

# Clasificación de Reseñas de Hoteles con LSTM y Embeddings Preentrenados

Franklin Manuel M

El proyecto se centra en la clasificación de texto para reseñas de hoteles utilizando redes LSTM (Long Short-Term Memory) y *embeddings* preentrenados, aplicando técnicas de Procesamiento del Lenguaje Natural (PLN). Comienza con la instalación de bibliotecas esenciales como **gensim** (para cargar embeddings Word2Vec), **nltk** (para tokenización y eliminación de *stopwords*) y **tensorflow** (para construir el modelo LSTM). Tras resolver conflictos de versiones de **numpy**, se importan herramientas para el preprocesamiento de texto, manejo de secuencias y evaluación de modelos. Los datos provienen de un archivo CSV que contiene reseñas de hoteles etiquetadas como positivas (1) o negativas (0). Se realiza una limpieza exhaustiva del texto: conversión a minúsculas, eliminación de números, puntuación y *stopwords* en inglés, junto con un análisis de longitudes de texto para entender la distribución de las reseñas.

En la segunda fase, se visualiza la distribución de longitudes de las reseñas para determinar parámetros como la longitud máxima de secuencias. Los datos se dividen en conjuntos de entrenamiento y prueba, seguido de tokenización y *padding* para uniformar las secuencias. Se cargan *embeddings* preentrenados de Word2Vec para inicializar la capa de *embedding* del modelo. La arquitectura del modelo incluye una capa de *embedding* con pesos preentrenados, una capa LSTM para capturar dependencias temporales y una capa densa de salida con activación sigmoide para clasificación binaria. El modelo se compila con el optimizador Adam y la función de pérdida binaria, entrenándose con los datos de entrenamiento. Finalmente, se evalúa el rendimiento con métricas como precisión, *recall* y F1-score en el conjunto de prueba, validando la eficacia del enfoque propuesto.