Proyecto Final

Resolución de un Problema de PLN

Tema: *Comparativa de enfoques de PLN para la detección de riesgo emocional en usuarios de Telegram con trastornos mentales en español*

Daniel Centurión, Manuel Núñez, Gustavo Báez
Universidad Comunera – Maestría en Ciencia de Datos

Resumen

Problema abordado: Detección automática de ansiedad en textos en español provenientes de Telegram.

Corpus utilizado: MentalRiskES 2023 — mensajes agrupados por usuario y etiquetados binariamente.

Modelos evaluados:

- **RoBERTa-base-bne**
- **SVM + TF-IDF**
- **FastText + TF-IDF**

Resultados clave:

- RoBERTa logró un **F1-score promedio de 0.72**
- El ajuste de umbrales y técnicas de balanceo fueron determinantes.
- El modelo profundo es más efectivo pero costoso; los modelos ligeros son viables con menor demanda computacional.

Introducción

- La salud mental es una problemática creciente a nivel global. Telegram contiene textos espontáneos que pueden revelar señales de malestar emocional.
- Aumento global de trastornos mentales post-COVID.
- Necesidad de herramientas automáticas de detección temprana.
- Los textos en redes sociales ofrecen oportunidades de análisis.

Problema y Objetivo

El aumento de trastornos de ansiedad exige herramientas automáticas para su detección temprana en textos reales.

Este trabajo propone abordar esa necesidad mediante **clasificación binaria por usuario**, a partir de mensajes de Telegram en español.

Se comparan tres enfoques representativos de PLN:

- **SVM + TF-IDF** (modelo tradicional)
- **FastText + TF-IDF** (modelo intermedio con embeddings)
- **RoBERTa-base-bne** (modelo profundo ajustado por fine-tuning)

Descripción del corpus

- MentalRiskES 2023: corpus diseñado para la detección temprana de trastornos mentales en español (IberLEF 2023).
- Contiene mensajes reales extraídos de grupos públicos de Telegram relacionados con salud mental.
- Los textos están agrupados por usuario, permitiendo clasificación a nivel persona (no por mensaje individual).
- Etiquetado binario realizado por 10 anotadores humanos mediante Doccano + Prolific.
- Cobertura de tres tipos de trastornos:
  - Ansiedad (ANX) — **foco de este estudio (altamente desbalanceado – control 11% del total)**
  - Depresión (DEP)
  - Trastornos de la conducta alimentaria (ED)
- El corpus tiene acceso restringido, solo disponible para investigación académica bajo licencia.

Metodología

Unificación por usuario: todos los mensajes de cada persona se concatenan en orden cronológico construyendo un único documento por usuario.

Clasificación binaria: 1: el usuario presenta señales de ansiedad, 0: usuario sin síntomas (control)

División del conjunto de datos (estratificada):

- Entrenamiento: 70%
- Validación: 20%
- Test: 10%

Pipeline general:

- Preprocesamiento textual
- Representación por usuario (según el modelo)
- Entrenamiento y calibración

Métricas de evaluación: F1-score por clase, Precisión (precision), Recall

Modelos Evaluados

Se compararon tres enfoques representativos de PLN, abarcando desde modelos tradicionales hasta redes neuronales profundas:

- **RoBERTa-base-bne**
  Modelo transformer preentrenado en español, ajustado mediante *fine-tuning*.
  Captura contexto y matices lingüísticos complejos.

- **FastText + TF-IDF**
  Combinación de embeddings preentrenados en español con ponderación léxica (TF-IDF).

- **SVM + TF-IDF**
  Clasificador lineal tradicional con vectorización TF-IDF.
  Incorporó calibración probabilística para mejorar la decisión final.

Estrategias de optimización

Para abordar el desbalance de clases y mejorar la capacidad de detección de casos “no sufre”, se aplicaron técnicas específicas por modelo:

- **RoBERTa-base-bne**
  - Ponderación de clases durante el entrenamiento
  - Ajuste de umbral de decisión (óptimo ≠ 0.5)

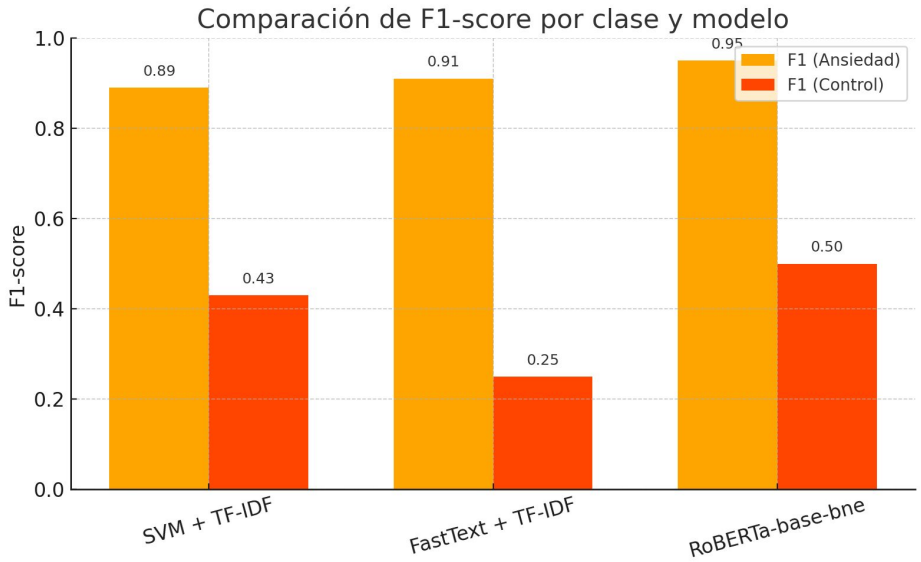
- **FastText + TF-IDF**
  - Ajuste de umbral mediante análisis de la curva precision-recall
  - Ponderación implícita vía TF-IDF

- **SVM + TF-IDF**
  - Oversampling con RandomOverSampler para equilibrar la clase minoritaria
  - Calibración probabilística (*CalibratedClassifierCV*)
  - Ajuste de umbral para maximizar macro-F1

Resultados

Rendimiento en conjunto de prueba: F1-score por clase y promedio ponderado en test

Modelo	F1 (Ansiedad)	F1 (Control)	F1 Promedio
SVM + TF-IDF	0.89	0.43	0.66
FastText + TF-IDF	0.91	0.25	0.58
RoBERTa-base-bne	**0.95**	**0.50**	**0.72**



Análisis Cualitativo

- Revisión manual de predicciones reveló patrones lingüísticos clave que influyen en el desempeño de los modelos.

Texto	Predicción	Prob. Ansiedad
No tengo ansiedad, solo algo de cansancio normal.	Sufre ansiedad	0.9844
Estoy harto, no sé cómo parar estos pensamientos.	Sufre ansiedad	0.9754
No puedo dormir, mi corazón late muy rápido por las noches.	Sufre ansiedad	0.9724
Todo está en calma, me siento estable emocionalmente.	Sufre ansiedad	0.9580
Me sudan las manos cuando estoy solo.	Sufre ansiedad	0.9191
La terapia me ha ayudado mucho, ya no tengo miedo.	Sufre ansiedad	0.8806
A veces me falta el aire sin razón aparente.	Sufre ansiedad	0.5346
Siento que algo malo va a pasar aunque no hay motivo.	Control (no sufre)	0.4739
Hoy me desperté feliz, con muchas ganas de salir.	Control (no sufre)	0.0569
Amo mi trabajo y estoy disfrutando la vida.	Control (no sufre)	0.0230

Limitaciones

Este estudio abordó exclusivamente el trastorno de **ansiedad**, dejando fuera otros presentes en el corpus (depresión, TCA), lo que reduce la generalización del enfoque.

El corpus presenta un fuerte **desbalance** de clases, con mayoría de usuarios etiquetados como “sufre”. Esto dificulta la detección precisa de la clase minoritaria (“no sufre”).

Se trabajó en Google Colab, lo que impuso **restricciones técnicas**, como poder de computo, memoria disponible y duración de las sesiones, limitando la capacidad de realizar experimentos más exhaustivos.

Trabajo Futuro

- Extender la metodología a depresión y TCA, ajustando parámetros según cada caso.
- Integrar herramientas como Weights & Biases (WandB) para trazabilidad y reproducibilidad.
- Explorar nuevas arquitecturas como BETO o modelos multilingües clínicos.
- Aplicar técnicas de interpretabilidad (LIME, SHAP, visualización de activaciones).
- Evaluar aumento o generación de datos sintéticos u otras técnicas, para compensar el desbalance de los datos.

Conclusiones

- RoBERTa-base-bne obtuvo el mejor rendimiento, especialmente en la detección de usuarios sin síntomas, aunque con un mayor costo computacional.
- SVM y FastText demostraron ser alternativas eficaces y livianas, especialmente útiles en entornos con recursos limitados.
- El enfoque propuesto es reproducible, escalable y adaptable a otros trastornos presentes en el corpus (como depresión y TCA).
- Este trabajo contribuye al desarrollo de herramientas automáticas de monitoreo emocional en español, con potencial aplicación en contextos clínicos y preventivos.

Bibliografía

[1] Ramos, L. V., García, C. M., Vázquez, J. M., & Álvarez, V. P. (2023). I2C-UHU at MentalRiskES 2023: Detecting and Identifying Mental Disorder Risks in Social Media using Transformer-Based Models. IberLEF 2023. Disponible en línea

[2] Mármol Romero, A. M., Moreno Muñoz, A., Plaza-del Arco, F. M., et al. (2024). MentalRiskES: A new corpus for early detection of mental disorders in Spanish. LREC-COLING 2024. https://aclanthology.org/2024.lrec-main.978

[3] Organización de MentalRiskES 2023. MentalRiskES 2023 – IberLEF Shared Task. Disponible en: https://sites.google.com/view/mentalriskes/home (Accedido: mayo 2025)