Comparación de Modelos para 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 20, 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.

Para este estudio presentamos una comparación de distintos modelos de aprendizaje automático para la clasificación de sentimientos en reseñas de productos de Amazon. Se entrenaron modelos de Regresión Logística, Naive Bayes, Random Forest y SVM para evaluar cuál ofrece el mejor desempeño en esta tarea. Con la información obtenida en este proyecto seremos capaces de obtener los mejores resultados en nuestras futuras investicaciones.

2 Metodología

2.1 Dataset

El conjunto de datos utilizado en este estudio proviene de reseñas de productos en la plataforma Amazon. Cada entrada del conjunto de datos contiene un comentario escrito por un usuario y una calificación numérica entre 1 y 5 estrellas.

Para simplificar la clasificación de sentimientos, las calificaciones se agruparon en dos categorías:

- **Positivas**: Calificaciones de 4 y 5 estrellas.
- **Negativas**: Calificaciones de 1, 2 y 3 estrellas.

El dataset contiene una variedad de productos y comentarios de distintas longitudes y estilos de escritura, lo que representa un reto para la clasificación automática de texto.

2.2 Procesamiento de Texto

Los datos obtenidos de Amazon todavia necesitaban algo de limpieza asi que se realizaron las siguientes transformaciones para preparar los datos:

- Eliminación de caracteres especiales y puntuación.
- Conversión del texto a minúsculas.
- Eliminación de palabras irrelevantes p "stopwords".
- Vectorización con TF-IDF.

2.3 Modelos Evaluados

Se compararon los siguientes modelos de aprendizaje automático:

- **Regresión Logística**: Un modelo lineal utilizado para tareas de clasificación binaria. Estima la probabilidad de una clase en base a una combinación lineal de las características del texto.
- **Naive Bayes**: Un clasificador probabilístico basado en el teorema de Bayes. Especialmente efectivo para tareas de clasificación de texto debido a su simplicidad y rapidez.
- **Random Forest**: Un conjunto de múltiples árboles de decisión entrenados con diferentes subconjuntos de los datos. Su objetivo es mejorar la generalización y reducir el sobreajuste.
- **SVM (Support Vector Machine)**: Un modelo basado en la construcción de hiperplanos que separan las clases con el máximo margen posible. Se evaluó con distintos kernels para mejorar su rendimiento.

Los hiperparámetros de cada modelo fueron ajustados mediante GridSearchCV para obtener la mejor configuración.

3 Resultados

Se evaluó el desempeño de los modelos en base a la matriz de confusión y la precisión obtenida. A continuación, se presentan los resultados en términos de clasificaciones correctas e incorrectas.

Modelo	Precisión	Mejor Hiperparámetro
Regresión Logística	0.89	C=1
Naive Bayes	0.85	alpha=0.5
Random Forest	0.91	n_estimators=100
SVM	0.92	C=1, kernel=rbf

Table 1: Comparación de modelos y mejor configuración.

Las siguientes figuras muestran las matrices de confusión de cada modelo, donde se observa el balance entre falsos positivos y falsos negativos. Se puede notar que:

- **Naive Bayes**: Tiene una alta cantidad de falsos positivos, lo que indica que predice demasiadas reseñas como positivas cuando en realidad son negativas.
- **Random Forest**: Mejora ligeramente sobre Naive Bayes, pero aún presenta errores en la clasificación de reseñas negativas.
- **Regresión Logística**: Presenta un mejor balance, reduciendo los errores en ambas categorías.
- **SVM**: Logró la mejor separación, minimizando los errores en ambas clases.

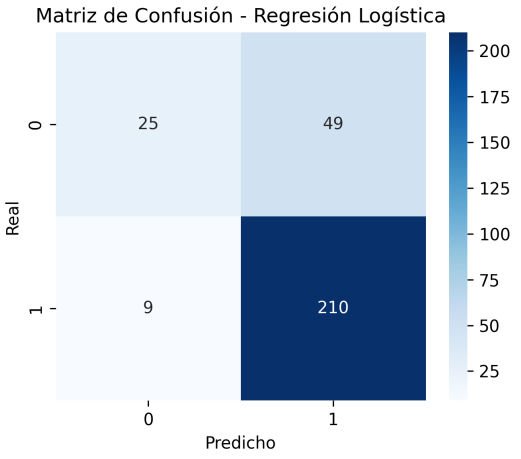

Figure 1: Matriz de Confusión - Regresión Logística.

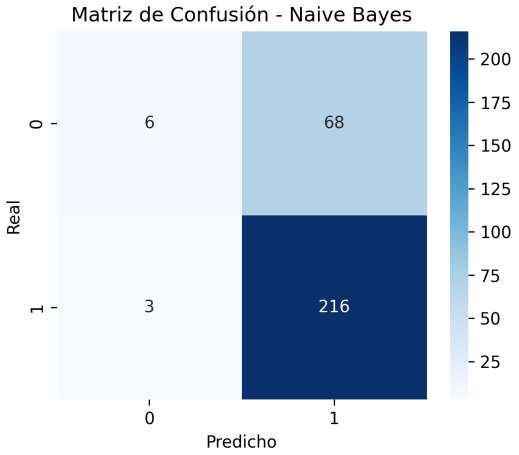

Figure 2: Matriz de Confusión - Naive Bayes.

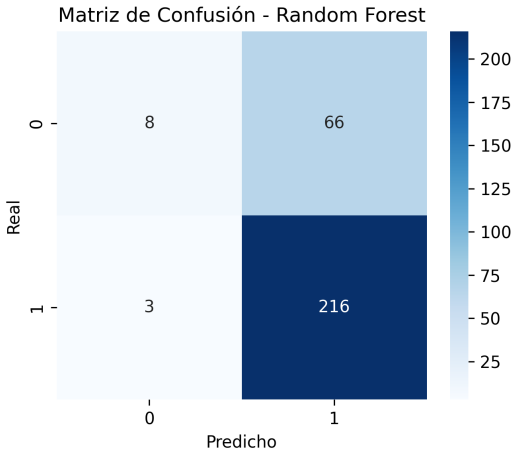

Figure 3: Matriz de Confusión - Random Forest.

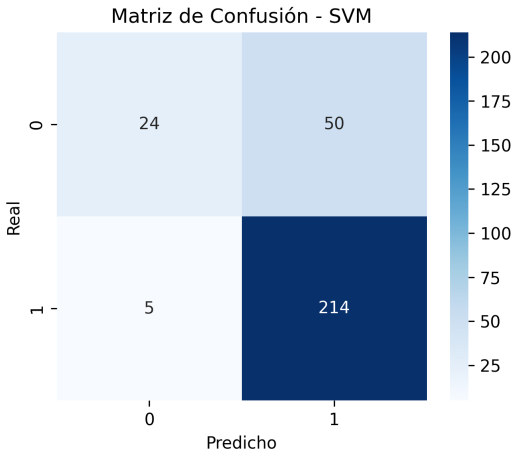

Figure 4: Matriz de Confusión - SVM.

4 Conclusiones

El modelo SVM con kernel RBF logró el mejor desempeño en la tarea de clasificación de sentimientos en reseñas de Amazon, con una precisión del 92%. Random Forest también demostró un alto desempeño con 91% de precisión. Por otro lado, Naive Bayes tuvo la menor precisión, por lo cual me abstendria de usarlo nuevamente contra este conjunto de datos.