MentalBERT on Reddit: Temporal Classification of Linguistic Signals of Depression and Anxiety

Miguel Molina-Calderón,MSc[1];Darwin Patiño-Pérez,Ph.D[2]; Alex Collantes-Farah,MSc[1]; Maylee Ordoñez-Valencia,MSc[1]; Ángel Ochoa-Flores,MSc[1]; José Córdova-Aragundi,MSc[3]; Zackery Patino-Munive,Std[4]

[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, miguel.molinac@ug.edu.ec, alex.collantesf@ug.edu.ec,maylee.ordonezv@ug.edu.ec,angel.ochoaf@ug.edu.ec,

[2]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Grupo de Investigación de Inteligencia Artificial, Ecuador, darwin.patinop@ug.edu.ec,

,[3]Universidad de Guayaquil, Facultad de Ciencias Administrativas, Ecuador, jose.cordovaa@ug.edu.ec,

[4]Kutztown University, Department of Psychology, PA, United State, zpati840@live.kutztown.edu

Abstract– Social media posts provide a wealth of linguistic material that can be analyzed using natural language processing. However, it is important to clarify that such analysis does not replace a clinical assessment. This study compared three text classification strategies applied to English-language Reddit posts: TF-IDF with logistic regression, BiLSTM with GloVe, and MentalBERT (a transformer model fine-tuned on mental health texts). The corpus—filtered for the 2020–2021 period and organized into four categories via weak labeling (depression, anxiety, a control group, and a comorbidity proxy class defined by author overlap)—comprised 674,612 posts. Data partitioning was performed by author, and balancing was restricted to the training set. MentalBERT achieved the best overall performance, with a macro-F1 score of 0.5716 and a macro-recall of 0.5718, outperforming BiLSTM (0.5495) and the classical model (0.5434). However, performance was not uniform: the comorbidity category showed a recall of only 0.2277, likely due to vocabulary overlap and noise inherent in weak labeling. Predictions were inspected using LIME. Subsequently, the same MentalBERT instance—without retraining—was evaluated on an independent corpus of 357,030 posts from 2025. The macro-F1 score remained at 0.5702, a value close to the original, although recall varied across categories.

Keywords- Anxiety, Depression, Explainable Artificial Intelligence, MentalBERT, Natural Language Processing.

MentalBERT en Reddit: Clasificación Temporal de Señales Lingüísticas de Depresión y Ansiedad

Miguel Molina-Calderón,MSc[1];Darwin Patiño-Pérez,Ph.D[2]; Alex Collantes-Farah,MSc[1];
Maylee Ordoñez-Valencia,MSc[1]; Ángel Ochoa-Flores,MSc[1]; José Córdova-Aragundi,MSc[3];
Zackery Patino-Munive,Std[4]
[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, miguel.molinac@ug.edu.ec,
alex.collantesf@ug.edu.ec,maylee.ordonezv@ug.edu.ec,angel.ochoaf@ug.edu.ec,
[2]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Grupo de Investigación de Inteligencia Artificial,
Ecuador, darwin.patinop@ug.edu.ec,
,[3]Universidad de Guayaquil, Facultad de Ciencias Administrativas, Ecuador, jose.cordovaa@ug.edu.ec,
[4]Kutztown University, Department of Psychology, PA, United State, zpati840@live.kutztown.edu

Resumen– ***Las publicaciones en redes sociales ofrecen un material lingüístico abundante que el procesamiento de lenguaje natural puede examinar. Con todo, conviene aclarar que dicho análisis no reemplaza una evaluación clínica. Este trabajo comparó tres estrategias de clasificación textual sobre publicaciones en inglés de Reddit: TF-IDF con regresión logística, BiLSTM con GloVe, y MentalBERT, un transformador ajustado en textos de salud mental. El corpus, filtrado para 2020–2021 y organizado en cuatro categorías mediante etiquetado débil —depresión, ansiedad, control relativo y una clase proxy de comorbilidad definida por participación cruzada de autores—, alcanzó 674.612 publicaciones. La partición se efectuó por autor, y el balanceo se restringió al entrenamiento. MentalBERT obtuvo el mejor desempeño global, con un F1 macro de 0,5716 y un recall macro de 0,5718, superando los 0,5495 de BiLSTM y los 0,5434 del modelo clásico. No obstante, el rendimiento distó de ser uniforme: la comorbilidad registró un recall de apenas 0,2277, probablemente por solapamiento del vocabulario y ruido del etiquetado débil. Las predicciones se inspeccionaron con LIME. Posteriormente, la misma instancia de MentalBERT —sin reentrenamiento— se evaluó sobre un corpus independiente de 357.030 publicaciones de 2025. El F1 macro se mantuvo en 0,5702, un valor cercano al original, aunque el recall varió entre categorías.***

Palabras clave- ***Ansiedad, Depresión, Inteligencia Artificial Explicable, MentalBERT, Procesamiento del Lenguaje Natural.***

I. Introducción

La depresión es uno de los trastornos de salud mental con mayor prevalencia a nivel mundial, y su impacto en la calidad de vida de las personas está ampliamente documentado en la literatura epidemiológica [1]. Durante los últimos años, el crecimiento explosivo de las redes sociales ha generado un volumen sin precedentes de datos textuales en los que los usuarios comparten experiencias, emociones y preocupaciones de manera más o menos espontánea. Este material, aunque heterogéneo y no estructurado, representa una fuente potencialmente valiosa para identificar indicios lingüísticos que, en condiciones controladas, podrían asociarse con estados de ánimo compatibles con la depresión o la ansiedad.

El procesamiento de lenguaje natural ofrece un conjunto de herramientas que permiten examinar sistemáticamente estos contenidos y extraer patrones que escapan al análisis manual [2]. Plataformas como Reddit resultan particularmente adecuadas para este tipo de estudios por su organización en comunidades temáticas —los denominados subreddits— y por la extensión media de las publicaciones, que suele ser superior a la de otras redes sociales. Sin embargo, es importante señalar que la mera pertenencia a una comunidad dedicada a la depresión o a la ansiedad no equivale a un diagnóstico clínico. Esta distinción constituye una premisa metodológica que atraviesa todo el trabajo.

Desde un punto de vista computacional, la clasificación de textos puede abordarse con modelos de muy distinta complejidad. Las representaciones TF-IDF, basadas en la frecuencia de los términos, ofrecen una línea de referencia eficiente y fácilmente interpretable. Por su parte, embeddings como GloVe incorporan relaciones semánticas entre palabras a partir de matrices de co-ocurrencia [3]. Cuando estos vectores se combinan con redes recurrentes, como las BiLSTM, es posible modelar el orden secuencial y las dependencias de largo alcance presentes en el lenguaje natural [4]. Estos enfoques, aunque menos sofisticados que los basados en atención, siguen siendo competitivos en muchas tareas de clasificación.

La irrupción de los transformadores supuso un cambio de paradigma en el PLN. El mecanismo de atención permite construir representaciones contextualizadas en las que el significado de una palabra depende de las demás palabras que la acompañan [5]. BERT consolidó este enfoque mediante un preentrenamiento bidireccional masivo [6], y posteriormente surgieron variantes especializadas para dominios específicos. MentalBERT es una de ellas: fue ajustado sobre textos provenientes de foros de salud mental, lo que podría proporcionarle una ventaja a la hora de reconocer el vocabulario y las expresiones propias de este ámbito [7] . No obstante, esta ventaja no puede darse por sentada sin una comparación controlada con un modelo generalista bajo las mismas condiciones experimentales.

A pesar de los avances en precisión, los modelos profundos plantean un problema bien conocido: pueden clasificar correctamente sin que el usuario entienda qué elementos del texto han influido en la decisión. Esta opacidad resulta particularmente problemática en ámbitos sensibles como la salud mental, donde la confianza en el sistema depende en parte de la capacidad para justificar sus predicciones. LIME aborda

esta cuestión aproximando localmente la contribución de cada término mediante perturbaciones controladas de la entrada [8]. Esta técnica permite inspeccionar ejemplos concretos, aunque no convierte al modelo en intrínsecamente interpretable ni sustituye una validación clínica con criterios externos [9].

Un segundo desafío, a menudo menos discutido, es el temporal. Cuando un modelo se entrena y evalúa con datos extraídos de un mismo periodo, su comportamiento puede cambiar de forma significativa al aplicarlo a textos de un momento posterior. El vocabulario evoluciona, los temas de conversación varían y la composición demográfica de los usuarios puede modificarse con el tiempo. Esta deriva conceptual —*concept drift*— es especialmente relevante en entornos como Reddit, donde las tendencias discursivas cambian con rapidez. Por ello, una evaluación únicamente estática ofrece una imagen incompleta del rendimiento real del clasificador.

El presente estudio se diseñó precisamente para abordar de forma simultánea estos dos desafíos: la interpretabilidad local y la validación temporal. Para ello, se optó por una tarea multiclase con cuatro categorías construidas mediante etiquetado débil, una división por autor que evita la contaminación entre particiones, una comparación controlada de tres familias de modelos y un análisis de explicaciones con LIME. Además, se incorporó una evaluación sobre un corpus de 2025, sin reentrenamiento, para observar el comportamiento del modelo en un periodo posterior. Esta estructura permite examinar no solo el rendimiento agregado, sino también su estabilidad y sus posibles sesgos.

En concreto, los objetivos fueron los siguientes: (i) comparar el rendimiento de regresión logística con TF-IDF, BiLSTM con GloVe y MentalBERT sobre un corpus multiclase con etiquetado débil; (ii) seleccionar el modelo con mejor F1 macro y analizar sus aciertos y errores por categoría; (iii) emplear LIME para examinar la influencia de términos específicos en las predicciones; (iv) evaluar el modelo seleccionado sobre datos de 2025 para explorar su estabilidad temporal; y (v) integrar estos componentes en un prototipo bilingüe con una capa de traducción automática, orientado a fines académicos y de investigación.

En las siguientes secciones se detalla la metodología empleada, se presentan los resultados obtenidos y se discuten sus implicaciones, limitaciones y posibles líneas de desarrollo futuro. Conviene insistir, desde ahora, en que todo el sistema aquí descrito se plantea como un recurso experimental para la investigación en clasificación de textos de salud mental, y no como una herramienta clínica ni como un sustituto del juicio profesional.

II. Materiales y métodos

La construcción de conjuntos de datos etiquetados en el ámbito de la salud mental a partir de redes sociales presenta dificultades inherentes. A diferencia de los corpus clínicos, donde los diagnósticos son establecidos por profesionales mediante instrumentos validados, las publicaciones en plataformas como Reddit carecen de una anotación de referencia. Una estrategia ampliamente adoptada consiste en utilizar la pertenencia a comunidades temáticas como fuente de etiquetas débiles [1]. Este enfoque, conocido como *weak supervision,* asume que los usuarios que publican en *subreddits* dedicados a la depresión o la ansiedad expresan, con cierta probabilidad, contenidos relacionados con esas condiciones. Sin embargo, esta suposición introduce ruido, ya que no todas las publicaciones reflejan el estado emocional asociado a la comunidad, y un mismo usuario puede participar en múltiples foros con temáticas solapadas [2]. Para mitigar el riesgo de contaminación entre particiones, la literatura recomienda realizar la división a nivel de autor en lugar de a nivel de publicación [3]. Este procedimiento impide que un mismo usuario contribuya simultáneamente a los conjuntos de entrenamiento y prueba, lo que reduce la posibilidad de que el modelo aprenda patrones idiosincráticos de un autor en lugar de características generalizables del lenguaje.

La elección de los clasificadores debe responder a un equilibrio entre complejidad, interpretabilidad y rendimiento. Los modelos basados en TF-IDF con regresión logística constituyen una línea de referencia clásica: su eficiencia computacional y su facilidad de interpretación los convierten en un punto de partida adecuado para cualquier tarea de clasificación textual [4]. No obstante, estas representaciones son dispersas y no capturan relaciones semánticas entre palabras. Los *embeddings* pre-entrenados como GloVe, basados en matrices de co-ocurrencia global, ofrecen una representación densa que incorpora información distribucional [5]. Cuando estos vectores se alimentan a una red BiLSTM, el modelo puede modelar dependencias secuenciales de largo alcance, lo que resulta especialmente útil en textos narrativos o expresivos [6]. Por su parte, los transformadores como BERT introducen un mecanismo de atención que permite representaciones contextualizadas, donde el significado de cada término depende de los demás términos que lo rodean [7]. MentalBERT, una variante ajustada sobre textos del dominio de la salud mental podría ofrecer ventajas adicionales al estar expuesto durante el preentrenamiento a un vocabulario y a unas estructuras sintácticas propias de foros de apoyo emocional [8]. Sin embargo, esta ventaja no es automática, y su efecto solo puede evaluarse mediante una comparación directa con modelos generalistas bajo las mismas condiciones experimentales.

Dos consideraciones metodológicas adicionales guiaron el diseño experimental: la interpretabilidad local y la evaluación temporal. La primera responde a la necesidad de comprender qué elementos del texto impulsan las predicciones del modelo, especialmente en dominios sensibles donde la opacidad puede erosionar la confianza en el sistema. LIME aborda este problema mediante la generación de explicaciones locales basadas en perturbaciones controladas de la entrada, lo que permite identificar los términos con mayor peso en una decisión concreta [9]. Esta técnica ha sido aplicada en estudios previos de clasificación de textos de salud mental, aunque su estabilidad y su fidelidad al modelo original siguen siendo objeto de debate

[10]. La segunda consideración se refiere a la deriva temporal (*concept drift*), un fenómeno bien documentado en sistemas de procesamiento de lenguaje natural desplegados en entornos dinámicos [11]. El vocabulario, los temas de conversación y la composición de los usuarios pueden cambiar con el tiempo, lo que puede degradar el rendimiento de un modelo entrenado con datos históricos. La evaluación sobre un corpus posterior, sin reentrenamiento, permite estimar el grado de estabilidad del clasificador frente a estos cambios [12]. Aunque esta validación no sustituye un monitoreo continuo, proporciona una evidencia inicial sobre la robustez temporal del modelo en condiciones controladas.

A. Diseño y flujo metodológico

Se aplicó un diseño cuantitativo de evaluación experimental comparativa. Los tres enfoques se entrenaron y evaluaron sobre las mismas particiones, conservando criterios comunes de medición. Como se muestra en la Fig. 1, el procedimiento se organizó en cinco etapas: preprocesamiento, balanceo, entrenamiento, predicción con interpretabilidad y desarrollo del prototipo.

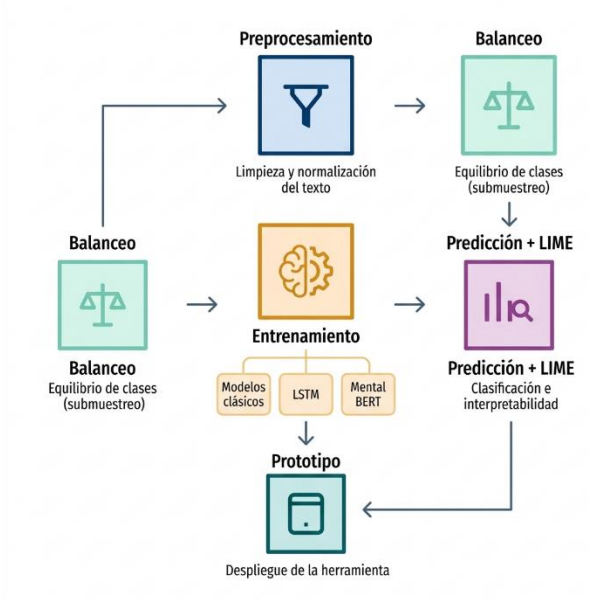


Fig. 1. Esquema del proceso metodológico de la investigación.

B. Conjunto de datos y construcción de clases

Se utilizó el *Reddit Mental Health Dataset* (RMHD), un recurso público con publicaciones en inglés recopiladas entre 2019 y 2022 [10]. Para el experimento se seleccionaron registros de 2020–2021 provenientes de r/*depression*, r/*anxiety* y r/*mentalhealth*. Después de eliminar entradas sin contenido textual útil, el corpus quedó conformado por 674.612 publicaciones. Las etiquetas se asignaron mediante criterios operacionales: las clases depresión y ansiedad se definieron según la comunidad de origen; la clase proxy de comorbilidad reunió publicaciones de autores con participación en ambas comunidades; y r/*mentalhealth* se utilizó como control relativo.

La Tabla I resume la composición y las particiones del corpus. Estas categorías corresponden a etiquetas débiles y no representan diagnósticos clínicos confirmados.

TABLA I
TABLA ICOMPOSICIÓN Y PARTICIONES DEL CORPUS

Elemento	Publicaciones	Observación
Muestra final	674.612	Publicaciones útiles de 2020–2021
Entrenamiento original	471.736	69,9 %; división por autor
Validación	100.821	14,9 %; distribución natural
Prueba	102.055	15,1 %; distribución natural
Entrenamiento balanceado	384.686	Submuestreo aplicado solo al entrenamiento

C. Preprocesamiento, partición y balanceo

El texto de cada registro se obtuvo al unir el título con el cuerpo de la publicación. El preprocesamiento se ajustó a las necesidades de cada modelo. Para TF-IDF y BiLSTM se aplicó conversión a minúsculas, eliminación de enlaces, menciones, caracteres no alfabéticos y registros vacíos. TF-IDF representó los documentos mediante n-gramas, mientras que BiLSTM utilizó secuencias numéricas y embeddings GloVe. En MentalBERT se mantuvo una normalización mínima para conservar información contextual; el tokenizador del *checkpoint* *mental/mental-bert-base-uncased* procesó secuencias de hasta 128 *tokens* con padding dinámico. La división se realizó por autor para impedir que un mismo usuario apareciera en más de una partición. Posteriormente, se limitó el entrenamiento a 120.000 publicaciones por clase y se calcularon pesos de clase, mientras que validación y prueba conservaron su distribución natural.

D. Modelos e hiperparámetros documentados

Se implementaron tres familias de clasificadores. El modelo clásico combinó TF-IDF con regresión logística; su mejor configuración fue LR-2, con C=0,5, solucionador SAGA, ponderación balanceada y semilla 42. La red recurrente utilizó una arquitectura BiLSTM con *embeddings* GloVe de 300 dimensiones; LSTM-2, entrenada con una tasa de aprendizaje de 5×10^{-4}, obtuvo el mejor resultado dentro de esta familia. Finalmente, MentalBERT se ajustó desde el *checkpoint* mental/mental-*bert*-base-*uncased*, y la configuración MB-1 con tasa de aprendizaje de 2×10^{-5} fue seleccionada. La Tabla II presenta los hiperparámetros verificados en los cuadernos de entrenamiento. Los experimentos neuronales se ejecutaron en entornos independientes. BiLSTM se entrenó con TensorFlow 2.10, CUDA 11.2 y cuDNN 8.1, mientras que MentalBERT se ajustó con PyTorch 2.2.2 y CUDA 12.1. En ambos casos se utilizó una GPU NVIDIA GeForce RTX 4060 de 8 GB. Solo se informan las versiones comprobadas directamente en los entornos de ejecución, con el propósito de mantener la trazabilidad de la configuración experimental.

TABLA II
HIPERPARÁMETROS

Parámetro	Modelo Clásico (LR-2)	BiLSTM (LSTM-2)	MentalBERT (MB-1)
Configuración seleccionada	LR-2	LSTM-2	MB-1
Representación / Checkpoint	TF-IDF	GloVe 6B, 300d	mental/mental-bert-base-uncased
Vocabulario / Longitud	50,000 características; n-gramas 1-3; min_df=5; max_df=0.95; sublinear_tf=True	30,000 palabras; secuencias de 150 tokens	128 tokens, padding dinámico
Arquitectura	Regresión Logística, 4 clases	Embedding(300) → SpatialDropout(0.3) → BiLSTM(128) → GlobalMaxPooling → Dense(128, ReLU) → Dropout(0.5) → Softmax(4)	MentalBERT + clasificador de 4 clases
Optimizador / Solver	SAGA	Adam	AdamW gestionado por Trainer
Tasa de aprendizaje	No aplica; regularización C=0.5	5×10^{-4}	2×10^{-5}
Configuraciones evaluadas	C = {1.0; 0.5; 5.0}	lr = $\{1 \times 10^{-3}; 5 \times 10^{-4}; 2 \times 10^{-3}\}$; además, ablación sin GloVe	lr = $\{2 \times 10^{-5}; 3 \times 10^{-5}; 5 \times 10^{-5}\}$
Tamaño de lote	No aplica	256	16 entrenamiento; 32 evaluación
Épocas	max_iter=1000	Máximo 10; mejor época=5	Máximo 6
Parada temprana	No aplica	Paciencia=3; monitor val_loss; restaura mejores pesos	Paciencia=2; monitor F1 macro; carga el mejor modelo
Regularización	class_weight='balanced'	SpatialDropout=0.3; dropout BiLSTM=0.4; dropout denso=0.5	weight_decay=0.01; fp16=True
Función de pérdida	Pérdida logística interna de scikit-learn	Entropía cruzada categórica dispersa	Entropía cruzada ponderada por clase
Pesos por clase	Sí; class_weight='balanced'	Sí, calculados con compute_class_weight	Sí; incorporados mediante WeightedTrainer
Semilla	42	42	42
Hardware de entrenamiento	CPU	NVIDIA GeForce RTX 4060 (8 GB)	NVIDIA GeForce RTX 4060 (8 GB)

E. Evaluación e interpretabilidad

La evaluación se concentró en F1 macro y recall macro, debido al desbalance y a la necesidad de considerar por igual las cuatro categorías. También se calcularon precisión, recall, F1 y soporte por clase a partir de la matriz de confusión. MentalBERT fue seleccionado para la etapa de interpretabilidad por alcanzar el mayor F1 macro. En el análisis individual, LIME utilizó 10 características y 1.000 perturbaciones; el análisis agregado empleó 10 características y 500 perturbaciones, por ejemplo. En el prototipo se configuraron 800 perturbaciones y una selección de 5 a 15 términos. El explicador operó sobre la versión inglesa del texto, conservó las contracciones como unidades completas y utilizó random_state=42. Los pesos positivos representaron términos que impulsan la clase explicada, mientras que los negativos indicaron términos que reducen su probabilidad [8].

F. Evaluación temporal y prototipo bilingüe

La evaluación temporal se construyó con Arctic Shift [11]. Se descargaron, en archivos JSON separados, las publicaciones de r/depression, r/anxiety y r/mentalhealth comprendidas entre el 1 de enero y el 31 de diciembre de 2025; los comentarios fueron excluidos. La integración inicial produjo 362.782 registros y, después de aplicar la misma limpieza y el mismo esquema de etiquetado débil del corpus original, se conservaron 357.030 publicaciones. La instancia de MentalBERT entrenada con datos de 2020–2021 se aplicó directamente, sin reentrenamiento. Debido a que Arctic Shift es un archivo externo, la muestra puede contener omisiones. Tampoco se efectuó una comparación exhaustiva de autores entre periodos, por lo que no puede descartarse un solapamiento residual.

El prototipo bilingüe se concibió como una interfaz de exploración académica. El usuario ingresa un texto en español o inglés; Langdetect identifica el idioma y, si es español, DeepL genera una versión en inglés para la inferencia. MentalBERT y LIME operan siempre sobre el texto inglés. La interfaz muestra la clase predicha, la probabilidad asociada y los términos que más contribuyeron a la decisión. Entre sus limitaciones destacan la dependencia de la traducción automática y la ausencia de validación clínica; por ello, se incluye una advertencia de uso exclusivamente investigativo.

G. Consideraciones éticas

El estudio utilizó corpus públicos y no estableció contacto con los autores de las publicaciones. Los identificadores de usuario se emplearon únicamente para construir las particiones y la clase proxy de comorbilidad; no se muestran en los resultados ni en el prototipo. Debido al carácter sensible del contenido, se evitó reproducir textos completos con información identificable. Las etiquetas representan proxies derivados de comunidades y no diagnósticos clínicos. El prototipo incluye una advertencia de uso académico y de investigación. Además, debido a que DeepL procesa temporalmente el texto enviado, se recomienda no introducir nombres, direcciones, teléfonos u otros datos personales.

III. RESULTADOS

A. Diseño y flujo metodológico

La Tabla III presenta las métricas macro obtenidas por los tres modelos en el conjunto de prueba. A diferencia de la exactitud global, el F1 macro y el *recall* macro asignan el mismo peso a cada categoría y permiten observar el rendimiento bajo una distribución desbalanceada.

TABLA III
DESEMPEÑO DE LOS MODELOS SOBRE EL CONJUNTO DE PRUEBA

Modelo	F1 macro	Recall macro
Baseline clásico (TF-IDF)	0,5434	0,5492
BiLSTM (Glove)	0,5495	0,5660
MentalBERT	0,5716	0,5718

La Fig. 2 complementa la comparación tabular y muestra la proximidad entre los resultados de los tres enfoques. MentalBERT alcanzó un F1 macro de 0,5716, frente a 0,5495 de BiLSTM y 0,5434 del modelo clásico. La mejora fue de 0,0221 y 0,0282, respectivamente.

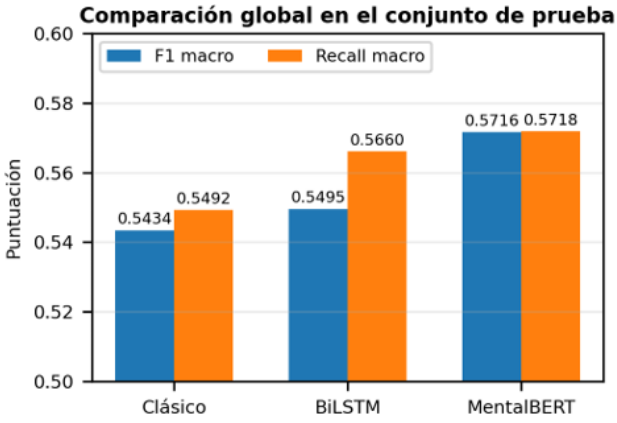


Fig. 2. Comparación de F1 macro y recall macro de los tres modelos.

Las diferencias en F1 macro y recall macro fueron moderadas, aunque consistentes a favor de MentalBERT.

Debido a que los experimentos no se repitieron con múltiples semillas ni se aplicaron pruebas de significancia, esta comparación se interpreta de manera descriptiva y no como evidencia de superioridad estadística.

El resultado global no fue uniforme entre las categorías. Como se presenta en la Tabla IV, depresión obtuvo precisión de 0,726, recall de 0,752 y F1 de 0,739; ansiedad alcanzó 0,698, 0,720 y 0,709, respectivamente. El control relativo mostró un comportamiento intermedio, con un F1 de 0,604. La clase proxy de comorbilidad registró los valores más bajos: precisión de 0,242, recall de 0,228 y F1 de 0,235. En consecuencia, el F1 macro debe analizarse junto con las métricas por clase, ya que el valor agregado no representa un desempeño homogéneo.

TABLA IV
DESEMPEÑO POR CLASE DE MENTALBERT

Clase	Precisión	Recall	F1	Soporte
Depresión	0,726	0,752	0,739	44.864
Ansiedad	0,698	0,720	0,709	18.220
Control relativo	0,621	0,588	0,604	25.744
Comorbilidad proxy	0,242	0,228	0,235	13.227

La clase proxy de comorbilidad concentró la mayor proporción de errores. En el corpus de 2025, solo el 20,9 % de las publicaciones etiquetadas como comorbilidad fueron clasificadas correctamente; el 34,9 % se confundió con depresión, el 28,8 % con ansiedad y el 15,4 % con control relativo. Este patrón sugiere que el modelo percibe la comorbilidad como una mezcla difusa de rasgos depresivos y ansiosos, sin una representación diferenciada. Un ejemplo típico son los textos que alternan desesperanza ("no le veo sentido a nada") con activación ansiosa ("me tiembla el cuerpo, no puedo respirar"). El esquema de etiqueta única fuerza una elección artificial; un enfoque multi-etiqueta permitiría probabilidades independientes y umbrales ajustables.

B. Explicaciones locales con LIME

Las explicaciones de LIME se complementaron con ejemplos concretos. En ansiedad, términos como *panic, worried y nervous* mostraron pesos positivos consistentes; en depresión, *hopeless, empty y tired* aparecieron con frecuencia; en comorbilidad, se observó una mezcla de ambos repertorios. Para estimar la estabilidad, se repitió el análisis cinco veces por instancia y se calculó el índice de Jaccard entre los conjuntos de términos principales. La coincidencia promedio fue moderada, lo que refuerza la cautela sobre el uso de LIME como evidencia única. Sería recomendable comparar estos resultados con SHAP o *Integrated Gradients* en futuras versiones.

C. Evaluación temporal sobre 2025

La evaluación temporal comparó el desempeño del mismo modelo sobre el conjunto de prueba original y el corpus independiente de 2025, sin actualizar sus parámetros. La Tabla V presenta la comparación temporal entre el conjunto de prueba de 2020–2021 y el corpus de 2025. El F1 macro pasó de 0,5716 a 0,5702, lo que corresponde a una variación de −0,0014. El recall aumentó para ansiedad (+0,0524) y depresión (+0,0172), pero disminuyó para control relativo (−0,0322) y comorbilidad proxy (−0,0188).

TABLA V
COMPARACIÓN TEMPORAL DEL DESEMPEÑO

Métrica	2020-2021	2025	Δ
F1 macro	0,5716	0,5702	−0,0014
Recall macro	0,5718	0,5765	+0,0047
Recall ansiedad	0,7199	0,7723	+0,0524
Recall depresión	0,7520	0,7692	+0,0172
Recall control relativo	0,5877	0,5555	−0,0322
Recall comorbilidad proxy	0,2277	0,2089	−0,0188

Para caracterizar con mayor precisión el cambio temporal, se calcularon medidas de distancia entre los corpus de 2020–2021 y 2025. La divergencia KL sobre la distribución de términos y la similitud coseno entre embeddings promedio mostraron variaciones moderadas, concentradas en vocabulario emergente y expresiones asociadas al contexto pospandemia. Estos indicios no invalidan la estabilidad observada en el F1 macro, pero ayudan a explicar por qué el recall de control relativo y comorbilidad disminuyó, mientras que el de ansiedad y depresión aumentó.

La Fig. 3 muestra la matriz de confusión obtenida sobre las publicaciones de 2025. Aunque el valor agregado permaneció similar, la variación por clase indica que el comportamiento del modelo no fue idéntico entre periodos. Debido a que no se aplicó una prueba de equivalencia ni se midieron directamente cambios de vocabulario, representación o composición de autores, el resultado evidencia estabilidad descriptiva dentro de la muestra analizada, pero no demuestra ausencia general de deriva.

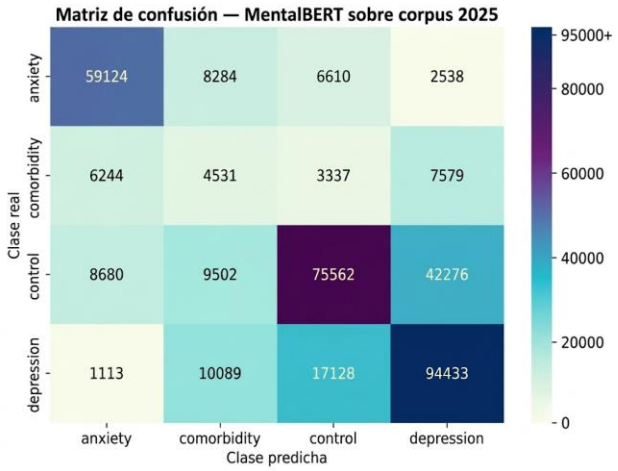


Fig. 3. Matriz de Confusión

IV DISCUSION

MentalBERT presentó el mejor desempeño global, aunque la mejora respecto de BiLSTM y del modelo clásico fue moderada. Este resultado coincide con revisiones que destacan el potencial de los transformadores en tareas de clasificación textual [2], [12]. Sin embargo, los estudios de predicción de depresión en redes sociales utilizan corpus, etiquetas, particiones y métricas muy diferentes [13]. En consecuencia, los valores obtenidos no pueden compararse de forma directa con trabajos binarios o con divisiones aleatorias por publicación.

La principal diferencia metodológica del estudio fue la partición por autor, que redujo el riesgo de que el modelo aprendiera patrones específicos de usuarios presentes simultáneamente en entrenamiento y prueba. Esta decisión fortalece el análisis experimental, aunque no permite afirmar superioridad frente a investigaciones desarrolladas con corpus o protocolos distintos.

El bajo rendimiento en la clase proxy de comorbilidad merece una mirada más detallada. En la matriz de confusión obtenida sobre el corpus de 2025 (Fig. 3), de las publicaciones realmente etiquetadas como comorbilidad, solo un 20,9 % fueron clasificadas correctamente; el 34,9 % se confundió con depresión, el 28,8 % con ansiedad y el 15,4 % con control relativo. Este patrón sugiere que el modelo percibe las señales de comorbilidad como una mezcla difusa de rasgos depresivos y ansiosos, sin lograr una representación diferenciada. A modo de ilustración, publicaciones que alternan expresiones de desesperanza ("no le veo sentido a nada") con síntomas de activación ansiosa ("me tiembla el cuerpo, no puedo respirar") son candidatas naturales para que sean absorbidas por una u otra clase principal.

En este contexto, el esquema multiclase de etiqueta única impone una restricción artificial: obliga al modelo a elegir una sola categoría cuando el texto puede reflejar simultáneamente ambas condiciones. Además, la etiqueta proxy se definió a partir de la participación cruzada del autor, lo que no garantiza que cada publicación contenga señales simultáneas de depresión y ansiedad, y el desbalance y el ruido del etiquetado débil agravan el problema. Por ello, trabajos posteriores deberían explorar esquemas multi-etiqueta capaces de asignar probabilidades independientes a depresión y ansiedad, con umbrales ajustables según el costo de los falsos negativos, así como modelos jerárquicos que primero detecten la presencia de cualquier señal y luego desambigüen la combinación específica, y corpus anotados con criterios independientes. Además, para aislar el efecto del preentrenamiento especializado, esas comparaciones deberían incluir BERT-base bajo las mismas condiciones experimentales, pues sin ese control la ventaja de MentalBERT no puede atribuirse únicamente al dominio de salud mental.

LIME permitió examinar la contribución local de los términos, pero sus explicaciones dependen de las perturbaciones generadas y pueden variar entre ejecuciones. Los vocablos destacados pueden representar señales semánticas, nombres directos de las categorías o expresiones frecuentes de los subreddits. Una evaluación posterior deberá ocultar términos explícitos como anxiety, depression y depressed para determinar si MentalBERT mantiene su comportamiento mediante patrones contextuales menos evidentes. También será necesario medir la estabilidad de las explicaciones y realizar un análisis sistemático de los errores de clasificación.

La evaluación temporal constituye una fortaleza porque aplicó el mismo modelo, sin reentrenamiento, sobre un periodo posterior. No obstante, la similitud del F1 macro no equivale a una evaluación completa de concept drift. El estudio no calculó distancias entre distribuciones, cambios de vocabulario, intervalos de confianza ni solapamiento histórico de autores. A esto se suma que la cobertura de Arctic Shift puede contener registros faltantes. Por estas razones, la conclusión temporal se limita al corpus y al procedimiento descritos.

El prototipo bilingüe facilita el ingreso de textos en español, pero MentalBERT continúa siendo un modelo ajustado sobre publicaciones en inglés. La traducción puede modificar matices, expresiones idiomáticas y la distribución de probabilidades. En consecuencia, los resultados deben interpretarse como una clasificación computacional experimental y no como una herramienta de diagnóstico o cribado clínico.

La mejora de MentalBERT sobre BiLSTM y el modelo clásico fue de 0,0221 y 0,0282 en F1 macro, respectivamente. Sobre un conjunto de prueba de 102.055 publicaciones, esa diferencia equivale a varios miles de textos clasificados correctamente, especialmente en depresión y ansiedad, donde detectar bien los casos importa más. Por otra parte, el costo computacional de los transformadores no es trivial. La literatura sobre Green AI advierte que entrenar modelos de este tipo puede consumir recursos considerables y generar emisiones significativas [14][5]. La decisión depende del contexto: si se dispone de recursos, se trabaja con volúmenes grandes y se prioriza reducir falsos negativos, MentalBERT puede justificarse; si se necesita rapidez, interpretabilidad y despliegue en máquinas modestas, el modelo clásico o BiLSTM siguen siendo opciones sensatas. En conjunto, la mejora es real pero acotada: no resuelve los problemas de fondo del etiquetado débil. Es un paso adelante, no la meta.

.

V. LIMITACIONES

El estudio presenta varias limitaciones. Las etiquetas se construyeron mediante criterios débiles y no cuentan con validación clínica. El control relativo procede de una comunidad dedicada a salud mental, mientras que la clase proxy de comorbilidad se definió a nivel de autor. El corpus de entrenamiento está en inglés y la traducción automática puede introducir variaciones. Los experimentos se ejecutaron con una sola semilla y no incluyeron pruebas de significancia. Para robustecer las conclusiones, se recomienda repetir el entrenamiento con múltiples semillas y reportar media y

desviación estándar; calcular intervalos de confianza al 95 % mediante *bootstrap;* aplicar la prueba de McNemar para comparaciones pareadas; y estimar el tamaño del efecto con la de Cohen. Estas técnicas permitirían distinguir una mejora trivial de una moderada y dar respaldo inferencial a las diferencias observadas.

Asimismo, las explicaciones de LIME son locales y aproximadas, y su estabilidad no se evaluó formalmente. La validación temporal se limitó a 2025. Aunque los hiperparámetros se verificaron en los scripts, la reproducción exacta también depende del hardware, las versiones de las bibliotecas y el estado de aleatoriedad del entorno.

VI. Conclusiones

MentalBERT obtuvo el mejor rendimiento global entre los tres enfoques evaluados, con un F1 macro de 0,5716 y un recall macro de 0,5718. La mejora respecto de BiLSTM y del modelo clásico fue moderada, pero consistente. La clase proxy de comorbilidad continuó siendo, con diferencia, la categoría más difícil de clasificar, con un recall de 0,2277. En el corpus temporal de 2025, el F1 macro se mantuvo cercano al resultado original (0,5702), aunque el recall experimentó variaciones apreciables entre categorías.

El bajo rendimiento en comorbilidad no debe atribuirse únicamente a una limitación del modelo, sino también al diseño del etiquetado: la clase proxy se definió a nivel de autor, el esquema de etiqueta única obliga a elegir una sola categoría y el desbalance agrava el problema. Este hallazgo refuerza la necesidad de explorar esquemas multi-etiqueta, modelos jerárquicos y corpus anotados con criterios independientes en futuras investigaciones.

Las explicaciones locales generadas con LIME permitieron identificar términos concretos que influyeron en las predicciones individuales, lo que supone un paso útil hacia la transparencia del modelo. Sin embargo, estas explicaciones no descartan la presencia de artefactos estadísticos ni sustituyen una validación clínica con criterios externos. La evaluación sobre el corpus de 2025, sin reentrenamiento, constituye uno de los aportes más valiosos del estudio, pues demuestra que el modelo puede aplicarse a un periodo posterior y mantener un F1 macro cercano al original. El prototipo bilingüe, por su parte, materializa la integración de clasificación, interpretabilidad y traducción automática en un flujo único.

No obstante, conviene ser prudentes. Los resultados dependen de un conjunto de datos con etiquetado débil, de una partición por autor que reduce, pero no elimina la contaminación, y de una ventana temporal acotada. Los experimentos se ejecutaron con una sola semilla y no se aplicaron pruebas de significancia formales, por lo que las diferencias observadas deben leerse como evidencia descriptiva, no concluyente. El sistema se concibe, por tanto, como un recurso experimental orientado a fines académicos y de investigación exploratoria, y no como un instrumento de diagnóstico ni de cribado clínico.

Para robustecer las conclusiones, futuras versiones deberían incorporar bootstrapping para estimar intervalos de confianza al 95 %, la prueba de McNemar para comparaciones pareadas, el tamaño del efecto (d de Cohen) y ejecuciones con múltiples semillas. En el plano temporal, una prueba de equivalencia (TOST) con un margen predefinido, por ejemplo ±0,02 en F1 macro, permitiría afirmar no solo que no hay diferencia significativa, sino que los rendimientos son equivalentes dentro de un rango aceptable. En cuanto a LIME, repetir el análisis varias veces por instancia y medir la coincidencia de los términos principales mediante el índice de Jaccard daría una idea de su estabilidad; calcular la similitud entre los vocabularios de las clases ayudaría a cuantificar el solapamiento léxico que complica la distinción entre comorbilidad, depresión y ansiedad. Estos análisis no cambiarían las conclusiones principales, pero elevarían su rigor y facilitarían que otros investigadores construyan sobre este trabajo con mayor confianza.

VII. Viabilidad del triaJE digital

El prototipo podría insertarse en un flujo de triaje no clínica: detección de señales de riesgo en foros de apoyo, priorización de casos para moderadores o derivación a recursos de ayuda. Su adopción requeriría alianzas con instituciones de salud mental, validación con corpus anotados por profesionales y un análisis de costos energéticos. A corto plazo, el sistema puede funcionar como herramienta de exploración académica; a mediano plazo, como apoyo a programas de salud digital, siempre con supervisión humana y advertencias claras sobre sus límites.

Referencias

[1] Organización Mundial de la Salud, “Trastorno depresivo (depresión).” Accessed: Jul. 01, 2026. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/depression

[2] W. A. Gadzama, D. Gabi, M. S. Argungu, and H. U. Suru, “The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review,” Jul. 01, 2024, *Elsevier Inc.* doi: 10.1016/j.pmip.2024.100125.

[3] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” 2014, *Association for Computational Linguistics*. doi: 10.3115/V1/D14-1162.

[4] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” *Physica D*, vol. 404, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[5] A. Vaswani *et al.*, “Attention Is All You Need,” *Advances in Neural Information Processing Systems 30*, Jun. 2017, Accessed: Jul. 22, 2026. [Online]. Available: https://arxiv.org/pdf/1706.03762

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019, *Association for Computational Linguistics, Stroudsburg, PA, USA*. doi: 10.18653/V1/N19-1423.

[7] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare,” 2022, *European Language Resources Association*. Accessed: Jul. 22, 2026. [Online]. Available: https://aclanthology.org/2022.lrec-1.778/

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," 2016, *Association for Computational Linguistics*. doi: 10.18653/V1/N16-3020.

[9] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1627078.

[10] S. Rani, K. Ahmed, and S. Subramani, "From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives," *Applied Sciences (Switzerland)*, vol. 14, no. 4, Feb. 2024, doi: 10.3390/app14041547.

[11] "Arctic Shift," *Download tool*, Accessed: Jul. 25, 2026. [Online]. Available: https://arctic-shift.photon-reddit.com/download-tool

[12] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural language processing with transformers: a review," 2024, *PeerJ Inc.* doi: 10.7717/PEERJ-CS.2222.

[13] D. Phiri, F. Makowa, V. L. Amelia, Y. V. A. Phiri, L. P. Dlamini, and M. H. Chung, "Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis," *J. Med. Internet Res.*, vol. 27, 2025, doi: 10.2196/59002.

[14] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, 2020.