Revisión sistemática de modelos de machine learning y deep learning aplicados a la detección temprana de depresión en redes sociales

Andrés Felipe Ruiz Delgado

Asesor

Julio Eduardo Mejía Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Dedicatoria

A Copito, Betty y Julio

Resumen

La Organización Mundial de la Salud estima que cerca de un 3,8 % de la población padece depresión. Tradicionalmente, su diagnóstico se realiza mediante cuestionarios aplicados por un profesional de la salud mental, pero las redes sociales han abierto nuevas oportunidades en su detección temprana gracias a la tendencia creciente de las personas de expresar sus emociones y compartir sus problemas a través de estas plataformas. Este estudio investiga y compara los métodos y técnicas de aprendizaje automático utilizados por diversos autores para detectar signos de depresión en redes sociales, incluyendo modelos clásicos como árboles aleatorios y Naive Bayes, enfoques de aprendizaje profundo como BERT y redes neuronales, y métodos de ensamble, con el objetivo de evaluar su efectividad en la detección temprana de la enfermedad y exponer sus principales limitaciones. Los resultados muestran que características como el contenido textual, el uso de emojis y el horario de publicación influyen en la identificación de tendencias depresivas, y que, entre los métodos revisados, los modelos basados en BERT alcanzan las métricas más altas.

***Palabras clave*:** Machine Learning, Depresión, Redes Sociales, Deep Learning, NLP

Abstract

The World Health Organization estimates that approximately 3.8% of the global population suffers from depression. Traditionally, diagnosis is carried out through questionnaires administered by mental health professionals, but social media has opened new opportunities for early detection due to the growing tendency of individuals to express their emotions and share their problems through these platforms. This study investigates and compares the machine learning methods and techniques used by various authors to detect signs of depression on social media, including classical models such as decision trees and Naive Bayes, deep learning approaches like BERT and neural networks, as well as ensemble methods, with the aim of evaluating their effectiveness in early detection and highlighting their main limitations. Results indicate that features like text, emoji usage and posting hours have an influence in the identification of depressive tendencies, and, among all models reviewed, those based on BERT reach the highest metrics.

Keywords**:** Machine Learning, Depression, Deep Learning, Social Media, NLP

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Introducción

La depresión es un problema que afecta cerca de 280 millones de personas en el mundo. Se caracteriza como una falta de ánimo y pérdida constante de interés que persiste durante un largo periodo de tiempo, diferenciándola de los cambios de humor y las fluctuaciones emocionales que normalmente experimenta cualquier individuo. Puede afectar todas las dimensiones de una persona, y, en el peor de los casos, llevar al suicidio. (Organización Mundial de la Salud, 2023). Varios factores contribuyen a su desarrollo: experimentar situaciones estresantes, la genética, trauma o abuso en la infancia, consumo de sustancias psicoactivas, cambio en los niveles hormonales, aislamiento y soledad, entre otros. (ManiMala et al., 2016).

Detectar la depresión requiere la participación del individuo. Implica una visita a la oficina de un profesional de la salud y un tratamiento, lo que suele estar acompañado de un costo. La barrera económica y el componente activo de participación provocan que muchos casos pasen sin diagnosticar (Rabie et al., 2025). El advenimiento de las redes sociales ha contribuido a la reducción del estigma asociado a las enfermedades mentales, y a que las personas expresen sus emociones, sentimientos y preocupaciones a través de dichas plataformas (Laguna y Araque, 2023). Las técnicas de aprendizaje automático y aprendizaje profundo pueden aprovechar dicha abundancia de información para contribuir en el diagnóstico de esta enfermedad.

En este estudio, se investigan los diferentes enfoques para la detección temprana de la depresión en las redes sociales. En particular, se revisarán modelos basados en técnicas de aprendizaje automático (Naive-Bayes, K-Vecinos más cercanos, árboles de decisión, árboles binarios, *Support Vector Machines* (SVM), entre otros), técnicas de aprendizaje profundo (redes neuronales y modelos de transformador), modelos de ensamble y modelos originales creados por los autores. Además, se considerarán aproximaciones que emplean el texto de las publicaciones,

estadísticas de uso de la red social y datos demográficos de los usuarios, ya sea de manera individual o combinada, para evaluar su eficacia en la predicción de síntomas depresivos.

En gran medida, los conjuntos de datos disponibles se encuentran en inglés, limitando su aplicabilidad a otros contextos lingüísticos y culturales, y expandir a otras fuentes mejora su generalización global (Baydili et al., 2025). Esta monografía busca incluir investigación en otros idiomas, incluyendo el arábigo, el español y el japonés. Esto permite determinar la efectividad de los modelos de aprendizaje bajo diferentes matices de la depresión influenciados por la cultura.

Planteamiento del Problema

La depresión es un trastorno del estado de ánimo que provoca una constante pérdida de interés y tristeza. Se presenta a través de diferentes síntomas incluyendo la falta de energía, pensamientos suicidas, bajo estado de ánimo, dificultades para conciliar el sueño, cambios en la alimentación y sentimientos de insuficiencia y culpa (Arif, H y Chand.S, 2023).

De acuerdo con la Organización Mundial de la Salud (2023), un 3.8% de la población ha experimentado depresión, aproximadamente 280 millones de personas sufren depresión y es más común en mujeres que en hombres. Un 75% de las personas en países de bajos recursos no reciben tratamiento.

El trauma, estrés, abuso de alcohol o sustancias psicoactivas, el desempleo, aislamiento, cambios hormonales, características genéticas y problemas de salud conforman algunos de los factores que llevan al trastorno. Si bien hay consenso en que la depresión es un problema del cerebro, el debate continúa entre los investigadores para encontrar las causas exactas del problema (ManiMala et al. 2016).

Las consecuencias de no abordar el problema son diversas, e incluyen una disminución en la autocompasión y autoaceptación, una tendencia creciente a compararse con otros, baja confianza y disposición de perdonar a otros, un incremento en la ira y dificultades para mantener las relaciones interpersonales (Hasler y Kupferberg, 2023). Esto a su vez, conlleva a dificultades en el trabajo, en el estudio y en la participación en las comunidades. En el peor de los casos, la depresión puede ocasionar el suicidio.

Este trabajo presentará las diferentes técnicas y métodos que se han abordado para presentar una solución al problema, ilustrando sus características, evaluando su efectividad y analizando sus limitaciones.

Justificación

Los métodos actuales para detectar la depresión conllevan demasiado tiempo e involucran un esfuerzo combinado entre el profesional de la salud mental y el paciente. Además, de acuerdo con la Organización Mundial de la Salud (2023), más del 75% de personas en países de bajos y medianos recursos no reciben un diagnóstico por la falta de inversión en salud mental o el estigma social. En el peor de los casos, la depresión puede llevar al suicidio, lo que hace imperativa la búsqueda de nuevos acercamientos que permitan detectar de forma temprana la enfermedad.

Las redes sociales se han convertido en un elemento importante para las personas con depresión, pues les permite discutir su enfermedad con individuos similares de manera cómoda y a distancia (Wang et al., 2020). Esto las convierte en una fuente de información útil para analizar patrones de lenguaje en las publicaciones de los usuarios y utilizar técnicas de procesamiento de lenguaje natural, desde algoritmos de aprendizaje automático hasta modelos de transformador para la identificación de la depresión (Qasim et al., 2025).

Normalmente, el reto de la detección de la depresión en redes sociales se presenta como un problema de clasificación de texto (Titla-Tlatelpa et al., 2021), pero el insumo para lograrlo no se limita a las publicaciones. Por ejemplo, el cuestionario PHQ-9 involucra diferentes preguntas relacionadas con los patrones de comportamiento del paciente, y la forma en la que actúan los usuarios en las redes ayudan a cubrir un rango mayor de síntomas de la depresión (Hemtanon et al., 2022).

Esto a su vez, revela las potenciales aplicaciones en la vida real: incrustarse directamente en la red social y monitorear el estado de la salud mental de sus usuarios (mientras se autorice su

uso) o servir como apoyo a los profesionales de la salud, para agilizar el proceso de diagnóstico de sus pacientes.

Objetivos

Objetivo General

Realizar una revisión sistemática de la literatura sobre técnicas de aprendizaje automático y profundo para la detección de depresión en redes sociales a través del método PRISMA 2020, evaluando su efectividad para la identificación temprana y señalando vacíos, sesgos y buenas prácticas de investigación reproducible

Objetivos Específicos

Identificar artículos científicos que aborden la detección de la depresión en redes sociales utilizando técnicas de aprendizaje automático y profundo mediante fuentes bibliográficas especializadas

Seleccionar estudios primarios en bases especializadas sobre detección de depresión en redes sociales utilizando aprendizaje automático y profundo aplicando criterios de inclusión y exclusión, como periodo, idioma y tipo de estudio.

Comparar los resultados reportados en los estudios seleccionados, a partir de las métricas de desempeño de los modelos de clasificación, para establecer su eficacia en la detección temprana de la depresión.

Marcos de Referencia

El marco de referencia establece las bases teóricas y conceptuales del estudio. El marco conceptual presenta una serie de definiciones relacionadas con el aprendizaje automático, aprendizaje profundo y métricas de evaluación. Por otra parte, el marco teórico presenta las teorías que respaldan la hipótesis que afirma la posibilidad de detectar la depresión utilizando las redes sociales. Finalmente, los antecedentes presentan a muy alto nivel, los acercamientos al problema expuestos por diferentes investigadores. Un análisis más profundo se desarrolla en secciones posteriores del estudio.

Marco Conceptual

Modelos Clásicos de Aprendizaje Automático

El aprendizaje automático o machine learning es un campo de las ciencias de la computación. De acuerdo con Burkov (2019), se puede definir como el proceso de resolver un problema aplicado en dos fases: recopilando un conjunto de datos y construyendo un modelo estadístico basado en esos datos por medio de algoritmos.

Los modelos de aprendizaje automático pueden ser supervisados, no supervisados o de refuerzo. Sin embargo, el problema de la detección temprana de la depresión en las redes sociales se ha considerado como un problema de aprendizaje supervisado, más específicamente, un problema de clasificación: los modelos de aprendizaje son entrenados utilizando un conjunto de observaciones agrupadas de acuerdo con ciertas características, y el objetivo es predecir la categoría o grupo a la que pertenecería una observación completamente nueva (Raschka y Mirjalili, 2017).

Algunos de los algoritmos más utilizados por los autores para la resolución del problema se explican brevemente a continuación.

Naive Bayes: es un algoritmo de clasificación eficaz para conjuntos de datos con muchas características (o multidimensionales). Está basado en los métodos de clasificación bayesianos, que, a su vez, se soportan en el teorema de Bayes. El objetivo es encontrar la probabilidad de que un dato pertenezca a una categoría (o etiqueta), basados en un conjunto de características. La versión gaussiana del algoritmo asume que los datos de cada etiqueta siguen una distribución normal; La versión multinomial asume una distribución multinomial de los datos (VanderPlas, 2017).

Máquina de vectores de soporte (SVM): es un algoritmo de clasificación que busca separar conjuntos de datos en categorías utilizando un hiperplano. Los puntos más cercanos al hiperplano se conocen como vectores de soporte, y la distancia entre ellos se conoce como margen. El objetivo del algoritmo es maximizar la margen entre el hiperplano y los vectores de soporte. (AlSagri y Ykhlef, 2020).

Árboles de decisión: este algoritmo clasifica los datos dependiendo del valor de sus atributos o características. Cada nodo en el árbol representa un atributo, y la condición bajo la que se separa el árbol en cada nodo se calcula utilizando la entropía (AlSagri y Ykhlef, 2020).

K-Vecinos más cercanos: Permite clasificar un dato en base a una métrica de distancia, un valor de K y un conjunto de datos de entrenamiento previamente agrupados. Consiste en encontrar los K-vecinos más cercanos o similares a una observación y clasificarla de acuerdo al grupo mayoritario entre los vecinos. Su costo computacional crece de manera lineal dependiendo del número de datos en el conjunto de entrenamiento (Raschka y Mirjalili, 2017).

Regresión logística: Este algoritmo permite clasificar un conjunto de datos utilizando la función sigmoide. El resultado de la función devuelve un valor entre 0 y 1, que será utilizado para clasificar un dato de acuerdo con un umbral (Raschka y Mirjalili, 2017).

Métodos de Ensamble

Los métodos de ensamble son un paradigma de aprendizaje que combina las predicciones de diferentes modelos y, utilizando un sistema de votos, generan una predicción final (Burkov, 2019).

En la detección temprana de la depresión, se han entrenado múltiples algoritmos para clasificar un usuario en deprimido / no deprimido. La clasificación final se determina encontrando la etiqueta en la que la mayoría de los algoritmos estuvieron de acuerdo (sistema de votos). Aquellos algoritmos con las mejores métricas tienen mayor peso en la votación.

Si bien los algoritmos que participan en el ensamble dependen del diseño del investigador, existen algunos métodos frecuentemente utilizados en detectar la enfermedad.

Bosques aleatorios: Un bosque aleatorio es un ensamble de árboles de decisión. Los árboles de decisión son susceptibles de generar resultados *sobreajustados* (es decir, se ajustan demasiado a los datos con los que fueron entrenados), pero, combinar y promediar sus predicciones genera una mejor clasificación (VanderPlas, 2017).

XGBoost: XGBoost es un ensamble de árboles de decisión. Se diferencia de los bosques aleatorios al procesar cada árbol de manera secuencial, de forma que cada componente del ensamble aprende de los errores del miembro anterior en la secuencia (Chen y Guestrin, 2016).

Modelos de Aprendizaje Profundo

El aprendizaje profundo es una rama del aprendizaje automático basada en redes neuronales profundas, es decir, redes neuronales con múltiples capas. En la detección temprana de la depresión, los modelos de aprendizaje profundo obtuvieron, en general, los mejores resultados, con la desventaja de ser computacionalmente costosos. Los más utilizados se describen a continuación.

Redes neuronales: Una red neuronal es un algoritmo diseñado para mimetizar el funcionamiento del cerebro humano y reconocer patrones. Está compuesta por una serie de capas compuestas por nodos (neuronas) conectados entre sí, que reciben una entrada y generan una salida a través de una serie de operaciones matemáticas. En las redes neuronales recurrentes, cada nodo tiene "memoria" recibiendo sus propias salidas como entradas (Islam et al., 2019). Las redes neuronales convolucionales son hábiles en el reconocimiento de imágenes y cuentan con una capa especial que aplica una convolución, escaneando las imágenes y determinando las figuras que la componen (Burkov. 2019).

Transformers: Son una arquitectura de red neuronal introducida en el artículo "Attention Is all you Need" de Vaswani et al. (2017), propuesta originalmente para tareas de traducción automática. Cuentan con un componente encodificador que recibe un texto de entrada y un decodificador que genera una salida. A su vez, cada uno de estos componentes posee una capa de incrustación, que transforma palabras a vectores numéricos, una capa posicional, que recuerda la posición de cada palabra o carácter (conocidos como *tokens*) en una oración, y un componente de atención, que analiza la relación entre tokens para entender el contexto de cada palabra que compone una frase.

BERT (Bidirectional Encoder Representations from Transformers): Es un modelo de representación de lenguaje introducido por Devlin et al. (2018). Gran parte de su arquitectura se basa en el componente encodificador de los transformadores. Fue entrenado utilizando un método de enmascaramiento (ocultar el 15% de los caracteres de una frase y permitir que el modelo prediga los caracteres faltantes), un método de predicción de frases (dadas dos frases A y B, permitir que el modelo indique si tiene sentido que B aparezca después de A), un conjunto de

800 millones de palabras (BookCorpus) y todo el contenido de Wikipedia en inglés (para un total de 16 GB de texto).

ALBERT (A Lite BERT) es una versión basada en BERT más ligera y menos costosa computacionalmente desarrollada por Lan et al. (2020).

RoBERTa (Robustly Optimized BERT Pretraining Approach) por Liu et al. (2019) es una versión de BERT entrenada con un conjunto de datos 10 veces más grande que en el estudio original (160 GB). Omite el método de predicción de frases y modifica dinámicamente las palabras que se ocultan en el método de enmascaramiento, además, se entrena con frases más largas.

Si bien las variantes de BERT se entrenaron utilizando palabras y frases en inglés, existen alternativas especializadas otros idiomas, por ejemplo, BETO, por Cañete et al. (2023) entrenado utilizando un conjunto de datos en español o AraBERT, por Antoun et al. (2021), entrenado utilizando un conjunto de datos en arábigo.

Métricas

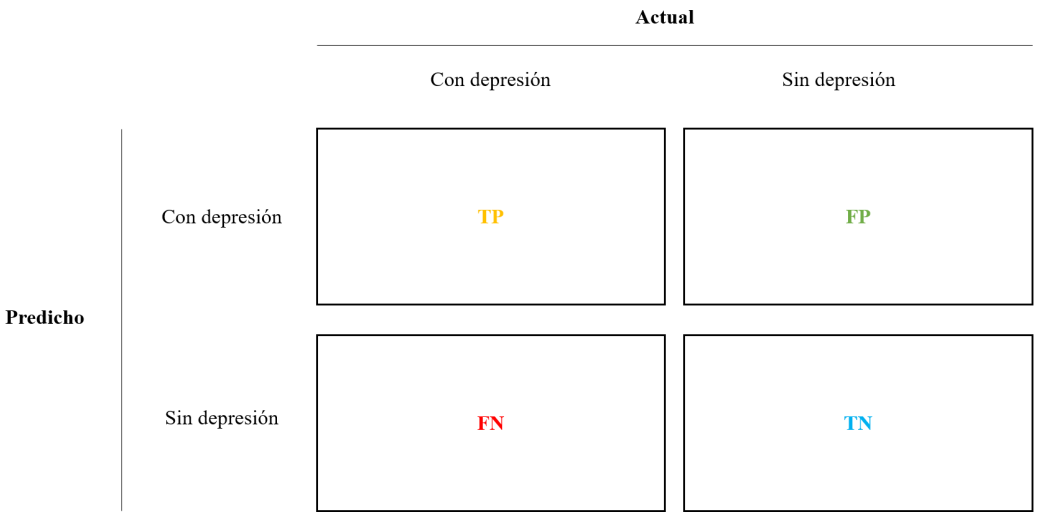
Las métricas son medidas cuantitativas que permiten determinar la efectividad de un modelo. En la detección de la depresión, el uso de métricas es fundamental para comparar cada uno de los acercamientos propuestos por los diferentes autores analizados en este estudio, y su detalle se describe a continuación.

Matriz de confusión: Es una matriz cuadrada que describe los resultados de aprendizaje de un modelo, presentando el número de instancias correctas o incorrectas predichas para cada clase (Kyriakides y Margaritis, 2019).

Por ejemplo, si se diseña un modelo de aprendizaje para clasificar usuarios de Reddit en usuarios con depresión / usuarios sin depresión, se define la matriz de confusión en la Figura 1.

Figura 1

Matriz de Confusión



- Los verdaderos positivos (*True Positives* o TP) hacen referencia a los usuarios con depresión, y que el modelo clasificó correctamente.
- Los verdaderos negativos (*True Negatives* o TN) se refieren a los usuarios sin depresión y que el modelo clasificó correctamente.
- Los falsos positivos (*False Positives* o FP) hacen referencia a los usuarios sin depresión y que el modelo clasificó como usuarios con depresión.
- Los falsos negativos (*False Negatives* o FN) hacen referencia a los usuarios con depresión y que el modelo clasificó como usuarios sin depresión.

Kyriakides y Margaritis (2019) describen otra serie de métricas importantes, explicadas a continuación.

Exactitud: Es la proporción de instancias correctamente clasificadas.

$$Exactitud = \frac{(TP+TN)}{TP+TN+FP+FN}$$

Precisión: Es la proporción de instancias correctamente clasificadas en una clase, relativas a todas las instancias predichas en la misma clase.

$$Precisión = \frac{TP}{TP+FP}$$

Exhaustividad: Es el porcentaje de instancias positivas clasificadas correctamente, divididas sobre el total de instancias positivas.

$$Exhaustividad = \frac{TP}{TP+FN}$$

Puntuación F1: Es la media armónica entre la precisión y exhaustividad. La media armónica previene que valores extremos en cualquiera de sus métricas afecte el resultado del cálculo, lo que la vuelve popular en el cálculo de la eficacia de modelos orientados a la detección de la depresión.

$$F1 = 2\frac{Precisión \cdot Exhaustividad}{Precisión+Exhaustividad}$$

Marco Teórico

En esta sección, se describen los estudios y herramientas que han soportado la premisa de encontrar la depresión a partir del uso de las redes sociales.

El estudio de De Choudhury et al. (2013) se considera pionero en la búsqueda de la resolución del problema, y reveló detalles clave sobre la forma en que un individuo con depresión se comporta en una red social. El estudio se aplicó a 476 personas, analizando el contenido de sus publicaciones y el uso de la red social Twitter. Se encontró:

- Un incremento del uso de pronombres personales, indicando una mayor atención a sí mismos.
- Alto uso de palabras relacionadas con síntomas de la depresión, mención a medicamentos antidepresivos y sentimientos de inseguridad y desesperanza.
- Un mayor volumen de actividad en horas de la noche.

- Distanciamiento social, expresado con bajos seguidores y poca interacción con otros usuarios.

Hemtanon et al. (2022) llevaron a cabo un estudio para identificar las diferencias en los comportamientos en Facebook entre usuarios deprimidos y no deprimidos. Además de confirmar un incremento en la actividad en horas de la noche y baja interacción social para los usuarios deprimidos, se descubrieron patrones irregulares de publicación, gracias a los estados de ánimo inestables provocados por la enfermedad.

Después de la clasificación, se encontró que el número de respuestas a las publicaciones, el número de acciones consecutivas en un lapso de 20 minutos y el número de publicaciones fueron algunos de los atributos no textuales con más influencia a la hora de clasificar un usuario como deprimido. Con respecto a las palabras, *inútil*, *morir*, y *estresado* resultaron ser las más representativas en las publicaciones de los individuos con depresión.

De acuerdo con Titla-Tlatelpa et al. (2021) la depresión se expresa de manera diferente dependiendo del género y la edad. Se encontró que algunas palabras tienen más peso a la hora de calcular la polaridad de un mensaje en Reddit. Por ejemplo, para el género masculino, las palabras *familia* y *sexual* ocurrían de manera más frecuente en publicaciones negativas, y *novio* para el género femenino Por otra parte, *virgen, peso* y *relación* permitían identificar publicaciones negativas en usuarios jóvenes (menores de 26 años).

En Twitter, se encontró que la palabra *calorías* y *ebrio* eran indicativos importantes a la hora de clasificar un tweet como negativo para un usuario femenino y masculino, respectivamente.

Los resultados del estudio de Laguna y Araque (2023) revelaron las palabras más importantes en los modelos de aprendizaje a la hora de clasificar publicaciones de Reddit y

Twitter en la escala de “alta severidad” de la depresión. Se encontró que los temas relacionados a medicamentos, ansiedad, miedo y suicidio tuvieron los mayores pesos en la clasificación final.

Diferentes estudios se han basado en el Manual de Trastornos Mentales (DSM) publicado por la Asociación Estadounidense de Psiquiatría para las tareas de clasificación. El manual describe 9 síntomas asociados con el trastorno depresivo mayor. Una persona puede ser diagnosticada de la enfermedad si cumple con al menos 5 de ellos en un periodo de 2 semanas:

- Estado de ánimo deprimido
- Pérdida de interés y placer en las actividades de la vida diaria
- Pérdida de peso significativa
- Insomnio
- Agitación psicomotriz
- Fatiga
- Sentimientos de culpa o inutilidad
- Pensamientos de muerte y suicidio
- Concentración reducida

Estudios como los desarrollados por Zogan et al. (2023), Baydili et al. (2025) y Rabie et al. (2025) utilizaron el manual diagnóstico para identificar dentro de los textos de las publicaciones en redes sociales, referencias o menciones a los síntomas descritos por el manual. De acuerdo con De Choudhury et al. (2013), las palabras relacionadas con los síntomas de la depresión suelen expresarse frecuentemente en las publicaciones, incluyendo detalles sobre el sueño o los hábitos alimenticios. Esto se corresponde con los datos presentados por el manual, volviéndolo una herramienta imprescindible en la resolución del problema.

Boyd et al. (2022) afirman que las palabras que utilizan las personas nos permiten conocer sobre su estado psicológico, incluyendo sus emociones, personalidades, creencias y experiencias. Esto forma la base del aplicativo LIWC, diseñado originalmente por James W. Pennebaker y Martha E. Francis. LIWC ha sido utilizado en aplicaciones de procesamiento de lenguaje natural y funciona comparando cada palabra de un texto de entrada contra un conjunto de diccionarios de diferentes temáticas. Alhamed et al. (2024) utilizaron la herramienta para analizar un conjunto de datos de Twitter. Se encontró una reducción en emociones positivas y un sesgo hacia enfocarse en el pasado cuando evaluaron las publicaciones de los usuarios deprimidos. Esto se corresponde con los síntomas de la depresión expuestos en el DSM y demuestra su utilidad en la extracción del sentimiento y la polaridad de una publicación.

En conclusión, es posible determinar la depresión de un individuo extrayendo información de sus redes sociales. El texto de las publicaciones y sus acciones, incluyendo las franjas horarias de uso y la relación que mantiene con otros usuarios permiten explicar su estado mental, y funcionan como atributos para alimentar y entrenar algoritmos de aprendizaje automático y aprendizaje profundo.

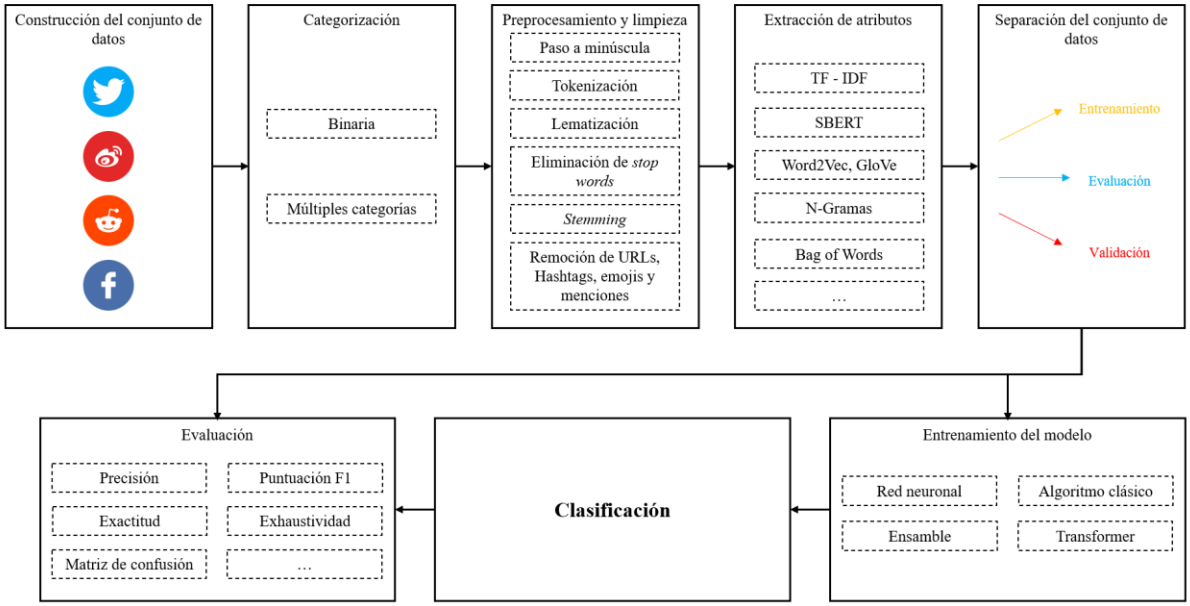
Antecedentes

A continuación, se exponen los diferentes acercamientos a la resolución del problema encontrados por diferentes autores, agrupados por la metodología utilizada. Si bien la mayoría de los estudios se han enfocado en conjuntos de datos en inglés, existen soluciones propuestas en otros idiomas, y se presentan más adelante.

A nivel general, los investigadores aplican el flujo de entrenamiento presentado en la Figura 2.

Figura 2

Flujo de Entrenamiento Típico en Modelos de Aprendizaje



Nota. Íconos de redes sociales tomados de Flaticon, creados por riajulislam (https://www.flaticon.com/authors/riajulislam).

- Construcción del conjunto de datos: Consiste en la extracción de publicaciones o estadísticas de uso de un individuo en una red social. Es posible obtener los datos a través de una API, la recopilación manual o el uso de un conjunto ya existente.
- Categorización: Consiste en agrupar el conjunto de datos en categorías. En algunos estudios, las publicaciones o los usuarios de una red social se dividen en dos grupos: deprimidos y no deprimidos. En otras ocasiones, la división corresponde a un espectro: depresión leve, moderada y severa. Esta tarea se puede llevar a cabo de forma manual empleando el conocimiento de un profesional en la salud mental o automática calculando el sentimiento, la polaridad o el conteo de palabras relacionadas con la depresión en una publicación.

- Preprocesamiento y limpieza: Consiste en la limpieza de datos, especialmente textuales, para eliminar ruido, mejorar su calidad y reducir la dimensionalidad de los vectores generados en el siguiente paso. Se convierten frases a minúscula, se eliminan URLs, emojis, hashtags y menciones a otros usuarios, se reducen las palabras a su raíz (“corriendo” se transforma a “correr”), se eliminan datos duplicados y se dividen las frases en palabras (o “tokens”). Existen otros métodos de limpieza y preprocesamiento, y su uso depende de los investigadores.

- Extracción de atributos: Debido a que los modelos de aprendizaje trabajan con números, es necesario transformar textos y otro tipo de datos. El resultado suele ser un vector numérico que representa una oración, o una matriz, donde cada fila es un vector numérico que representa una palabra. Idealmente, las palabras u oraciones similares deberían estar cercanas en el espacio vectorial.

- Separación del conjunto de datos: Consiste en tomar el conjunto de datos y apartar un porcentaje para entrenar los modelos, otro para verificar los resultados del aprendizaje y, en ocasiones, un último grupo para evaluar el rendimiento del modelo.

- Entrenamiento del modelo: Consiste en entrenar los modelos de aprendizaje utilizando el conjunto de datos recientemente generado con el fin de aprender a clasificar. Existe la posibilidad de controlar la forma en que el modelo entrena modificando una serie de valores llamados hiperparámetros.

- Clasificación: Es el resultado final del proceso de aprendizaje. El modelo clasifica el conjunto de datos de entrada en categorías. En ocasiones, la clasificación es binaria (deprimido / no deprimido), o puede constituirse por varias categorías (niveles de intensidad de la depresión).

- Evaluación: La detección de la depresión en redes sociales es un problema de clasificación supervisado, es decir, ya se conoce de antemano la forma en que el conjunto de datos se encuentra etiquetado. Utilizando algunas métricas como la puntuación F1, la depresión, la exactitud y la exhaustividad es posible conocer el número de clasificaciones correctas e incorrectas.

Aprendizaje Automático

Los algoritmos clásicos de aprendizaje automático, como SVM o árboles de decisión han sido utilizados ampliamente en la tarea de la detección de la depresión. En esta sección se exploran algunos estudios que han aprovechado las capacidades de estos algoritmos.

Titla-Tlatelpa et al. (2021): Clasificaron usuarios de Twitter y Reddit en deprimidos y no deprimidos. Aprovechando el algoritmo de SVM y un ensamble de árboles de decisión, y segmentando las clasificaciones por edad y género se alcanzó una puntuación F1 de 89% en Twitter y 71 % en Reddit.

Laguna y Araque (2023): Emplearon los algoritmos de K-vecinos más cercanos, SVM y árboles aleatorios para clasificar publicaciones de Reddit (en inglés) y Tweets (en español) en tres niveles de depresión. Enriqueciendo el análisis con el tema y la emoción de la publicación se obtuvo una puntuación F1 de 73% para el conjunto de Reddit utilizando árboles aleatorios y 93% para el conjunto de Twitter utilizando SVM.

Qasim et al. (2025): Compararon algoritmos de aprendizaje automático (XGBoost, Naive Bayes, árboles de decisión) y modelos de aprendizaje profundo (BERT, DeBERTa y RoBERTa) para clasificar publicaciones de Reddit en 3 niveles de severidad. Estudiaron la influencia de la extracción de atributos en la clasificación final, obteniendo una puntuación F1 de 91% para el modelo RoBERTa.

AlSagri y Ykhlef (2020): Evaluaron la influencia de utilizar diferentes variables de uso en la red social (franja horaria de publicación, sentimiento de publicaciones, número de seguidores, entre otros) en la clasificación de usuarios de Twitter en deprimidos o no deprimidos. SVM obtuvo la mejor puntuación F1 (79 %), y se concluyó que incorporar variables de uso de la red junto con el texto de las publicaciones mejora la clasificación.

Aprendizaje Profundo

Múltiples investigadores han tratado de abordar el problema de la detección temprana de la depresión a través de redes sociales utilizando las redes neuronales y modelos de transformador.

Bokolo et al. (2023): Compararon modelos clásicos, como regresión logística y bosques aleatorios y modelos de aprendizaje profundo como RoBERTa y DistilBERT para clasificar tweets en depresivos y no depresivos. Si bien RoBERTa alcanzó una puntuación F1 de 98%, todos los modelos superaron un 90%, demostrando la efectividad de ambas estrategias.

Zogan et al. (2022): Combinaron un perceptrón multicapa con una red de atención jerárquica (HAN) para clasificar usuarios de Twitter en deprimidos y no deprimidos. Se incorporaron al estudio atributos como el conteo de emojis, la valencia-excitación-dominancia (VAD) de las publicaciones y su frecuencia horaria. Se logró una puntuación F1 de 91.2 % resaltando el enfoque multidimensional del estudio.

Kerasiotis et al. (2024): Clasificaron publicaciones de Reddit en 4 niveles de depresión, utilizando el modelo DistilBERT y un perceptrón multicapa, incluyendo el texto, la emoción, el sentimiento de las publicaciones y la mención de uso de medicamentos. Los investigadores alcanzaron una puntuación F1 de 84.15%, a pesar de estar limitados por recursos de CPU y un conjunto de datos desbalanceado.

Kumar et al. (2024): Emplearon un conjunto de datos compuesto por tweets y publicaciones de Reddit para entrenar diversas variantes de memoria a corto-largo plazo (LSTM) y unidades recurrentes controladas (GRU). Las versiones clásicas alcanzaron una puntuación F1 de 95%, demostrando la efectividad de estas arquitecturas.

Chen et al. (2023): Clasificaron usuarios de Reddit en deprimidos y no deprimidos utilizando el modelo SBERT para generar vectores numéricos a partir de textos. La red neuronal convolucional encargada de la clasificación logró una puntuación F1 de 87%, superando en al menos 7 puntos porcentuales otros estudios realizados con el mismo conjunto de datos

Alhamed et al. (2024): Clasificaron tweets publicados por usuarios antes y después de un diagnóstico de depresión usando BERT, RoBERTa, MentalBERT y LLMs como ChatGPT y Gemini. BERT alcanzó una puntuación F1 de 97%, mientras que los LLMs no superaron el 36%, exponiendo sus debilidades en tareas especializadas.

Ensambles

Los modelos de ensamble aprovechan las capacidades combinadas de múltiples algoritmos y modelos de aprendizaje profundo y aprendizaje automático para consolidar un único resultado a través de un voto. Utilizar exclusivamente un modelo tiende a resultar en clasificaciones sobre-ajustadas, y limita su generalización (Ogunleye et al., 2024).

Rizwan et al. (2024) crearon un ensamble compuesto por 4 clasificadores: el algoritmo FastText desarrollado por Facebook, regresión logística, naive bayes y árboles de decisión para clasificar tweets en 3 categorías de severidad. La puntuación F1 alcanzada fue de 98%

Tasnim et al. (2024): Con el objetivo de clasificar un conjunto de tweets en 4 niveles de depresión, se diseñó un ensamble compuesto por modelos de transformador, a saber, BERT,

AlBERT y BERTweet. La puntuación F1 alcanzada fue de 85.02%, pero resultó en un incremento de uso de recursos computacionales y mayor tiempo de inferencia.

Baydili et al. (2025): Se construyó un ensamble compuesto por SVM, árboles de decisión, K-Vecinos más cercanos, regresión logística y una red neuronal para clasificar 6 diferentes conjuntos de datos compuestos por publicaciones de diversas redes sociales. La clasificación en uno de ellos alcanzó una puntuación F1 de 99.61%.

Finalmente, Ogunleye et al. (2024) diseñaron un ensamble de 3 modelos base (AdaBoost, regresión logística y potenciación del gradiente) y un perceptrón multicapa como meta-modelo para clasificar publicaciones de Reddit en diferentes niveles de la depresión. Se incluyó en el estudio el sentimiento de la publicación. Se alcanzó una puntuación F1 máxima de 76% en los resultados.

Detección en Otros Idiomas

Existe un gran número de conjuntos de datos en inglés disponibles para el procesamiento de lenguaje natural. Además, modelos como BERT fueron entrenados exclusivamente en un corpus de habla inglesa. Gracias a esto, la mayoría de estudios se han enfocado en entrenar modelos de aprendizaje en inglés.

Sin embargo, algunos autores han presentado un acercamiento a la resolución del problema en otros idiomas utilizando diversas estrategias, como traducir directamente el corpus al idioma objetivo o utilizar un modelo de transformador entrenado específicamente en otra lengua.

Almars (2022). Buscó clasificar tweets en arábigo en depresivo y no depresivo utilizando una red neuronal de memoria a largo plazo y AraVec para transformar palabras en árabe en vectores numéricos. El estudio obtuvo una puntuación F1 de 70%.

Rabie et al. (2024), Realizaron un estudio multidimensional: utilizaron ARABERT para clasificar tweets en depresivos y no depresivos y Multilingual BERT para detectar 9 síntomas diferentes de la depresión en un tweet, logrando exactitudes de 93% y 97.8% respectivamente.

Wang et al. (2020), clasificaron blogs de Weibo en 4 niveles de la depresión utilizando los modelos de transformador BERT y RoBERTa entrenados en un corpus en chino. Se lograron puntuaciones F1 micro promediadas de 85.6% y macro promediada de 42.4% respectivamente, exponiendo los retos que presenta un lenguaje con caracteres ambiguos, como es el caso del chino. Por otra parte, Li et al. (2022), tradujeron publicaciones de Twitter en inglés a chino y blogs de Weibo para determinar qué tan propenso es un usuario a sufrir depresión. Utilizando un enfoque de *Bagging,* se alcanzaron exactitudes superiores a 70%.

Angskun et al. (2022): Clasificaron tweets en tailandés en 3 niveles de depresión (ninguna, leve, moderada) apoyándose de cuestionarios demográficos y médicos. Los bosques aleatorios lograron una puntuación F1 de 91.1%, subrayando la importancia de los datos contextuales. Adicionalmente, Hemtanon et al. (2022) clasificaron 160 personas en Tailandia en deprimidas o no deprimidas de acuerdo a sus publicaciones en Facebook y sus estadísticas de uso (como la franja horaria de publicación), alcanzando una puntuación F1 de 100% utilizando el algoritmo de K-Vecinos más cercanos.

Cha et al. (2022) Buscaron clasificar un conjunto de tweets multilingües (japonés, inglés, coreano) en depresivos y no depresivos utilizando BERT, KoBERT y tohoku-BERT, alcanzando puntuaciones F1 superiores al 99%. El resultado bajó a 64% cuando se trató de generalizar su uso a una red social diferente.

Pool-Cen et al. (2023): El estudio tuvo como objetivo clasificar tweets en español a partir de un conjunto de datos en inglés, utilizando una traducción directa y un método de destilación

del conocimiento. SVM logró una puntuación F1 de 85% con la traducción; la regresión logística obtuvo 93% con destilación.

En conclusión, la detección automática de la depresión a través de redes sociales en idiomas distintos al inglés enfrenta diversos desafíos, como la escasez de conjuntos de datos y las ambigüedades y características lingüísticas únicas de cada idioma. No obstante, los modelos de transformador ofrecen un potencial prometedor en la solución del problema, especialmente cuando se entrenan utilizando un corpus adaptado al idioma objetivo.

Propuestas Originales

Si bien la mayoría de los estudios se han enfocado en utilizar arquitecturas de aprendizaje profundo o algoritmos de aprendizaje automático ya existentes, algunos investigadores ofrecen un acercamiento innovador al problema, presentando modelos completamente nuevos.

Zogan et al. (2021): Propusieron *DepressionNet*, una arquitectura híbrida que combina BiGRU para análisis conductual (VAD, uso de emojis, patrones de publicación, mención de medicamentos) y CNN+BiGRU para análisis textual. Con un mecanismo de atención incluido, el modelo alcanzó una puntuación F1 de 91.2 %, superando enfoques previos.

Burbano et al. (2025): Diseñaron *DEENT*, un modelo de transformador similar a BERT para clasificar tweets en depresivos o no depresivos. Los dos sabores propuestos: *DEENT-Generic* y *DEENT-BERT* lograron puntuaciones F1 de 78.4 % y 81.8 % respectivamente al entrenarse con un conjunto de datos de Twitter.

Ibrahimov et al. (2025)**:** Presentaron *DepressionX*, un modelo de aprendizaje compuesto por un codificador, un mecanismo de atención y un grafo de conocimiento entrenado en artículos de Wikipedia. Al evaluar su comportamiento utilizando dos conjuntos de datos de publicaciones de Reddit, se alcanzaron puntaciones F1 de 82.5% y 90.9%.

Metodología

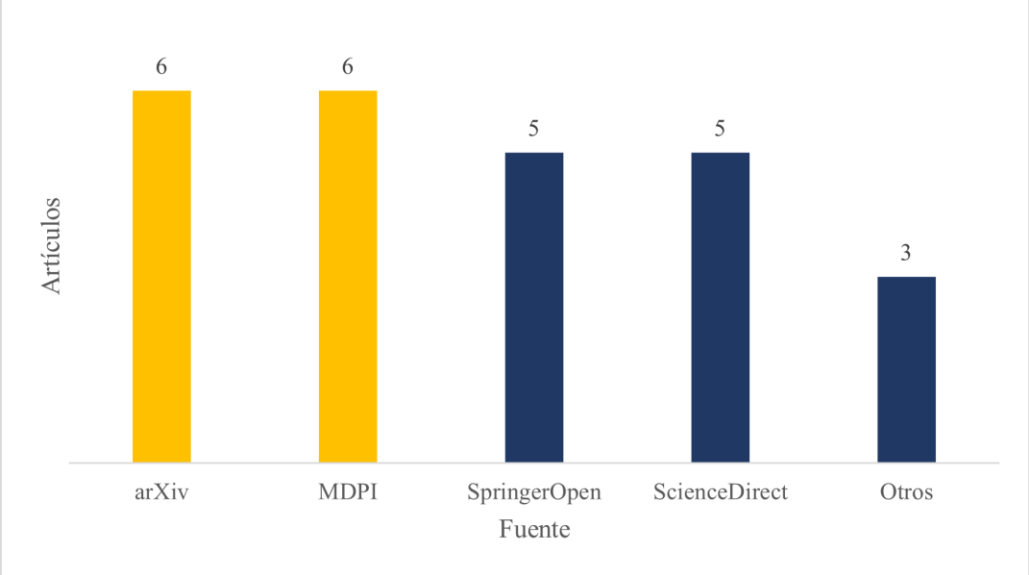
Para llevar a cabo la revisión sistemática se definieron una serie de condiciones que debe cumplir cada artículo para ser considerado:

- El estudio debe especificar la red social base del estudio y el tamaño de la muestra.
- Especificarse el tipo de clasificación final: binaria o categórica.
- Detallar el proceso de construcción del modelo, incluyendo preprocesamiento de datos, entrenamiento del modelo y obtención del resultado.
- Presentar las métricas del problema.
- Antigüedad no mayor a 5 años.

Se realizó una consulta en bases de datos especializadas en inglés, con artículos cuya antigüedad no supere los 5 años. Se han seleccionado 25 estudios, cuya distribución por siti especializado se muestra en la Figura 3.

Figura 3

Distribución de Artículos Investigados por Base



Se han utilizado una combinación de operadores booleanos para refinar la búsqueda, y se describe a continuación:

("depression" OR "depressive symptoms" OR "mental health")

AND

("social network" OR "social media" OR "Twitter" OR "Facebook" OR "Weibo" OR "Reddit")

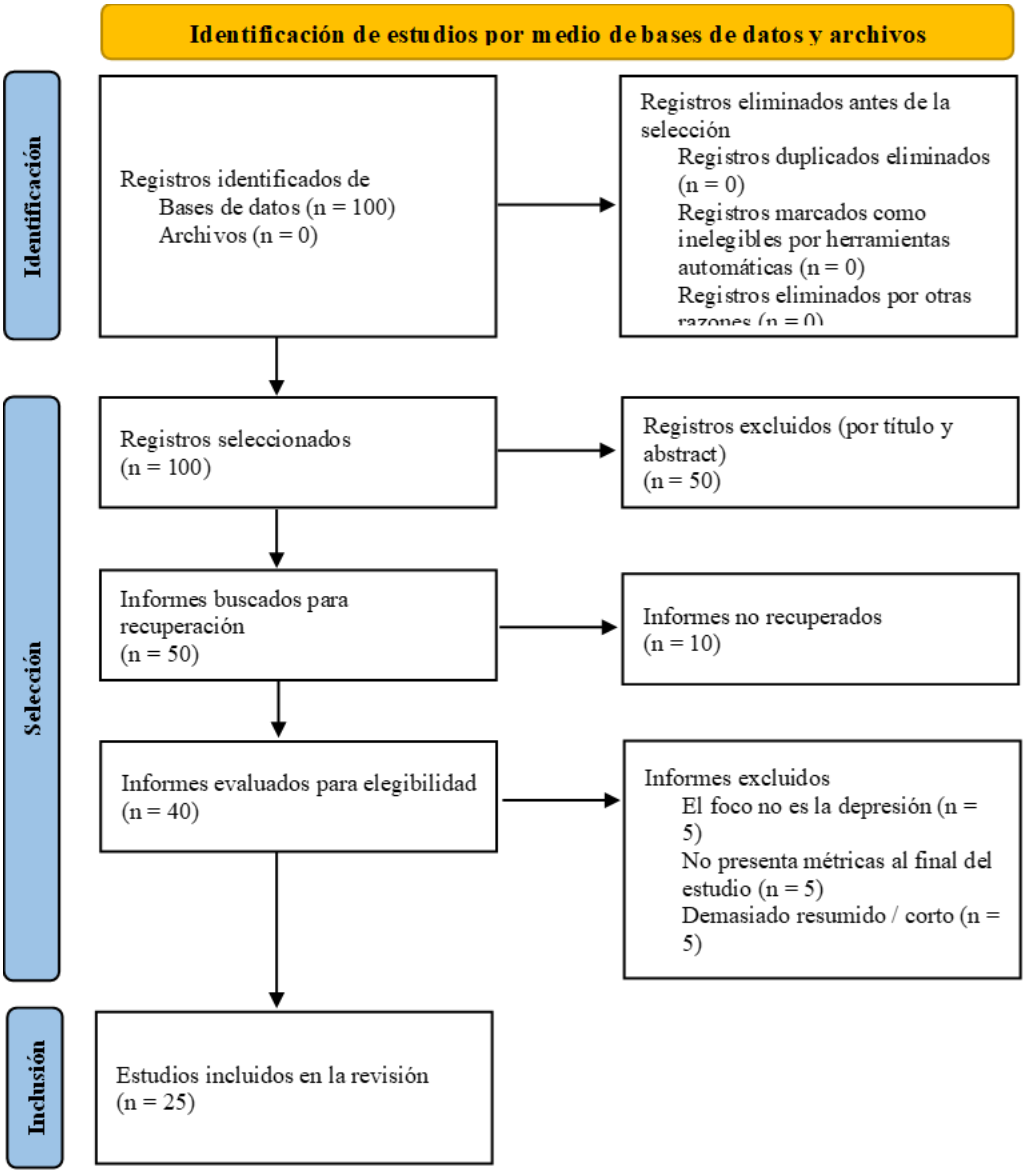
AND

("detection" OR "prediction" OR "machine learning" OR "deep learning" OR "BERT")

El diagrama de flujo, que describe el proceso de selección de los artículos se muestra a continuación. Varios artículos fueron descartados por ser de pago, no cumplir con las condiciones especificadas al inicio de la sección o no tener a la depresión como foco principal del estudio.

Figura 4

Diagrama de Datos - PRISMA 2020



Se realizó una síntesis de cada artículo por medio de una tabla comparativa, presentando las siguientes características:

- Título
- Base de datos

- Año
- Mix de búsqueda
- Palabras clave
- Resumen
- URL Estable
- Nombre de revista
- Citas recibidas
- Autores
- Universidad
- Instrumentos de recolección por parte de los autores
- Técnicas utilizadas para el análisis de datos
- Limitaciones
- Tamaño de la muestra
- Tipo de muestreo
- Objetivo General y objetivos específicos
- Hipótesis / Pregunta de investigación
- Variables utilizadas
- Conclusiones y principales hallazgos

Las siguientes limitantes influyeron en la elección de los artículos y las bases de datos a consultar.

- Artículos de pago o *paywalled*.
- Presenta una métrica, o los resultados de forma muy resumida.

- Antigüedad mayor a 5 años, pero permitiendo estudios excepcionales, como el artículo De Choudhury et al. (2013)
- Menciona la depresión y las técnicas para evaluarla utilizando machine learning, pero no es el foco del artículo.

Análisis de Resultados

Resultados Generales

La Tabla 1 compara los resultados de cada estudio. Para cada uno de ellos, se ha seleccionado el modelo que ha presentado las mejores métricas. Si un estudio evaluó diferentes conjuntos de datos, la métrica presentada es la más alta entre todos los conjuntos. No todas las métricas se encontraban disponibles para todos los estudios. Se presenta la puntuación F1, exactitud (A), precisión (P) y exhaustividad (R).

Tabla 1

Métricas y Modelos Superiores Alcanzados por Cada Estudio

Autores	Red social	Idioma	Mejor Modelo	F1	A	P	R
Ahmed et al. (2024)	Twitter	Inglés	Ensamble (BERT)	0.85	0.84	0.85	0.85
Alhamed et al. (2024)	Twitter	Inglés	BERT	0.98	0.98	0.98	0.98
Almars (2021)	Twitter	Arábigo	Bi-LSTM con mecanismo de atención	0.81		0.78	0.83
AlSagri y Ykhlef (2020)	Twitter	Inglés	SVM Lineal	0.79	0. 82	0.73	0.85
Angskun et al. (2022)	Twitter	Tailandés	Bosques aleatorios			0.91	

Autores	Red social	Idioma	Mejor Modelo	F1	A	P	R
Baydili et al. (2025)	Reddit Twitter	Inglés	SVM	0.99	0.99	0.99	0.99
Bokolo y Liu (2023)	Twitter	Inglés	RoBERTa	0.98	0.98	0.99	0.98
Burbano et al. (2025)	Twitter	Inglés	DEENT-BERT	0.81	0.84	0.83	0.80
Cha et al. (2022)	Twitter Everytime	Coreano Japonés Inglés	BERT	0.98	0.99	0.99	0.99
Chen et al. (2023)	Reddit	Inglés	SBERT - CNN	0.86	0.86	0.85	0.87
Hemtanon et al. (2022)	Facebook	Tailandés	KNN	1	1	1	1
Ibrahimov et al. (2025)	Reddit	Inglés	DepressionX	0.90		0.91	0.90
Kerasiotis et al. (2024)	Reddit	Inglés	DistilBERT + MLP	0.84		0.84	0.84

Autores	Red social	Idioma	Mejor Modelo	F1	A	P	R
Laguna y Araque (2023)	Reddit Twitter	Inglés Español	SVM Lineal	0.93			
Li et al. (2022)	Weibo Twitter	Chino	Bagging		0. 69		
Ogunleye et al. (2024)	Reddit	Inglés	Ensamble + SBERT	0.76	0.83	0.77	0.74
Pool-Cen et al. (2023)	Twitter	Español Inglés	Análisis Discriminante Cuadrático	0.93	0.93	0.97	0.89
Qasim et al. (2025)	Reddit	Inglés	DeBERTa	0.90			
Rabie et al. (2025)	Twitter	Arábigo	Multilingual BERT	0.94	0.97	0.97	0.97
Rizwan et al. (2024)	Twitter	Inglés	Ensamble	0.89	0.93	0.89	0.89
Kumar et al. (2024)	Twitter Reddit	Inglés	Unidad recurrente cerrada	0.95	0.95	0.96	0.95

Autores	Red social	Idioma	Mejor Modelo	F1	A	P	R
Titla-Tlatelpa et al. (2021)	Twitter Reddit	Inglés	Bagging y árboles de decisión	0.89			
Xiofeng et al. (2020)	Weibo	Chino	BERT	0.85			
Zogan et al. (2021)	Twitter	Inglés	DepressionNET	0.91	0.9	0.9	0.9
Zogan et al. (2022)	Twitter	Inglés	Perceptrón multicapa + HAN	0.89	0.89	0.90	0.89

Nota. Cada estudio es diferente y consideró un conjunto de datos, idioma y modelo de aprendizaje distinto, por tanto, el objetivo principal de la tabla no es determinar el mejor acercamiento, sino esclarecer el panorama.

En general, es posible observar cómo los modelos basados en transformadores (es decir, BERT y sus variantes) han presentado los mejores resultados en las clasificaciones a lo largo de diferentes estudios. Esto se puede atribuir al mecanismo de atención que hace parte de su arquitectura, y permite a cada palabra observar las demás palabras en la oración, enriqueciendo el entendimiento del contexto.

Además, existen versiones de BERT especializadas para algunos lenguajes, como AraBERT y Multilingual BERT utilizados por Rabie et al. (2025) para la clasificación de tweets

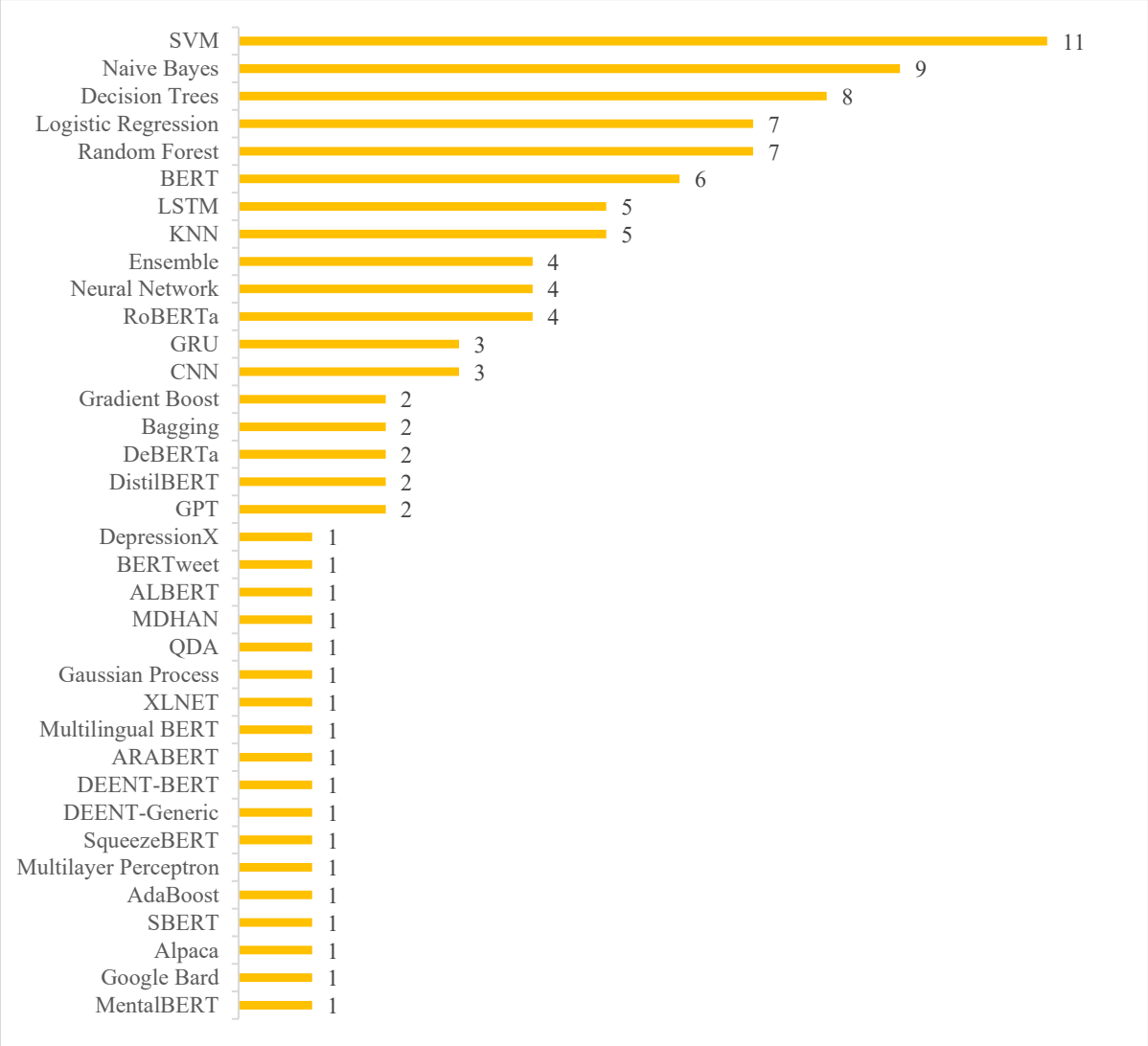
en arábigo, o BETO empleado por Pool-Cen et al. (2023) para obtener representaciones vectoriales de tweets traducidos del español al inglés.

El resultado del acercamiento propuesto por Hemtanon et al. (2022) salta a la vista por haber obtenido valores en 1 para todas las métricas. El conjunto de usuarios que participaron en el estudio se recopiló manualmente y su clasificación se realizó basándose en el cuestionario PHQ-9. Se extrajo información de uso de la red social como el número de publicaciones, comentarios, y respuestas durante diferentes periodos de tiempo, y el número de acciones durante ciertos intervalos, además de los atributos extraídos de las publicaciones en tailandés utilizando la herramienta LexToPlus, creada especialmente para el idioma. Esta combinación de técnicas y disciplinas es, probablemente, lo que llevó a obtener estos resultados.

La figura 5 demuestra la distribución de modelos de clasificación en todos los estudios incluidos. En general, es constante la presencia de modelos tradicionales, como SVM, regresión logística, árboles de decisión, Naive Bayes y bosques aleatorios, que sirven como punto de partida para comparar contra modelos más sofisticados, sin embargo, los modelos de transformador como BERT y sus variantes se encuentran entre los modelos más usados; Incluso cuando no se utilizan como el modelo clasificador final, se implementan en pasos intermedios para generar vectores de incrustación semántica, como en el trabajo de Ogunleye et al. (2024).

Figura 5

Distribución de Modelos de Clasificación



Es posible observar que la mayoría de los estudios fueron realizados en inglés. Esto se debe a que los investigadores entrenaron sus modelos con conjuntos de datos preexistentes, en vez de recopilarlos manualmente. Desafortunadamente, la mayoría de los conjuntos de datos encontrados en la red están en ese idioma, lo que ha forzado a investigadores de otros países a

buscar alternativas para solucionar el problema en su propia lengua. Los detalles se describirán en la sección de limitaciones.

Existe una gran cantidad de acercamientos al problema que utilizan Twitter y Reddit para entrenar los modelos. El estudio realizado por Bucur et al. (2025) reveló que estas dos redes sociales han sido las más utilizadas para construir conjuntos de datos gracias a sus APIs, que permiten extraer una gran cantidad de publicaciones rápidamente. Es posible que esto cambie en el futuro, pues se han actualizado los términos de servicio de las redes sociales, limitando la cantidad de datos que se pueden obtener.

Limitaciones

Conjuntos de Datos en Inglés

Una de las principales limitaciones que han presentado los estudios es la escasez de conjuntos de datos diferentes al inglés. Esto ha forzado a los investigadores a buscar alternativas para entrenar sus modelos.

Pool-Cen et al. (2023) trabajaron con un conjunto de datos en español. Se evaluó la estrategia de traducir los textos directamente a inglés para entrenar los modelos o utilizar una destilación del conocimiento para llevar los textos en español al mismo espacio vectorial del inglés. Esta última estrategia generó los mejores resultados.

Por otra parte, Li et al. (2022) utilizaron la herramienta *Youdao* para traducir el conjunto de datos en chino obtenido de la red social Weibo. Este fue un paso necesario para extraer atributos de las publicaciones y entrenar el modelo, pero resultó en métricas bajas, al compararse con los demás estudios.

Almars (2021), utilizó la API de Twitter para extraer publicaciones en arábigo, sin embargo, las políticas de la red social permitieron obtener aproximadamente 6000 tweets,

reduciendo el tamaño del conjunto de datos, especialmente cuando se compara con aquellos en inglés.

Desbalanceo de los Datos

Diferentes investigadores encontraron un problema con un desbalanceo de los datos. En general, se cuenta con un mayor número de muestras de usuarios no deprimidos.

Burbano et al. (2025) trabajaron con un conjunto de datos desbalanceado con un 56.87% de tweets no depresivos y 43.13% depresivos. Se utilizó la técnica de SMOTE para generar datos sintéticos, y balancear ambas clases. De igual manera, Kumar et al. (2023) utilizaron la misma técnica para balancear un conjunto de textos de Reddit con un 53% de las publicaciones perteneciendo a la clase deprimida.

Wang et al. (2023) entrenaron modelos de aprendizaje con blogs de Sina Weibo, categorizados en 4 escalas de la depresión. 2367 blogs pertenecieron a la clase 0 (sin indicios de depresión), pero solo 26 a la clase 3 (depresión severa), lo que afectó directamente a los resultados de la clasificación.

Ogunleye et al. (2024) recopilaron dos conjuntos de datos de publicaciones de Reddit separadas en escalas de la depresión. El primero de ellos, con más de 3000 textos pertenecientes a las clases sin depresión y depresión moderada, pero sólo 968 para la categoría de depresión severa. En el caso del segundo conjunto de datos, 2587 publicaciones pertenecieron a la clase de depresión mínima, pero ninguna de las demás clases, a saber, leve, moderada y severa, tuvieron más de 400 instancias. Los modelos tradicionales de aprendizaje automático tuvieron dificultades en las predicciones gracias a estos desbalances.

Limitación en las Redes Sociales

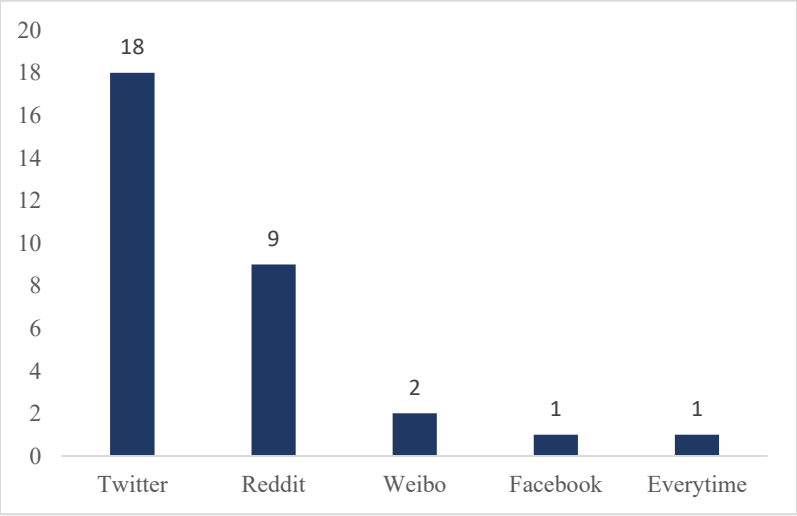
El uso de las APIs permite obtener una gran cantidad de publicaciones y datos de usuarios de redes sociales como Reddit y Twitter. Sin embargo, entrenar estos modelos con sólo estas dos herramientas limitan su generalización.

Cha et al. (2022) buscaron detectar la depresión en la red social universitaria surcoreana *Everytime*. Para empezar, se entrenó un modelo BERT utilizando publicaciones exclusivamente de dicha red social, obteniendo una puntuación F1 de 0.99. Sin embargo, cuando se entrenó el modelo con datos de Twitter y se aplicaron los resultados a *Everytime*, la puntuación F1 bajó a 0.64, demostrando los limitantes de entrenar un modelo con una sola red social.

La figura 6 ilustra el desbalanceo en el uso de las redes sociales en los estudios seleccionados, Twitter y Reddit se posicionan como los orígenes de datos más frecuentes.

Figura 6

Distribución de Redes Sociales Utilizadas como Fuente de Datos



La adquisición de Twitter por parte de Elon Musk en 2023 trajo consigo nuevos retos para la investigación científica. La API de investigación académica fue abruptamente

reemplazada por un modelo de pago, con una versión *Enterprise* de un costo de 42.000 dólares o más dependiendo de su uso. El estudio de Murtfeldt et al. (2024) reveló las disciplinas y temas más afectados por estas nuevas restricciones, entre las que se incluyen metodologías de investigación de Big Data y el estudio del comportamiento a humano, interesados, respectivamente, en el análisis de sentimientos y el análisis de la salud mental. Así pues, estas nuevas políticas de privatización representan un obstáculo en el camino de la investigación y la detección temprana de la depresión.

Por otro lado, Meta otorga acceso de manera gratuita a su API de librería de contenidos, sin embargo, los requerimientos de elegibilidad pueden representar una barrera para los investigadores. Es necesaria una afiliación a una institución académica o de investigación certificada y sin ánimos de lucro, y la postulación es posteriormente evaluada por Meta (Meta, s.f).

Adicionalmente, existen restricciones para la descarga de datos personales por parte de los investigadores. Meta provee un entorno seguro de investigación, más específicamente, una instancia modificada de Jupyter para trabajar con sus datos. Esto con el fin de garantizar la privacidad de los usuarios (Meta, s.f).

Prueba de Concepto

Con el fin de evaluar la capacidad de los modelos de lenguaje para detectar automáticamente la depresión se entrenó un modelo BERT (bert-base-uncased) utilizando un conjunto de Tweets recopilados del sitio Kaggle (Shinde, 2022).

Conjunto de Datos

El conjunto de datos recopilado de Kaggle se compone de 20.000 tweets separados en etiquetas *deprimido* y *no deprimido*. Incluye campos como el texto de la publicación, el número de favoritos, retweets y fecha de creación. Si bien los estudios han demostrado que la inclusión de dichas variables tiene una influencia en la efectividad del modelo, se ha seleccionado exclusivamente el texto del tweet y la etiqueta para realizar el entrenamiento de la prueba.

Limpieza y Pre-procesamiento

Se ha creado una función que permite eliminar elementos de los tweets como menciones a usuarios, URLs, hashtags (se conserva la palabra), caracteres no alfanuméricos, espacios múltiples y emojis. La Tabla 2 presenta los tweets antes y después, el siguiente código describe la limpieza que se llevó a cabo.

```
def limpiar_tweet(text):
    # Eliminar menciones (@usuario)
    text = re.sub(r"@\w+", "", text)

    # Eliminar URLs
    text = re.sub(r"http\S+|www\S+", "", text)

    # Quitar hashtags, pero conservar la palabra
    text = re.sub(r"#", "", text)

    # Eliminar caracteres no alfanuméricos (excepto espacios)
    text = re.sub(r"[^a-zA-Z0-9áéíóúÁÉÍÓÚñÑ ]", " ", text)

    # Quitar espacios múltiples
    text = re.sub(r"\s+", " ", text).strip()

    return text
```

Tabla 2

Comparativo: Tweet Original y Procesado

Texto original	Texto procesado
Completed on my house. Got the keys. What a fab day! Whoop! 🏠🎉💕	Completed on my house Got the keys What a fab day Whoop
RT @BBCBreaking: 7 people confirmed dead after plane crashed into several vehicles on A27 during #Shoreham Airshow http://t.co/iSmQGJY1eB	RT 7 people confirmed dead after plane crashed into several vehicles on A27 during Shoreham Airshow
RT @bxllaneira: how dare this outfit I planned in my head not look good on my body. Disrespectful	RT how dare this outfit I planned in my head not look good on my body disrespectful

Entrenamiento del Modelo

El entrenamiento del modelo se realizó en Google Collab, utilizando una GPU NVIDIA Tesla T4, con 12.67 GB de memoria RAM y procesador x86_64. El entorno de ejecución utilizó Python 3.12.12, junto con las librerías PyTorch 2.8.0 y Transformers 4.57.1.

El conjunto de datos cargado como un dataframe, se transforma a un dataset de HuggingFace, una estructura optimizada para modelos de procesamiento de lenguaje natural. Se utiliza bert-base-uncased (Devlin et al. 2019) para tokenizar el texto de los tweets y llevar a cabo el entrenamiento.

Se utilizó una tasa de aprendizaje de 1×10^{-5}, junto con 4 épocas. Durante las primeras épocas, la pérdida de validación disminuyó, revelando un aprendizaje efectivo, sin embargo, a

partir de la tercera época, la pérdida de validación aumentó, lo que sugiere un sobreajuste. Este comportamiento indica que el modelo alcanzó su mejor desempeño alrededor de la segunda época. La Tabla 3 ilustra este comportamiento.

Tabla 3

Desempeño del Entrenamiento por Época

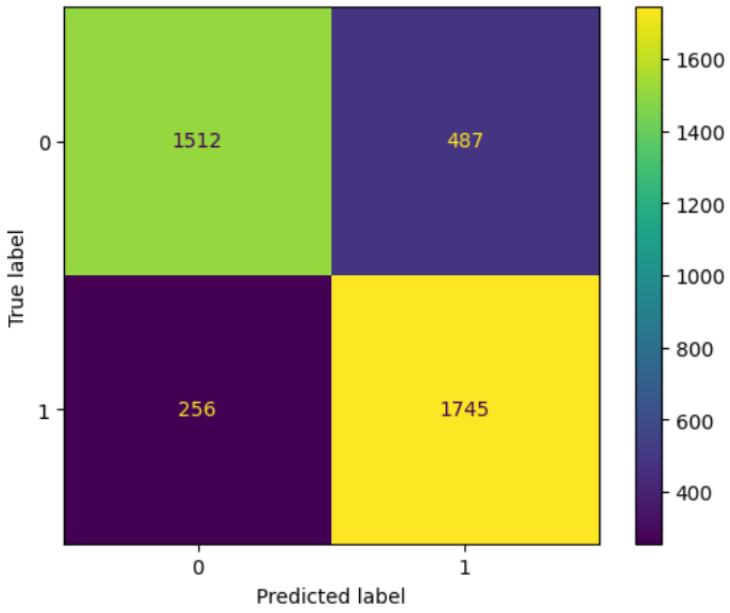
Época	Pérdida de entrenamiento	Pérdida de validación
1	0.448000	0.357054
2	0.347000	0.371085
3	0.291900	0.569526
4	0.158900	0.652130

Resultados

La Figura 7 ilustra el resultado del modelo. El conjunto de datos fue dividido en 80% entrenamiento y 20% validación.

Figura 7

Matriz de Confusión para Clasificación Binaria



El modelo obtuvo una puntuación F1 de 0.82, precisión de 0.7818, recall de 0.8721 y exactitud de 0.8143. Los resultados demuestran la efectividad del modelo BERT para entender lenguaje natural y clasificar correctamente las publicaciones de los usuarios.

Recomendaciones

Con el fin de mejorar el desempeño del modelo, es posible modificar el número de épocas, la tasa de aprendizaje o el tamaño de los lotes. Adicionalmente, es posible incluir las variables adicionales del conjunto de datos, como la hora de publicación del tweet o el número de seguidores, pues de acuerdo con el estudio de De Choudhury et al. (2013), publicaciones en altas horas de la noche y baja interacción con otros usuarios suelen ser característicos de individuos con depresión.

Conclusiones

Utilizar el aprendizaje automático en las redes sociales representa una estrategia prometedora para la detección temprana de la depresión, pues, basados en la evidencia y la investigación, los comportamientos en línea y la forma de expresarse de los usuarios suelen reflejar su estado emocional y mental.

Los modelos de aprendizaje tradicionales como árboles de decisión, máquinas de soporte vectorial o bosques aleatorios han demostrado su eficacia en la tarea de clasificación, sin embargo, se han utilizado principalmente como punto comparativo para modelos más sofisticados, como ensambles o BERT.

Los modelos de transformador, en especial BERT y sus variantes destacan por su capacidad superior para detectar la depresión en redes sociales. Esto se atribuye a su mecanismo de atención y su habilidad para entender el contexto de una frase leyendo en ambas direcciones simultáneamente.

A pesar de generar los mejores resultados, los modelos de aprendizaje basados en redes neuronales y transformadores suelen ser una “caja negra”, ofuscando los detalles de cómo llegaron a la clasificación. Por tal motivo, es crucial incluir un estudio de *explicabilidad* en los modelos para garantizar su transparencia.

Aunque los avances en el área son significativos, los sistemas de detección no están listos para el despliegue en un entorno médico real. Esto se atribuye al alto uso de recursos computacionales requeridos por algunos modelos y los desafíos asociados a la protección y privacidad de los datos de los usuarios.

Recomendaciones

El objetivo de los estudios es integrar los modelos de aprendizaje en un entorno médico real, sin embargo, el uso de recursos computacionales de algunos modelos como BERT dificultan su acoplamiento, pues es posible que existan limitaciones de hardware en los hospitales y clínicas. Es necesario explorar modelos más ligeros, como ALBERT o DistilBERT, que puedan aprovechar el potencial de los modelos de transformador consumiendo una menor cantidad de recursos.

Twitter y Reddit se posicionan como las redes sociales más utilizadas en la búsqueda de la solución del problema, sin embargo, la depresión puede expresarse de diferentes maneras en otras páginas web. Incluir en los conjuntos de datos publicaciones de diferentes fuentes mejora su generalización y permite a los modelos aprender las diferentes expresiones de la depresión en los individuos.

La mayoría de los modelos, en especial las redes neuronales y transformadores se presentan como una caja negra. A pesar de lograr excelentes resultados, el proceso de toma de decisión se mantiene como un misterio para el investigador y el profesional de la salud mental. Es necesario llevar a cabo estudios de *explicabilidad*, que permitan dilucidar los pasos y razones que utilizó el modelo para llegar a una conclusión, describiendo qué atributos tuvieron más peso en la clasificación y porqué.

La detección de la depresión se ha llevado a cabo utilizando mayoritariamente conjuntos de datos en inglés. Además, los modelos de transformador, ampliamente utilizados en la resolución del problema, también fueron entrenados en dicho idioma. Es importante construir y compartir con la comunidad conjuntos de datos en otras lenguas que permitan expandir la

funcionalidad de los modelos de aprendizaje. Los modelos BETO (español) y AraBERT(árabe) ya han dado el primer paso.

Los conjuntos de datos para entrenar los modelos suelen ser recopilados a través de una API de forma masiva, reuniendo publicaciones de redes sociales sin el consentimiento de cada usuario. Es necesario establecer un marco que defina claramente cómo se van a gestionar los datos sensibles de los individuos, y cómo se va a respetar su privacidad.

Referencias Bibliográficas

Ahmed, T., Ivan, S., Munir, A., & Ahmed, S. (2024). Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable AI. *Natural Language Processing Journal*, 7, 100079. https://doi.org/10.1016/j.nlp.2024.100079

Alhamed, F. Ive, J., & Specia, L. (2024). Classifying Social Media Users Before and After Depression. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) Diagnosis via their Language Usage: A Dataset and Study*. 3250–3260

AlSagri, H. & Ykhlef, M. (2020). Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features. *IEICE Transactions On Information And Systems*, E103.D(8), 1825-1832. https://doi.org/10.1587/transinf.2020edp7023

Almars, A. M. (2021). Attention-Based Bi-LSTM Model for Arabic Depression Classification. *Computers, Materials & Continua*, 71(2), 3091-3106. https://doi.org/10.32604/cmc.2022.022609

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5 (R))*. American Psychiatric Association Publishing.

Angskun, J., Tipprasert, S. & Angskun, T. (2022). Big data analytics on social networks for real-time depression detection. *Journal Of Big Data*, 9(1). https://doi.org/10.1186/s40537-022-00622-2

Baydili, İ., Tasci, B. & Tasci, G. (2025). Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. *Behavioral Sciences*, 15(3), 352. https://doi.org/10.3390/bs15030352

Bokolo, B. G. & Liu, Q. (2023). Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques. *Electronics*, 12(21), 4396. https://doi.org/10.3390/electronics12214396

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Bucur, A., Moldovan, A., Parvatikar, K., Zampieri, M., KhudaBukhsh, A. R., & Dinu, L. P. (2025). *Datasets for Depression Modeling in Social Media: An Overview*. arXiv.org. https://arxiv.org/abs/2503.21513

Burbano, R. N., Rendon, O. M. C. & Astudillo, C. A. (2025). An Encoder-Only Transformer Model for Depression Detection from Social Network Data: The DEENT Approach. *Applied Sciences*, 15(6), 3358. https://doi.org/10.3390/app15063358

Burkov, A. (2019). *The hundred-page machine learning book*

Cha, J., Kim, S., & Park, E. (2022). A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Humanities And Social Sciences Communications*, 9(1). https://doi.org/10.1057/s41599-022-01313-2

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1603.02754

Chen, Z., Yang, R., Fu, S., Zong, N., Liu, H. & Huang, M. (2023). Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN. *2022 IEEE 10th International Conference On Healthcare Informatics (ICHI)*. https://doi.org/10.1109/ichi57859.2023.00035

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. *Proceedings Of The International AAAI Conference On Web And Social Media*, *7*(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

De Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R. M., Montes-Y-Gómez, M. & Villaseñor-Pineda, L. (2021). A profile-based sentiment-aware approach for depression detection in social media*. EPJ Data Science*, 10(1). https://doi.org/10.1140/epjds/s13688-021-00309-3

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.04805

Hemtanon, S., Aekwarangkoon, S. & Kittiphattanabawon, N. (2022). Proactive depression detection from Facebook text and behavior data. *International Journal Of Power Electronics And Drive Systems/International Journal Of Electrical And Computer Engineering*, 12(5), 5027. https://doi.org/10.11591/ijece.v12i5.pp5027-5035

Ibrahimov, Y., Anwar, T., & Yuan, T. (2025). DepressionX: Knowledge Infused Residual Attention for Explainable Depression Severity Assessment. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.14985

Islam, M., Chen, G., & Jin, S. (2019). An Overview of Neural Network. *American Journal Of Neural Networks And Applications*, *5*(1), 7. https://doi.org/10.11648/j.ajnna.20190501.12

Kupferberg, A., & Hasler, G. (2023). The social cost of depression: Investigating the impact of impaired social emotion regulation, social cognition, and interpersonal behavior on social functioning. Journal Of Affective Disorders Reports, 14, 100631. https://doi.org/10.1016/j.jadr.2023.100631

Kerasiotis, M., Ilias, L. & Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis And Mining*, 14(1). https://doi.org/10.1007/s13278-024-01360-4

Kumar, K., Anoop, R., Koolagudi., S, Rao. & T, Kodipalli. A. (2024). Stratification of Depressed and Non-Depressed Texts from Social Media using LSTM and its Variants. *Procedia Computer Science*, 235, 1353-1363. https://doi.org/10.1016/j.procs.2024.04.127

Kyriakides, G., & Margaritis, K. G. (2019). *Hands-On Ensemble Learning with Python*. Packt Publishing.

Laguna, A. & Araque, O. (2023). A Cost-aware Study of Depression Language on Social Media using Topic and Affect Contextualization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.17564

Li, C., Liu, H., Yin, B. & Yang, J. (2022). Weibo Depression Posts Detection by Natural Language Processing. *Highlights In Science Engineering And Technology*, 16, 430-437. https://doi.org/10.54097/hset.v16i.2605

ManiMala, S., Gautam, S. & Reddy G, B. (2016). An Overview on Depression. *Research & Reviews: Journal of Pharmacology and Toxicological Studies*, 4 (3), 119-124. https://www.rroij.com/open-access/an-overview-on-depression-.php?aid=79969

Meta. (s.f.). *Content Library & API: Get access*. Meta for Developers. https://developers.facebook.com/docs/content-library-and-api/get-access/

Meta. (s.f.). *Content Library API*. Meta for Developers. https://developers.facebook.com/docs/content-library-and-api/content-library-api/

Murtfeldt, R., Paik, S., Alterman, N., Kahveci, I., & West, J. D. (2024). *RIP Twitter API: A eulogy to its vast research contributions*. arXiv.org. https://arxiv.org/abs/2404.07340

Ogunleye, B., Sharma, H., & Shobayo, O. (2024). Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection. *Big Data And Cognitive Computing*, 8(9), 112. https://doi.org/10.3390/bdcc8090112

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1). https://doi.org/10.1186/s13643-021-01626-4

Pool-Cen, J., Carlos-Martínez, H., Hernández-Chan, G. & Sánchez-Siordia, O. (2023). Detection of Depression-Related Tweets in Mexico Using Crosslingual Schemes and Knowledge Distillation. *Healthcare*, 11(7), 1057. https://doi.org/10.3390/healthcare11071057

Qasim, A., Mehak, G., Hussain, N., Gelbukh, A. & Sidorov, G. (2025). Detection of Depression Severity in Social Media Text Using Transformer-Based Models. *Information*, 16(2), 114. https://doi.org/10.3390/info16020114

Rabie, E. M., Hashem, A. F. & Alsheref, F. K. (2025). Recognition model for major depressive disorder in Arabic user-generated content. *Beni-Suef University Journal Of Basic And Applied Sciences*, 14(1). https://doi.org/10.1186/s43088-024-00592-9

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow*.

Rizwan, M., Mushtaq, M. F., Rafiq, M., Mehmood, A., De la Torre Diez, I., Villar, M. G., Garay, H. & Ashraf, I. (2024). Depression Intensity Classification from Tweets Using

FastText Based Weighted Soft Voting Ensemble. Computers, *Materials & Continua*, 78(2), 2047-2066. https://doi.org/10.32604/cmc.2024.037347

Shinde, V. (2022). *Depression: Twitter Dataset + Feature Extraction*. Kaggle. https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media

Vanderplas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

Wang, X., Chen, S., Li, T., Li, W., Zhou, Y., Zheng, J., Chen, Q., Yan, J. & Tang, B. (2020). Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis. *JMIR Medical Informatics*, 8(7), e17958. https://doi.org/10.2196/17958

World Health Organization. (2023). *Depressive disorder (depression)*. https://www.who.int/news-room/fact-sheets/detail/depression

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1), 281-304. https://doi.org/10.1007/s11280-021-00992-2

Zogan, H., Razzak, I., Jameel, S. & Xu, G. (2021). DepressionNet: Learning Multi-modalities with User Post Summarization for Depression Detection on Social Media. *Proceedings Of The 45th International ACM SIGIR Conference On Research And Development In Information Retrieval*, 133-142. https://doi.org/10.1145/3404835.3462938