Modelos de Clasificación con Reseñas de IMDB

Dora Alicia Guevara Villalpando
Matrícula: 1551003

Universidad Autónoma de Nuevo León)
Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

dora.guevaravll@uanl.edu.mx

***Resumen*—Este estudio presenta una comparación de diferentes modelos de clasificación para el análisis de sentimiento en reseñas de películas, se emplea el conjunto de datos IMDB Movie Reviews. Se aplicó un preprocesamiento exhaustivo, seguido de una vectorización con TF-IDF y la evaluación de cuatro modelos de aprendizaje supervisado: Regresión Logística, Support Vector Machines (SVM), Random Forest y Redes Neuronales. Los resultados muestran que la Regresión Logística obtuvo el mejor desempeño con una exactitud del 88.73 %.**

I. Introducción

El análisis de sentimiento es una tarea fundamental en el campo de la minería de texto y el aprendizaje automático. Este tipo de análisis permite identificar las emociones o actitudes expresadas en textos, como positivas, negativas o neutrales. En el contexto de las reseñas de usuarios, es especialmente útil para comprender la percepción del público hacia productos, servicios o contenidos.

En este proyecto, se empleó el conjunto de datos *IMDB Movie Reviews*, que contiene 50,000 reseñas de películas etiquetadas como positivas o negativas. Este dataset es ampliamente utilizado en la comunidad de aprendizaje automático debido a su balance entre clases y la complejidad del lenguaje natural empleado en las reseñas.

El objetivo principal fue desarrollar un modelo capaz de predecir el sentimiento de las reseñas con un alto grado de precisión. Para lograr esto, se exploraron técnicas clásicas de representación de texto y clasificación, evaluando tres enfoques: *Logistic Regression*, *Support Vector Machines (SVM)* y *Random Forest*. Este trabajo también buscó comparar el rendimiento de los modelos para recomendar el más adecuado según las características del dataset y las necesidades del análisis.

II. Descripción del Conjunto de Datos

El conjunto de datos utilizado en este estudio es el **Large Movie Review Dataset v1.0**, desarrollado por Andrew L. Maas et al. (2011) y publicado en la conferencia **ACL-HLT 2011**. Este dataset es ampliamente utilizado como referencia en tareas de clasificación de sentimientos y contiene reseñas de películas etiquetadas con polaridad positiva o negativa.

II-1. Características del dataset:

- Contiene 50,000 reseñas etiquetadas de manera balanceada en 25,000 positivas y 25,000 negativas. Se divide en:
  - 25,000 reseñas de entrenamiento (12,500 positivas y 12,500 negativas).
  - 25,000 reseñas de prueba (12,500 positivas y 12,500 negativas).
- Cada reseña proviene de una película diferente para evitar sesgos de correlación.

El dataset está estructurado en carpetas train/ y test/, con subdirectorios pos/ y neg/ que contienen los archivos de texto con las reseñas y sus respectivas etiquetas.

Sin embargo, a pesar que el conjunto de datos originalmente viene separado en *train* y *test* no se considera esto para realizar los análisis presentados a continuación; se opta por juntar ambos conjuntos y trabajar con una base completa con las 50,000 reseñas.

III. Metodología

III-A. Preprocesamiento

El preprocesamiento de los datos incluyó:

1. Limpieza del texto: Se eliminaron etiquetas HTML, caracteres no alfanuméricos y espacios extra.
2. Normalización: Las reseñas se convirtieron a minúsculas para garantizar consistencia.
3. Eliminación de *stop words*: Durante la vectorización, se excluyeron palabras comunes que no aportan información significativa al análisis (como *the*, *and*, entre otras).

III-B. Conjunto de datos

El conjunto de datos utilizado, IMDB Movie Reviews, contiene 50,000 reseñas divididas equitativamente en clases positivas y negativas. A pesar de que el dataset original separa los datos en entrenamiento y prueba, en este estudio se optó por combinarlos y realizar una nueva división en un 80 % para entrenamiento y 20 % para prueba.

III-C. Vectorización

Para transformar las reseñas en un formato numérico adecuado para los modelos de clasificación, se utilizó el método **TF-IDF (Term Frequency-Inverse Document Frequency)**. Esta técnica mide la importancia de una palabra dentro de un documento en relación con el corpus completo. Se seleccionó un máximo de 5000 características, lo que permitió capturar las palabras más relevantes sin sobrecargar el modelo con dimensionalidad excesiva.

III-D. Modelos Evaluados

- **Logistic Regression:** Este modelo lineal es eficiente y altamente interpretable. Es adecuado para tareas donde las características tienen una relación lineal con las etiquetas de salida.
- **Support Vector Machines (SVM):** Utilizando un kernel lineal, este modelo busca maximizar los márgenes entre clases en espacios de alta dimensión, como los generados por TF-IDF.
- **Random Forest:** Este enfoque basado en conjuntos utiliza múltiples árboles de decisión para capturar patrones no lineales y manejar datos complejos.
- **Redes Neuronales:** Modelo basado en capas densamente conectadas.

Cada modelo fue entrenado utilizando el conjunto de datos de entrenamiento y evaluado en el conjunto de validación. Las métricas clave, como precisión, recall, F1-score y exactitud, fueron utilizadas para comparar su desempeño.

IV. Resultados

Logistic Regression

- **Precisión**: 0.89 (clase negativa), 0.88 (clase positiva).
- **Recall**: 0.88 (clase negativa), 0.90 (clase positiva).
- **F1-Score**: 0.88 - 0.89.
- **Exactitud Global**: 0.88.

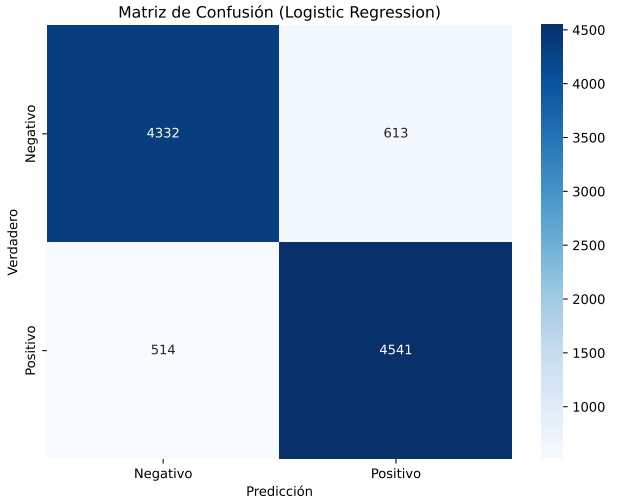

Figura 1. Matriz de confusion *logistic regression*.

Support Vector Machines (SVM)

- **Precisión**: 0.89 (clase negativa), 0.88 (clase positiva).
- **Recall**: 0.88 (clase negativa), 0.89 (clase positiva).
- **F1-Score**: 0.88 - 0.89.
- **Exactitud Global**: 0.88.

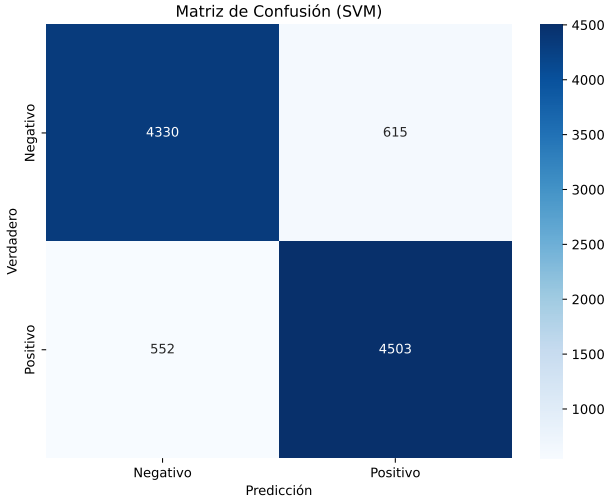

Figura 2. Matriz de confusion *SVM*.

Random Forest

- **Precisión**: 0.83 (clase negativa), 0.85 (clase positiva).
- **Recall**: 0.85 (clase negativa), 0.83 (clase positiva).
- **F1-Score**: 0.84.
- **Exactitud Global**: 0.84.

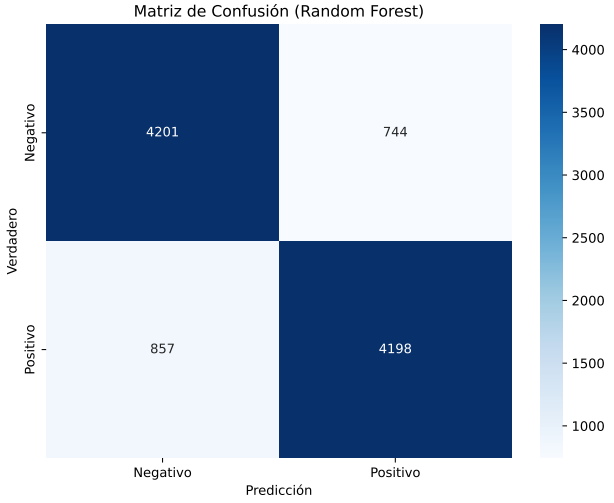

Figura 3. Matriz de confusion *Random Forest*.

Red Neuronal

- **Precisión**: 0.86 (clase negativa), 0.87 (clase positiva).
- **Recall**: 0.87 (clase negativa), 0.86 (clase positiva).
- **F1-Score**: 0.86.
- **Exactitud Global**: 0.86.

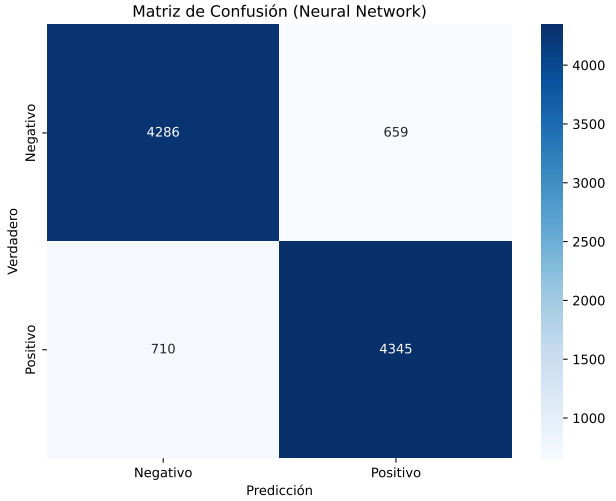

Figura 4. Matriz de confusion *Neural Network*.

V. Análisis comparativo

Acorde a lo que se observa en la tabla I se puede concluir lo siguiente:

- Se observa que la Regresión Logística obtuvo el mejor desempeño en términos de exactitud, seguido de SVM y Redes Neuronales. El modelo Random Forest tuvo el peor desempeño en esta tarea.

Modelo	Exactitud	F1 Score Promedio
Logistic Regression	88 %	0.88
SVM	88 %	0.88
Random Forest	84 %	0.84
Neural Network	86 %	0.86

Tabla I
Análisis comparativo

En la Figura 5 se muestra un gráfico comparativo entre la exactitud de los modelos.

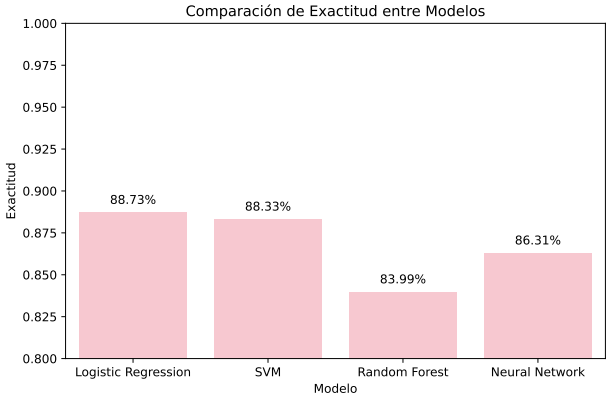

Figura 5. Comparación de modelos.

VI. Conclusiones

Este estudio demostró la efectividad de la Regresión Logística en la clasificación de sentimientos en reseñas de IMDB. Aunque SVM y Redes Neuronales también lograron desempeños competitivos, la simplicidad y eficiencia computacional de la Regresión Logística la hacen una opción viable para aplicaciones en análisis de sentimientos. En trabajos futuros, podría explorarse el uso de modelos de aprendizaje profundo para mejorar los resultados obtenidos.

Tomemos en cuenta los sigueintes puntos:

1. **Modelo Recomendado**:
   - **Logistic Regression** es la mejor opción si se prioriza simplicidad y velocidad.
   - **SVM** es una alternativa igualmente válida si se busca robustez en otras configuraciones.
2. **Modelo Descartado**:
   - **Random Forest**, aunque flexible, no fue competitivo en este escenario.
3. **Futuras Mejoras**:
   - Probar modelos basados en *embeddings* preentrenados como BERT para mejorar el desempeño.
   - Incrementar el conjunto de datos para explorar el impacto en modelos más complejos.

Referencias

- Dataset: IMDB Movie Reviews (https://ai.stanford.edu/~amaas/data/sentiment/)
- Métodos: Scikit-learn Library (TF-IDF, Logistic Regression, SVM, Random Forest)
- Análisis de texto (text mining) con Python. (https://cienciadedatos.net/documentos/py25-text-mining-python)