ECORBI-UPV at EXIST 2025: Large Language Models and Embedding Strategies for Sexism Detection in Tweets

Notebook for the EXIST Lab at CLEF 2025

Eurídice Corbí Verdú[1], Marc Franco-Salvador[2*]

[1]Universitat Politècnica de València, Spain
[2]United Nations International Computing Centre (UNICC)*, Valencia, 46930, Spain

Abstract

This paper presents our participation in Task 1 of the EXIST 2025 shared task, which focuses on the identification and characterization of sexism in tweets across three subtasks: (1) binary sexism detection, (2) categorization of sexist content, and (3) identification of the source and target of the sexist message. We leverage semantic embeddings generated using pre-trained models from Google's Generative AI suite, evaluated in both frozen and fine-tuned forms. Classification is carried out using traditional machine learning models such as Random Forest, SVM, and MLP. Experiments are conducted in both English and Spanish, with results evaluated using 10-fold cross-validation. Our findings demonstrate that fine-tuned Gemini embeddings consistently outperform generic representations. Challenges remain particularly in subtask 1.3 due to label ambiguity and sparsity.

Keywords

Sexism Identification, Large Language Models, Embeddings, Text Classification, Multilingual NLP

1. Introduction

The EXIST 2025 shared task [1] focuses on the automatic detection and characterization of sexism across different social media platforms and languages. Since its introduction in 2021, EXIST has evolved to address increasingly complex and socially relevant subtasks, reflecting the urgent need for automated tools to tackle gender-based discrimination online.

Sexism remains a persistent and deeply rooted issue in contemporary society, primarily affecting women. It is defined as discrimination based on sex [3], and it manifests through a variety of social and linguistic behaviors. Two particularly pervasive forms are hostile and benevolent sexism. While the former is overt and aggressive, benevolent sexism presents itself as seemingly positive or protective attitudes that reinforce traditional gender roles. Although less explicit, this second form can be equally damaging by limiting women's social roles and opportunities [4].

Twitter (now rebranded as *X*), as a major platform for public conversation, reflects and amplifies these dynamics. While it serves as a space for social advocacy and visibility, it also facilitates the dissemination of sexist narratives, often normalized under the guise of humor, opinion, or cultural references. Given the speed and scale at which content is generated on such platforms, automatic detection systems are essential to identify, categorize, and ultimately help mitigate the spread of gender-based discrimination online.

In recent years, research on automatic sexism detection has evolved from traditional machine learning approaches to neural-based architectures. Early systems relied on TF-IDF or bag-of-words representations with classifiers such as SVM or logistic regression. More recent systems have leveraged contextual embeddings from pre-trained language models such as BERT, RoBERTa, and BERTweet [8, 9], achieving state-of-the-art results in offensive language and bias detection. In the EXIST shared tasks, multilingual models like XLM-RoBERTa have shown strong performance in capturing culturally nuanced expressions of sexism [10].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*This publication does not reflect the position or views of UNICC and represents only the authors' views.
euridicecorbi@gmail.com (E. C. Verdú); francom@unicc.org (M. Franco-Salvador)

Several approaches to automatic sexism detection have emerged in recent editions of the EXIST shared task. In 2024, top-performing systems leveraged transformer-based models such as RoBERTa, XLM-RoBERTa, and BERTweet, often combined with data augmentation and disagreement-aware strategies. For example, the NYCU-NLP team [5] employed diverse fine-tuned transformer encoders and annotator demographic features, achieving strong results across subtasks. Similarly, the UO-LIA team [6] explored different pooling strategies and granular modeling of annotation variability. A detailed overview of all contributions can be found in the CLEF 2024 Working Notes [7].

The 2025 edition of EXIST covers content from Twitter, TikTok, and memes, structured into three main tasks: (1) identifying whether content is sexist, (2) classifying the type of sexism, and (3) identifying the source and target of the sexist discourse. This work focuses exclusively on the Twitter data, involving English and Spanish languages, and adopts the *Learning with Disagreement* annotation scheme that preserves label diversity from multiple annotators [2].

In parallel, large-scale embedding models have gained traction as lightweight yet powerful alternatives to end-to-end fine-tuned systems. Google's Gemini embedding models have shown promising performance in multilingual benchmarks such as the Massive Text Embedding Benchmark (MTEB) [14] and are designed to support downstream tasks like classification and semantic similarity through frozen or parameter-efficient fine-tuning strategies [11]. In this work, we investigate the use of Gemini-based embeddings as a flexible and efficient backbone for sexist language detection. Gemini provides high-quality multilingual semantic representations that allow us to bypass the computational and financial costs associated with training large language models from scratch. By leveraging Google's third-party cloud API at Google Cloud Platform (GCP), we significantly reduce development time and infrastructure complexity, enabling a faster deployment cycle. Additionally, Gemini embeddings have demonstrated strong performance on a range of natural language understanding and processing tasks, as evidenced by their results on the Massive Text Embedding Benchmark (MTEB), making them a practical and reliable choice for building a scalable, production-ready detection system.

To address this challenge, we generated embeddings using two pre-trained models from Google's Generative AI suite: `embedding-001`, a general-purpose encoder, and `gemini-embedding-exp-03-07`, a model optimized for classification tasks. These embeddings were either kept frozen or fine-tuned through lightweight neural layers and later fed into traditional classifiers. Our pipeline treats English and Spanish data independently, respecting linguistic differences in sexist expression. This architecture aims to balance effectiveness, interpretability, and computational efficiency [1].

The remainder of this paper is organized as follows. Section 2 introduces the proposed architecture and design considerations. Section 3 presents the dataset and task setup. Section 4 details the classification models and embedding strategies. Section 5 describes the experimental setup and evaluation protocol. Results and rankings are presented in Section 6, followed by a discussion in Section 7. Finally, Section 8 concludes the paper and outlines future work.

2. Proposed Approach

Our approach leverages semantic embeddings generated by large language models (LLMs) provided by Google's Generative AI suite. Specifically, we utilize two embedding models: `embedding-001`, a general-purpose encoder, and `gemini-embedding-exp-03-07`, an experimental model optimized for classification tasks.

The pipeline consists of generating fixed-size vector embeddings (768-dimensional) for each tweet using these models via API calls. We explore two strategies with these embeddings: (1) using them as frozen features directly fed into classical classifiers, and (2) fine-tuning the embeddings with lightweight trainable layers in a parameter-efficient way.

Figure 1 illustrates the full pipeline, including embedding generation, usage strategy, and classification.

While Gemini embeddings serve as the core representation model in our system, the pipeline remains model-agnostic and could support embeddings derived from other neural encoders such as XLM-RoBERTa or multilingual BERT. Gemini offers high-quality multilingual semantic vectors, allowing for

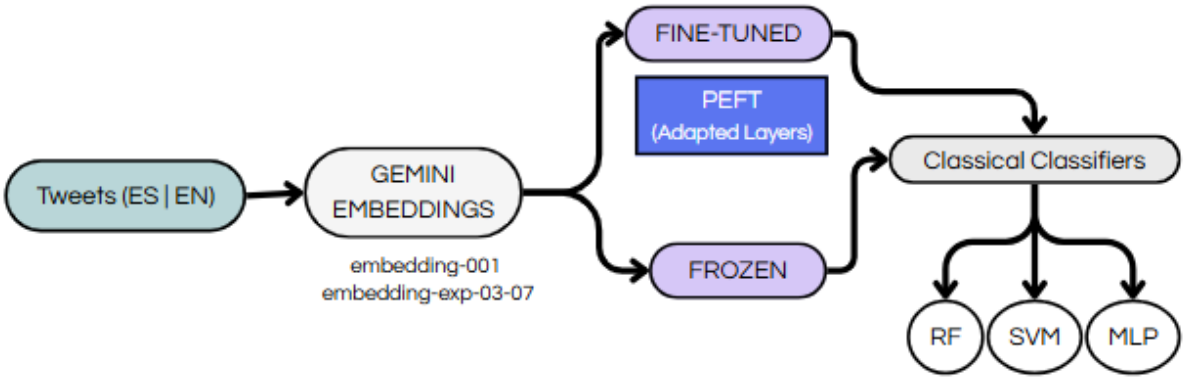

Figure 1: Overview of the proposed sexism detection pipeline.

robust feature extraction without the need to train task-specific deep models from scratch.

To adapt embeddings to the downstream task, we apply parameter-efficient fine-tuning (PEFT) techniques. Specifically, we freeze the base embedding model and introduce a lightweight classification head (typically a linear or shallow MLP layer), which is trained using cross-entropy loss on each subtask. Fine-tuning is performed separately for each subtask (1.1, 1.2, and 1.3) and language (EN/ES), using only the EXIST 2025 training data for that subtask. This strategy ensures low computational cost while improving task alignment.

The embeddings—either frozen or fine-tuned—are then passed to classical classifiers, including Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP), trained independently for each configuration. The modularity of this architecture allows for efficient experimentation and rapid switching between embedding strategies and classifiers.

By combining large language model embeddings with lightweight fine-tuning and classical classifiers, our system captures complex semantic nuances across English and Spanish tweets while maintaining interpretability and efficiency.

3. Task Description and Dataset

We worked with the official EXIST 2025 dataset for Task 1, which includes tweet-level annotations in both English and Spanish. Each tweet was annotated by multiple annotators, resulting in a set of potentially divergent labels per instance. To resolve this disagreement for training purposes, we applied majority voting to the label lists provided for each subtask. Tweets for which a clear majority label could not be determined were discarded from the training set.

Task 1 is divided into three subtasks:

- **Subtask 1.1 – Sexism Identification:** A binary classification task in which each tweet is labeled as either sexist or non-sexist.
- **Subtask 1.2 – Sexism Categorization:** A multi-class classification task where each sexist tweet must be assigned to one of the following categories: *ideological and inequality, stereotyping and dominance, objectification, misogyny and violence,* or *non-sexist.*
- **Subtask 1.3 – Source and Target Identification:** A multi-label classification task where tweets may be annotated with zero, one, or more labels indicating the source (e.g., individual, group, media) and target (e.g., individual woman, feminist movement) of the sexist message. This task introduces additional complexity due to sparse label distributions and overlapping roles.

For Subtask 1.1, we computed the most frequent label among the `labels_task1_1` field. For Subtasks 1.2 and 1.3, which include multiple possible labels, we discarded tweets with empty or unresolved

annotations. The labels were treated as single-label (1.2) or multi-label (1.3) targets, depending on task formulation.

We maintained the original `lang` field to process English and Spanish tweets separately, allowing for language-specific model training and evaluation. The label distributions across tasks and languages were notably imbalanced, particularly in Subtask 1.3, where many tweets lacked clear source or target annotations, increasing classification difficulty.

No aggressive preprocessing was applied to the tweet text. Specifically, we chose to retain original casing, punctuation, user mentions, and emojis. This decision was based on the fact that many linguistic and pragmatic cues in Twitter (now X)—such as emphasis, sarcasm, or gendered tone—are often conveyed through informal conventions, including emojis. Since our semantic embeddings were generated using large language models capable of interpreting such features, we considered their removal potentially detrimental to performance.

The final processed dataset was stored in `.parquet` format and included tweet IDs, raw text, resolved label(s), and language metadata. We applied a consistent processing pipeline across all three subtasks and used 10-fold stratified cross-validation for evaluation, prioritizing data efficiency and robustness over the predefined development/test splits. This preprocessed dataset served as the foundation for all subsequent embedding and classification experiments described in the following sections.

4. Methodology

To establish a strong baseline and evaluate the benefits of using semantic embeddings, we employed several classical machine learning models widely used in text classification tasks.

4.1. Baseline Models

Logistic Regression (LR) and **XGBoost** were chosen as baseline classifiers due to their robustness, interpretability, and proven effectiveness in various NLP tasks. Logistic Regression uses a linear decision boundary and probabilistic outputs, making it suitable for binary and multiclass tasks with relatively fast training times. XGBoost is a powerful gradient boosting framework that builds an ensemble of decision trees, enabling it to capture complex feature interactions and nonlinear relationships often missed by linear models. Both models were trained on vector representations derived from word n-grams (1,2) and character n-grams (3,5), capturing both lexical and sub-lexical features relevant in social media text.

4.2. Embedding-Based Models

Our main approach utilizes semantic embeddings generated from large language models (LLMs) provided by Google's Generative AI suite, specifically the `embedding-001` and `gemini-embedding-exp-03-07` models. These embeddings convert raw tweet texts into dense, fixed-length vectors encoding rich contextual and semantic information beyond traditional sparse representations.

We experimented with two embedding strategies:

- **Frozen embeddings:** Embeddings are computed once and kept fixed during training. Classical classifiers such as SVM and MLP operate on these static representations, relying on the pretrained semantic knowledge encoded by the LLMs. This approach significantly reduces training time and computational requirements, as only the classifier parameters are optimized.
- **Fine-tuned embeddings:** We applied parameter-efficient fine-tuning (PEFT) techniques, such as adapter modules or low-rank adaptations, to update only a small subset of the embedding model's parameters. This allows the embeddings to adapt to the sexism detection task without requiring full backpropagation through the entire LLM, thus balancing efficiency and task-specific specialization.

In our case, PEFT was applied using the same labeled training data used for the downstream classification tasks, and optimized using categorical cross-entropy as the training objective. This enabled task-specific adaptation of the embeddings while keeping the majority of the model parameters frozen, thus making training feasible under limited resources.

Unless otherwise stated, results reported in subsequent sections correspond to classifiers trained on fine-tuned embeddings, as this configuration consistently yielded higher performance.

4.3. Classification Algorithms

The embedding representations were fed into the following classifiers:

- **Support Vector Machine (SVM):** SVM with a linear kernel is effective for high-dimensional feature spaces and is commonly used in text classification. It attempts to find the optimal hyperplane separating classes with maximum margin, providing good generalization performance.
- **Multi-Layer Perceptron (MLP):** A feed-forward neural network with a single hidden layer and nonlinear activations (e.g., ReLU), capable of modeling complex patterns in the embedding space. MLPs can learn nonlinear decision boundaries, which may be beneficial for nuanced tasks like sexism detection.
- **Random Forest (RF):** RF constructs an ensemble of decision trees trained on random subsets of features and data samples, providing robustness to overfitting and ease of use. However, despite its theoretical advantages, RF showed lower predictive performance and higher computational costs compared to LR and XGBoost in our experiments. Consequently, RF was excluded from the final system runs and submissions.

4.4. Pipeline Overview

The full pipeline begins with raw tweets being converted to embeddings through API calls. Depending on the experiment, these embeddings are either kept frozen or fine-tuned via PEFT. The resulting embeddings are then used to train the classifiers independently for each subtask and language, enabling models to adapt to task-specific label distributions and linguistic variations.

This approach allows us to leverage the rich semantic power of LLM embeddings while maintaining model efficiency and interpretability through classical classifiers. Overall, this modular and language-specific architecture allowed flexible experimentation across subtasks and supported robust performance comparisons across different embedding and classifier configurations.

5. Experimental Setup

All experiments were conducted using Python 3.10 within Google Cloud's Vertex AI environment, which provided scalable GPU resources suitable for efficient training and evaluation of classification models.

Semantic embeddings were generated via the Google Generative AI API using two models: `embedding-001` and `gemini-embedding-exp-03-07`. Each tweet was encoded into a 768-dimensional vector using the `CLASSIFICATION` mode, which is optimized for downstream classification tasks. To mitigate API rate limits and network instability, we implemented a checkpointing mechanism with periodic '.parquet' dumps and exponential backoff retry logic to ensure robustness in large-scale embedding extraction.

For classification, we trained Support Vector Machine (SVM) classifiers with linear kernels and default regularization parameters, selected for their strong performance in high-dimensional vector spaces. Multi-Layer Perceptron (MLP) classifiers were configured with a single hidden layer of 256 neurons and ReLU activation, optimized using early stopping based on validation loss. All models were implemented using scikit-learn and TensorFlow.

Parameter-efficient fine-tuning (PEFT) was applied to adapt the embeddings to the sexism detection task, using the same training data and classification objectives described in Section 4. This setup enabled semantic adaptation without backpropagating through the full LLM backbone.

To handle the known class imbalance in the EXIST 2025 dataset—especially severe in Subtask 1.3—we used stratified 10-fold cross-validation for model selection and evaluation. This approach preserved label proportions across folds, ensuring balanced comparisons. Evaluation metrics included macro-averaged F1 (primary), micro F1, precision, recall, and accuracy.

Hyperparameter tuning was conducted via manual grid search, with macro F1 on validation folds as the selection criterion. The best-performing configuration per subtask and language was retained for final predictions. Final predictions were obtained directly from these models, with no ensemble methods or post-processing applied.

This setup allowed us to systematically explore the interaction between embedding strategies, model architectures, and multilingual label distributions under controlled and reproducible conditions.

6. Results

We report macro-averaged F1-scores obtained via 10-fold cross-validation for each subtask, classifier, and embedding type. All the results presented in this section correspond to our best-performing configurations using **fine-tuned embeddings**, as they consistently outperformed their frozen counterparts across tasks and languages.

The results are shown for each subtask separately and are broken down by language and system version. These include:

- **Version 1:** A monolithic setup using the same embedding-classifier combination for both English and Spanish.
- **Version 2:** An alternative combination using a different embedding-classifier pair for both languages.
- **Version 3:** A hybrid system that selects the best-performing configuration independently for each language.

Subtask 1.1: Sexism Identification

Table 1 shows macro F1-scores for Subtask 1.1. The best performance in Spanish was achieved using MLP with `embedding-001`, while SVM with `gemini-embedding-exp-03-07` performed better on English.

Table 1
Subtask 1.1 (fine-tuned): Macro F1-scores by system version and language.

Version	Model	Embedding	Language	Macro F1
Version 1	MLP	embedding-001	EN	0.782
			ES	**0.877**
Version 2	SVM	gemini-embedding-exp-03-07	EN	0.828
			ES	0.839
Version 3	SVM	gemini-embedding-exp-03-07	EN	0.828
	MLP	embedding-001	ES	**0.877**

Subtask 1.2: Sexism Categorization

This task is more complex due to class imbalance and semantic overlap between categories. Table 2 shows macro F1-scores for each version. The best configuration involved combining MLP for English and SVM for Spanish, using fine-tuned embeddings.

Table 2
Subtask 1.2 (fine-tuned): Macro F1-scores by system version and language.

Version	Model	Embedding	Language	Macro F1
Version 1	SVM	embedding-001	EN	0.608
			ES	**0.808**
Version 2	SVM	gemini-embedding-exp-03-07	EN	0.695
			ES	0.692
Version 3	MLP	gemini-embedding-exp-03-07	EN	0.701
	SVM	embedding-001	ES	**0.808**

Subtask 1.3: Source and Target Identification

Table 3 presents results for Subtask 1.3, the most difficult task due to sparse labels and label co-occurrence. The highest macro F1 scores were achieved using fine-tuned MLP with Gemini embeddings in both languages.

Table 3
Subtask 1.3 (fine-tuned): Macro F1-scores by system version and language.

Version	Model	Embedding	Language	Macro F1
Version 1	MLP	embedding-001	EN	0.684
			ES	**0.834**
Version 2	SVM	embedding-001	EN	0.647
			ES	0.818
Version 3	MLP	gemini-embedding-exp-03-07	EN	0.768
			ES	**0.782**

EXIST 2025 Official Rankings

Table 4 summarizes our team's performance in the official evaluation across all three subtasks, including soft and hard evaluations. The total number of participating systems is indicated for each case.

Table 4
Official EXIST 2025 rankings (position / total systems).

Subtask	Type	Rank	Total
1.1	Soft	61	67
	Hard	158	189
1.2	Soft	24	56
	Hard	95	138
1.3	Soft	34	53
	Hard	105	138

These rankings demonstrate that while our system did not reach the top tier, it remained competitive, particularly in Subtask 1.2. Future work will explore improvements in class imbalance handling and multi-label modeling for better ranking in complex subtasks.

7. Analysis

Our experiments reveal that the selection of embeddings, classifiers, and fine-tuning strategies significantly influences performance across all three subtasks. In particular, models using **fine-tuned embeddings** consistently outperformed those based on frozen representations, confirming the effectiveness of adapting semantic features to the task.

The binary sexism identification task (Subtask 1.1) shows the highest performance, particularly when using MLP with fine-tuned `embedding-001` for Spanish tweets and SVM with fine-tuned `gemini-embedding-exp-03-07` for English tweets. This suggests that adapting embeddings to specific languages and subtasks captures linguistic nuances effectively.

The multi-class and multi-label subtasks (1.2 and 1.3) present greater challenges due to increased complexity, class imbalance, and ambiguity in labels. Macro F1 scores are generally lower for these subtasks, reflecting the difficulty in fine-grained categorization and source-target identification. Nonetheless, parameter-efficient fine-tuning (PEFT) provides measurable improvements in generalization and task alignment, especially in low-resource or imbalanced settings.

Additionally, the strategy of selecting the best model configuration for each language and task independently leads to improved overall performance. This highlights the importance of flexible, task-specific pipelines in handling the diverse expressions of sexism in multilingual social media data.

Future directions include exploring more advanced metric learning techniques and label adaptation methods to better manage label scarcity and annotator disagreement, potentially enhancing robustness and accuracy across subtasks.

Overall, our findings underscore the necessity of embedding adaptation, careful classifier choice, and task-tailored tuning to effectively tackle sexism detection and characterization in real-world social media data.

8. Conclusion and Future Work

In this work, we presented our approach to sexism identification and characterization in tweets for the EXIST 2025 shared task. By leveraging semantic embeddings from large language models in combination with classical classifiers and parameter-efficient fine-tuning techniques, we achieved competitive results across multiple subtasks and languages.

Our experiments confirmed that **fine-tuning embeddings** consistently improved performance over frozen counterparts, particularly in multilingual and nuanced classification settings. Moreover, tailoring model configurations per task and language proved beneficial in capturing linguistic diversity and reducing error propagation across subtasks.

Despite these promising outcomes, several challenges persist—most notably in the fine-grained categorization and source-target identification subtasks, where label sparsity and semantic ambiguity hinder classifier performance.

For future work, we plan to explore advanced aggregation architectures such as Deep Averaging Networks (DAN), as well as improved implementations of zero-shot and few-shot learning techniques. These approaches may enhance generalization in low-resource scenarios and better capture subtle linguistic cues.

Additionally, incorporating domain adaptation techniques and modeling annotator disagreement more explicitly may improve system robustness and fairness. Expanding our experiments to other modalities such as memes and TikTok videos, as featured in the broader EXIST 2025 setup, is also a compelling next step.

Overall, our findings contribute to the ongoing development of interpretable, adaptable, and resource-efficient systems for sexism detection in multilingual social media environments.

Acknowledgments

We would like to thank the organizers of the EXIST shared task and the CLEF community for providing the dataset and evaluation framework that made this work possible.

This work has been developed as part of the author's Bachelor's Thesis at Universitat Politècnica de València. The project benefited from institutional support and resources provided by the university.

We also thank the developers of Google's Generative AI suite for enabling access to powerful embedding models that were fundamental for our experiments.

Declaration on Generative AI

The authors used Generative AI tools (e.g., ChatGPT) in a limited capacity to assist with phrasing and text refinement. All content was written, reviewed, and edited primarily by the authors, who take full responsibility for the final version of the manuscript.

References

[1] Laura Plaza, Jorge Carrillo-de-Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, Roser Morante (2025). Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*. Jorge Carrillo-de-Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, Nicola Ferro (Eds.)

[2] Laura Plaza, Jorge Carrillo-de-Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, Roser Morante (2025). Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview). *CLEF 2025 Working Notes*. Guglielmo Faggioli, Nicola Ferro, Paolo Rosso, Damiano Spina (Eds.)

[3] Real Academia Española (2014). *Diccionario de la lengua española* (23.ª ed.). Madrid: Espasa-Calpe.

[4] Healthline Media (2023). ¿Qué es el sexismo? Tipos, ejemplos y cómo enfrentarlo. *Healthline en español*.

[5] C.-H. Li, Y.-C. Liao, J.-C. Huang, C.-S. Wang (2024). NYCU-NLP at EXIST 2024: Leveraging Transformers with Diverse Annotator Information for Sexism Detection. *CLEF 2024 Working Notes*, CEUR Workshop Proceedings.

[6] J. M. Perea-Ortega, M. T. Martín-Valdivia, L. A. Ureña-López (2024). UO-LIA at EXIST 2024: Exploring Pooling Strategies and Disagreement-aware Architectures for Sexism Identification. *CLEF 2024 Working Notes*, CEUR Workshop Proceedings.

[7] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (eds.) (2024). *CLEF 2024 Working Notes. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. Grenoble, France, 9–12 September, 2024. CEUR Workshop Proceedings, Vol. 3740.

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[9] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv preprint arXiv:2010.12421*.

[10] Plaza, L., Carrillo-de-Albornoz, J., Amigó, E., Rosso, P., Morante, R., & Spina, D. (2021). Overview of EXIST 2021: Explainable sexism identification in social networks. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of CLEF 2021* (pp. 406–426). Springer, Cham.

[11] S. Cheng, K. Lee, Y. Lin, et al. (2024). Evaluating Gemini Embeddings on Massive Multilingual Tasks. *arXiv preprint* arXiv:2503.07891.

[12] N. Muennighoff, S. Magister, J. Reimers, et al. (2023). MTEB: Massive Text Embedding Benchmark. *arXiv preprint* arXiv:2210.07316.

[13] Muller, T., Perez-Torró, G., & Franco-Salvador, M. (2024). Few-Shot Learning with Siamese Networks and Label Tuning. *arXiv preprint* arXiv:2203.14655.

[14] Muennighoff, N., Magister, S., Reimers, N., et al. (2023). MTEB: Massive Text Embedding Benchmark. *arXiv preprint* arXiv:2210.07316.