Facultad de Ingeniería

Tópicos Avanzados de Analítica

CamilaAndrea Arias Vargas
Felipe Clavijo Acosta
Joel Alfredo Márquez Álvarez
Juan Camilo Ramírez Restrepo
Juan Pablo Cuellar Solano

Contenido

MOVIE GENRE CLASSIFICATION

1. Entendimiento del Negocio

En la industria cinematográfica y del entretenimiento, la correcta clasificación de películas por género es fundamental para el éxito de las plataformas de streaming y servicios que ofrecen contenido audiovisual. Esta clasificación no solo facilita a los usuarios la búsqueda y selección de películas que se ajustan a sus gustos y preferencias, sino que también es esencial para el desarrollo de algoritmos de recomendación más efectivos y personalizados. Con el auge de las plataformas de streaming y la producción masiva de contenido a nivel global, el volumen de películas disponibles ha crecido exponencialmente. Este incremento hace que la clasificación manual de cada película sea impracticable, lenta y propensa a errores, lo que puede llevar a clasificaciones inconsistentes y afectar negativamente la experiencia del usuario.

Además, una clasificación inexacta puede tener implicaciones significativas en las estrategias de marketing, ya que dificulta la segmentación adecuada del mercado y la promoción dirigida a audiencias específicas. En un entorno altamente competitivo, donde las preferencias de los usuarios cambian rápidamente, las empresas necesitan herramientas eficientes y necesarias para mantenerse relevantes y competitivas.

La automatización de la clasificación de géneros mediante técnicas avanzadas de análisis de datos y aprendizaje automático ofrece una solución efectiva a estos desafíos, que reduce significativamente la dependencia de la intervención humana, minimiza errores y garantiza una consistencia en la clasificación. Esto no solo mejora la organización interna y la eficiencia operativa de las plataformas, sino que también enriquece la experiencia del usuario al proporcionar recomendaciones más precisas y relevantes.

2. Objetivos de Negocio

Mejorar la satisfacción del usuario incrementando el Net Promoter Score (NPS) de la plataforma de streaming. Esto se logrará mediante la optimización de la clasificación de películas, facilitando su identificación y acceso, lo que proporcionará una experiencia de usuario más eficiente y agradable. Este enfoque se centra tanto en la mejora directa del NPS como en la creación de una experiencia intuitiva para los usuarios, lo que puede llevar a una mayor retención y fidelización.

Indicador:

- ***Descripción:*** *El NPS mide la disposición de los usuarios para recomendar la plataforma a otros.*
- ***Fórmula:*** *% de Promotores (usuarios que puntúan 9-10) - % de Detractores (usuarios que puntúan 0-6).*
- ***Frecuencia de medición:*** *Mensual o trimestral.*
- ***Meta:*** *Incremento de 10 puntos porcentuales en el NPS dentro de los próximos 6 meses.*

3. Objetivo de Minería

Diseñar un modelo de aprendizaje supervisado para la clasificación automática de los géneros de películas a partir de sus descripciones, garantizando un rendimiento mínimo con un AUC de 0.89. Esto asegurará la capacidad del modelo para distinguir correctamente los géneros correspondientes a cada película.

Indicador:

- ***Descripción:*** *El AUC mide la capacidad del modelo para distinguir entre clases (géneros), siendo 1.0 un modelo perfecto y 0.5 un modelo aleatorio.*
- ***Fórmula:*** *Calculada a partir de la curva ROC (tasa de verdaderos positivos frente a la tasa de falsos positivos).*
- ***Frecuencia de medición:*** *Tras cada iteración de validación o actualización del modelo.*
- ***Meta:*** *Mantener un AUC* ≥ 0.89.

4. Entendimiento de los Datos

Se realizó exploración inicial de la base de datos de entrenamiento, la cual contiene información sobre diferentes películas. Esta base de datos está compuesta por 7,895 registros y 5 atributos en total. A continuación, se presenta una descripción detallada de cada una de las variables que la componen:

- year: Variable numérica que representa el año de lanzamiento de la película.
- title: El nombre de la película. Este atributo es único para cada registro y facilita la identificación de cada película en el conjunto de datos.
- plot: Descripción de la trama de la película. Este es un campo textual que contiene información detallada sobre la historia de la película, y es la principal fuente de datos para la tarea de clasificación de géneros
- genres: Los diferentes géneros bajo los cuales está clasificada la película. Este atributo puede contener múltiples valores para una misma película, ya que una película puede pertenecer a más de un género. Será nuestra variable objetivo en el modelo de clasificación.
- rating: La puntuación que recibe cada película. Esta variable numérica indica la evaluación de la película por parte de los espectadores y críticos, y podría proporcionar información adicional sobre la calidad o popularidad de ciertos géneros.

Se exploraron las variables de la base de datos mediante gráficos, lo que permitió obtener un entendimiento más claro de su estructura y comportamiento. Este enfoque facilitó la identificación de información clave para el análisis y la clasificación de géneros.

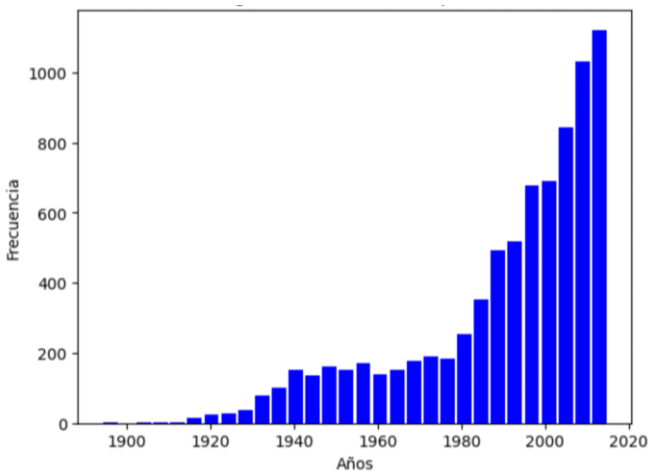

Imagen 1. Histograma de la variable año

La imagen 1 muestra un histograma de los diferentes años de lanzamiento de las películas, en él se muestra cómo ha crecido la cantidad de películas lanzadas a lo largo del tiempo. Desde el año 1890, se observa un crecimiento progresivo en la cantidad de películas producidas, con un crecimiento importante a partir de 1980. Este crecimiento se vuelve exponencial en las últimas décadas, alcanzando su pico en el 2015.

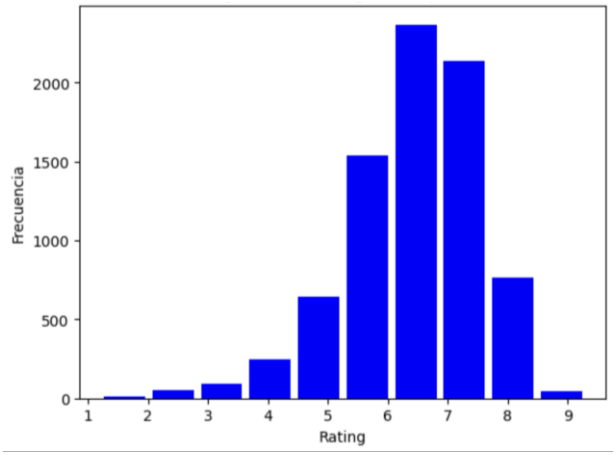

Imagen 2. Histograma de la variable rating

En la Imagen 2, se presenta la distribución de las puntuaciones de las películas, que oscilan en una escala de 1 a 9. La mayoría de las películas tienen una puntuación entre 6 y 7, lo que sugiere que la calificación promedio es moderadamente alta. Las películas con una calificación de 8 son menos frecuentes, mientras que las películas con puntuaciones por debajo de 5 son escasas. Esto indica que las películas con bajas calificaciones son menos comunes en la base de datos

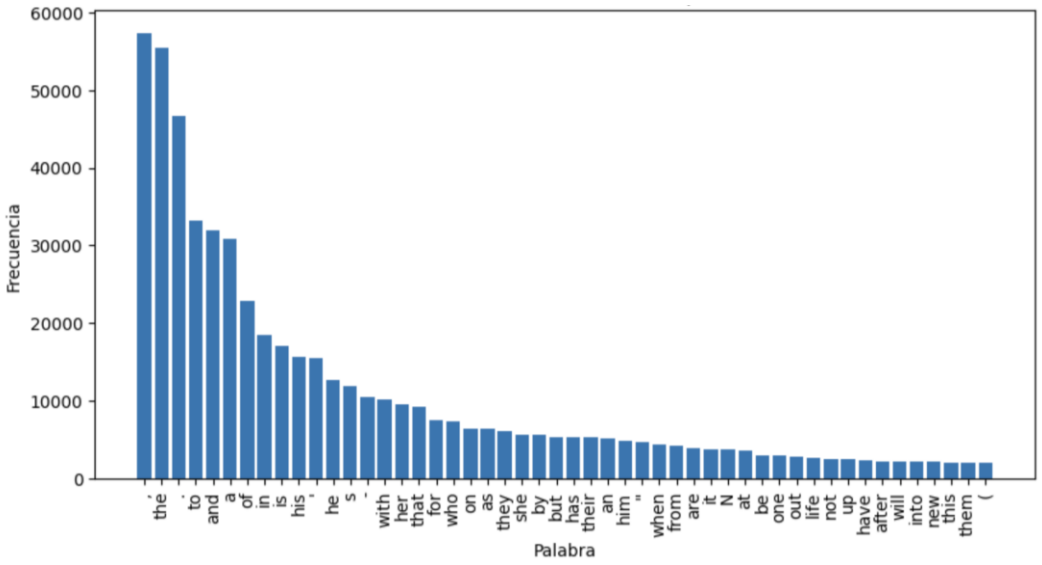

Imagen 3. Frecuencias de palabras

En la tercera imagen, se presenta un gráfico de frecuencias que muestra las palabras más comunes en las descripciones de las películas. Las palabras más frecuentes incluyen artículos, pronombres y preposiciones, tales como "the", "to", "and", "of", y "in", las cuales son conocidas como "stop words", suelen ser muy comunes, pero no aportan un significado relevante para el modelo. Este análisis es fundamental para depurar el texto, ya que eliminando las "stop words" nos podemos enfocar en términos específicos que realmente sean relevantes en las descripciones de las películas.

Además, este análisis de frecuencia de palabras puede servir como punto de partida para ajustar modelos de NLP, permitiendo detectar patrones de uso de palabras que podrían estar asociados a géneros específicos. Al eliminar estas palabras comunes y centrarnos en términos más relevantes, podemos mejorar la precisión de los modelos

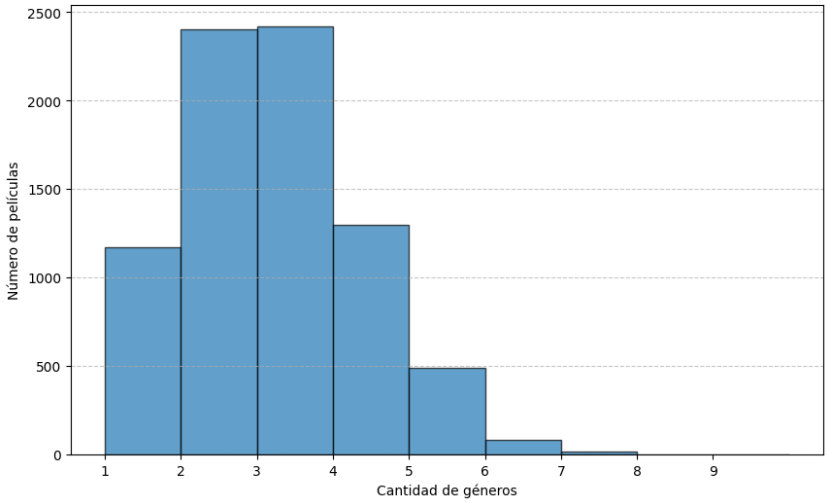

Imagen 4. Distribución de la cantidad de géneros

Finalmente, en la imagen 4, se visualiza la distribución de la cantidad de géneros asociados a cada

película. La mayoría de las películas están clasificadas entre 2 a 4 géneros. Las películas con un solo género son menos comunes, lo que sugiere que la mayoría de las películas se pueden encontrar calcificadas bajo géneros. Las películas con 5 o más géneros son muy raras y poco frecuentes dentro de la base de datos

5. Preparación de los Datos

Para la preparación de los datos, se inició con la carga de los conjuntos de datos de entrenamiento y prueba desde un repositorio en GitHub, utilizando archivos comprimidos en formato ZIP. Los datos se importaron un DataFrame llamados dataTraining.

El preprocesamiento del texto se centró en la columna plot, variable que contiene las descripciones de las películas. Se aplicaron las siguientes transformaciones para limpiar y preparar el texto:

- Se combino el titulo con la descripción de las películas
- Todo el texto se convirtió a minúsculas para evitar distinciones entre palabras iguales
- Se removieron signos de puntuación y caracteres no alfanuméricos para limpiar el texto.
- Se eliminaron los números para enfocarse únicamente en las palabras.
- Se eliminaron los espacios para el formato del texto.
- Se realizó tokenización para dividir el texto en palabras individuales y facilitar su análisis
- Se removieron palabras comunes “stopwords” que no aportan significado significativo al análisis
- Se realizo lematización para agrupar términos similares y reducir la dimensionalidad.

En la imagen 5 y 6 , se observa un gráfico de frecuencias de las palabras más comunes y una nube de las palabras más comunes después de eliminar las stopwords de las descripciones. Se aprecia un cambio notable en la distribución, destacándose palabras clave relevantes para las películas, como: "life", "one", "find", "get" y "love". Estas palabras proporcionan una mejor representación de los temas y patrones que predominan en el corpus de datos, permitiendo un análisis más preciso del contenido textual.

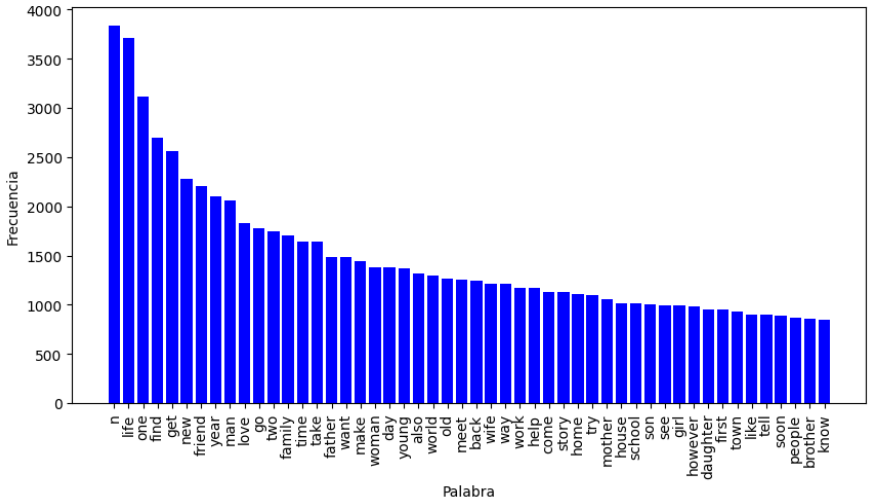

Imagen 5. Frecuencia de palabras

Imagen 6. Nube de palabras más frecuentes

Vectorización

A partir de los datos transformados, se aplicaron diferentes técnicas de vectorización para transformar el texto en representaciones numéricas:

- Bag of Words: Se utilizó la función CountVectorizer para crear una matriz de frecuencias de cada término.
- TF-IDF: Se empleó TfidfVectorizer para ponderar la importancia de las palabras en función de su frecuencia en el documento y en el corpus.
- Word2Vec: Se generaron vectores de palabras que capturan relaciones semánticas utilizando el modelo de gensim.
- Embeddings preentrenados: Integrando embeddings como GloVe para aprovechar representaciones vectoriales de palabras previamente entrenadas en grandes corpus.

- Etiquetas: Se procesó la columna genres para convertir la lista de géneros de cada película en una representación binaria. Se utilizó MultiLabelBinarizer para transformar los géneros en una matriz donde cada columna representa un género y los valores indican la presencia (1) o ausencia (0) de dicho género en cada película. cada una de las 24 labels es una generos

Finalmente, se realizó la partición de la base de datos de entrenamiento en conjuntos de train y test, con un 80% de los datos para entrenamiento y un 20% para prueba.

6. Modelos

Para abordar la clasificación de géneros de películas, se utilizaron distintos algoritmos de clasificación y técnicas de vectorización de texto. A continuación, se describe brevemente cada técnica empleada en los modelos mencionados:

Algoritmos de Clasificación:

Dado que la variable de respuesta presenta 24 clases diferentes (géneros de películas), se ha optado por implementar el enfoque ***One vs. Rest*** para cada uno de los modelos aplicados. Este enfoque consiste en entrenar un modelo binario para cada género, donde la clase de interés se considera como "positiva" y todas las demás como "negativas". A continuación, se detallan los modelos que se han implementado utilizando esta estrategia:

- Regresión Logística: Algoritmo de clasificación binaria que predice la probabilidad de que una entrada pertenezca a una clase específica, utilizando la función sigmoide para generar resultados entre 0 y 1.

- Random Forest: Algoritmo de ensamble basado en múltiples árboles de decisión que clasifica los datos mediante una votación mayoritaria entre los árboles.

- XGBoost: Algoritmo de boosting que optimiza secuencialmente los árboles de decisión, corrigiendo errores anteriores.

Estos algoritmos y técnicas se combinaron en los modelos implementados, como se muestra a continuación:

Sin Procesamiento de Datos			
Modelo	**Vectorización**	**Algoritmo**	**AUC**
Modelo 1	Bag of Words	Regresión Logística Random Forest XGBoost	0.82335 0.79688 0.82430
Modelo 2	TF-IDF	Regresión Logística Random Forest XGBoost	0.86463 0.80228 0.81132
		Regresión Logística	0.86463

Modelo 3	Word2Vec	Random Forest XGBoost	0.80228 0.81132
Modelo 4	GloVe	Regresión Logística Random Forest XGBoost	0.61247 0.55887 0.60012

Con Procesamiento de Datos			
Modelo	**Vectorización**	**Algoritmo**	**AUC**
Modelo 1	Bag of Words	Regresión Logística Random Forest XGBoost	0.85378 0.82568 0.85159
Modelo 2	TF-IDF	Regresión Logística Random Forest XGBoost	0.89590 0.81564 0.84634
Modelo 3	Word2Vec	Regresión Logística Random Forest XGBoost	0.61823 0.56792 0.59897
Modelo 4	GloVe	Regresión Logística Random Forest XGBoost	0.62198 0.57246 0.59473

7. Evaluación y Selección

La evaluación de los modelos y la selección del mejor se llevó a cabo utilizando la métrica AUC, que es adecuada para problemas de clasificación binaria y, con las adaptaciones necesarias, también aplicable a problemas de "clasificación multiclase o multietiqueta". Los resultados mostraron que el modelo de "regresión logística" obtuvo el mejor rendimiento, alcanzando un AUC de 0.89590, superando a otros modelos evaluados, como Random Forest y XGBoost.

TOP 3 MEJORES MODELOS	
Modelo	**Métrica**
Regresión Logística (TF-IDF)	**0.89590**
Regresión Logística (BoW)	0.85378
XGBoost (BoW)	0.85159

En este contexto, la predicción de géneros de películas se aborda como un problema de "clasificación multietiqueta", dado que una película puede pertenecer a varios géneros al mismo tiempo. El alto desempeño del modelo de regresión logística en esta tarea tiene importantes repercusiones para el negocio, pues una predicción precisa de los géneros permite a las plataformas de streaming, como Netflix, Prime Video, DGO, Paramount, Max, entre otras, así como a páginas web y blogs de reseñas, organizar y categorizar el contenido de manera más eficiente. Esto mejora notablemente la experiencia del usuario, facilitando la búsqueda y permitiendo una recomendación más personalizada.

Por otro lado, la correcta clasificación multietiqueta de los géneros es clave para las estrategias de mercadeo y publicidad de las productoras cinematográficas, ya que al identificar con precisión los géneros predominantes y sus combinaciones en una película, las productoras pueden orientar sus campañas promocionales hacia las audiencias más relevantes, aumentando así las probabilidades de éxito comercial. Esto contribuye a optimizar los recursos maximizando el retorno sobre la inversión.

8. Conclusiones

- La implementación de un modelo de regresión logística utilizando TF-IDF obtuvo un AUC de 0.89590, lo que refleja una alta precisión en la predicción de géneros de películas. Esta clasificación más precisa facilita la organización del catálogo, mejorando la accesibilidad y personalización para los usuarios. Al ofrecer una experiencia de búsqueda más intuitiva y eficiente, se espera que esto contribuya al incremento del Net Promoter Score (NPS) y a una mayor fidelización de los usuarios.

- El modelo seleccionado superó el rendimiento mínimo de AUC 0.89 fijado como objetivo, demostrando su eficacia para clasificar géneros de manera precisa. Este resultado asegura que la plataforma si implementa este modelo cuenta con un sistema confiable para la asignación automática de géneros, optimizando así la navegación y recomendaciones para los usuarios.

- Aunque el rendimiento del modelo es satisfactorio, existen oportunidades para seguir mejorando. La exploración de modelos más avanzados, como los transformers, podría ayudar a capturar patrones más complejos en los datos textuales, elevando aún más la precisión de las predicciones. Estas mejoras no solo aumentarían la calidad del servicio, sino que también podrían fortalecer las estrategias de recomendación y personalización, generando un mayor impacto en la satisfacción y retención de los usuarios.