Clasificación de Sexismo en tweets

MIARFID - ALC - Laboratorio 1

Shiyi Cheng - Pablo Segovia Martínez

26 de febrero de 2026

1. Introducción

Este proyecto aborda la tarea de detección automática de contenido sexista en publicaciones de redes sociales, utilizando el dataset de la competición EXIST 2025. El objetivo es desarrollar y evaluar diferentes aproximaciones de clasificación binaria (sexista/no sexista) mediante modelos clásicos de machine learning y modelos de lenguaje pre-entrenados con fine-tuning.

2. Metodología

2.1. Preprocesamiento de Datos

Se han implementado dos estrategias de preprocesamiento:

- **Tweet Original**: Texto sin modificaciones, manteniendo URLs, menciones y emojis.
- **Text Clean**: Limpieza avanzada eliminando URLs, menciones, hashtags, y normalizando el texto.

2.2. Modelos Evaluados

Se ha experimentado con tres familias de modelos:

2.2.1. Modelos Clásicos

Modelos de machine learning tradicionales utilizando vectorización TF-IDF y Bag-of-Words:

- Regresión Logística
- Support Vector Machines (LinearSVC, SVC-RBF)
- Árboles de Decisión y Random Forest
- Gradient Boosting y AdaBoost
- Naive Bayes (Multinomial, Complement, Bernoulli)
- K-Nearest Neighbors
- Ensambles (Voting, Bagging, Stacking)

2.2.2. Modelos de Lenguaje Fine-tuned

Modelos transformer pre-entrenados ajustados con LoRA (Low-Rank Adaptation):

- **F2LLM-4B**: Modelo de 4B parámetros especializado en español
- **KaLM**: Modelo multilingüe con soporte para español

2.2.3. Modelos Generativos

Modelos grandes de lenguaje evaluados en modo zero-shot e inference con fine-tuning:

- **Ministral-3-8B-Instruct**: Modelo instruccional de 8B parámetros con cuantización FP8

3. Resultados

3.1. Métricas en Conjunto de Validación (DEV)

La Tabla 1 muestra los resultados de los modelos de lenguaje fine-tuned, mientras que la Tabla 2 presenta los mejores modelos clásicos.

Modelo	Accuracy	Precision	Recall	F1-Score
F2LLM-4B (tweet)	**0.8604**	0.8294	0.8642	**0.8464**
Ministral-3-8B (FT)	0.8451	**0.8587**	0.7802	0.8176
KaLM (tweet)	0.8429	0.8164	0.8346	0.8254
F2LLM-4B (clean)	0.8341	0.7581	**0.9210**	0.8317
KaLM (clean)	0.8363	0.8137	0.8198	0.8167
Ministral-3-8B (ZS)	0.8264	0.7892	0.8321	0.8101

Tabla 1: Resultados de Modelos de Lenguaje (DEV)

Modelo	Accuracy	Precision	Recall	F1-Score
Stacking	**0.7824**	0.7867	0.7012	**0.7415**
Gradient Boosting	0.7736	**0.8373**	0.6099	0.7057
LinearSVC	0.7736	0.7901	0.6691	0.7246
Logistic Regression	0.7725	0.7845	0.6741	0.7251
Bagging (LR)	0.7703	0.7849	0.6667	0.7210

Tabla 2: Mejores Modelos Clásicos (DEV) - TF-IDF

3.2. Análisis Comparativo

3.2.1. Rendimiento por Familia de Modelos

- **Modelos de Lenguaje**: Los modelos transformer fine-tuned superan significativamente a los modelos clásicos, con mejoras de hasta **+10.5 puntos** en F1-Score (F2LLM-4B vs Stacking). El fine-tuning de Ministral-3-8B mejora todas las métricas respecto a su versión zero-shot, alcanzando la mayor precision (0.8587) de todos los modelos evaluados.

- **Modelos Clásicos**: El ensamble mediante Stacking alcanza el mejor rendimiento (F1=0.7415), demostrando que la combinación de múltiples clasificadores mejora los resultados individuales.
- **Impacto del Preprocesamiento**:
  - Los textos originales (tweet) obtienen mejor *accuracy* y *precision*
  - Los textos limpios (clean) mejoran el *recall* en +5.7 puntos (F2LLM-4B)
  - El trade-off entre precision y recall depende de la estrategia de preprocesamiento

3.2.2. Mejores Configuraciones

1. **Mayor F1-Score**: F2LLM-4B con tweet original (0.8464)
2. **Mayor Recall**: F2LLM-4B con texto limpio (0.9210)
3. **Mayor Precision**: Ministral-3-8B fine-tuned (0.8587)
4. **Mejor modelo clásico**: Stacking con TF-IDF (0.7415)

4. Conclusiones

1. Los modelos de lenguaje pre-entrenados con fine-tuning superan ampliamente a los métodos clásicos de ML, especialmente en tareas de comprensión contextual como la detección de sexismo.
2. El modelo F2LLM-4B con texto original alcanza el mejor equilibrio entre precision y recall (**F1=0.8464**), constituyendo la mejor solución para esta tarea.
3. La estrategia de preprocesamiento debe seleccionarse según el objetivo: texto limpio maximiza recall (detectar más casos positivos), mientras que el texto original mejora precision (evitar falsos positivos).
4. Los ensambles de modelos clásicos (Stacking) son competitivos y computacionalmente más eficientes que los LLMs, representando una alternativa válida para entornos con recursos limitados.
5. El fine-tuning con LoRA permite adaptar modelos grandes (4B-8B parámetros) con recursos limitados, manteniendo alta calidad en las predicciones.

4.1. Trabajo Futuro

- Explorar ensambles de modelos de lenguaje (votación entre F2LLM, KaLM y Ministral)
- Análisis de errores para identificar patrones lingüísticos desafiantes
- Optimización de hiperparámetros mediante búsqueda sistemática
- Evaluar arquitecturas más recientes y técnicas de prompting avanzadas