DOCUMENTACION DEL PROPROCESAMIENTO LLEVADO A CABO

1. Contexto y Justificación de la Elección del Dataset

Durante la fase inicial del proyecto, se llevó a cabo una búsqueda exhaustiva de datasets que contuvieran tanto etiquetas de Reconocimiento de Entidades Nombradas (NER) como de Análisis de Sentimiento (SA). Sin embargo, tras explorar fuentes como Hugging Face, Kaggle y otras plataformas de datos abiertas, se observó que no existía un dataset con ambos tipos de anotaciones que cumpliera los requisitos del proyecto:

- Estructura tipo secuencia (sentence-level).
- Etiquetas de entidades y sentimiento en paralelo.
- Calidad y tamaño suficiente para entrenamiento o evaluación.

Ante esta limitación, se decidió optar por una solución híbrida y modular, seleccionando un dataset exclusivamente de NER, y aplicando posteriormente un modelo de análisis de sentimiento ya entrenado.

2. Elección del Dataset

Se seleccionó el siguiente dataset de Kaggle:
https://www.kaggle.com/datasets/naseralqaydeh/named-entity-recognition-ner-corpus/data

Este dataset destaca por:

- Contener más de 48.000 frases anotadas.
- Incluir las siguientes columnas: Sentence#, Sentence, POS, Tag.
- Ofrecer etiquetado granular tipo BIO en el campo Tag (usado para NER).
- Ser apto para tareas de token classification o sequence-level NER.

3. Limpieza y Reducción de Dimensionalidad

El dataset original incluía información irrelevante para el propósito final, por lo que se procedió a filtrar y renombrar columnas:

Se eliminó la columna POS, que contenía etiquetas de partes del discurso (Part-of-Speech), ya que no se usarían. También se descartó el índice original Sentence#, considerando que la secuencia ya estaba representada por la frase misma (Sentence). La columna Tag, que contiene etiquetas NER por token, fue renombrada a NER Tag para mayor claridad semántica.

4. Aplicación de Modelo de Sentiment Analysis y Generacion del CSV Final

Una vez preparado el dataset, se aplicó un modelo preentrenado de análisis de sentimiento sobre cada oración para estimar su polaridad emocional. Para esta tarea se seleccionó el modelo cardiffnlp/twitter-roberta-base-sentiment, disponible en Hugging Face. La elección de este modelo se basó en varios factores clave. En primer lugar, está basado en RoBERTa, una arquitectura potente y probada en tareas de NLP modernas. En segundo lugar, ha sido entrenado específicamente sobre texto de Twitter, lo que le da una gran capacidad para captar matices en frases breves e informales, similares a los titulares o noticias cortas que componen nuestro dataset.

Además, el modelo tiene una salida triclase que encaja perfectamente con los objetivos del proyecto: clasificar cada oración como negativa (0), neutra (-1) o positiva (1). A nivel técnico, se utilizó el pipeline de análisis de sentimiento de la librería transformers, junto con un pequeño mapeo para traducir las etiquetas del modelo (LABEL_0, LABEL_1, LABEL_2) a los valores numéricos requeridos.

Durante el análisis, se aplicó una estrategia de truncado a 512 tokens como medida de precaución, ya que es el máximo que RoBERTa puede procesar. Aunque la mayoría de las frases estaban por debajo de ese umbral, esta limitación fue gestionada con un corte preventivo en el texto, garantizando así que todas las entradas fueran válidas para el modelo.

Finalmente, el resultado de cada análisis de sentimiento se añadió como una nueva columna llamada Sentiment al DataFrame original. Esto dio como resultado un dataset final con tres columnas: Sentence, que contiene la frase original, NER Tag, que mantiene la lista de etiquetas por token, y Sentiment, que refleja la polaridad emocional estimada por el modelo. Este conjunto de datos fue guardado como ner_with_sentiment.csv y representa ahora una base enriquecida sobre la que se pueden desarrollar tareas conjuntas de NER y análisis de sentimiento, tanto para entrenamiento como para evaluación o visualización.

Gracias a este proceso, se ha logrado generar un recurso único que combina lo mejor de dos mundos: un corpus etiquetado manualmente para NER y un sistema automático y preciso para asignar sentimiento, ideal para modelos multitarea o análisis exploratorios del lenguaje natural.