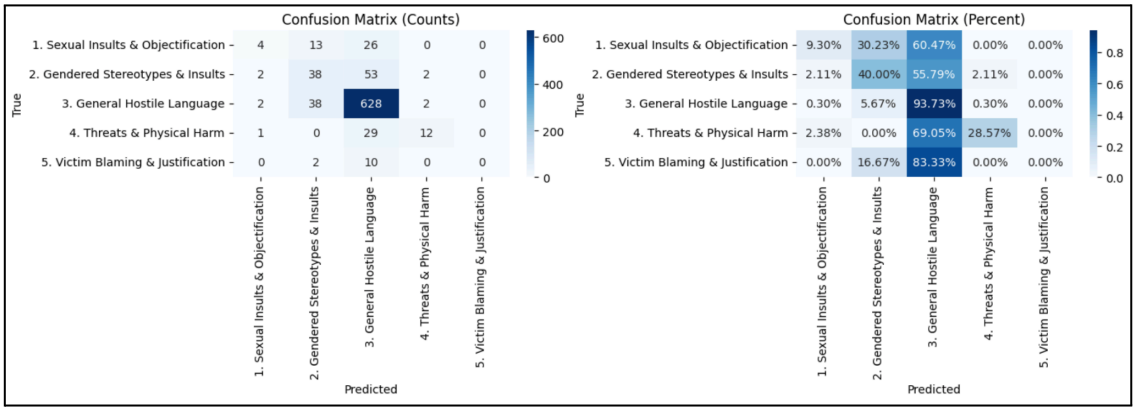
7.1.4 4o-mini: Binary Classification

OpenAI's GPT 4o mini, used via the ChatGPT API for binary classification, underperformed compared to fine tuned transformer models, achieving high recall (0.86) but low precision (0.49) due to frequent overprediction of the "sexist" label. Its zero shot, prompt only setup without training data or contextual grounding limited its ability to handle nuanced cases, making it less suitable for real world moderation. To improve performance, future work should explore fine tuning or few shot prompting, as well as testing more advanced models like GPT 4 turbo.



7.2 Multiclass

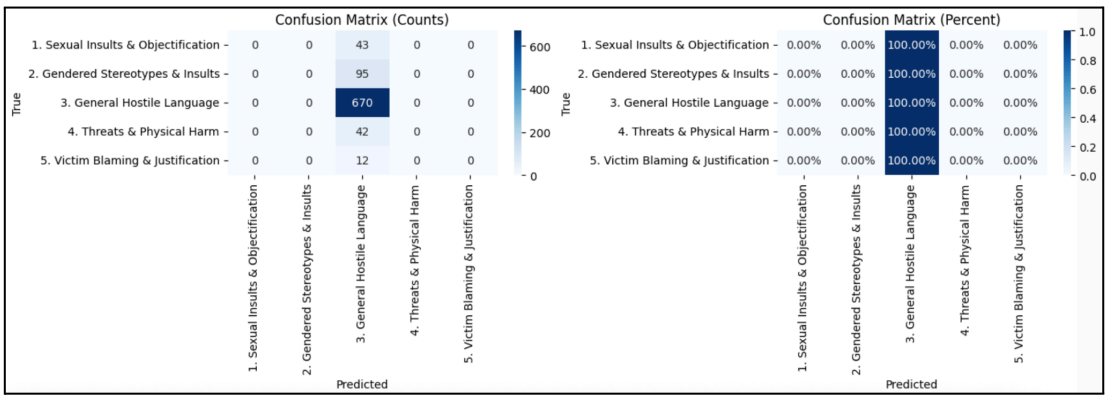
7.2.1 BETO - Multi Class



We fine-tuned the Spanish-language BETO model for multi-class classification of sexist tweets into five predefined categories. A grid search was used to optimize hyperparameters,

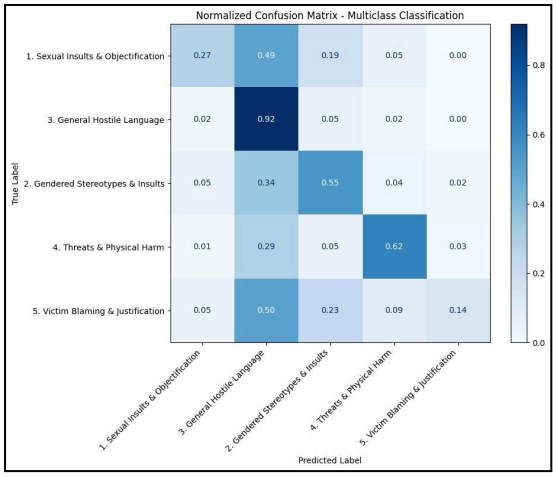
and the final model was trained with a learning rate of 5e-5, batch size of 16, four epochs, and a weight decay of 0.1. On the validation set of 862 tweets, the model achieved 79.12% accuracy, a macro F1 score of 0.37, and a weighted F1 score of 0.76. While performance was strong on the dominant class, General Hostile Language, the model struggled with minority classes like Victim Blaming and Sexual Insults. The confusion matrix showed frequent misclassification of minority categories into the majority class, revealing a bias driven by class imbalance. Training loss decreased steadily, but rising validation loss by epoch four suggested some overfitting. Although weighted scores remained stable, the gap between macro and weighted F1 highlights the model's limited generalization across all classes. Future work should consider rebalancing techniques or alternative models better suited to handling linguistic nuance in low-support categories.

7.2.2 BERTIN - Multi Class



To assess BERTIN's performance in multi-class classification, we fine-tuned by grid search to optimize hyperparameters, and the final model was trained with a learning rate of 5e-5, batch size of 16, four epochs, and a weight decay of 0.1. On the validation set of 862 tweets, BERTIN achieved 77.73% accuracy, a macro F1 score of 0.17, and a weighted F1 score of 0.68. It failed to correctly classify any instances from the minority classes, all of which scored zero on precision, recall, and F1.

7.3.2 XLM-T: Multi Classification



To perform multi-class classification of sexist tweets, the XLM-T model (cardiffnlp/twitter-xlm-roberta-base) was fine-tuned using a learning rate of 5e-5, batch size of 16, four epochs, and a weight decay of 0.1. The model was evaluated on a validation set of 1,723 tweets, where it achieved 82% accuracy, a weighted F1 score of 0.82, and a macro F1 score of 0.51. It performed exceptionally well on the most frequent category, but struggled with underrepresented classes like Victim Blaming and Sexual Insults, revealing the impact of class imbalance. Training progressed smoothly with no overfitting observed, though misclassifications in minority classes suggest the need for additional strategies such as focal loss, class weighting, or

data augmentation to improve performance on less frequent forms of sexist language.

8. Results & Discussion

Following the development and evaluation of several transformer-based models, this section presents a comprehensive analysis of our findings, structured across three dimensions: model performance, technical and linguistic insights, and ethical considerations. Together, these perspectives not only justify our final model selections but also reveal the deeper implications of applying AI to detect sexist content in Spanish-language social media.

8.1 Performance Comparison

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
BETO - Binary	0.83	0.78	0.82	0.82	0.83
Josefina - Binary	0.81	0.80	0.75	0.75	0.87
XLM-T - Binary	0.83	0.77	0.81	0.82	0.83
roBERTa - Binary	0.63	0.40	0.63	0.78	0.26
BERTIN - Binary	0.75	0.63	0.72	0.74	0.75
ChatGPT - Binary	0.71	0.63	0.63	0.49	0.86
XLM-T - Multi-Class	0.82	0.51	0.82	0.82	0.83
BETO - Multi-Class	0.79	0.37	0.76	0.60	0.57
BERTIN - Multi-Class	0.77	0.17	0.68	0.15	0.20

After evaluating a diverse range of transformer-based models across binary and multiclass classification tasks, we identified BETO, Josefina, and XLM-T as the top performers aligned with the project's objectives. For binary classification, BETO delivered the most balanced results, accuracy of 0.83, macro F1 of 0.78, and precision/recall of 0.82/0.83, making it ideal for consistent and high-precision moderation. Josefina, while slightly less precise (0.72), achieved the highest recall (0.84) of all binary models, making it well-suited for risk-averse use cases where capturing subtle or borderline sexist content is critical. Interestingly, XLM-T, designed for multilingual applications, matched BETO's binary performance almost identically, confirming its adaptability to Spanish-language toxicity detection.

In the multiclass task, XLM-T stood out as the most robust model. It achieved a macro F1 of 0.51 and weighted F1 of 0.82, alongside precision and recall of 0.82/0.83, showing strong generalization across the five refined subcategories of sexism. By contrast, other models like BERTIN and BETO underperformed in the multiclass setting, particularly in handling class imbalance. Ultimately, BETO and Josefina were selected for their complementary strengths in binary detection, while XLM-T was chosen for its nuanced, fine-grained classification capabilities, together forming a well-rounded architecture for broad and targeted moderation strategies.

8.2 Insights

While the previous section focused on performance, our modeling process also yielded important technical and linguistic insights. A key challenge was annotation quality: many tweets labeled "non-sexist" included sarcasm, implicit bias, or ambiguous tone, making them difficult to classify confidently, even for high-performing models like BETO or Josefina. This

reflects the inherent subjectivity in labeling sexism and highlights how inconsistencies can cap model learning. In the multiclass setting, this issue became even more pronounced. Despite XLM-T's strong accuracy (0.82) and recall (0.83), its lower macro F1 (0.51) exposed how overlapping or subtle category boundaries, especially between "Victim Blaming" and "Stereotypes", complicate multiclass prediction.

Language variability across Spanish dialects further shaped model behavior. Josefina's higher recall may stem from sensitivity to regional slang or implicit cues, though it also led to greater prediction variance. Additionally, our decision to subsample the dataset (50% for training, 20% for tuning) enabled broader experimentation under GPU constraints but introduced class imbalance sensitivity, particularly in underrepresented multiclass categories. These insights emphasize the need for improved annotation protocols, dialectal diversity in training data, and careful sampling to enhance future model reliability.

8.3 Fairness & Ethics

Developing AI systems for detecting sexism entails not only technical rigor but also deep ethical reflection. BETO, Josefina, and XLM-T operate within the high-stakes context of online content moderation, where misclassification has real-world consequences. In Spanish-language spaces, where moderation tools are scarce or inconsistent, these models can help fill urgent detection gaps. However, training on social media data also risks reproducing embedded biases, and our team's lack of formal gender studies expertise underscores the need for caution in designing taxonomies and interpreting labels.

To address these concerns, we applied manual auditing, removed misaligned content (e.g., racially motivated hate speech misclassified as sexist), and selected models with distinct strengths, Josefina's high recall for subtle harms, and BETO/XLM-T's balance and interpretability. Still, long-term fairness demands more than technical tuning. We recommend integrating expert input, formal bias audits, and ongoing stakeholder feedback to ensure models remain culturally relevant, ethically robust, and aligned with evolving standards of equity and harm prevention.

9. Conclusion

This project presents a comprehensive approach to detecting sexist language in Spanish social media, combining technical rigor with a clear social mission. Building on the Universitat Politècnica de València's work with the Josefina model, we curated a multilingual dataset, refined a fragmented label structure, and benchmarked several transformer-based models for binary and multiclass classification.

A key contribution was consolidating 16 overlapping subcategories into 5 clear, semantically distinct classes, improving both annotation consistency and model performance. Using Hugging Face and Optuna, we fine-tuned models, such as BETO, BERTIN, XLM-T, and Josefina within limited computing resources and delivered a scalable, reproducible pipeline.

BETO performed strongly in binary detection with high precision, Josefina achieved the highest recall, and XLM-T excelled in multiclass tasks with balanced F1 scores across all five sexism categories. Together, these models offer complementary strengths for broad, sensitive, and fine-grained moderation needs. Beyond accuracy, we prioritized fairness, contextual understanding, and the protection of marginalized voices. Our work in label consolidation,

minority class evaluation, and model selection reflects a commitment to ethical and effective automated moderation.

10. Future Work

This project, developed in collaboration with the United Nations International Computing Centre (UNICC), lays the foundation for scalable and ethical AI tools aimed at improving digital safety. By aligning with Sustainable Development Goal (SDG) 5.2, which advocates for the elimination of violence against women and girls, our goal was not only technical innovation but also social impact, particularly in Spanish-language online spaces.

Building on this foundation, future work should address key challenges identified during model development. A major priority is improving annotation quality and reducing label ambiguity, especially for borderline or context-dependent content. Partnering with gender studies and sociolinguistics experts can help refine labeling standards and increase dataset representativeness, especially in underrepresented dialects. Technical efforts should also focus on addressing class imbalance in multiclass settings and scaling up to more powerful models through data augmentation, low-resource tuning, and infrastructure optimization.

10.1 Policy Compliance and Moderation Tools

Social media platforms are increasingly obligated by law to moderate hate speech, including technology-facilitated gender-based violence. As part of future work, we propose adapting our model into a compliance-aligned tool to help platforms identify, flag, and potentially block posts that violate terms of service or legal standards. The model could be used internally to support human moderators or offered as an automated moderation service via API. This would be especially valuable in Spanish-language contexts where culturally specific expressions of sexism are often missed by global moderation systems. With further validation and transparency, such tools could help institutions uphold international human rights standards online.

10.2 Public-Facing Web Application

We also see strong potential for a public-facing web tool that enables individuals to assess sexist language in online content. It could serve as both an educational and monitoring resource, promoting awareness and accountability in digital spaces. As prototyped in Appendix K, this tool would provide features, such as real-time trend visualizations and filters by language, date range, or platform.

This concept is inspired by the success of previous public awareness platforms such as the World Emissions Clock, the Doomsday Clock, and the Panama Papers project, all of which demonstrated that clear, accessible data visualization can drive public engagement and policy conversations. By making NLP insights interpretable for non-technical users, this tool can help generate a sense of urgency and action, especially in regions where online abuse is underreported or normalized.

To implement this application, the following steps are required:

1. Connect to the Twitter API (X API) to stream or search recent tweets

2. Define effective filtering parameters to surface tweets with higher likelihood of containing sexist or gendered language (e.g. using keyword heuristics or initial scoring models)
3. Run our binary classification model on the retrieved tweets in real time
4. Update and refresh the dashboard dynamically, providing users with the latest labeled content

10.3 Expansion to Multilingual and Regionally Diverse Data

Currently, our dataset and models are centered on Latin American Spanish. However, to expand reach and fairness, future efforts should incorporate data from other Spanish-speaking regions. This could involve tailoring the taxonomy to local norms and collecting new annotated corpora that reflect diverse sociolinguistic realities. Expanding linguistic coverage would ensure greater inclusivity and strengthen generalization across different online communities.

10.4 Integration with Large Language Models (LLMs)

With the availability of larger domain-specific datasets, future work could involve fine-tuning large language models such as GPT-4-turbo, GPT-4, or LLaMA 4 Maverick for sexism detection. Given sufficient scale, it may be possible to train high-performing models capable of capturing subtleties in language use, tone, and context that smaller models cannot handle. Although we tested LLMs on a subset of the data and saw promising results, we lacked the local computational resources, and the need for hundreds of thousands of tweets for fine tuning the model, therefore we needed to run more advanced Ollama-powered models effectively. Additionally, we were unable to experiment with pro-level models that support in-context learning or fine-tuning on our own dataset, limiting our ability to fully evaluate their potential. Developing a production-grade system will require access to greater infrastructure, ethical oversight, and rigorous testing to avoid risks such as overfitting or bias amplification. By addressing these challenges, future efforts could evolve this work from a proof-of-concept into a scalable, real-world tool for promoting safer online environments and advancing gender equity through NLP.

11. Limitations

Despite meaningful results, this project also faced several constraints that shape how the findings should be interpreted. First, the training data was predominantly sourced from Latin America, while some key models, such as BETO and Josefina, were pre-trained on corpora from Spain. This mismatch may have reduced their effectiveness in capturing region-specific expressions, slang, or sociolinguistic subtleties.

Second, limited GPU access influenced our ability to scale experiments. We had to subsample training data (50%) and tune models on just 20% during grid search. While efficient, this approach introduced instability, particularly in the multiclass task, and may have limited our ability to fine-tune larger, more expressive models, such as LLMs. Our inability to run pro-level fine-tuning on domain-specific LLMs constrained the scope of our evaluations and kept certain model comparisons exploratory.

Lastly, while we approached the project with a socially responsible lens, most team members were young men without formal training in gender or race studies. This shaped how we

interpreted labels, refined categories, and excluded content. Although we aimed for transparency and took steps to mitigate bias, these limitations underscore the importance of diverse, interdisciplinary collaboration in future iterations.

12. Appendices

Appendix A

Below are three tables of descriptions of the EXIST, HatEval and ManRosSexism datasets:

Aspect	EXIST Data Details
Language(s)	English and Spanish
Platform	Twitter and Gab
Size	6,977 tweets (train), 3,366 tweets + 982 gabs (test)
Annotation Type	Manual, informed by domain experts
Labels	- Ideological and Inequality- Stereotyping and Dominance- Objectification- Sexual Violence- Misogyny and Non-Sexual Violence
Classification Task	Multi-class classification
Balance	Balanced across both languages
Use Case	Fine-grained sexism classification in multilingual social media texts

Aspect	HatEval Data Details
Language(s)	English and Spanish
Platform	Twitter
Size	~10,000 English tweets, ~5,000 Spanish tweets
Annotation Type	Manual annotation using consistent multilingual guidelines
Labels (Subtask A)	HATE vs. NOHATE
Labels (Subtask B)	- Target: Individual vs. Generic- Aggressiveness: Aggressive vs. Non-Aggressive
Target Groups	Women and Immigrants
Use Case	Binary and fine-grained hate speech detection across multiple languages

Aspect	ManRosSexism_Twitter_MeTwo Data Details

Language(s)	Spanish
Platform	Twitter
Size	3,077 tweets
Annotation Type	Manual, two-level annotation scheme
Labels (Level 1)	Sexist vs. Non-Sexist
Labels (Level 2)	- Ideological and Inequality- Stereotyping and Dominance- Objectification- Sexual Violence- Misogyny and Non-Sexual Violence
Alignment	Aligned with EXIST taxonomy
Use Case	Nuanced, multi-label sexism detection in Spanish-language tweets

Appendix B

Captured here are screenshots from our data unification process:

```
[ ]
for name, df in loaded_data.items():
    # Inicializar columnas vacías
    df['subcategory_general'] = None
    df['subcategory_specific'] = None
    df['subcategory_combined'] = None

    if name.startswith("edos_labelled_"):
        if {'label_category', 'label_vector', 'target_column'}.issubset(df.columns):
            # Solo mantener subcategorías si target_column == 1
            mask = df['target_column'] == 1
            df.loc[mask, 'subcategory_general'] = df.loc[mask, 'label_category'].str.strip()
            df.loc[mask, 'subcategory_specific'] = df.loc[mask, 'label_vector'].str.strip()
            df.loc[mask, 'subcategory_combined'] = (
                df.loc[mask, 'target_column'].astype(str) + " | " +
                df.loc[mask, 'subcategory_general'] + " | " +
                df.loc[mask, 'subcategory_specific']
            )
            print(f"✅ {name}: subcategorías extraídas de 'label_category' y 'label_vector'")
        else:
            print(f"⚠️  {name}: faltan columnas 'label_category', 'label_vector' o 'target_column'")

    elif name.startswith("hateval2019") and {'TR', 'AG', 'target_column'}.issubset(df.columns):
        # Solo mantener subcategorías si target_column == 1
        mask = df['target_column'] == 1
        df.loc[mask, 'subcategory_general'] = df.loc[mask, 'TR'].map({1: "targeted", 0: "untargeted"})
        df.loc[mask, 'subcategory_specific'] = df.loc[mask, 'AG'].map({1: "aggressive", 0: "non-aggressive"})
        df.loc[mask, 'subcategory_combined'] = (
            df.loc[mask, 'target_column'].astype(str) + " | " +
            df.loc[mask, 'subcategory_general'] + " | " +
            df.loc[mask, 'subcategory_specific']
        )
        print(f"✅ {name}: subcategorías construidas desde 'TR' y 'AG'")

    else:
        print(f"ℹ️  {name}: no tiene subcategorías disponibles, columnas quedan vacías")

    loaded_data[name] = df
```

```
# Columnas creadas que queremos analizar
columnas_objetivo = [
    "target_column",
    "subcategory_general",
    "subcategory_specific",
    "subcategory_combined"
]

print("\n--- Análisis completo de columnas clave por dataset ---\n")

for name, df in loaded_data.items():
    total_filas = len(df)
    print(f"📘 Dataset: {name}")
    print(f"   Total de filas: {total_filas}")

    for col in columnas_objetivo:
        if col not in df.columns:
            print(f"   ⚠️  {col} no existe en este dataset")
            continue

        n_nulos = df[col].isna().sum()
        pct_nulos = (n_nulos / total_filas) * 100
        unicos = df[col].nunique(dropna=True)

        print(f"   🔹 {col}:")
        print(f"      - Nulos: {n_nulos} ({pct_nulos:.2f}%)")
        print(f"      - Valores únicos (no nulos): {unicos}")

        # Análisis especial para target_column
        if col == "target_column":
            n_1 = (df[col] == 1).sum()
            n_0 = (df[col] == 0).sum()
            n_05 = (df[col] == 0.5).sum()
            pct_1 = (n_1 / total_filas) * 100
            pct_0 = (n_0 / total_filas) * 100
            pct_05 = (n_05 / total_filas) * 100
            print(f"      - Valor 1 (sexista): {n_1} ({pct_1:.2f}%)")
            print(f"      - Valor 0 (no sexista): {n_0} ({pct_0:.2f}%)")
            print(f"      - Valor 0.5 (doubtful): {n_05} ({pct_05:.2f}%)")
```

Appendix C

In this section we provide our notes and logic around themes in the dataset that led to our remapping of the subcategories:

subcategory_combined	Notes	Groupings	New Proposed Categories
1 | 3. animosity | 3.3 backhanded gendered compliments	general insults leaning sexual	sexual	1. Sexual Insults & Objectification
1 | 2. derogation | 2.3 dehumanising attacks & overt sexual objectification	sexual insults, comments on appearance	sexual	1. Sexual Insults & Objectification
1 | 2. derogation | 2.1 descriptive attacks	general insults	general	3. General Hostile Language
1 | 2. derogation | 2.2 aggressive and emotive attacks	general insults	general	3. General Hostile Language
1 | 4. prejudiced discussions | 4.2 supporting systemic discrimination against women as a group	opposite of women empowerment	gendered insults	2. Gendered Stereotypes & Insults
1 | 1. threats, plans to harm and incitement | 1.2 incitement and encouragement of harm	sexual harm	sexual harm	4. Threats & Physical Harm
1 | 4. prejudiced discussions | 4.1 supporting mistreatment of individual women	victim blaming	victim blaming	5. Victim Blaming & Justification
1 | 3. animosity | 3.2 immutable gender differences and gender stereotypes	general insults with a tendency toward gender affirming roles or referring to feminism	gendered insults	2. Gendered Stereotypes & Insults
1 | 3. animosity | 3.1 casual use of gendered slurs, profanities, and insults	general insults	general	3. General Hostile Language
1 | 1. threats, plans to harm and incitement | 1.1 threats of harm	physical harm	physical harm	4. Threats & Physical Harm
1 | 3. animosity | 3.4 condescending explanations or unwelcome advice	sexualization, comments on physical appearance, child bearing abilities	sexual	1. Sexual Insults & Objectification
1 | untargeted | non-aggressive	mislabelling, lots of anti immigrant speech	remove	-
1 | targeted | aggressive	lots of name calling (bitch)	aggressive general	3. General Hostile Language
1 | untargeted | aggressive	lots of anti immigrant we should remove from training data	remove	-
1 | targeted | non-aggressive	name calling (whore), more slut shaming	aggressive general	3. General Hostile Language

Appendix D

Bulleted below are example tweets for each multiclass category:

- **Category 1 (Sexual Insults & Objectification):** *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.*
- **Category 2 (Gendered Stereotypes & Insults):** *A man makes the table.*
- **Category 3 (General Hostile Language):** *The View. Empty head emotional woman's daily bitchcraft lesson.*
- **Category 4 (Threats & Physical Harm):** *aim for the head so she dosn't come back*
- **Category 5 (Victim Blaming & Justification):** *It was rape because she regretted it afterwards, duh. There's no way she could have sex with a nerd. Sheesh.*

Appendix E

Provided below are subcategory counts of tweets for the previous sixteen categories and five new multivariable categories:

original_subcategory	tweet_count
1 | targeted | aggressive	2296
1 | targeted | non-aggressive	1459
1 | 2. derogation | 2.1 descriptive attacks	1025
1 | 2. derogation | 2.2 aggressive and emotive...	961
1 | 3. animosity | 3.1 casual use of gendered ...	910
1 | 3. animosity | 3.2 immutable gender differ...	596
1 | 4. prejudiced discussions | 4.2 supporting...	368
1 | 1. threats, plans to harm and incitement |...	363
1 | 2. derogation | 2.3 dehumanising attacks &...	287
1 | 4. prejudiced discussions | 4.1 supporting...	107
1 | 3. animosity | 3.3 backhanded gendered com...	91
1 | 1. threats, plans to harm and incitement |...	80
1 | 3. animosity | 3.4 condescending explanati...	69

final_class	tweet_count
3. General Hostile Language	6651
2. Gendered Stereotypes & Insults	964
1. Sexual Insults & Objectification	447
4. Threats & Physical Harm	443
5. Victim Blaming & Justification	107

Appendix F

The table below lists all the models researched in the development of this project:

Model	Language Focus	Origin	Strengths	Unique Traits / Notes
roBERTa	English	Facebook AI	Robust English classification	Not multilingual; solid benchmark model
mBERT	Multilingual (100+)	Google	Language-agnostic baseline	Lower Spanish performance than dedicated models
BETO	Spanish	Spanish NLP community	Strong baseline for Spanish tasks	First major Spanish-only BERT
RoBERTuito	Spanish (Twitter)	Argentina	Excellent with informal/social media text	Optimized for slang, hashtags, short content
mDeBERTa	Multilingual	Microsoft	Strong generalization, better than mBERT	Uses disentangled attention — great for multi-class
BERTIN	Spanish	Hugging Face + Spain	Trained on large, high-quality Spanish corpora	Good benchmark rival to BETO
XLM-RoBERTa	Multilingual (100+)	Facebook AI	Very strong multilingual model	Excellent for cross-lingual zero-shot tasks
MarIA	Spanish (formal)	Barcelona Supercomputing Center	Legal, medical, and general Spanish corpora	Excellent for formal/professional Spanish texts
LatAm-BERT	Latin American Spanish	Regional researchers	Dialect-specific strength	Ideal for apps involving regional Latin American varieties
LLaMA 3	Multilingual	Meta (Facebook)	Few-shot / zero-shot via prompting	Lightweight LLM good for in-context classification
Mistral-7B	Multilingual	Mistral AI	Open-weight, efficient	Smaller than GPT-4 but very performant
Qwen2.5	Multilingual (Chinese+)	Alibaba	LLM with multilingual capacity	Chinese strengths; decent generalization via prompting
XLM-T	Multilingual (with Twitter data)	Meta (Facebook AI)	Strong performance on informal text; robust on Spanish tweets	Pretrained on Twitter data; ideal for noisy, short, social media content
GPT 4o mini	Multilingual (high coverage)	OpenAI	Fast, cost efficient API; good recall in binary tasks	Prompt only zero shot model; low precision without fine tuning

Appendix G

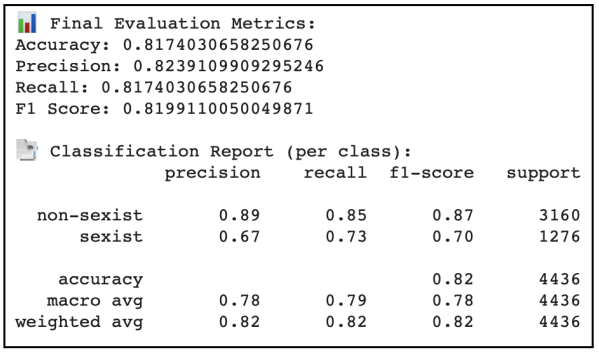
The screenshot below is an example of a prompt when using an LLM.

```
prompt = f"""
You are an expert in detecting sexist language in Spanish text.

Analyze the following Spanish text and determine if it contains sexist content.

Text to analyze: "{text}"

Consider sexist content as:
- Language that discriminates, stereotypes, or demeans people based on their gender
- Expressions that promote gender-based inequality or harmful stereotypes
- Derogatory language specifically targeting women or men
- Content that objectifies or reduces people to their gender

Respond ONLY with a JSON object in this exact format:
{{
    "label": "sexist" or "non-sexist"
}}
"""
```

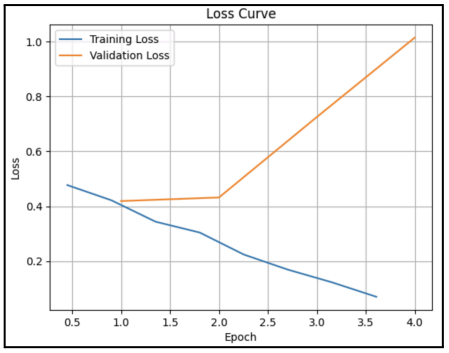
Appendix H

BETO Binary:

The screenshot below shows the final evaluation metrics and loss curves for the BETO binary classification model after training.

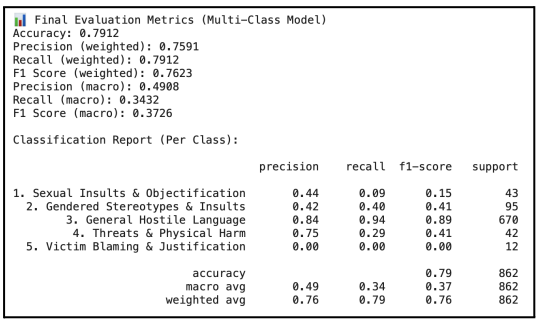


```
📊 Final Evaluation Metrics:
Accuracy: 0.8174030658250676
Precision: 0.8239109909295246
Recall: 0.8174030658250676
F1 Score: 0.8199110050049871

📄 Classification Report (per class):
              precision    recall  f1-score   support

  non-sexist       0.89      0.85      0.87      3160
      sexist       0.67      0.73      0.70      1276

    accuracy                           0.82      4436
   macro avg       0.78      0.79      0.78      4436
weighted avg       0.82      0.82      0.82      4436
```

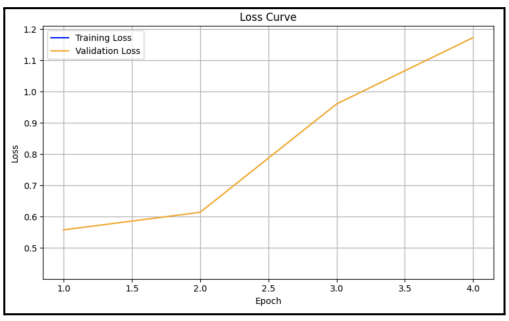


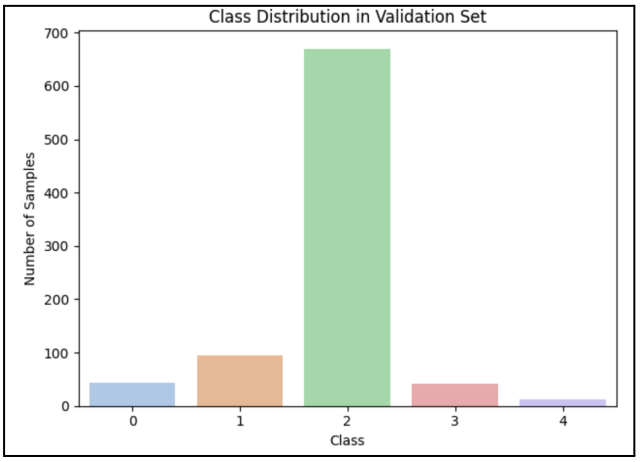
BETO Multi Class:

The screenshots below show the performance metrics, loss curve, and class distribution for the BETO multi-class model, illustrating its difficulty in identifying minority classes due to data imbalance.



```
📊 Final Evaluation Metrics (Multi-Class Model)
Accuracy: 0.7912
Precision (weighted): 0.7591
Recall (weighted): 0.7912
F1 Score (weighted): 0.7623
Precision (macro): 0.4908
Recall (macro): 0.3432
F1 Score (macro): 0.3726

Classification Report (Per Class):

                                precision    recall  f1-score   support

1. Sexual Insults & Objectification       0.44      0.09      0.15        43
  2. Gendered Stereotypes & Insults       0.42      0.40      0.41        95
       3. General Hostile Language        0.84      0.94      0.89       670
          4. Threats & Physical Harm      0.75      0.29      0.41        42
  5. Victim Blaming & Justification       0.00      0.00      0.00        12

                           accuracy                           0.79       862
                          macro avg       0.49      0.34      0.37       862
                       weighted avg       0.76      0.79      0.76       862
```



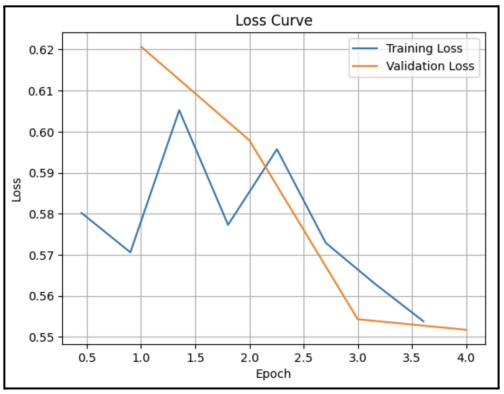


Appendix I

BERTIN Binary:

The figure below presents BERTIN's binary classification results, including evaluation metrics and the loss curve across epochs. While training was stable, the model struggled to recall sexist content despite high accuracy on the non-sexist class.

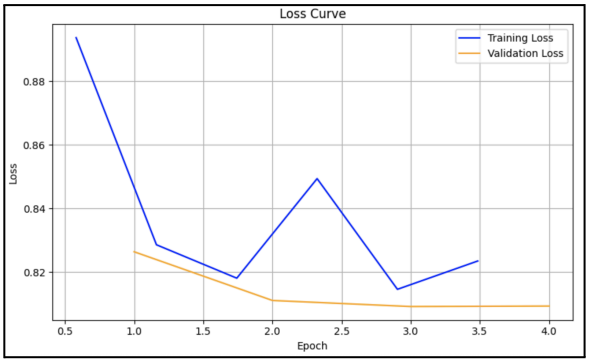
```
Final Evaluation Metrics:
Accuracy: 0.757439134355275
Precision: 0.7432201557292801
Recall: 0.757439134355275
F1 Score: 0.7218701595007933

Classification Report (per class):
              precision    recall  f1-score   support

  non-sexist       0.77      0.94      0.85      3160
      sexist       0.68      0.29      0.41      1276

    accuracy                           0.76      4436
   macro avg       0.72      0.62      0.63      4436
weighted avg       0.74      0.76      0.72      4436
```

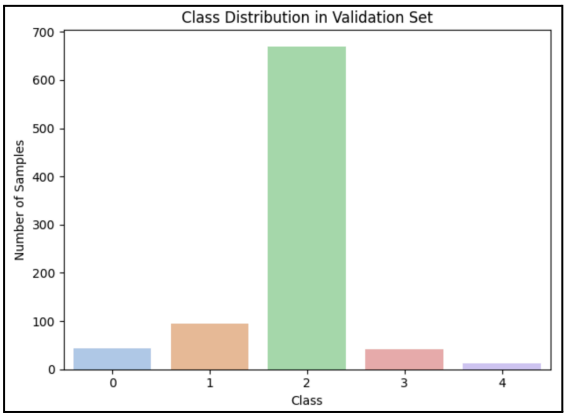


BERTIN Multi Class:

The visual below shows BERTIN's performance on the multiclass task. While overall accuracy was high, the model failed to detect minority classes, as seen in the flat recall scores for all but the majority class (General Hostile Language).

```
Final Evaluation Metrics (Multi-Class Model)
Accuracy: 0.7773
Precision (weighted): 0.6041
Recall (weighted): 0.7773
F1 Score (weighted): 0.6799
Precision (macro): 0.1555
Recall (macro): 0.2000
F1 Score (macro): 0.1749

Classification Report (Per Class):

                                precision    recall  f1-score   support

1. Sexual Insults & Objectification       0.00      0.00      0.00        43
  2. Gendered Stereotypes & Insults       0.00      0.00      0.00        95
       3. General Hostile Language        0.78      1.00      0.87       670
          4. Threats & Physical Harm      0.00      0.00      0.00        42
  5. Victim Blaming & Justification       0.00      0.00      0.00        12

                           accuracy                           0.78       862
                          macro avg       0.16      0.20      0.17       862
                       weighted avg       0.60      0.78      0.68       862
```





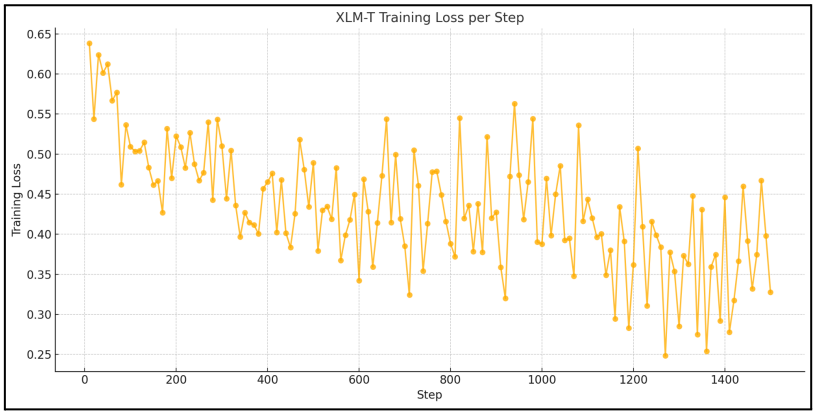
Appendix J

XLM-T Binary:

The graphic below displays XLM-T's binary classification performance, highlighting its strong accuracy and balanced precision–recall for both classes. The training loss curve indicates steady convergence with moderate variance.

Classification Report (Per Class Metrics) – Binary Classification:

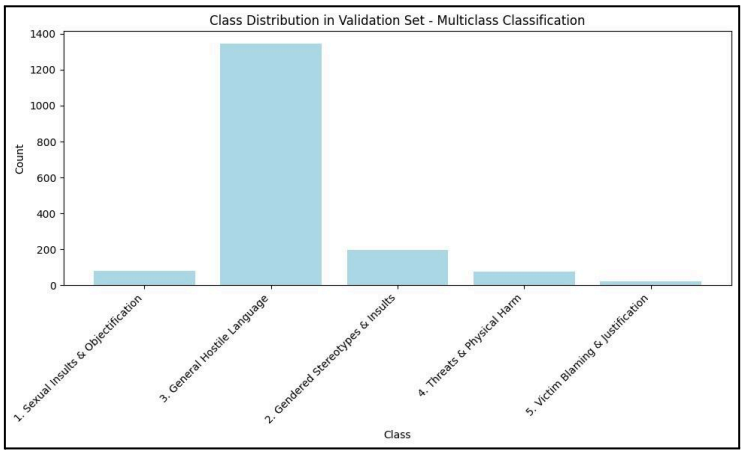
	precision	recall	f1-score	support
Non-sexist	0.85	0.90	0.88	6353
Sexist	0.70	0.61	0.66	2518
accuracy			0.82	8871
macro avg	0.78	0.76	0.77	8871
weighted avg	0.81	0.82	0.81	8871



XLM-T Multi Class:

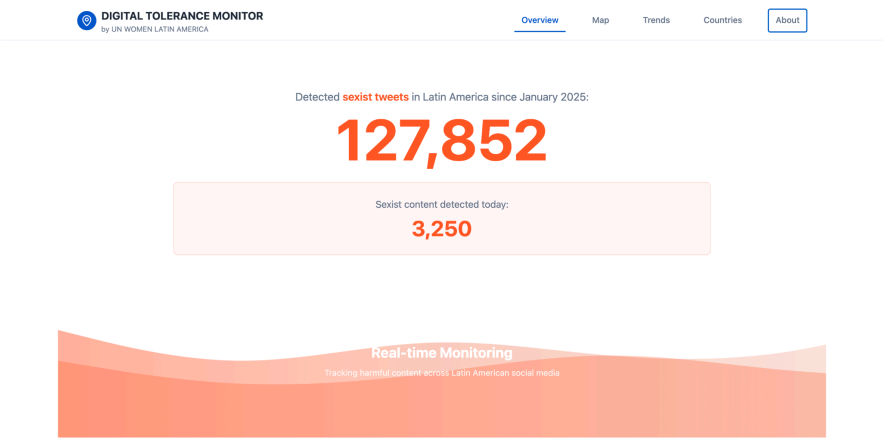
The figure below shows XLM-T's multiclass classification performance. While it achieves high accuracy, precision and recall vary across classes, reflecting challenges tied to class imbalance shown in the validation distribution.

```
Classification Report - Multiclass Classification:
                                precision    recall  f1-score   support

1. Sexual Insults & Objectification       0.39      0.27      0.32        81
       3. General Hostile Language        0.90      0.92      0.91      1347
  2. Gendered Stereotypes & Insults       0.55      0.55      0.55       196
         4. Threats & Physical Harm       0.56      0.62      0.59        77
  5. Victim Blaming & Justification       0.33      0.14      0.19        22

                          accuracy                           0.82      1723
                         macro avg       0.55      0.50      0.51      1723
                      weighted avg       0.81      0.82      0.82      1723
```



Appendix K

The screenshot below illustrates a prototype of the Digital Tolerance Monitor, showcasing the volume of sexist tweets detected across Latin America. Tools like this can power public-facing dashboards to track trends and promote digital accountability.



13. References

Cañete, J., Chaperon, G., Fuentes, R., & Pérez, J. (2020). *BETO: Spanish BERT model – dccuchile/bert-base-spanish-wwm-cased* [Model]. Hugging Face. https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

Codrea-Rado, A. (2017, October 16). *#MeToo movement floods social media with stories of harassment and assault*. *The New York Times*. https://www.nytimes.com/2017/10/16/technology/metoo-twitter-facebook.html

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … & Stoyanov, V. (2020). *XLM-RoBERTa: A multilingual masked language model – FacebookAI/xlm-roberta-base* [Model]. Hugging Face. https://huggingface.co/FacebookAI/xlm-roberta-base

De la Rosa, M., & Román, J. (2021). *BERTIN: RoBERTa-base model for Spanish – bertin-project/bertin-roberta-base-spanish* [Model]. Hugging Face. https://huggingface.co/bertin-project/bertin-roberta-base-spanish

Oxford University Press. (n.d.). *Sexism*. In *Oxford Learner's Dictionaries*. Retrieved July 6, 2025, from https://www.oxfordlearnersdictionaries.com/definition/english/sexism

UNICC. (n.d.). *Goal 5: Gender equality*. UNICC. https://www.unicc.org/what-we-do/unicc-for-the-sustainable-development-goals/goal-5-gender-equality/

UNICC (2024). *UPV Capstone: Women Safer Online – Data folder*. GitHub: https://github.com/UN-ICC/UPV-capstone-Women-Safter-Online/tree/main/Data