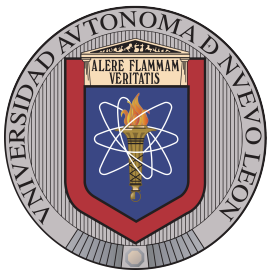

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MAESTRÍA EN CIENCIA DE DATOS

CLASIFICACIÓN DE TEXTOS SOBRE RESEÑAS DE NETFLIX

POR:

ACT. DAMIÁN ATILANO MARTÍNEZ ALVARADO

SAN NICOLÁS DE LOS GARZA, N.L. 2024

1. Introducción

Este estudio se centra en clasificar las reseñas de usuarios de *Netflix* como positivas o negativas. Para ello, se utilizaron dos modelos de aprendizaje automático populares: *Regresión Logística* y *Random Forest*. Se exploraron diferentes maneras de transformar el texto en datos que los modelos puedan entender, como *n-grams* y *TF-IDF*.

El objetivo es encontrar el mejor modelo que pueda clasificar las reseñas de manera precisa. Esto es importante porque una clasificación precisa puede ayudar a mejorar los sistemas de recomendación de películas y series, y a entender mejor las opiniones de los usuarios.

Este análisis también compara modelos con sus configuraciones predeterminadas y ajustadas para ver si ajustar los parámetros mejora el rendimiento. A través de este estudio, se busca proporcionar una guía sobre cómo elegir y ajustar modelos de clasificación de texto para obtener los mejores resultados.

2. Descripción de los datos

Se utilizó un conjunto de datos de reseñas de usuarios de *Netflix*, enfocándose en dos columnas principales:

1. content (Contenido): Texto de las reseñas de los usuarios.
2. score (Puntuación): Puntuación numérica de 1 a 5 asignada por los usuarios.

Para simplificar, la puntuación se convirtió en una variable binaria:

- Positivo: Puntuaciones de 4 o 5.
- Negativo: Puntuaciones de 1, 2 o 3.

3. Metodología

El estudio siguió una serie de pasos metodológicos detallados. Primero, se realizó un preprocesamiento del texto, que incluyó la limpieza de caracteres especiales, la conversión de los textos a minúsculas, la tokenización de las reseñas en palabras individuales, la eliminación de palabras comunes que no aportan significado (*stopwords*) y la lematización para reducir las palabras a su forma base.

A continuación, se aplicaron diversas técnicas de vectorización del texto. Se utilizaron conteos de *n-grams* para capturar patrones complejos en las reseñas, y se aplicó *TF-IDF* (*Term Frequency-Inverse Document Frequency*) para medir la relevancia de las palabras en el contexto del conjunto de datos. Además, se combinó la técnica de *n-grams* con *TF-IDF* para mejorar la capacidad del modelo de capturar relaciones contextuales.

Se utilizaron dos modelos de clasificación principales: *Regresión Logística* y *Random Forest*. La *Regresión Logística* se utilizó para la clasificación binaria de las reseñas, mientras que *Random Forest* se empleó para evaluar su rendimiento en comparación con la *Regresión Logística*.

También se ajustaron los hiperparámetros de los modelos para buscar obtener mejores resultados.

Finalmente, se evaluaron los modelos en términos de exactitud, matriz de confusión y reporte de clasificación, que incluía precisión, *recall* y *F1-Score*. Se compararon los modelos con y sin ajuste de hiperparámetros para determinar cuál ofrecía el mejor rendimiento en la clasificación de las reseñas de Netflix.

4. Resultados

4.1. Visualización de datos

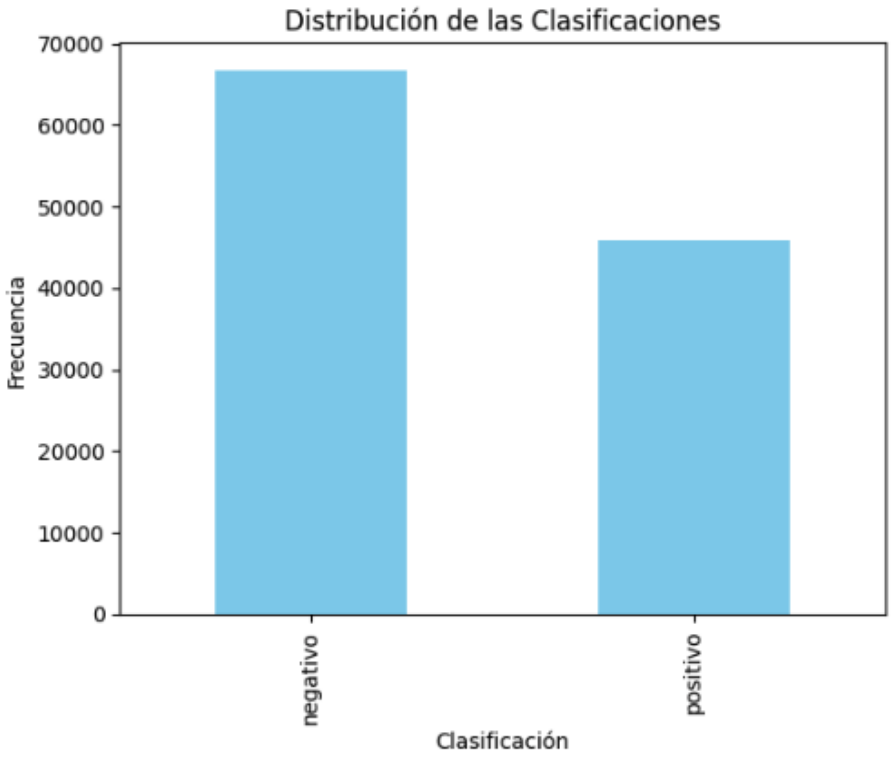

Figura 1: Distribución de las Clasificaciones

La primera gráfica muestra la distribución de las clasificaciones de las reseñas en dos categorías (negativo y positivo). Se observa que hay una mayor cantidad de reseñas clasificadas como negativas en comparación con las positivas. Específicamente, la frecuencia de reseñas negativas supera las 60,000, mientras que las positivas están cerca de las 40,000. Esta diferencia en las frecuencias sugiere que los usuarios tienden a dar más reseñas negativas que positivas.

4.2. Distribución de la Longitud de las Reseñas

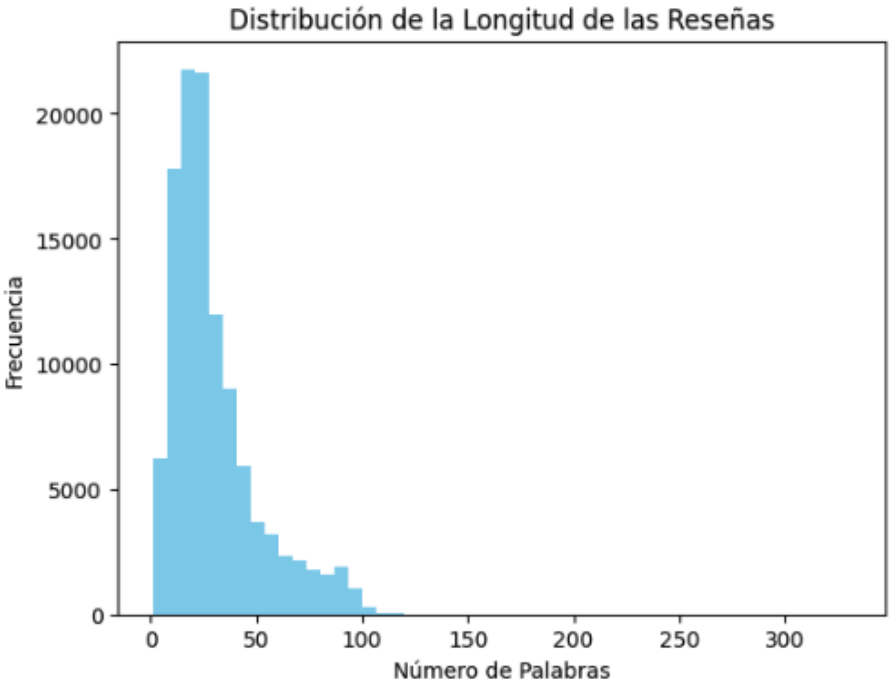

Figura 2: Distribución de la Longitud de las Reseñas

La segunda gráfica muestra la distribución de la longitud de las reseñas, medida en número de palabras. La mayoría de las reseñas tienen entre 0 y 50 palabras, con un pico significativo alrededor de las 10 a 20 palabras. La frecuencia disminuye gradualmente a medida que aumenta el número de palabras, con muy pocas reseñas que superan las 100 palabras. Esto indica que los usuarios tienden a escribir reseñas relativamente cortas, concentrándose principalmente en reseñas breves y concisas.

Estas visualizaciones ayudan a entender mejor la naturaleza de los datos de reseñas, destacando tanto la distribución de las clasificaciones como la variabilidad en la longitud de las reseñas.

4.3. Modelos de Clasificación y Técnicas de Vectorización utilizados

En este estudio, se utilizaron dos modelos de clasificación principales:

- *Regresión Logística*: Este modelo se utiliza para la clasificación binaria de las reseñas. Predice la probabilidad de que una reseña pertenezca a una clase (positiva o negativa) basada en una combinación lineal de las características de entrada.
- *Random Forest*: Este modelo consiste en múltiples árboles de decisión entrenados con diferentes subconjuntos del conjunto de datos. La predicción final se obtiene promediando las predicciones de todos los árboles. Es robusto contra el sobreajuste y maneja bien la complejidad de los datos.

Para transformar el texto en características numéricas que los modelos pudieran procesar, se aplicaron las siguientes técnicas de vectorización:

- *Count n-grams*: Esta técnica cuenta las apariciones de secuencias de palabras consecutivas (*n-grams*) en el texto, capturando patrones complejos que pueden mejorar la precisión del modelo.
- *TF-IDF* (*Term Frequency-Inverse Document Frequency*): Mide la relevancia de las palabras en el contexto del conjunto de datos, ponderando la frecuencia de términos en el documento frente a su frecuencia en el corpus completo. Esto ayuda a reducir la importancia de palabras muy comunes y a resaltar términos más distintivos.
- *TF-IDF n-grams*: Combina las técnicas de *n-grams* y *TF-IDF* para mejorar la capacidad del modelo de capturar relaciones contextuales y patrones complejos en el texto.

  Cada modelo se evaluó utilizando estas técnicas de vectorización para determinar su rendimiento en la clasificación de las reseñas de Netflix.

4.4. Resultados de los Modelos

Modelo	Vectorización	Exactitud
Logistic Regression	Count n-grams	0.823
Random Forest	Count n-grams	0.596
Logistic Regression	TF-IDF	0.864
Random Forest	TF-IDF	0.610
Logistic Regression	TF-IDF n-grams	0.814
Random Forest	TF-IDF n-grams	0.596

Cuadro 1: Exactitud de los Modelos sin Ajuste de Hiperparámetros

- La *Regresión Logística con TF-IDF* es el modelo con mejor rendimiento, alcanzando una exactitud del 86.4
- Los modelos *Random Forest* no lograron un rendimiento comparable, obteniendo una exactitud significativamente menor en todas las técnicas de vectorización, con un máximo de 61.0
- La inclusión de *n-grams* en la vectorización (*Count n-grams* y *TF-IDF n-grams*) no mejoró significativamente el rendimiento de los modelos en comparación con la simple técnica de *TF-IDF*.

4.4.1. Reporte de Clasificación

El reporte de clasificación proporciona tres métricas clave para evaluar el rendimiento del mejor modelo *Regresión Logística con TF-IDF*: precisión, recall y F1-Score.

- *Precisión*: Es el número de verdaderos positivos dividido por el número total de elementos que el modelo ha etiquetado como positivos. Indica qué tan preciso es el modelo cuando predice la clase positiva.
- *Recall*: Es el número de verdaderos positivos dividido por el número total de elementos que realmente son positivos. Indica qué tan bien el modelo captura todos los ejemplos positivos.
- *F1-Score*: Es la media armónica de la precisión y el recall. Proporciona un balance entre precisión y recall, especialmente útil cuando se necesita un equilibrio entre ambos.

- **Negativo**:
    - Precisión: 86.61 %
    - Recall: 91.35 %
    - F1-Score: 88.92 %
- **Positivo**:
    - Precisión: 86.13 %
    - Recall: 79.18 %
    - F1-Score: 82.51 %

Estas métricas indican que el modelo *Regresión Logística con TF-IDF* tiene un rendimiento superior en la clasificación de las reseñas de Netflix. El modelo muestra un buen equilibrio entre precisión y recall en ambas clases, positivas y negativas. La alta precisión en la clase negativa (86.61 %) significa que el modelo es muy efectivo para identificar correctamente las reseñas negativas, mientras que el alto recall (91.35 %) en la misma clase indica que el modelo captura la mayoría de las reseñas que son realmente negativas.

En la clase positiva, la precisión es del 86.13 %, lo que indica que la mayoría de las reseñas que el modelo predice como positivas son efectivamente positivas. El recall del 79.18 % en esta clase sugiere que el modelo también es competente en capturar las reseñas que son realmente positivas, aunque hay margen para mejorar en la identificación completa de todas las reseñas positivas.

El F1-Score, que es la media armónica de la precisión y el recall, es alto para ambas clases (88.92 % para negativas y 82.51 % para positivas), lo que confirma que el modelo mantiene un buen balance entre precisión y recall, siendo fuerte en sus predicciones.

4.4.2. Matriz de Confusión

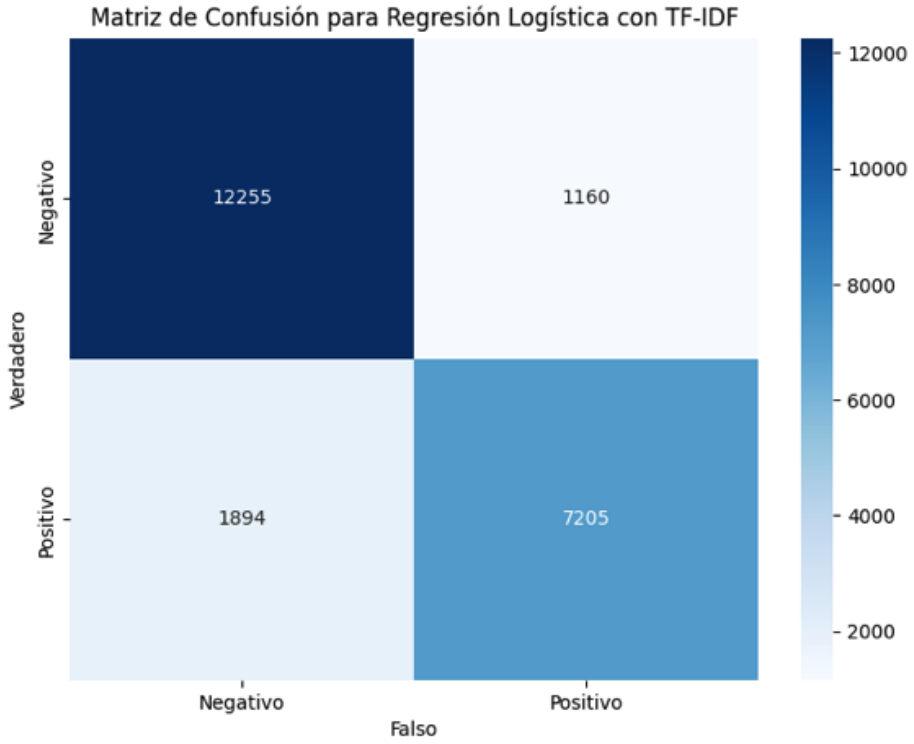

Figura 3: Matriz de Confusión para *Regresión Logística con TF-IDF*

- **Verdaderos Negativos (TN)**: La cantidad de reseñas que fueron correctamente clasificadas como negativas (12,255).
- **Falsos Positivos (FP)**: La cantidad de reseñas que fueron incorrectamente clasificadas como positivas, cuando en realidad eran negativas (1,160).
- **Falsos Negativos (FN)**: La cantidad de reseñas que fueron incorrectamente clasificadas como negativas, cuando en realidad eran positivas (1,894).
- **Verdaderos Positivos (TP)**: La cantidad de reseñas que fueron correctamente clasificadas como positivas (7,205).

En la matriz de confusión generada, los valores de TN, FP, FN y TP se pueden visualizar claramente, mostrando que el modelo es eficaz en minimizar los errores y en clasificar correctamente la mayoría de las reseñas. Una baja cantidad de FP y FN indica que el modelo tiene un buen rendimiento general en la clasificación de las reseñas de *Netflix*.

En este estudio se ajustaron los hiperparámetros de los modelos de *Regresión Logística* y *Random Forest* con el objetivo de mejorar su rendimiento en la clasificación de reseñas de *Netflix*. Los ajustes realizados incluyeron la modificación del parámetro `C` y el solver *liblinear* para la *Regresión Logística*, y la modificación del número de árboles, la profundidad máxima, el número mínimo de muestras para dividir un nodo y el número mínimo de muestras en un nodo hoja para el *Random Forest*.

A pesar de estos esfuerzos, los resultados mostraron que el modelo ajustado no logró superar el rendimiento del modelo sin ajuste. El modelo *Regresión Logística con TF-IDF* sin ajuste de hiperparámetros demostró ser el más efectivo, con una precisión ligeramente superior en ambas clases y un mejor balance general entre precisión y *recall*.

La exactitud del modelo sin ajuste fue del 86.4 %, mientras que la del modelo ajustado fue del 86.2 %. Aunque el ajuste de hiperparámetros mejoró ligeramente algunas métricas específicas, como el *recall* para la clase negativa, no produjo un aumento significativo en el rendimiento global del modelo.

Estos hallazgos subrayan la importancia de evaluar múltiples ajustes y técnicas al abordar problemas de clasificación de textos. La capacidad de un modelo para generalizar bien a datos no vistos no siempre mejora con el ajuste de hiperparámetros, y a veces el modelo original sin ajustes puede ofrecer un rendimiento más robusto y confiable.

5. Conclusiones

En conclusión, el modelo *Regresión Logística con TF-IDF* sin ajuste de hiperparámetros fue el más efectivo para la tarea de clasificación de reseñas de *Netflix* en este estudio. Este resultado destaca la relevancia de un enfoque cuidadoso y sistemático en la selección y evaluación de modelos y técnicas de preprocesamiento de datos en proyectos de *machine learning*.

6. Referencias

Ashish Kumar - Netflix Reviews - https://www.kaggle.com/datasets/ashishkumarak/netflix-reviews-playstore-daily-updated