Tecnológico de Monterrey

# Reporte de resultados obtenidos

Desarrollo de aplicaciones avanzadas de ciencias computacionales

**Presenta**

Alan Fernando Razo Peña A01703350

Grupo 301

25 de mayo del 2025

# Introducción

En este proyecto, el uso del aprendizaje automático (Machine Learning) para analizar y sintetizar información sobre artículos de noticias resulta sumamente útil. Combinado con una búsqueda eficiente, nos permite revisar grandes volúmenes de texto, identificar patrones lingüísticos sospechosos y discriminar de forma rápida entre afirmaciones basadas en hechos y desinformación, lo que ahorra una gran cantidad de tiempo en comparación con una revisión manual por parte de verificadores humanos.

Sin embargo, aunque el aprendizaje automático es una herramienta poderosa para esta tarea, no es infalible. Puede presentar errores, especialmente en sus versiones gratuitas o limitadas, al analizar matices del lenguaje, sarcasmo, o noticias que imitan el estilo de fuentes legítimas. Por ello, una vez que se han clasificado los casos sospechosos de noticias falsas, es fundamental revisar manualmente el contenido implicado, contrastando con fuentes confiables y utilizando criterio profesional, sin depender ciegamente de los resultados generados por el modelo.

# Obtener, generar o aumentar un set de datos

El Dataset utilizado para el proyecto fue obtenido en la plataforma Kaggle. Posteriormente se almacenó en una carpeta de Google Drive para su manipulación dentro del modelo de Machine Learning (ML). En este caso escogí el dataset de “Fake-News-Detection-Dataset” debido a que me pareció un tema bastante interesante, relevante a la actualidad.

Entender la diferencia entre una noticia falsa de una real es crucial para que no etén circulando libremente en medios de difusion masiva como la prensa escrita, la radio, la televisión, el cine y, más recientemente, Internet y las redes sociales.

# Separación de los sets de prueba y entrenamiento

Usando Scikit-learn, conocido como sklearn, que es una biblioteca de código abierto para aprendizaje automático en Python, se usó la herramienta de “train_test_split” para separar la información contenida en el dataset para entrenar y probar el modelo de Machine Learning.

La división de los datos se realizó de la siguiente manera:

- Entrenamiento: 80%
- Prueba: 20%

# Técnicas de escalamiento y preprocesado de datos

Los métodos que se utilizaron para analizar los datos fueron: el escalamiento estándar de sklearn y la vectorización de texto con TF-IDF con TfidfVectorizer. Estas técnicas estadisticas son

fundamentales para el desarrollo de nuestro modelo del ML. Los beneficios más importantes son que se puede clasificar y ponderar la importancia de las palabras en los archivos de texto.

El software utilizado para ejecutar estas herramientas fue Jupyter Notebook (junto con Python), que sirve para crear documentos de cuaderno interactivos que pueden contener código en vivo, ecuaciones, visualizaciones, medios y otros resultados computacionales.

# Implementación del modelo

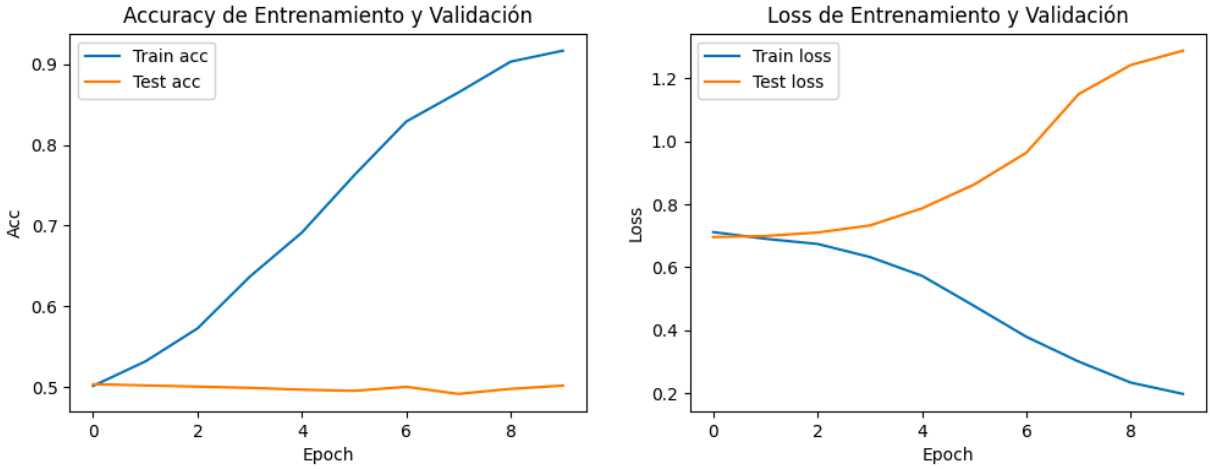
Para la primera implementación del modelo se utilizó un modelo de red neuronal donde las capas se apilan secuencialmente (una tras otra). Este se llama **Sequential** y fue visto durante las clases de Inteligencia Artificial.

Para la segunda implementación se ocupará un modelo de regresión logística para resolver de una manera más optimizada el problema de la clasificación binaria. Esta selección está respaldada por un artículo del estado del arte. En este caso se utilizó el artículo de **Analysis and Detection of Fake News Using Machine Learning.**

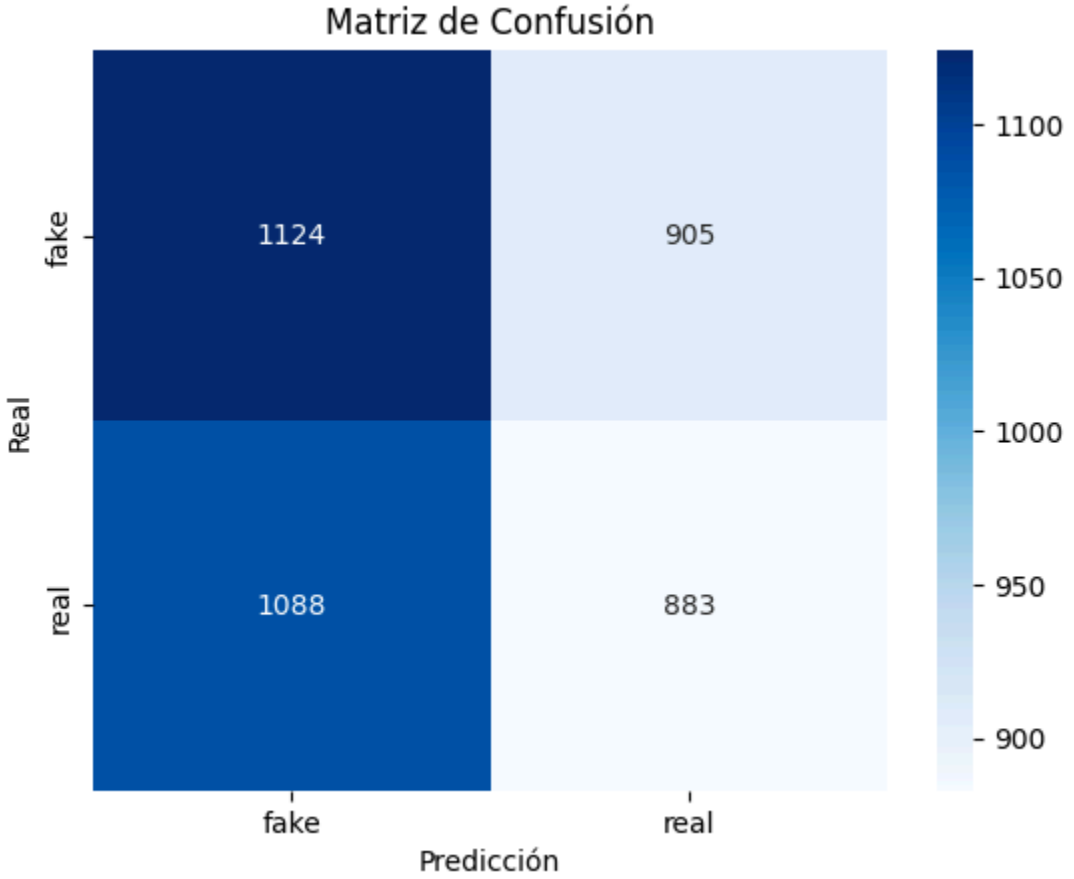
Se utilizan algoritmos de preprocesamiento natural (NLP) y características como la lematización, la tokenización, las palabras vacías (stop-words) y la vectorización TF-IDF para entrenar el modelo de aprendizaje automático. Luego, se utilizan algoritmos de aprendizaje automático como la regresión logística y el bosque aleatorio para clasificar el conjunto de datos, lo que da una precisión del 98% y un puntaje de precisión del 99%.

# Evaluación inicial del modelo

Para esta primera evaluación del modelos, se demostró por medio de las métricas de “accuracy” y “loss” que el modelo está sobreajustado (Overfitting). Esto se debe a una variación alta de los datos y un sesgo bajo.



Usando una matriz de confusión, se pudo observar que para la primera implementación del modelo mostraba mayormente un falso positivo en las predicciones que deberian ser verdaderas.




Para resolver esta situación se debe hacer una regularización y simplificación del modelo. También se va a realizar una implementación del modelo de regresión logística para comparar su desempeño en contraste con el Secuencial de redes neuronales.

# Bibliografías

[1]
P. Kumar, P. Suthanthiradevi, C. A. Stephen, E. Abishek B, S. Sivakumar, and M. Mathiyarasu, “Analysis and Detection of Fake News Using Machine Learning,” May 2024, doi: 10.1109/aiiot58432.2024.10574761