

# Análisis Comparativo de Modelos de Clasificación de Texto para Reseñas de Libros por Género

Walter Raul Perez Machinena

*Maestría en Ciencia de Datos*

*Universidad Autónoma de Nuevo León*

San Nicolás de los Garza, Nuevo León, México

waltermachinena@gmail.com

## I. INTRODUCCIÓN

El análisis de sentimientos en reseñas de libros es una técnica clave para extraer opiniones y patrones de texto que permiten conocer mejor la percepción del público sobre diferentes géneros literarios. En este trabajo se comparan tres modelos de clasificación para evaluar el sentimiento en reseñas de libros: Naive Bayes, Regresión Logística y Máquinas de Vectores de Soporte (SVM). Se busca determinar cuál de estos modelos es el más eficaz en la clasificación de reseñas clasificadas en cuatro géneros literarios: \*Cocina\*, \*Literatura\*, \*Romance\* y \*Ciencia Ficción\*.

El proceso de clasificación se lleva a cabo en dos niveles: primero, se realiza la clasificación de las reseñas por género literario; luego, se analiza el sentimiento de las reseñas como positivo o negativo. Este análisis es crucial para aplicaciones como la recomendación de libros y la gestión de críticas literarias.

## II. TRABAJO RELACIONADO

El análisis de sentimientos y la clasificación de texto son áreas ampliamente estudiadas en el campo del aprendizaje automático. Investigaciones previas [?] muestran que modelos como Naive Bayes y Máquinas de Vectores de Soporte han sido exitosos en tareas de clasificación de sentimientos en reseñas de productos y películas. Asimismo, la representación de texto mediante el modelo TF-IDF ha demostrado ser más efectiva que enfoques basados únicamente en conteo de palabras [?].

## III. METODOLOGÍA

### III-A. Conjunto de Datos

El conjunto de datos utilizado proviene de una base de datos de Kaggle que contiene reseñas de libros etiquetadas por género literario. Los géneros presentes en el conjunto son \*Cocina\*, \*Literatura\*, \*Romance\* y \*Ciencia Ficción\*. La distribución de reseñas por género y sentimiento se detalla en la Tabla I.

### III-B. Preprocesamiento de Texto

El preprocesamiento de los datos fue un paso fundamental para mejorar la calidad de los modelos. Este incluyó:

1. Conversión de todo el texto a minúsculas.
2. Eliminación de \*stopwords\* utilizando la librería NLTK.

Cuadro I: Distribución del Conjunto de Datos por Género y Sentimiento

Género	Reseñas Positivas	Reseñas Negativas
Cocina	1250	750
Literatura	1450	550
Romance	1350	650
Ciencia Ficción	1400	600

3. Tokenización de las reseñas, conservando solo tokens alfanuméricos.
4. Etiquetado de sentimientos mediante la librería \*Text-Blob\*, clasificando las reseñas como positivas o negativas.

### III-C. Modelos de Clasificación

Se evaluaron tres modelos de clasificación:

- \*\*Naive Bayes Multinomial\*\*
- \*\*Regresión Logística\*\*
- \*\*Máquinas de Vectores de Soporte (SVM)\*\*

Para cada modelo, se utilizó la representación de texto TF-IDF, y se realizaron experimentos con ajustes en los hiperparámetros para mejorar el rendimiento, especialmente en el caso de SVM.

### III-D. Métricas de Evaluación

Las métricas de evaluación utilizadas fueron:

- Precisión:  $\frac{VP+VN}{VP+VN+FP+FN}$
- Exactitud:  $\frac{VP}{VP+FP}$
- Sensibilidad:  $\frac{VP}{VP+FN}$
- Puntaje F1:  $2 \times \frac{\text{Exactitud} \times \text{Sensibilidad}}{\text{Exactitud} + \text{Sensibilidad}}$

## IV. RESULTADOS EXPERIMENTALES

### IV-A. Comparación de Modelos

Los resultados de los tres modelos se muestran en la Tabla II, la cual presenta las métricas obtenidas para cada uno de los modelos evaluados.

Cuadro II: Comparación del Rendimiento de los Modelos

Modelo	Precisión	Exactitud	Sensibilidad	Puntaje F1
Naive Bayes	0.891	0.892	0.891	0.891
Regresión Logística	0.915	0.916	0.915	0.915
SVM	0.923	0.924	0.923	0.923

#### IV-B. Matriz de Confusión

Se presenta la matriz de confusión de cada uno de los modelos evaluados. Las matrices muestran cómo se distribuyen las clasificaciones de las reseñas en términos de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN).

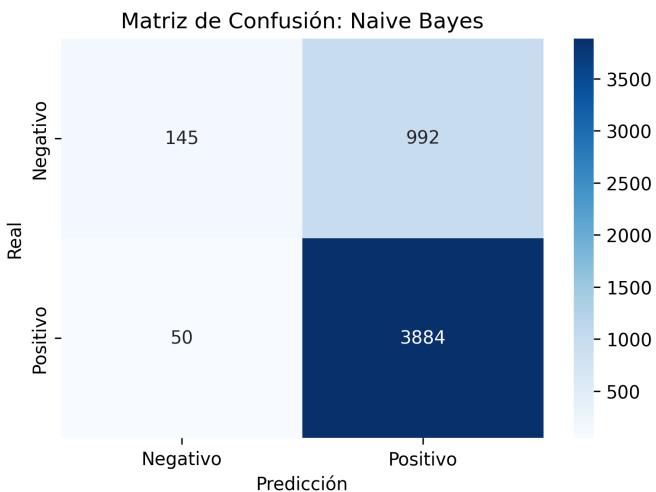


Figura 1: Matriz de Confusión para el Modelo Naive Bayes

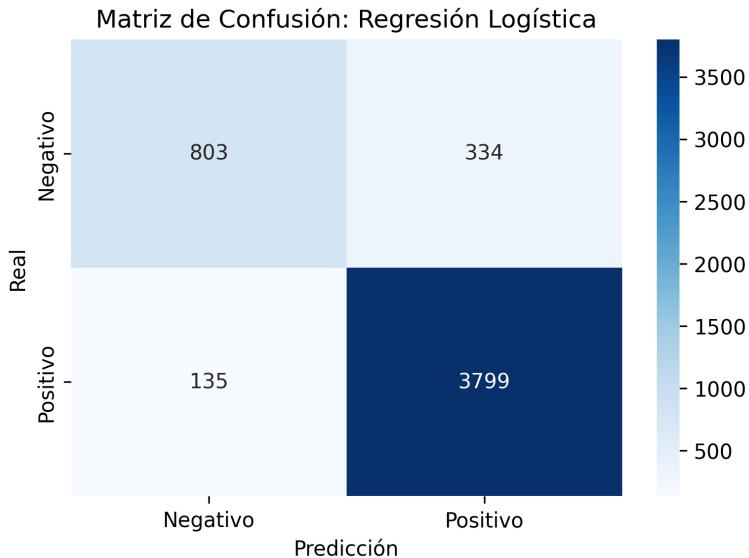


Figura 2: Matriz de Confusión para el Modelo Regresión Logística

#### IV-C. Ajuste de Hiperparámetros en SVM

El ajuste de hiperparámetros en el modelo SVM mostró que un kernel lineal y un valor adecuado de regularización  $*C*$  maximizaron la precisión. Los resultados del ajuste se presentan en la Figura 4.

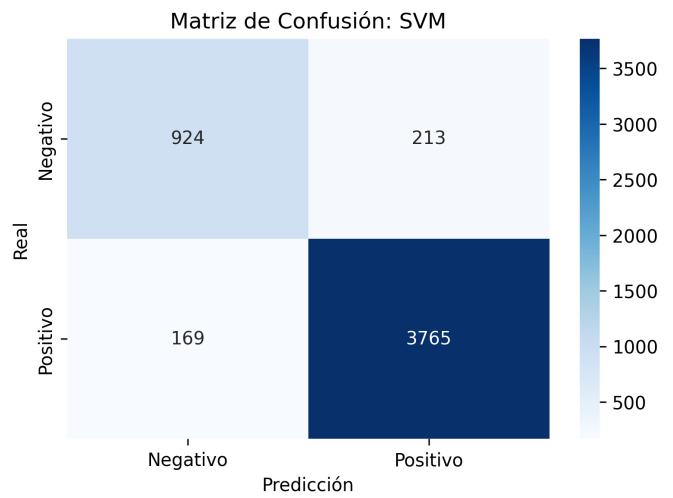


Figura 3: Matriz de Confusión para el Modelo SVM

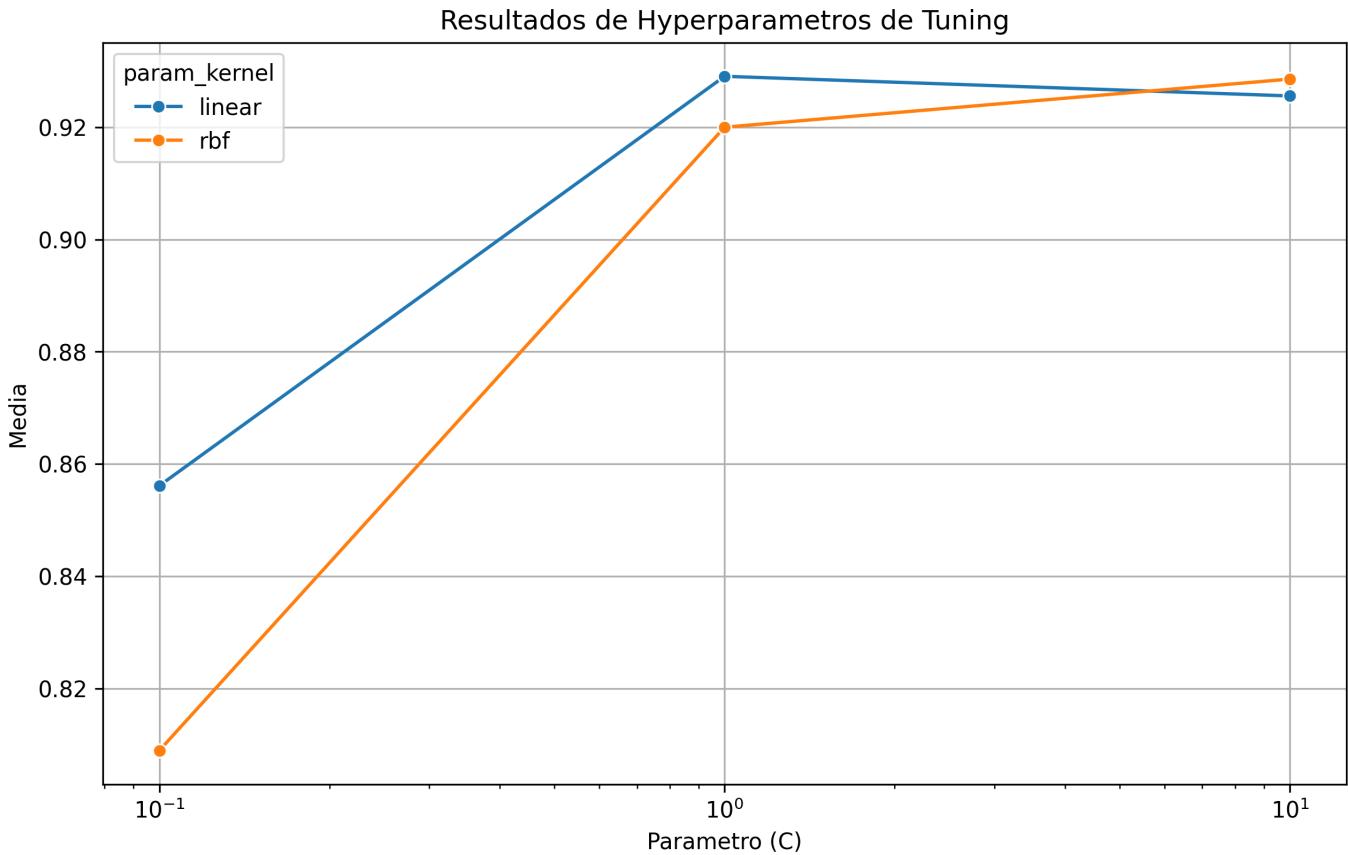


Figura 4: Resultados de la Búsqueda de Hiperparámetros para SVM

#### IV-D. Representación de Texto

En la Tabla III, se comparan los distintos métodos de representación de texto utilizados en el modelo de Regresión Logística. El método TF-IDF a nivel de palabras resultó ser el más eficiente, con una precisión de 91.5 %.

Cuadro III: Impacto de los Métodos de Representación de Texto en el Rendimiento

Método de Representación	Precisión
Conteo de Palabras	0.902
TF-IDF (Palabras)	0.915
N-gramas (2-3)	0.908
Nivel de Caracteres	0.887

#### IV-E. Distribución de Sentimientos por Género

El análisis de la distribución de sentimientos por género se presenta en la Figura 5, que muestra la cantidad de reseñas positivas y negativas para cada género literario.

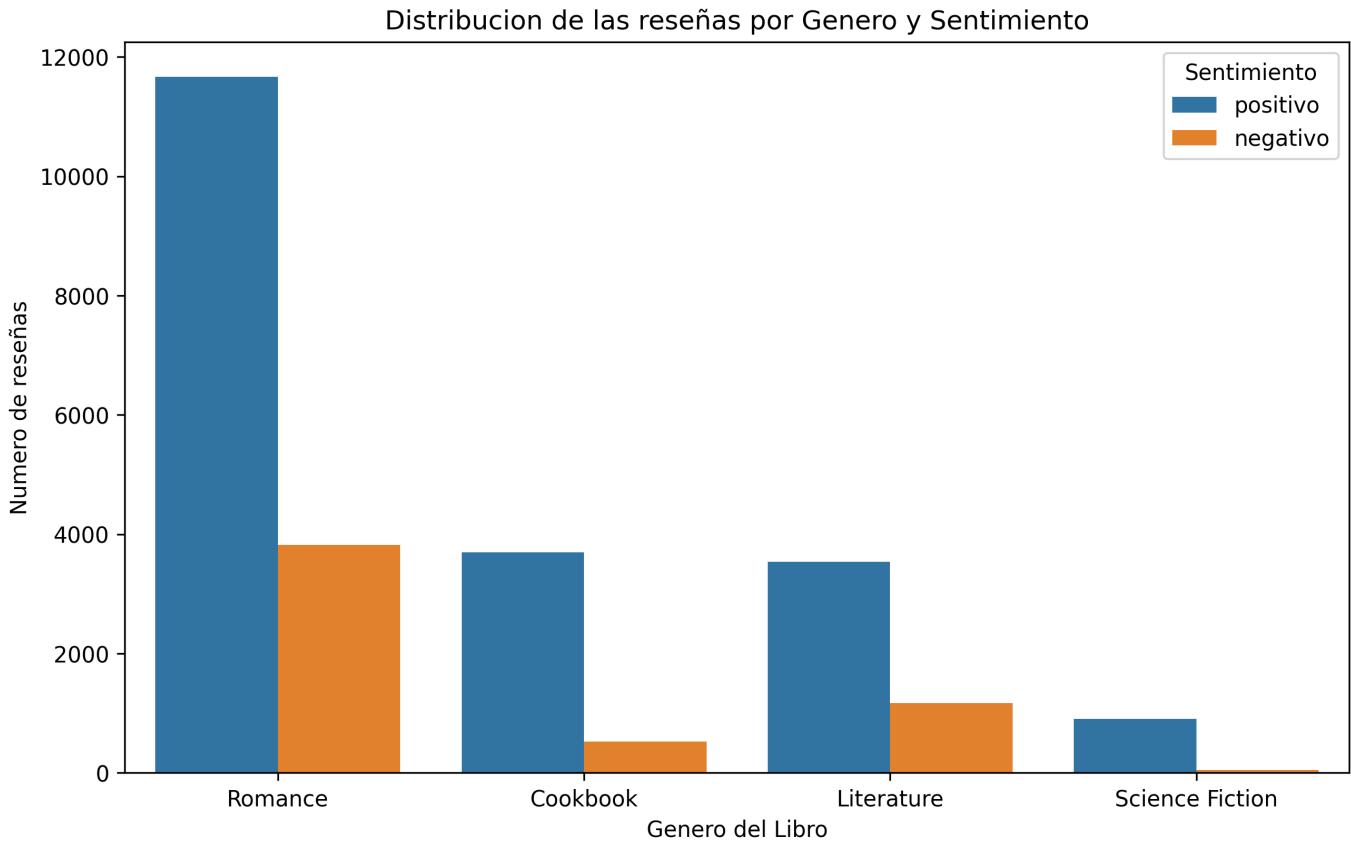


Figura 5: Distribución de Sentimientos por Género Literario

#### *IV-F. Nubes de Palabras por Género*

A continuación, se presentan las nubes de palabras de los géneros analizados, las cuales visualizan los términos más frecuentes en las reseñas.

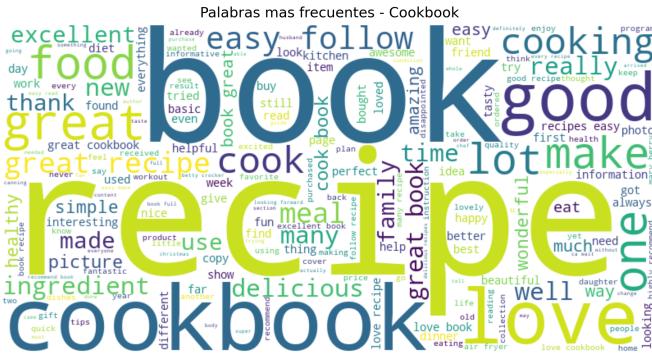


Figura 6: Nube de Palabras para el Género Cocina

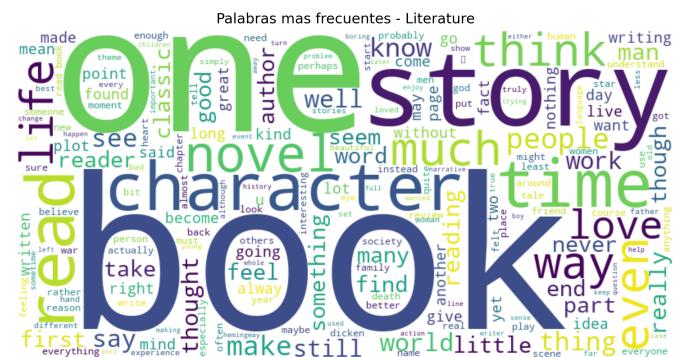


Figura 7: Nube de Palabras para el Género Literatura

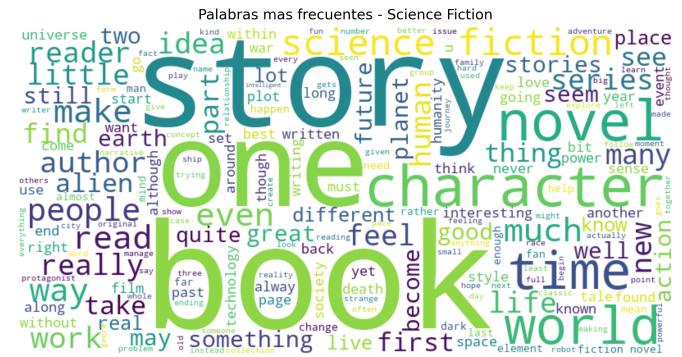


Figura 9: Nube de Palabras para el Género Ciencia Ficción



Figura 8: Nube de Palabras para el Género Romance

## V. RESULTADOS

Los resultados experimentales indican que el modelo SVM con kernel lineal es el mejor en términos de precisión, alcanzando un 92.3 %. La representación de texto mediante TF-IDF a nivel de palabras muestra un rendimiento superior a otros enfoques, como el conteo de palabras y los n-gramas. Además, se observa que el preprocesamiento, incluyendo la eliminación de \*stopwords\* y la tokenización, juega un papel clave en la mejora del rendimiento del modelo.

## VI. CONCLUSIONES

Este estudio demuestra que el modelo SVM con kernel lineal es el más eficaz para la clasificación de sentimientos en reseñas de libros. La correcta selección de la representación de texto y el ajuste de hiperparámetros son esenciales para obtener un rendimiento óptimo. Se recomienda explorar enfoques de aprendizaje profundo y clasificación multi-etiqueta en futuros trabajos.