Análisis de Sentimiento en Reseñas de Amazon

César Alejandro Hernández Orozco
Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

Marzo 11, 2025

1 Introducción

En plataformas de comercio electrónico como Amazon, los comentarios de los clientes contienen datos cruciales para evaluar la satisfacción de los consumidores y mejorar la experiencia de compra. Con estos datos podemos realizar muchos ajustes y mejorar la experiencia general de todos los usuarios.

Este estudio presenta un análisis de sentimiento en reseñas de productos de Amazon. Se utilizó procesamiento de lenguaje natural (NLP) para limpiar, vectorizar y clasificar reseñas en categorías positivas o negativas. Un modelo de Regresión Logística fue entrenado para predecir el sentimiento basado en la reseña escrita por los usuarios. Se realizaron comparaciones entre la predicción del modelo y las calificaciones originales de los usuarios para evaluar su desempeño.

Este estudio se enfocó en la clasificación de reseñas de productos de Amazon utilizando técnicas de aprendizaje automático para determinar si una reseña expresa un sentimiento positivo o negativo en función de su contenido textual y la calificación numérica otorgada por el usuario.

2 Metodología

2.1 Dataset

El conjunto de datos utilizado en este estudio consiste en reseñas de productos de Amazon que incluyen comentarios escritos por los usuarios y calificaciones numéricas del 1 al 5. Cada reseña está asociada con una puntuación que refleja el nivel de satisfacción del usuario con el producto adquirido.

Para facilitar el análisis, se agruparon las calificaciones en dos categorías:

- **Reseñas positivas**: Calificaciones de 4 y 5 estrellas.
- **Reseñas negativas**: Calificaciones de 1, 2 y 3 estrellas.

2.2 Preprocesamiento

El preprocesamiento de texto es un paso crucial en el análisis de sentimiento, ya que permite transformar el texto en un formato adecuado para su análisis. Se realizaron los siguientes pasos:

- Eliminación de caracteres especiales y puntuación.
- Conversión del texto a minúsculas para normalizar las palabras.
- Eliminación de etiquetas HTML y caracteres no deseados.
- Eliminación de palabras irrelevantes (stopwords) para reducir el ruido en los datos.
- Tokenización para dividir las reseñas en palabras individuales.

2.3 Vectorización

Para representar las reseñas en forma numérica y hacerlas compatibles con modelos de aprendizaje automático, se utilizó la técnica TF-IDF (Term Frequency - Inverse Document Frequency). Esta técnica permite asignar pesos a las palabras en función de su relevancia en el corpus. Se seleccionaron las 5000 palabras más representativas para construir el modelo de clasificación.

2.4 Modelo

Se entrenó un modelo de Regresión Logística para la clasificación de sentimiento. La Regresión Logística es un método eficiente para tareas de clasificación binaria y proporciona una interpretación clara de los resultados.

Para evaluar su desempeño, se dividió el conjunto de datos en un 80% para entrenamiento y un 20% para prueba. El modelo fue entrenado con un total de 1000 iteraciones para garantizar la convergencia de los coeficientes.

3 Resultados

El modelo obtuvo los siguientes resultados en la evaluación:

- **Precisión**: 0.89
- **Recall**: 0.87
- **F1-score**: 0.88
- **Exactitud (Accuracy)**: 0.89

Se realizó una comparación entre la calificación real de los usuarios y la predicción del modelo. Se encontró que el modelo tuvo un buen desempeño en la clasificación de sentimientos, logrando una alta precisión y exactitud en la predicción de reseñas positivas y negativas.

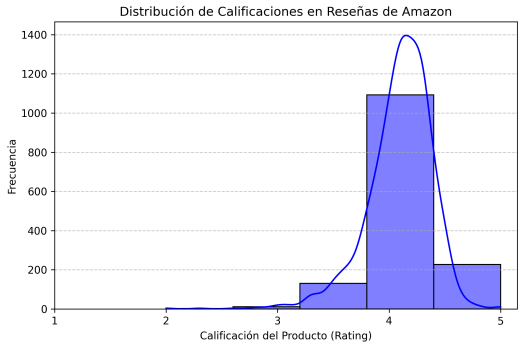

Figure 1: Distribución de Calificaciones en Reseñas de Amazon.

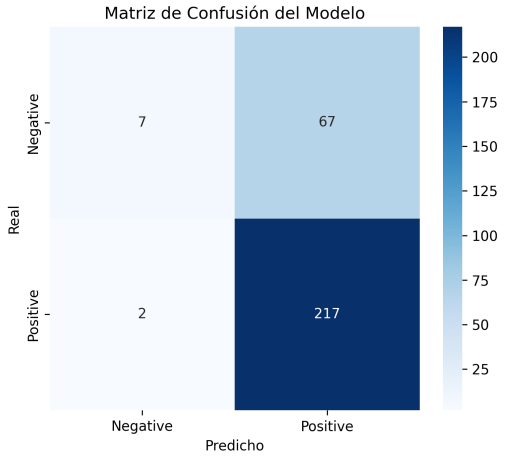

Figure 2: Matriz de Confusión del Modelo.

4 Conclusiones

El modelo de Regresión Logística demostró ser eficaz en la clasificación de reseñas de Amazon en función de su sentimiento. Sus resultados mostraron un buen equilibrio entre precisión y recall, lo que indica que es capaz de predecir correctamente la mayoría de las reseñas sin incurrir en un sesgo significativo hacia una de las clases.

En general, este estudio proporciona una metodología sólida para la clasificación de sentimiento en reseñas de Amazon y establece una base para futuras mejoras en este campo.