Sentiment Analysis IMBD

Desafío y Elección del Proyecto

El desafío principal que se planteó en este proyecto fue realizar un análisis de sentimientos en el conjunto de datos IMDB, que contiene reseñas de películas. Inicialmente, se probó entrenar un modelo BERT que no estaba diseñado para la tarea específica de análisis de emociones. Este experimento resultó interesante porque ofrecía la posibilidad de explorar las limitaciones y capacidades de un modelo de propósito general aplicado a una tarea específica. Sin embargo, también se consideró utilizar un modelo preentrenado como una alternativa eficiente. Este tema nos pareció interesante porque el análisis de sentimientos es una de las aplicaciones más relevantes en el procesamiento de lenguaje natural (NLP), con amplias aplicaciones en la industria para entender la opinión del público sobre productos, servicios y más. Además, trabajar con modelos avanzados, ya sean entrenados desde cero o preentrenados, ofrece la oportunidad de aplicar técnicas sofisticadas sin la necesidad de grandes recursos computacionales o tiempo extenso.

Detalles Técnicos y Decisiones Tomadas

1. Exploración de Datos y Preparación:
    - Inicialmente, se llevó a cabo un análisis exploratorio de datos (EDA) utilizando herramientas como ydata_profiling para entender mejor la estructura y las características del conjunto de datos. Esto reveló que algunas filas estaban duplicadas y que había una cantidad significativa de stopwords que no aportaban valor predictivo.
    - Decisión: Eliminar las filas duplicadas y proceder con la limpieza del texto para eliminar stopwords y etiquetas HTML. Esta decisión se tomó para mejorar la calidad del texto y, en última instancia, el rendimiento del modelo.
2. Particionamiento del Conjunto de Datos:
    - Aunque el conjunto de datos estaba inicialmente dividido en entrenamiento y prueba, se decidió combinar ambos conjuntos para realizar un nuevo particionamiento, con una proporción del 80% para entrenamiento y 20% para prueba. Esto se hizo para asegurar una mayor cantidad de datos en el entrenamiento, lo que podría mejorar la capacidad del modelo para generalizar.
3. Eliminación de Stopwords:
    - Se utilizó la librería NLTK para eliminar stopwords, lo cual es una práctica común para reducir el ruido en el texto y mejorar la relevancia de las palabras restantes.