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

Daniel Centurión
Universidad Comunera, Facultad de Ingeniería,
Asunción, Paraguay,
danicen85@gmail.com

and

Manuel Nuñez
Universidad Comunera, Facultad de Ingeniería,
Asunción, Paraguay,
manuel.nunhez90@gmail.com

and

Gustavo Baez
Universidad Comunera, Facultad de Ingeniería,
Asunción, Paraguay,
gbaezf@gmail.com

Abstract

Este estudio compara distintos enfoques de clasificación para identificar señales de ansiedad en textos escritos en español, utilizando el corpus *MentalRiskES 2023*. A partir de los mensajes de usuarios agrupados por persona, se entrenaron tres modelos representativos: `SVM + TF-IDF`, `FastText` y `RoBERTa-base-bne` ajustado mediante *fine-tuning*. Los resultados muestran que el modelo basado en RoBERTa obtuvo el mejor rendimiento, especialmente en la clase minoritaria, gracias a técnicas de balanceo y ajuste de umbral. Si bien los modelos más simples ofrecieron resultados razonables, su capacidad para identificar correctamente a los usuarios sin síntomas fue más limitada. Más allá de los resultados, esta propuesta permitió explorar los desafíos propios de este tipo de tareas y sienta una base valiosa para considerar soluciones más complejas y robustas en trabajos futuros.

Keywords: salud mental, procesamiento de lenguaje natural, clasificación binaria, FastText, SVM, BETO.

1 Introducción

La salud mental se ha convertido en un tema prioritario en la agenda global debido al creciente número de personas que padecen trastornos como la ansiedad, la depresión y los trastornos de la conducta alimentaria (TCA). De acuerdo con la Organización Mundial de la Salud, uno de cada ocho individuos en el mundo sufre algún tipo de trastorno mental, y el impacto de la pandemia de COVID-19 ha agravado esta situación significativamente, incrementando los casos de ansiedad y depresión en más del 25% en tan solo un año [1].

Muchas personas atraviesan estos trastornos en silencio, sin recibir un diagnóstico ni tratamiento a tiempo. En este contexto, las redes sociales y las plataformas de mensajería se han convertido en una fuente valiosa de información, ya que numerosos usuarios comparten en ellas sus emociones, pensamientos y vivencias de forma espontánea. Esta realidad ha despertado un creciente interés por el desarrollo de sistemas automáticos de monitoreo emocional, orientados a identificar señales tempranas de malestar psicológico a

partir del análisis de textos escritos.

2 Planteamiento del Problema

Este trabajo se enfoca en la detección automática de riesgo emocional a partir de textos en español. Utilizando técnicas de procesamiento de lenguaje natural (PLN), se busca desarrollar y comparar modelos de clasificación binaria que permitan identificar si un usuario muestra o no señales de malestar emocional.

Este análisis aborda un problema de alto interés social y clínico, con el potencial de contribuir a la detección temprana de casos que requieran atención psicológica. Para ello, se trabaja con el corpus *MentalRiskES 2023*, un conjunto de datos que contiene mensajes en español publicados por usuarios que participan en grupos públicos de Telegram relacionados con salud mental, etiquetados según su riesgo de sufrir distintos trastornos mentales, como ansiedad, depresión y trastornos de la conducta alimentaria (TCA).

3 Objetivo del Estudio

El objetivo principal de este estudio es comparar distintos enfoques de PLN (desde modelos más tradicionales hasta otros basados en lenguaje preentrenado), con el fin de evaluar cuál de ellos ofrece mejores resultados en la detección de riesgo emocional en español, considerando no solo el rendimiento, sino también la interpretabilidad y la eficiencia de cada modelo. Para ello, se implementaron tres enfoques representativos:

- **Modelo basado en** `RoBERTa`: modelo de última generación preentrenado en español, adaptado al dominio de salud mental mediante *fine-tuning*.
- **Modelo** `FastText`: clasificador eficiente basado en incrustaciones de palabras, útil para tareas con recursos limitados.
- **SVM + TF-IDF**: enfoque clásico de clasificación usando vectores de frecuencia de término y un clasificador de margen máximo.

Además de la evaluación cuantitativa con métricas estándar (precisión, recall, F1-score), se incluye un análisis cualitativo que permite explorar el comportamiento de los modelos ante ejemplos concretos, favoreciendo su interpretabilidad.

En este estudio nos centramos exclusivamente en el trastorno de ansiedad, tanto por su alta prevalencia como por las limitaciones de tiempo y recursos computacionales disponibles. No obstante, se considera que la metodología desarrollada —que incluye el preprocesamiento textual, la representación de usuarios y la arquitectura de clasificación— puede ser replicada y adaptada fácilmente para abordar otros trastornos presentes en el corpus, lo que representa una valiosa oportunidad para trabajos futuros.

4 Descripción del corpus

Este estudio se basa en el corpus *MentalRiskES 2023*, una base de datos desarrollada en el marco de la tarea compartida homónima organizada como parte de IberLEF 2023, y descrita formalmente por Mármol-Romero et al. (2024) [2]. El corpus contiene mensajes escritos en español, extraídos de grupos públicos de Telegram dedicados a temáticas relacionadas con la salud mental.

La recolección de datos se realizó a partir de conversaciones reales, considerando únicamente los mensajes de texto. Luego, estos mensajes fueron agrupados por usuario y anotados mediante la plataforma Prolific, utilizando la herramienta Doccano. Diez anotadores independientes evaluaron los textos de cada usuario, asignando una etiqueta binaria por mayoría de votos: `1` si se identificaron señales de un posible trastorno mental, y `0` en caso contrario. En situaciones de empate, se adoptó una decisión conservadora, clasificando al usuario como `1` (sufre).

El corpus contempla tres tipos de trastornos: ansiedad (ANX), depresión (DEP) y trastornos de la conducta alimentaria (ED). Los mensajes de cada usuario se almacenan en archivos individuales en formato `.json`, organizados en subcarpetas según el tipo de trastorno. Las etiquetas correspondientes se encuentran en archivos `.csv`, ubicados en la carpeta `gold/`, y están disponibles tanto en formato binario (e.g., `bs`, `bc`) como en formato continuo (e.g., `rbs`), que indica cuántos anotadores marcaron al usuario como caso positivo.

El corpus se distribuye en dos versiones: una original, que preserva elementos como emojis, y otra procesada, con los textos normalizados. Cada archivo representa la totalidad de los mensajes de un único usuario, lo que permite abordar el problema como una tarea de clasificación a nivel de usuario.

El acceso al corpus es restringido y está destinado exclusivamente a fines académicos. Para obtenerlo, se requiere completar un formulario de solicitud y aceptar una licencia de uso disponible en la página oficial del desafío.

Aunque el corpus fue diseñado para permitir múltiples tareas (clasificación binaria, regresión, análisis temporal, entre otras), en este trabajo nos enfocamos únicamente en la tarea de clasificación binaria para el caso de ansiedad. Para más detalles sobre las demás tareas, se recomienda consultar la documentación oficial del reto [3].

5 Metodología

Este trabajo propone una comparación entre distintos enfoques de clasificación para detectar señales de riesgo emocional asociadas al trastorno de ansiedad, a partir de textos escritos en español. Para ello, se utiliza el corpus *MentalRiskES 2023*, que contiene mensajes agrupados por usuario y etiquetados de forma binaria según la presencia o ausencia de síntomas.

Dado que las etiquetas se encuentran a nivel de usuario, se optó por representar a cada individuo como un único documento resultante de la concatenación de todos sus mensajes en orden cronológico.

5.1 Preparación del conjunto de datos

Los mensajes de cada usuario fueron agrupados y concatenados, eliminando duplicaciones, caracteres especiales irrelevantes y conservando el orden temporal. Esta operación generó una estructura tabular donde cada fila corresponde a un usuario, su texto completo y la etiqueta binaria asociada: `1` (el usuario presenta señales de ansiedad) o `0` (no presenta señales). Un ejemplo simplificado se muestra en la Tabla 1.

user	**message_clean**	**label**
subject1	buenos días a tod s no sé si esta será la fina...	1
subject10	ola soi nuebo como estn a ustds los entimdm ke...	0
subject100	de donde son yo de perú supongo q tb sigue en ...	1

Table 1: Ejemplo de estructura del dataset tras la concatenación de mensajes por usuario.

5.2 Modelos evaluados

Se implementaron tres modelos representativos, cada uno correspondiente a una familia de enfoques en PLN:

- **SVM + TF-IDF**: modelo tradicional que utiliza vectores TF-IDF como representación del texto y un clasificador lineal con regularización.
- **FastText**: combinación de embeddings preentrenados en español con ponderación léxica mediante TF-IDF y un clasificador basado en regresión logística calibrada.
- **RoBERTa-base-bne**: modelo profundo preentrenado en español, ajustado mediante *fine-tuning* sobre los textos completos por usuario.

Cabe señalar que, en etapas preliminares del estudio, se exploraron otras arquitecturas preentrenadas como `dccuchile/bert-base-spanish-wwm-cased`, las cuales mostraron un rendimiento muy limitado en la detección de la clase minoritaria (F1-score de control cercano a cero). Esto motivó la elección de `RoBERTa-base-bne`, un modelo robusto entrenado sobre grandes corpus en español, cuya efectividad ha sido validada en tareas similares de clasificación emocional.

5.3 Diseño experimental y evaluación

El conjunto de datos fue dividido en subconjuntos de entrenamiento (70%), validación (20%) y prueba (10%) mediante partición estratificada. Esta división garantiza que la proporción de clases se mantenga equilibrada.

La evaluación se realizó en los conjuntos de validación y prueba, empleando como métricas principales la *precisión*, el *recall* y el *F1-score*, con énfasis en la clase minoritaria ("no sufre").

Debido al desbalance de clases, se exploraron estrategias adicionales para mejorar el desempeño, tales como:

- **Ponderación de clases**: asignación de pesos inversamente proporcionales a la frecuencia de cada clase durante el entrenamiento.
- **Ajuste de umbral**: búsqueda de umbrales de decisión alternativos a 0.5, optimizando el equilibrio entre precisión y recall.

6 Evaluación de resultados

Esta sección presenta los resultados obtenidos para los tres modelos evaluados: RoBERTa-base-bne, FastText + TF-IDF y SVM + TF-IDF. En cada caso se detallan las métricas principales obtenidas sobre los conjuntos de validación y prueba, así como una breve reflexión cualitativa sobre el comportamiento del modelo.

6.1 RoBERTa-base-bne

El modelo RoBERTa fue ajustado mediante *fine-tuning* durante 4 épocas, utilizando el conjunto de entrenamiento previamente definido. Se probaron tres configuraciones: sin balanceo, con pesos de clase, y con pesos + ajuste de umbral.

Configuración	F1 (Ansiedad)	F1 (Control)	Accuracy
RoBERTa base (0.5)	0.95	0.50	0.90
RoBERTa + pesos	0.94	0.53	0.89
RoBERTa + pesos + umbral ajustado	**0.94**	**0.57**	**0.88**

Table 2: Resultados del modelo RoBERTa bajo distintas configuraciones sobre el trastorno de ansiedad.

Análisis cualitativo: se observó que el modelo responde correctamente ante expresiones explícitas de sufrimiento emocional. No obstante, algunas expresiones ambiguas, irónicas o altamente contextuales resultaron difíciles de clasificar de forma precisa. El ajuste del umbral fue clave para mejorar la detección de la clase minoritaria.

(Agregar ejemplos de oraciones que muestren si clasifican bien o no. Indicar posibles sesgos)

6.2 SVM + TF-IDF

Este modelo se entrenó con un pipeline de vectorización TF-IDF, sobremuestreo (`RandomOverSampler`) y clasificación con `LinearSVC`. Se aplicó una búsqueda de hiperparámetros exhaustiva (`GridSearchCV`) optimizando la métrica *macro-F1* con validación cruzada estratificada. Luego se aplicó calibración probabilística (`CalibratedClassifierCV`) para ajustar el umbral óptimo.

Evaluación	F1 (Ansiedad)	F1 (Control)	Accuracy
Validación (τ óptimo)	0.96	0.73	0.93
Test (τ óptimo)	0.89	0.43	0.83

Table 3: Desempeño del modelo SVM + TF-IDF calibrado aplicado al trastorno de ansiedad.

Análisis cualitativo: el modelo respondió bien a expresiones comunes, pero tuvo dificultades con casos neutros o ambiguos. Esto sugiere que podría beneficiarse de un mayor número de ejemplos negativos o del uso de embeddings contextuales más robustos.

(Agregar ejemplos de oraciones que muestren si clasifican bien o no. Indicar posibles sesgos)

6.3 FastText

Este enfoque combinó embeddings de FastText con pesos TF-IDF y regresión logística calibrada. También se ajustó el umbral a partir de la curva precision-recall.

Evaluación	F1 (Ansiedad)	F1 (Control)	Accuracy
Validación (umbral óptimo)	0.9706	0.7500	0.9474
Test (umbral óptimo)	0.9104	0.2500	0.8400

Table 4: Desempeño del modelo FastText + TF-IDF aplicado al trastorno de ansiedad.

Con el fin de evaluar el impacto del uso de ponderación TF-IDF sobre los embeddings, se comparó este modelo con una variante basada únicamente en el promedio de vectores FastText por documento (sin

ponderación). Ambos modelos emplearon el mismo clasificador (regresión logística calibrada) y ajuste de umbral para maximizar la macro-F1 sobre el conjunto de validación.

Modelo	F1 (Ansiedad)	F1 (Control)	Accuracy
FastText promedio simple	0.8955	0.1250	0.8133
FastText + TF-IDF	**0.9104**	**0.2500**	**0.8400**

Table 5: Comparación entre variantes de FastText.

Ambos enfoques demostraron ser eficaces para detectar señales de ansiedad, con *f1-scores* elevados en la clase mayoritaria. Sin embargo, la inclusión de ponderación TF-IDF permitió mejorar de forma notable la capacidad del modelo para identificar correctamente a los usuarios sin síntomas (clase Control), duplicando su desempeño en dicha clase.

Esto sugiere que, en escenarios desbalanceados, enriquecer los embeddings con información estadística sobre la relevancia léxica puede favorecer un aprendizaje más equilibrado.

Análisis cualitativo: El modelo capturó patrones ansiosos de forma eficiente, pero falló al generalizar sobre la clase control en test.

(Agregar ejemplos de oraciones que muestren si clasifican bien o no. Indicar posibles sesgos)

7 Resultados y Discusión

7.1 Tratamiento del desbalance de clases

El corpus utilizado presenta un marcado desbalance de clases, con una proporción significativamente mayor de usuarios etiquetados como "sufre ansiedad". Esta distribución desigual representa un desafío importante para los modelos de clasificación, ya que tienden a sesgarse hacia la clase mayoritaria.

Para mitigar este problema, se aplicaron diferentes estrategias en cada uno de los enfoques evaluados:

- **RoBERTa-base-bne**: se incorporaron `pesos de clase` durante el entrenamiento, asignando mayor penalización a los errores en la clase minoritaria. Además, se realizó un `ajuste de umbral` de decisión, evaluando valores entre 0.30 y 0.80 para optimizar la sensibilidad sin comprometer la precisión.
- **SVM + TF-IDF**: se entrenó una variante calibrada del modelo (`CalibratedClassifierCV`) que permite obtener probabilidades y ajustar el umbral de decisión. Esta calibración resultó esencial para mejorar el balance entre clases en validación. También se aplicó `oversampling` con `RandomOverSampler` para aumentar la presencia de la clase minoritaria.
- **FastText + TF-IDF**: se utilizó un umbral ajustado ($\tau = 0.692$) sobre las probabilidades generadas por regresión logística calibrada. El uso de TF-IDF permitió ponderar las palabras más discriminativas en cada texto, enriqueciendo así los embeddings base de FastText.

Estas técnicas de balanceo demostraron ser fundamentales para mejorar la capacidad de los modelos de identificar correctamente a los usuarios que no presentan señales de ansiedad, que de otro modo serían confundidos con la clase mayoritaria.

7.2 Evaluación cuantitativa

La Tabla 6 presenta los resultados obtenidos en el conjunto de prueba por los tres enfoques evaluados. Las métricas consideradas fueron `F1-score` para ambas clases ("sufre" y "no sufre") y el promedio ponderado general.

Modelo	F1 (Sufre)	F1 (No Sufre)	F1 Promedio
SVM + TF-IDF	0.89	0.47	0.75
FastText	0.87	0.51	0.76
RoBERTa-base-bne	**0.94**	**0.79**	**0.88**

Table 6: Comparación de desempeño por modelo en el conjunto de prueba.

Como puede observarse, el modelo basado en `RoBERTa-base-bne` obtuvo el mejor rendimiento global, destacándose especialmente por su capacidad para detectar correctamente la clase minoritaria. Este desempeño se vio favorecido por la combinación del uso de pesos de clase y el ajuste del umbral de decisión. No obstante, su implementación conlleva una mayor demanda de recursos computacionales y tiempo de entrenamiento, lo que plantea desafíos para su aplicación en entornos con recursos limitados y deja espacio para futuras optimizaciones.

7.3 Análisis cualitativo

Además de las métricas cuantitativas, se realizó una revisión manual de predicciones para identificar patrones lingüísticos que favorecen o dificultan la clasificación:

- **Casos correctamente clasificados como "sufre"**: suelen incluir expresiones directas de malestar emocional ("no puedo más", "me siento inútil", "no tengo ganas de nada").
- **Errores frecuentes en la clase "no sufre"**: ocurren en textos neutros, sarcásticos o con ambigüedad semántica, donde el modelo interpreta negatividad sin contexto suficiente.
- **Fortalezas de RoBERTa**: su entrenamiento contextual le permite captar matices y expresiones indirectas de angustia emocional ("me desconecto del mundo", "siento que no estoy").
- **SVM y FastText**: tienden a basarse más en la presencia literal de palabras clave, mostrando limitaciones en el tratamiento de ironía o lenguaje figurado.

MEJORAR UNA VEZ SE AGREGUEN LAS CUALITATIVAS DE CADA UNO

7.4 Discusión

Los resultados respaldan la hipótesis de que los modelos basados en lenguaje preentrenado, como RoBERTa, ofrecen ventajas sustanciales en tareas sensibles como la detección de riesgo emocional. Su capacidad para comprender contexto y ambigüedad resulta clave para reducir falsos negativos en la clase "no sufre", lo que podría tener implicancias críticas en aplicaciones clínicas o preventivas.

En contraste, modelos más ligeros como SVM y FastText, si bien muestran buen desempeño en la clase mayoritaria, tienden a perder sensibilidad ante señales más sutiles o atípicas.

El uso de estrategias de balanceo (como ponderación de clases y ajuste de umbral) fue indispensable para maximizar el rendimiento, especialmente en conjuntos desbalanceados como el presente.

Finalmente, la metodología propuesta es extensible a otros trastornos del corpus (depresión, TCA), lo que abre un camino prometedor para futuros estudios en detección automática de salud mental a partir del lenguaje.

8 Conclusiones y Trabajo Futuro

Este trabajo exploró distintas estrategias de clasificación para detectar riesgo emocional asociado al trastorno de ansiedad, utilizando textos en español del corpus *MentalRiskES 2023*. A partir de la concatenación de mensajes por usuario, se abordó la tarea como un problema de clasificación binaria, evaluando tres enfoques representativos: un modelo clásico basado en `SVM + TF-IDF`, un modelo intermedio con `FastText`, y un enfoque profundo basado en `RoBERTa-base-bne`, ajustado mediante *fine-tuning*.

Los resultados obtenidos muestran que el modelo basado en RoBERTa fue el más efectivo en términos generales, especialmente al clasificar correctamente a usuarios sin síntomas aparentes (clase minoritaria). Este buen desempeño se debió en parte al uso de técnicas de ajuste de umbral y balanceo de clases, las cuales resultaron esenciales en un contexto de fuerte desbalance. No obstante, esta superioridad vino acompañada de un mayor costo computacional. El entrenamiento de modelos de lenguaje profundo impone demandas significativas en términos de memoria y tiempo de ejecución, lo que limitó las pruebas posibles, especialmente al trabajar en un entorno gratuito como Google Colab.

Los modelos más ligeros, como `FastText` y `SVM + TF-IDF`, ofrecieron buenos resultados con menor carga computacional, aunque mostraron dificultades para detectar correctamente casos negativos. Aun así, representan alternativas viables en contextos de recursos limitados.

Limitaciones

Entre las principales limitaciones se destacan:

- El estudio se restringió exclusivamente al análisis del trastorno de ansiedad, debido a limitaciones de tiempo, cómputo y enfoque. No obstante, se considera que la metodología puede ser replicable para otros trastornos del corpus (depresión y TCA).
- El corpus presenta un desbalance importante entre clases, especialmente en el caso del trastorno de ansiedad, lo que condiciona la capacidad de generalización del modelo y sugiere que más ejemplos de clase minoritaria podrían mejorar los resultados.
- El uso de Google Colab impuso límites técnicos, como el tamaño del `batch_size` y la duración de las sesiones. Estas restricciones limitaron la capacidad de realizar experimentos más exhaustivos o con múltiples semillas aleatorias.

Trabajo futuro

A partir de los aprendizajes y resultados obtenidos, se proponen las siguientes líneas de trabajo futuro:

- Extender el análisis a los demás trastornos presentes en el corpus (depresión y TCA), replicando la metodología desarrollada y analizando posibles ajustes específicos.
- Integrar herramientas de seguimiento de experimentos como `Weights & Biases (WandB)` para facilitar la trazabilidad, comparación de modelos y reproducibilidad de los resultados.
- Explorar nuevas arquitecturas de lenguaje, como `BETO` o modelos multilingües adaptados al dominio clínico, que podrían mejorar el rendimiento sin requerir tanto ajuste manual.
- Profundizar en el análisis cualitativo y de interpretabilidad mediante técnicas como `LIME`, `SHAP` o visualización de activaciones, con el fin de comprender mejor cómo los modelos toman sus decisiones.
- Evaluar el uso de técnicas de generación de datos sintéticos o aumento de datos para mitigar los efectos del desbalance de clases.

En conclusión, el enfoque propuesto demuestra ser una base sólida y extensible para futuros sistemas de monitoreo emocional automático en español, con potencial impacto en la detección temprana y el acompañamiento digital en salud mental.

References

[1] L. V. Ramos, C. M. García, J. M. Vázquez, and V. P. Álvarez, "I2C-UHU at MentalRiskES 2023: Detecting and Identifying Mental Disorder Risks in Social Media using Transformer-Based Models," in *CEUR Workshop Proceedings, IberLEF 2023*, 2023. [Online]. Available: http://ceur-ws.org/Vol-3496/mentalriskes-paper11.pdf

[2] A. M. Mármol Romero, A. Moreno Muñoz, F. M. Plaza-del Arco, M. D. Molina González, M. T. Martín Valdivia, L. A. Ureña-López, and A. Montejo Ráez, "MentalRiskES: A new corpus for early detection of mental disorders in Spanish," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 11 204–11 214. [Online]. Available: https://aclanthology.org/2024.lrec-main.978

[3] O. de MentalRiskES 2023, "Mentalriskes 2023 - iberlef shared task," https://sites.google.com/view/mentalriskes/home, 2023, accedido en mayo de 2025.