Clasificación de sexismo en tweets

MIARFID - ALC - Laboratorio 1

Shiyi Cheng - Pablo Segovia Martínez

28 de febrero de 2026

1. Introducción

Detección automática de contenido sexista en tweets usando el dataset EXIST 2025. Clasificación binaria (sexista/no sexista) mediante modelos clásicos y LLMs con fine-tuning.

2. Metodología

2.1. Preprocesamiento

V1: tweet original + text_clean básico (lowercase, elimina URLs/menciones/hashtags), 13 features.

V2: Mejoras en text_clean: eliminación de stopwords, preservación de negaciones {no, nunca, jamás, nada, nadie, tampoco, ni, sin} con marcado NEG_, normalización de acentos/elongaciones, 16 features. Reducción de texto: ∼52 %.

2.2. Modelos evaluados

Clásicos (TF-IDF): Logistic Regression, LinearSVC, Random Forest, Gradient Boosting, Naive Bayes, Stacking.

LLMs con LoRA: F2LLM-4B (4B params, español), KaLM (multilingüe), Ministral-3-8B-Instruct (8B params, FP8).

3. Resultados

3.1. Comparativa V1 vs V2

Observaciones: F2LLM-4B (tweet) V2 mejora +5.43 pp recall y +0.68 pp F1. KaLM V2 empeora. Ministral-3-8B FT V2 falla drásticamente (F1=0.0847).

Evolución V2: Stacking -0.26 pp F1, LogReg +0.91 pp F1. Impacto mixto de eliminación de stopwords.

3.2. Métricas finales V2 (conjunto de validación)

3.3. Análisis comparativo

Conclusión: Tweet original supera a text_clean en LLMs (F2LLM-4B: -0.0215 en V2). Los transformers se benefician del contexto completo.

Puntos clave: F2LLM-4B mejora por varianza aleatoria (no cambios sistemáticos). KaLM empeora. Modelos clásicos, mejora en LogReg y empeora en stacking (impacto mixto).

Modelo	Texto	Ver	Acc	Prec	Rec	F1
F2LLM-4B	tweet	V1	0.8604	0.8294	0.8642	0.8464
	tweet	V2	**0.8593**	0.7966	**0.9185**	**0.8532**
	clean	V1	0.8473	0.8122	0.8543	0.8327
	clean	V2	0.8341	0.7581	**0.9210**	0.8317
KaLM	tweet	V1	0.8363	0.8137	0.8198	**0.8167**
	tweet	V2	0.8143	0.7682	0.8346	0.8000
	clean	V1	0.8363	0.7963	**0.8494**	**0.8220**
	clean	V2	—	—	—	—
Ministral-3-8B	ZS	V1	0.8264	0.7892	0.8321	**0.8101**
	ZS	V2	0.8143	0.7500	**0.8741**	0.8073
	FT	V1	**0.8451**	**0.8587**	0.7802	**0.8176**
	FT	V2	0.5725	0.9000	0.0444	0.0847

Tabla 1: Comparativa completa V1 vs V2 - Modelos LLM

Modelo	Ver	Acc	Prec	Rec	F1
Stacking	V1	**0.7846**	**0.7895**	**0.7037**	**0.7441**
Stacking	V2	0.7824	0.7867	0.7012	0.7415
LogReg + TF-IDF	V1	0.7637	0.7699	0.6691	0.7160
LogReg + TF-IDF	V2	**0.7725**	**0.7845**	**0.6741**	**0.7251**

Tabla 2: Comparativa V1 vs V2 - Modelos clásicos

3.4. Rendimiento por familia

- **LLMs**: Superan modelos clásicos +10.93 pp F1 (F2LLM-4B: 0.8532 vs Stacking: 0.7415).
- **Clásicos**: Stacking mejor V2 (F1=0.7415), computacionalmente eficientes.
- **Preprocesamiento**: Tweet original superior en LLMs (Δ=-0.0215). Preservación de negaciones crítica en text_clean.
- **Ensemble Top 5**: F1=0.8532, idéntico a F2LLM-4B individual. No aporta mejora.

3.5. Mejores configuraciones V2

1. **Mejor modelo general**: F2LLM-4B (tweet) - F1=0.8532, Acc=0.8593, Recall=0.9185
2. **Mejor clásico**: Stacking (TF-IDF) - F1=0.7415
3. **Ensemble**: Idéntico a F2LLM-4B individual, sin mejora

4. Conclusiones

1. **LLMs superan modelos clásicos**: +10.93 pp F1 (F2LLM-4B: 0.8532 vs Stacking: 0.7415).
2. **Modelo óptimo: F2LLM-4B (tweet) V2**: F1=0.8532, Recall=0.9185, Acc=0.8593. Mejora V1→V2 por**varianza aleatoria** (hiperparámetros y texto idénticos, thresholds diferentes: 0.3812→0.2214).

Modelo	Accuracy	Precision	Recall	F1-Score
F2LLM-4B (tweet)	**0.8593**	0.7966	**0.9185**	**0.8532**
F2LLM-4B (clean)	0.8341	0.7581	0.9210	0.8317
KaLM (tweet)	0.8143	0.7682	0.8346	0.8000
Ministral-3-8B (ZS)	0.8143	0.7500	0.8741	0.8073
Ministral-3-8B (FT)	0.5725	0.9000	0.0444	0.0847

Tabla 3: Resultados de Modelos de Lenguaje V2 (DEV)

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

Tabla 4: Mejores Modelos Clásicos V2 (DEV) - TF-IDF

3. **Preprocesamiento V2**: Preservación de negaciones crítica para text_clean. Eliminación de 313 stopwords: impacto mixto (LogReg +0.91 pp, Stacking -0.26 pp).
4. **Tweet original mejor que text_clean en LLMs**: F2LLM-4B: 0.8532 vs 0.8317, Δ=-2.15 pp. Transformers manejan ruido y aprovechan contexto completo.
5. **Ensemble sin mejora**: Votación mayoritaria = F2LLM-4B individual (F1=0.8532). F2LLM-4B domina predicciones.
6. **LoRA fine-tuning**: Éxito en F2LLM-4B/KaLM. Ministral-3-8B FT ha colapsado (F1=0.0847, Recall=0.0444).
7. **Reproducibilidad**: Fijar semillas aleatorias y ejecutar múltiples runs para distinguir mejoras reales de varianza estocástica.

4.1. Trabajo Futuro

- **Análisis de errores**: Identificar patrones fallidos (ironía, sarcasmo, sexismo implícito).
- **Ensambles avanzados**: Stacking con meta-aprendizaje, ponderación por confianza.
- **Optimización hiperparámetros**: Grid/Random Search (LoRA rank/alpha, LR, dropout).
- **Modelos recientes**: Llama 3, Mixtral 8x7B, RoBERTa-es/BETO, MarIA/BERTIN.
- **Augmentación de datos**: Back-translation, parafraseo con LLMs, generación sintética.
- **Explicabilidad**: LIME/SHAP, visualización de atención, rationales automáticos.