Detección Automática de Depresión en Redes Sociales mediante Arquitecturas Transformer

Automatic Depression Detection in Social Media Using Transformer Architectures

Wendy Morales Castro[1], Rafael Guzmán Cabrera[2], José Ruiz Pinales[2], Tirtha Prasad Mukhopadhyay[2]
[1]Departamento de Estudios Multidisciplinares (Sede Yuriria). Campus Sur; Av. Universidad S/N; Colonia Yacatitas; Yuriria, Gto. (México). w.moralescastro@ugto.mx
[2] División de Ingenierías Campus Irapuato-Salamanca. Autopista Salamanca - Valle de Santiago Km. 3,5 + 1,8; Comunidad Palo Blanco; Salamanca, Gto. (México)

Resumen

Este trabajo de investigación estudia el desempeño de una arquitectura de red neuronal Transformer para la detección de depresión mediante el análisis de textos no estructurados provenientes de redes sociales, específicamente de Twitter. La relevancia del estudio radica en que la depresión es una de las principales causas de discapacidad y suicidio a nivel global.

Diversos estudios han señalado que este trastorno puede influir en la forma en que las personas expresan sus pensamientos y emociones. En este sentido, las redes sociales pueden convertirse en un espacio donde los usuarios manifiestan su estado de ánimo, comparten experiencias personales o buscan apoyo, favorecidos en algunos casos por el nivel de anonimato que ofrecen estas plataformas.

Durante los últimos años, la detección automática de indicadores relacionados con la depresión ha recibido una atención creciente, principalmente debido al desarrollo de técnicas de Inteligencia Artificial y Procesamiento de Lenguaje Natural. Dentro de este contexto, los modelos Transformer han adquirido especial importancia por su capacidad para identificar relaciones y patrones presentes en textos, incluso cuando estos son sutiles o dependen del contexto. En este trabajo se analiza su aplicación sobre tweets escritos en español y se compara su desempeño bajo diferentes escenarios de clasificación. Los resultados obtenidos muestran un comportamiento competitivo respecto a metodologías previamente reportadas en el área.

Palabras clave: Depresión, Red Tranformer, Inteligencia Artificial, Procesamiento de Lenguaje Natural, Twitter.

Introducción

La depresión según la Organización Mundial de la Salud es un trastorno mental afectivo común y tratable, que ha logrado afectar a gran parte de la población mundial, este trastorno es caracterizado por alteraciones en el estado de ánimo, acompañadas con síntomas cognitivos y físicos. A pesar de ser un problema de salud mental frecuente a nivel global, no ha recibido la atención necesaria. En particular, la depresión puede convertirse en un problema de salud llegando a afectar la vida laboral, social, escolar, o familiar de las personas que la padecen. En casos muy extremos, incluso puede conducir al suicidio (Secretaría de Salud de México; Organización Panamericana de la Salud; Organización Mundial de la Salud, 2021).

Los métodos tradicionales empleados por los expertos para detectar síntomas suelen basarse en encuestas o preguntas personales, que lleva a los pacientes a responder de forma incorrecta o evitar este tipo de cuestionamientos, debido a que aun en la actualidad la depresión ha sido tratada como tema tabú, del cual las personas no suelen hablar, por temor o vergüenza, lo que conlleva a obtener resultados imprecisos (Retamal, 1998). Por esta razón, se han concentrado esfuerzos en la búsqueda de alternativas que contribuyan a abordar esta problemática. En este contexto, diversos autores han recurrido al uso de la Inteligencia Artificial como una herramienta complementaria para la identificación de la depresión. Aprovechando la diversidad de áreas de la misma, una de las áreas que ha adquirido mayor auge para el desarrollo de este tipo de herramientas es el Procesamiento de Lenguaje Natural (PLN)(Coll-Florit *et al*., 2020).

El PLN, como se mencionó anteriormente, ha adquirido una gran relevancia en los últimos años debido al creciente nivel de interacción entre humanos y computadoras. En este contexto, el PLN desempeña un papel fundamental, ya que permite a los sistemas interpretar y comprender lenguaje humano. A esto se suma la enorme cantidad de información que se genera diariamente en redes sociales, la cual representa una fuente valiosa para investigaciones en diversos campos, como en el ámbito social, el sector empresarial, para la evaluación de productos o servicios, y por supuesto en el área médica, específicamente para identificar casos de depresión. Sin embargo, en el idioma español aún se dispone de poca información relevante. Algunos estudios en esta área han empleado herramientas como cuestionarios, chatbots (Arrabales, 2020) o metodologías relacionadas (Coll-Florit *et al*., 2020; Del Pilar *et al*., 2018; Leis *et al*., 2019; Ramirez-Esparza *et al*., 2008). Estas técnicas presentan desafíos, como la falta de precisión en los resultados o los sesgos derivados de la traducción de datos.

Cabe destacar que esta tarea ha avanzado considerablemente en países como Estados Unidos, Reino Unido y China (Colombo-Mendoza *et al*., 2020). No obstante, cada idioma tiene características particulares, por lo que es fundamental procesar adecuadamente las oraciones para comprender su contexto en el idioma.

Para obtener información relevante, diversas investigaciones han recurrido al análisis de redes sociales como Twitter, Facebook, Instagram, blogs, foros, etc., debido a que se han convertido en una pieza fundamental en la comunicación de las personas. A través de estas plataformas, los usuarios interactúan entre sí, mediante publicaciones, mensajes privados o participación en foros, donde comparten contenidos en formato de texto, audio, o video. Se estima, que, en promedio, las personas dedican un promedio de 3 horas y 49 minutos al día al uso de estas plataformas (Uso Diario Global de Las Redes Sociales Statista, 2025.). Esta alta frecuencia de uso convierte a las redes sociales en un espacio significativo en el que los individuos suelen manifestar sus pensamientos, emociones y estado de ánimo, sin temor a ser juzgados, lo que permite la obtención de información valiosa usada en el presente trabajo.

Además del uso de herramientas propias del Procesamiento de Lenguaje Natural, este trabajo también explora las ventajas de implementar una arquitectura neuronal de última generación conocida como Transformers. Este tipo de red ha sido diseñada específicamente para manejar secuencias de datos, como texto o audio, de manera más eficiente y precisa. A diferencia de las redes neuronales recurrentes (RNN) o convolucionales (CNN), los modelos Transformer no requieren procesar los datos en orden secuencial, lo que les permite identificar y aprender relaciones a largo plazo dentro del texto con mayor eficacia(Gonzalo, 2020).

Metodología Propuesta

La metodología propuesta para el presente trabajo de investigación se presenta en la Figura 1.

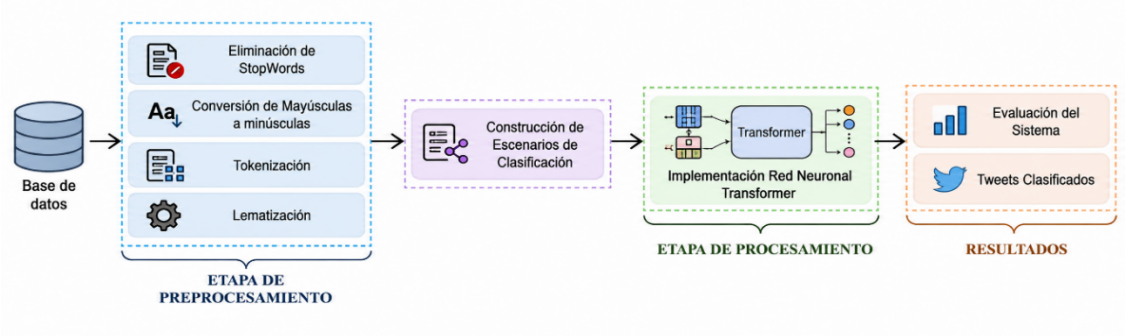


Figura 1. Diagrama de la Metodología.

A continuación, se describen cada una de las etapas de la metodología propuesta.

La primera etapa es la de la base de datos o corpus, para la cual se generó un nuevo corpus combinando dos conjuntos de datos previamente etiquetados por expertos en la materia. El primero, denominado "conjunto de datos de usuarios depresivos[1]", incluye 1000 tweets creados por 90 usuarios. Estos tweets fueron analizados y clasificados con la etiqueta "depresión" por el contenido expresado en ellos. El segundo conjunto, denominado "Análisis del sentimiento de Twitter en los Tweets en español[2]", se incorporó al trabajo por su enfoque y contenido diferente al primer conjunto, con el propósito de representar una clase diferente y asignarle la etiqueta de "no depresivo". Este corpus contenía un total de 2590 tweets, pero solo se seleccionaron 1000 para garantizar un equilibrio entre las clases en el corpus final.

Siguiendo la metodología propuesta, una vez creado el corpus, el siguiente paso será su preprocesamiento. Para ello, se aplicarán técnicas de Procesamiento del Lenguaje Natural diseñadas para limpiar y optimizar el corpus(González *et al*., 2023). Las técnicas utilizadas incluyen:

- Eliminación de palabras vacías.
- Conversión de mayúsculas y minúsculas.
- Lematización.
- Tokenización.

Una vez preprocesados los datos, se dividirán en dos escenarios de clasificación: Cross-Validation (CV) (Pérez-Planells *et al*., 2024.) con 10 folds y Training and Test Sets (TTS) (Águila *et al*., 2017), asignando el 80% de los datos al entrenamiento y el 20% a la prueba. A continuación, se procederá al tratamiento de los datos. Cada escenario se evaluará individualmente mediante el uso de la Red Neuronal Transformer cuya arquitectura se encuentra presentado en la figura 2.

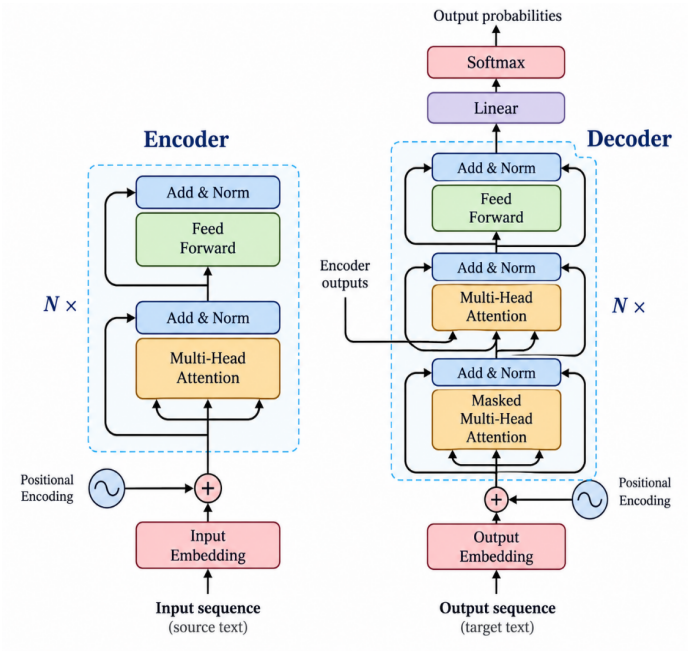


Figura 2. Arquitectura Red Neuronal Transformer

[1]Spanish tweets suggesting depression
[2]Twitter Sentiment Analysis in Spanish Tweets

Para el presente trabajo de investigación es usado el modelo de encoder de la izquierda, dentro del cual los tweets entran ya preprocesados a la red, se realiza el procedimiento mostrado en el diagrama en el que los tweets son convertidos en vectores, y se les va realizando suma de posiciones en cada embeding, para encontrar la relación entre las palabras que existen en el corpus, para después realizar las sumas residuales y la normalización, que permitirá mejorar el flujo del gradiente y estabilizar el entrenamiento, el corpus va transformándose dentro de capas densas, este proceso se va realizando repetidas veces hasta lograr afinar la red. Y finalmente es usada la capa de softmax del bloque derecho para ayudar a calcular la clase a la que pertenece. La red en este sentido estaría ayudando a ajustar los pesos de los embedings, atención y capas densas, del mismo modo se minimizaría la perdida entre la predicción y la etiqueta real.

Continuando con la metodología propuesta, una vez teniendo el cálculo de las clases, se somete a una evaluación mediante las métricas de evaluación, F1-score, Recall, Precisión y Accuracy (Borja-Robalino *et al*., 2025.), para observar que tan buena resulto la red tranformer en la realización de la tarea.

Resultados

La metodología fue ejecutada 20 veces, con la finalidad de observar cómo se iba comportando la red y tener un promedio de las métricas de evaluación. Los resultados para cada escenario de clasificación se presentan en las siguientes gráficas.

En la figura 3 podemos apreciar los resultados de la red usando el escenario de clasificación TTS.

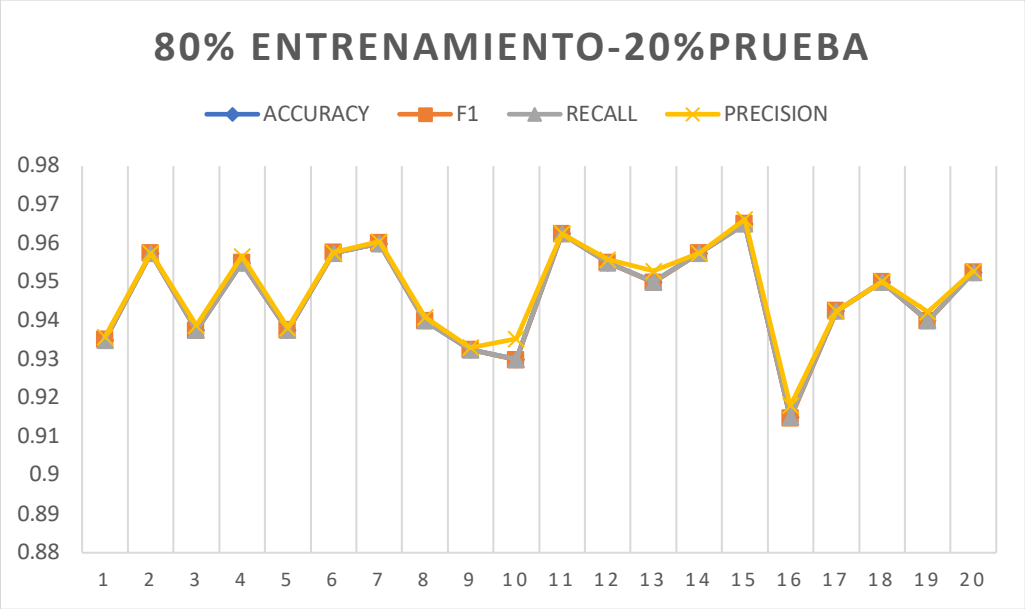


Figura 3. Grafica de resultados con escenario TTS

A continuación, se presentan los resultados obtenidos para el escenario CV.

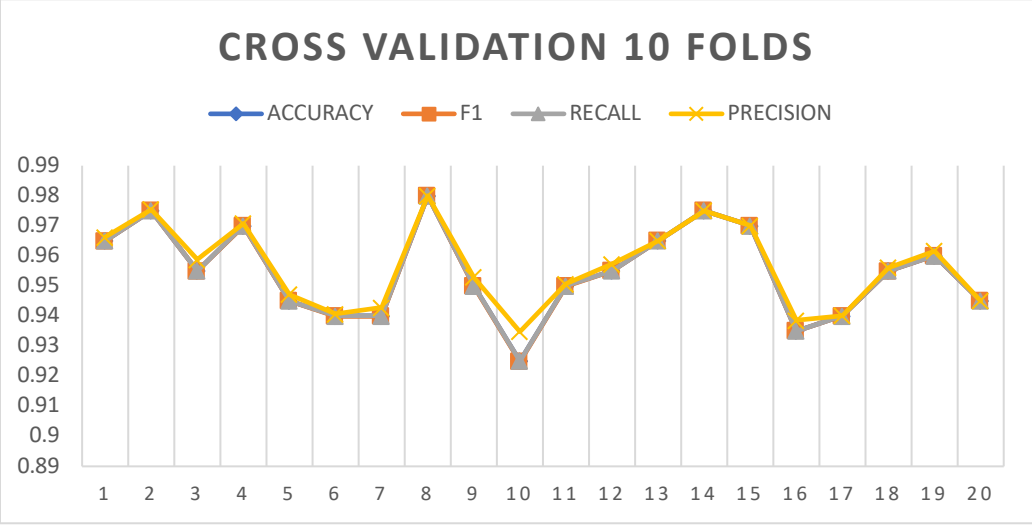


Figura 4. Grafica de resultados con escenario CV

Presentar de manera objetiva, clara y concisa los datos obtenidos mediante la metodología propuesta. Es fundamental contrastar y discutir estos hallazgos con trabajos previos de la literatura científica para demostrar la superioridad, eficiencia o innovación del diseño propuesto. Los datos complejos deben apoyarse firmemente en gráficas y tablas debidamente rotuladas.

Resultados

Con los resultados obtenidos, se pudo observar que la Red Neuronal Trasformer es eficaz para la tarea propuesta de la identificación de depresión, aumentando la eficacia de las métricas de evaluación obtenidas por metodologías clásicas mostradas en el estado del arte, al ejecutar cada escenario de clasificación 20 veces nos permitió observar la estabilidad que se puede obtener con la red, los resultados obtenidos en ambos escenarios, muestran un desempeño destacado, con valores promedio del 94% o superiores, esto evidencia la capacidad de los Transformer para identificar patrones sutiles en el lenguaje que pueden ser asociados con síntomas de depresión.

Asimismo, este estudio pone de manifiesto el potencial del Procesamiento de Lenguaje Natural y la IA como herramientas complementarias a los métodos tradicionales de evaluaciones psicológicas, superando limitaciones como la autoevaluación sesgada o el estigma social. La implementación del modelo en idioma español, un área aun poco explorada, representa un aporte relevante en el campo, abriendo camino para futuras investigación es y aplicaciones clínicas o sociales centradas en la salud mental digital.

Referencias

Águila, J., Trabajo, M., y Solanas Gómez, A. (2017). Aprendizaje supervisado en conjuntos de datos no balanceados con redes neuronales artificiales: métodos de mejora de rendimiento para modelos de clasificación. *Universitat Oberta de Catalunya*. https://openaccess.uoc.edu/handle/10609/64768

Arrabales, R. (2020). Perla: un Agente Conversacional para la Detección de Depresión en Ecosistemas Digitales. Diseño, Implementación y Validación. *Conscious-Robots.* https://www.conscious-robots.com/papers/Perla-Paper-Aug2020_v1_ES.pdf

Borja-Robalino, R., y Monleón-Getino, A. (2020). Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *ResearchGate*. https://www.researchgate.net/profile/Antonio-Monleon-Getino/publication/342009715_Estandarizacion_de_metricas_de_rendimiento_para_clasificadores_Machine_y_Deep_Learning/links/5ede3d0392851cf13869078e/Estandarizacion-de-metricas-de-rendimiento-para-clasificadores-Machine-y-Deep-Learning.pdf

Coll-Florit, M., Climent, S., Sanfilippo, M., y Hernández-Encuentra, E. (2020). Metáforas de la depresión. Estudiar relatos en primera persona de la vida con depresión publicados en blogs. *Metáfora y Símbolo*, *36*(1), 1–19. https://doi.org/10.1080/10926488.2020.1845096

Colombo-Mendoza, A. E., Sánchez-Cervantes, J. L., Alor-Hernández, G., Rodríguez-Mazahua, L., Del P. Salas-Zarate, M., y Alor-Hernández, G. (2020). Análisis comparativo de herramientas y APIs para la identificación y detección de depresión. *Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología*, UNAM.

Gonzalo, J. S. (2020). *Análisis del estado del arte de la generación de texto con redes neuronales mediante modelos de Transformers* [Tesis de grado, Universidad de Valladolid]. https://uvadoc.uva.es/handle/10324/43030

González, S., Reyes, J., Cedeño, D., y Azcuy, R. (2023). Herramienta de PNL para la detección de ambigüedades en requisitos de software escritos en español [Inédito/Preprint]. *EasyChair*. https://easychair.org/publications/preprint_download/RgRDR

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., y Sanz, F. (2019). Detecting signs of depression in tweets in Spanish: Behavioral and linguistic analysis. *Journal of Medical Internet Research*, *21*(6), e14199. https://doi.org/10.2196/14199

Pérez-Planells, L., Delegido, J., y Moreno, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de Teledetección*, (43), 43–54. http://ojs.upv.es/index.php/raet/article/view/4153

Ramirez-Esparza, N., Chung, C., Kacewic, E., y Pennebaker, J. W. (2008). La psicología del uso de palabras en foros de depresión en inglés y en español: Probando dos enfoques analíticos de texto. *Proceedings of the International AAAI Conference on Web and Social Media*, *2*(1), 114–121. https://ojs.aaai.org/index.php/ICWSM/article/view/18623

Retamal, P. (1998). *Depresión*. Editorial Universitaria. https://books.google.com.mx/books?hl=esylr=yid=1kwVmA7st_cCyoi=fndypg=PA3ydq=depresion

Secretaría de Salud de México, Organización Panamericana de la Salud, y Organización Mundial de la Salud. (2021). *WHO MiNDbank - IESM-OMS Informe sobre el sistema de salud mental en México (WHO-AIMS Report on Mental Health System in Mexico)*. WHO MiNDbank. https://extranet.who.int/mindbank/item/448

Statista. (2025). *Uso diario global de las redes sociales 2025*. Statista. https://www.statstatista.com/statistics/433871/daily-social-media-usage-worldwide/

Zárate, M. D. P., y Rodríguez González, D. A. (2018). Detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español. *Procesamiento del Lenguaje Natural*, *60*, 79–82. https://doi.org/10.26342/2018-60-10