*Marco Ramirez 21032 – Josué Mórales 21116*

*Laboratorio No. 4*

Data Science

Tabla de contenidos

[Modelo simple 3](#_Toc177932519)

[Descripción 3](#_Toc177932520)

[Embedding Layer 3](#_Toc177932521)

[LSTM Layer 3](#_Toc177932522)

[Dense Layer (Output) 3](#_Toc177932523)

[Optimización y Función de Pérdida 3](#_Toc177932524)

[Entrenamiento y Evaluación 3](#_Toc177932525)

[Resultados 4](#_Toc177932526)

[Modelo modificado 5](#_Toc177932527)

[Descripción 5](#_Toc177932528)

[Preprocesamiento y Configuración de Datos 5](#_Toc177932529)

[Capas LSTM 5](#_Toc177932530)

[Entrada de Características Adicionales 6](#_Toc177932531)

[Concatenación y Capas Densas 6](#_Toc177932532)

[Capa de Salida 7](#_Toc177932533)

[Optimización y Función de Pérdida 7](#_Toc177932534)

[Entrenamiento y Evaluación 7](#_Toc177932535)

[Resultados 8](#_Toc177932536)

[Análisis de los Modelos 9](#_Toc177932537)

[Características y Selección de Parámetros 9](#_Toc177932538)

[Arquitectura y Justificación de los Modelos 9](#_Toc177932539)

[Resultados y Comparaciones 11](#_Toc177932540)

# Modelo simple

## Descripción

### Embedding Layer

* Convierte palabras en vectores de 128 dimensiones usando las 20,000 palabras más frecuentes.

### LSTM Layer

* 128 unidades.
* Dropout de 0.2 para reducir el sobreajuste, aplicado a las entradas.
* Recurrent Dropout de 0.2 para evitar el sobreajuste en las conexiones recurrentes.

### Dense Layer (Output)

* Una sola neurona con activación sigmoid para clasificación binaria.

### Optimización y Función de Pérdida

* Optimizador: Adam, adecuado para este tipo de tareas.
* Función de pérdida: binary\_crossentropy, ideal para clasificación binaria.

### Entrenamiento y Evaluación

* Entrenado por 15 épocas con un tamaño de lote de 32.
* Evaluación final en el conjunto de prueba para determinar la precisión y la pérdida.

### Resultados

A screenshot of a computer program

Description automatically generated

A computer screen with numbers and letters

Description automatically generated with medium confidence

# Modelo modificado

## Descripción

### Preprocesamiento y Configuración de Datos

* **Embedding Layer**: Transforma las palabras en vectores de 128 dimensiones, usando un vocabulario de las 50,000 palabras más frecuentes del dataset de IMDB.
* **Secuencias de Entrada**: Cada secuencia tiene una longitud máxima de 150 palabras, ajustada mediante padding.

### Capas LSTM

* **Primera Capa LSTM**:
  + **Unidades**: 128.
  + **Dropout**: 0.3, aplicado a las entradas de la capa para reducir el overfitting.
  + **Return Sequences**: True, devuelve secuencias completas a la siguiente capa para mantener la temporalidad entre los datos.
* **Segunda Capa LSTM**:
  + **Unidades**: 64.
  + **Dropout**: 0.3, similar a la primera capa para mantener la coherencia en la regularización.

### Entrada de Características Adicionales

* **Características**: Longitud de la secuencia y proporción de palabras positivas a negativas.
* **Proceso**: Estas características se extraen del texto y se escalan usando StandardScaler para normalizar los datos antes de su uso en el modelo.

### Concatenación y Capas Densas

* **Concatenación**:
  + Une las salidas de la segunda capa LSTM y las características adicionales escaladas para formar una entrada compuesta que se pasa a las siguientes capas densas.
* **Capas Densas**:
  + **Primera Capa Densa**:
    - **Unidades**: 64.
    - **Activación**: relu.
    - **Dropout**: 0.5, para reducir el riesgo de sobreajuste dado el aumento en la complejidad del modelo.
  + **Segunda Capa Densa**:
    - **Unidades**: 32.
    - **Activación**: relu.

### Capa de Salida

* **Neurona Final**: Una neurona con activación sigmoid para realizar la clasificación binaria, discriminando entre críticas positivas y negativas.

### Optimización y Función de Pérdida

* **Optimizador**: Adam, conocido por su eficiencia en el manejo de grandes volúmenes de datos y ajustes dinámicos de la tasa de aprendizaje.
* **Función de Pérdida**: binary\_crossentropy, adecuada para problemas de clasificación binaria.

### Entrenamiento y Evaluación

* **Entrenamiento**: Configurado para correr durante 15 épocas con un tamaño de lote de 32, incluyendo validación con el conjunto de test para monitorear el overfitting y ajustar el entrenamiento si es necesario.
* **Evaluación**: Se realiza post entrenamiento para verificar la precisión (accuracy) y la pérdida (loss) en el conjunto de datos de prueba.

### Resultados

A screenshot of a computer

Description automatically generated

# Análisis de los Modelos

## Características y Selección de Parámetros

* **Características Adicionales Utilizadas**:
  + **Longitud de la Crítica**: Se optó por incluir la longitud de las críticas debido a la premisa de que textos más largos podrían contener más matices, afectando así la polaridad del sentimiento expresado.
  + **Ratio Positivo/Negativo de Palabras**: Al calcular la proporción entre palabras positivas y negativas, se buscó capturar de manera explícita la polaridad inherente en el texto, crucial para tareas de análisis de sentimientos.

## Arquitectura y Justificación de los Modelos

* **Modelo Simple**:
  + **Entrada de Secuencias**: Utiliza una capa de embedding que transforma las palabras en vectores de 128 dimensiones, limitado a las 20,000 palabras más frecuentes.
  + **LSTM**: Una sola capa LSTM de 128 unidades, sin dropout adicional, facilitando un modelo más básico y menos propenso a ajustes finos.
  + **Capa de Salida**: Una sola neurona con activación sigmoid para clasificación binaria.
* **Modelo Modificado**:
  + **Capas LSTM Mejoradas**: Dos capas LSTM, donde la primera tiene 128 unidades con return\_sequences=True y la segunda 64 unidades con dropout de 0.3, permitiendo capturar relaciones temporales más complejas y mitigar el sobreajuste.
  + **Entrada de Características Adicionales**: Incorpora la longitud de las críticas y el ratio positivo/negativo, ambas normalizadas, proporcionando una visión más holística de los datos.
  + **Capas Densas Adicionales**: Incluye dos capas densas con activación relu y dropout de 0.5, aumentando la capacidad del modelo para aprender representaciones no lineales más ricas.
  + **Concatenación y Procesamiento**: La salida de la última capa LSTM se concatena con las características adicionales, permitiendo que el modelo integre información contextual y específica de la crítica.

## Resultados y Comparaciones

* **Modelo Simple**:
  + **Precisión**: Oscila alrededor del 80.64% en el conjunto de prueba.
  + **Pérdida**: Presenta una pérdida de aproximadamente 1.2009, indicativa de una buena generalización, pero con margen de mejora.
* **Modelo Modificado**:
  + **Precisión**: Alcanza un 83.95%, mostrando una mejora significativa respecto al modelo simple.
  + **Pérdida**: Reduce la pérdida a 0.8289, lo que sugiere una mejor capacidad para modelar la complejidad de los datos y emitir predicciones más precisas.