SEMANA7.ipynb

Escuela de ingeniería y ciencias.

Campus **Monterrey**

Unidad de formación **TC3006C.102**

Inteligencia artificial avanzada para la ciencia de datos II

Semana 7: RNN

Equipo **3**

Grupo: **102**

Regina Reyes Juárez - **A01275790**
Nadia Salgado Alvarez - **A01174509**
Gilberto Angel Camacho Lara - **A01613895**
Santiago Miguel Lozano Cedillo - **A01198114**

11 de noviembre

SEMANA7.ipynb

Introducción

El objetivo de este notebook es desarrollar un modelo de Red Neuronal Recurrente (RNN) capaz de clasificar automáticamente reseñas de productos de Amazon según su sentimiento (positivo o negativo). A partir del texto escrito por los usuarios, el modelo busca identificar patrones lingüísticos que reflejen la opinión del consumidor, permitiendo predecir la percepción general de un producto con base en su descripción textual.

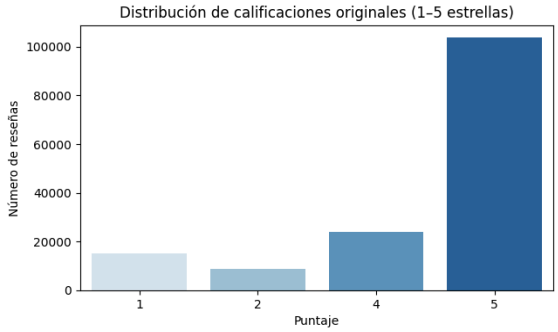
Descripción del dataset

El dataset utilizado es “Amazon Fine Food Reviews”, obtenido de Kaggle, que contiene más de 560,000 reseñas de productos alimenticios vendidos en Amazon. El conjunto incluye 10 columnas, entre ellas el identificador de usuario, producto, texto de la reseña, puntuación (Score) y resumen (Summary).

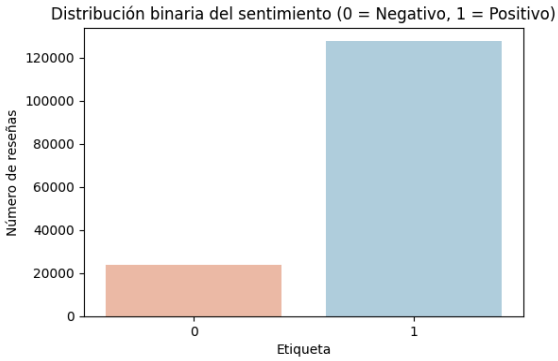
La base de datos presenta 164,346 entradas, de las cuales solo tenemos valores faltantes en ProfileName y Summary. Cinco de las columnas que tenemos son int64 y otras 5 son tipo object.

EDA

En la base de datos contamos con un número considerablemente mayor de reseñas positivas que negativas claramente observable en la Imagen 2, sin embargo, siempre resulta mayor el número de calificaciones al límite de la opinión (puntaje de 1 o 5) en comparación al más cercano al neutro (puntaje de 3) como se observa en la Imagen 1.

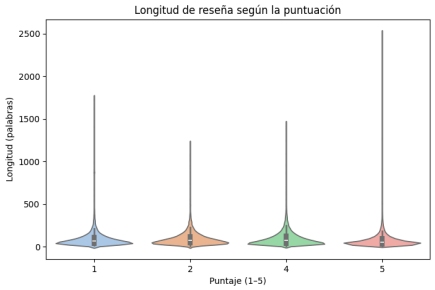


Imágen 1. Distribución de calificaciones originales.



Imágen 2. Distribución binaria de sentimiento.

Por otro lado, en la Imagen 3 se puede observar que, analizamos la longitud de la reseña conforme al puntaje dado, dándonos cuenta que casi a nivel general mientras un puntaje vaya siendo mejor, la longitud de la reseña va aumentando.
Sin embargo esta diferencia es muy mínima considerando la longitud más común por puntaje.



Imágen 3. Longitud de reseña según la puntuación.

Decidimos graficar las palabras más frecuentes en las reseñas en la Imagen 4, esto para poder ver si encontramos algún tipo de insight, como por ejemplo, podemos anticipar que la mayoría de las reseñas negativas se derivan por el sabor (taste) mientras que las positivas suelen contener la palabra great en general para el precio, el sabor, el producto del café y demás.

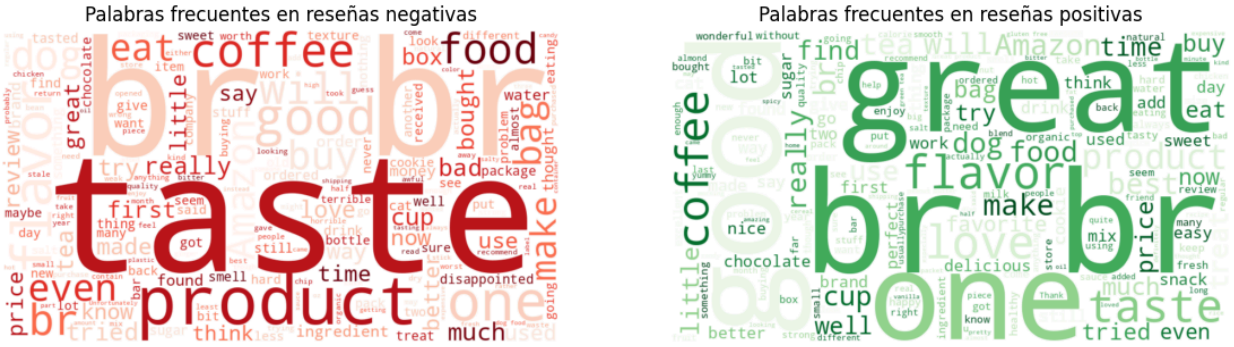


Imagen 4. Palabras frecuentes por distribución binaria de reseñas.

Limpieza.

Para el análisis se conservaron únicamente las columnas Score, Summary y Text, eliminando los valores nulos y las reseñas con puntuación igual a 3, consideradas neutras.

Se eliminaron las stopwords guardándolo en un DataFrame limpio. Se generó una nueva etiqueta binaria (label) donde 1 representa una reseña positiva (score ≥ 4) y 0 una negativa (score ≤ 2).

Gracias a esto obtenemos un promedio de 40.26 palabras por reseña, con un mínimo de 1 y un máximo de 1677 palabras.

SEMANA7.ipynb

El dataset final quedó con 525 789 reseñas, de las cuales el 84% son positivas y el 16% negativas, lo que refleja un ligero desbalance de clases. En promedio, las reseñas contienen 83 palabras, con una longitud máxima de más de 2,000.

Durante la limpieza se aplicaron transformaciones básicas: conversión a minúsculas, eliminación de símbolos, URLs y espacios extra, normalización de acentos y eliminación de stopwords (manteniendo negaciones como “not” o “never”). El texto limpio quedó almacenado en la columna text_clean, lista para la etapa de tokenización y modelado.

Modelo

Hiperparámetro	Modelo 1	Modelo 2
vocab_size	20,000	20,000
max_len	100	150
embed_dim	64	64
rnn-units	64	128
dropout	0.4	0.4
batch_size	64	64
epochs	5	5
learning_rate	1e-3	1e-3
Accuracy	0.9477	0.9452

Resultados

Como se observa en los resultados, la modificación del parámetro max_len que representa la cantidad máxima de tokens o palabras procesadas por el modelo y el aumento del número de unidades rnn_units (de 64 a 128 en comparación entre el modelo 1 y el modelo 2) no generaron una mejora significativa en la métrica de accuracy.

Lo más destacable fue el incremento considerable en el tiempo de entrenamiento, lo que indica que esta configuración es excesiva para el tamaño y la complejidad del dataset utilizado. En este caso, el aumento de max_len de 100 a 150 no resulta

beneficioso, dado que la longitud promedio de los textos, después del proceso de limpieza, es de aproximadamente 40.

Algo similar ocurre con el parámetro rnn_units, que controla la cantidad de neuronas internas (o “puertas”) de la RNN encargadas de retener u olvidar información. Si bien un mayor número de unidades incrementa la capacidad de memoria y representación del modelo, en un conjunto de datos relativamente corto, esto solo aumenta la complejidad computacional, reflejándose en un entrenamiento más lento y sin una mejora tangible en el rendimiento.

En conclusión, para este tipo de dataset, una configuración más ligera (por ejemplo, max_len = 100 y rnn_units = 64) resulta más eficiente, manteniendo un equilibrio adecuado entre desempeño y costo de entrenamiento.

Conclusión

El modelo logró un accuracy sobresaliente de 94.7%, demostrando que configuraciones ligeras de RNN pueden ser altamente efectivas para textos cortos y balanceados. El análisis pudo mostrar que aumentar los parámetros como la longitud máxima de secuencia o el número de unidades recurrentes no necesariamente mejora el rendimiento, sino que incrementa el costo computacional. En general, el pipeline de preprocesamiento implementado incluyendo limpieza, normalización y eliminación de stopwords permitió obtener representaciones limpias y consistentes. Los resultados confirman que un enfoque simple bien ajustado y eficiente puede ofrecer un excelente desempeño en tareas de análisis de sentimientos con datos de reseñas.

Referencias

McAuley, J., & Leskovec, J. (2013). *Amazon Fine Foods reviews* [Dataset]. Stanford Network Analysis Project.
https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews