Universidad del Valle de Guatemala
Data Science
Sección 20

Informe de Laboratorio

Análisis de Sentimientos con LSTM y Características adicionales

Javier Alejandro Azurdia Arrecis - 21242
Angel Sebastían Castellanos Pineda - 21700

Informe de Laboratorio

Análisis de Sentimientos con LSTM y Características adicionales

Introducción

Este laboratorio busca mejorar la precisión en el análisis de sentimientos sobre críticas de películas usando una red neuronal recurrente (RNN) con unidades LSTM y la incorporación de características adicionales. Para ello, se utilizó el conjunto de datos de IMDB y se incrementó el número de palabras consideradas de 20,000 a 50,000. También se implementaron nuevas características extraídas de las críticas, con el objetivo de complementar la información textual.

Arquitectura del Modelo

El modelo diseñado cuenta con una arquitectura sencilla pero efectiva. Se compone de:

1. **Capa de Embedding:** convierte las palabras en vectores de 128 dimensiones, lo que permite al modelo aprender representaciones vectoriales útiles de las palabras. Esta capa tiene 6,400,000 parámetros entrenables.
2. **Capa LSTM:** con 128 unidades, se encarga de procesar las secuencias de palabras y capturar las dependencias temporales. Se aplicaron dropout (0.2) y recurrent_dropout (0.2) para evitar el sobreajuste. Esta capa cuenta con 131,584 parámetros entrenables.
3. **Capa Densa:** con una sola unidad y activación sigmoide, produce la salida final para la clasificación binaria (positivo/negativo). Tiene 129 parámetros entrenables.

El resumen del modelo es el siguiente:

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 500, 128)	6,400,000
lstm (LSTM)	(None, 128)	131,584
dense (Dense)	(None, 1)	129

Total de parámetros:

- Total params: 19,595,141
- Trainable params: 6,531,713
- Optimizer params: 13,063,428

La elección de esta arquitectura se basa en su capacidad para manejar secuencias largas y capturar contextos complejos en el texto. Aunque no se añadieron capas LSTM adicionales, esta configuración es eficiente y evita el sobreajuste.

Características Adicionales

Se incorporaron las siguientes características adicionales:

1. **Longitud de la crítica**: Para reflejar la relación entre la extensión del texto y el sentimiento expresado. Críticas más largas pueden contener más matices y, por lo tanto, afectar la clasificación.
2. **Proporción de palabras positivas**: Calculada como el número de palabras positivas dividido entre el total de palabras en la crítica. Ayuda a cuantificar el tono positivo del texto.
3. **Proporción de palabras negativas**: Similar al anterior, pero enfocada en palabras con connotación negativa.

Estas características se concatenaron con las secuencias de palabras para proporcionar información adicional al modelo.

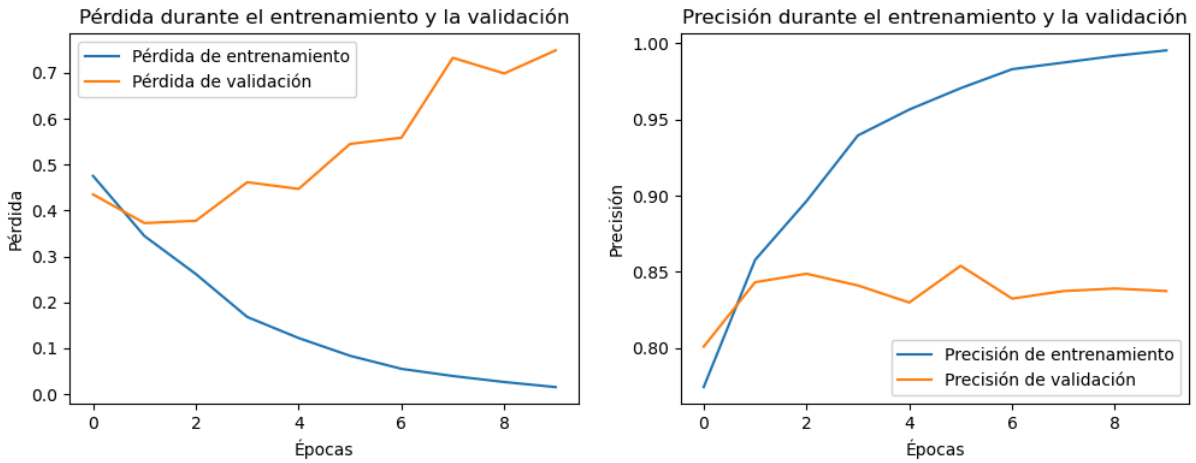
Resultados

El modelo se entrenó durante 10 épocas y mostró una mejora progresiva en las primeras épocas. Los resultados en el conjunto de prueba fueron:

- Precisión: 83.74%
- Pérdida: 0.7489

La precisión final en el conjunto de prueba fue de **83.74%** con una pérdida de **0.7489**. Observamos que, aunque la precisión en entrenamiento alcanzó valores muy altos (hasta 99.62%), la precisión en validación se mantuvo alrededor del 83-85%, indicando posible sobreajuste.

Al evaluar el modelo contra la validación, se obtuvieron los siguientes resultados:



Se encontró diferencia clara entre el entrenamiento y la validación. Comparado con el modelo simple del ejercicio anterior, que tenía una precisión menor, este modelo muestra una mejora gracias a la mayor capacidad de representación y las características adicionales.

Conclusión

El experimento ha demostrado que la incorporación de un vocabulario más amplio y la adición de características adicionales como la longitud de las críticas y la proporción de palabras positivas y negativas mejoran significativamente el rendimiento del modelo en el análisis de sentimientos. El modelo LSTM, junto con estas mejoras, logró una precisión notable del 83.74%, lo que resalta la efectividad de este enfoque. A futuro, se podrían explorar arquitecturas más complejas y una mayor integración de características adicionales, lo que promete seguir elevando el desempeño del modelo.

Recomendaciones

A futuro, se debería considerar los riesgos del sobreajuste en una arquitectura como la que se ha planteado. Fue evidente que, usando menos épocas, por ejemplo 8, se hubiera obtenido un mejor resultado en cuanto a accuracy y loss.