Proyecto NPL– Tópicos Avanzados en Analítica

Predicción Género de Película

Cesar Enrique Acosta Sequeda [a,c], Roger Fernando Palomeque Lopez [a,c] , Michel Felipe Pedraza Cardenas[a,c], Juan Pablo Prada Barrios [a,c], Carlos Andrés Rodríguez Rodríguez [a,c]

Sergio Alberto Mora Pardo [b,c]

[a]*Estudiante de Maestría en Analítica para la Inteligencia de Negocios*
[b]*Profesor, Facultad de Ingeniería*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

1. Entendimiento del Negocio.

El procesamiento del lenguaje natural (NLP) es un campo de la inteligencia artificial que se enfoca en el análisis, comprensión y generación del lenguaje humano por parte de las computadoras. Esta tecnología ha encontrado numerosas aplicaciones en diversos sectores, incluyendo el entretenimiento y la industria cinematográfica. Una de las tareas clave del NLP en este contexto es la asignación automática de géneros a las películas basada en su sinopsis, resumen o descripción.

La correcta asignación de géneros a las películas es crucial para diversos aspectos de la industria cinematográfica. En primer lugar, permite una mejor organización y categorización de las películas en plataformas de streaming, sitios web de reseñas y bases de datos de películas. Esto facilita la búsqueda y recomendación de películas para los espectadores según sus preferencias de género. Además, la información de género es fundamental para las estrategias de marketing y publicidad, ya que permite dirigir las campañas promocionales a las audiencias objetivo-adecuadas.

Varias empresas líderes en el sector del entretenimiento y la tecnología han implementado soluciones de NLP para la asignación automática de géneros a las películas. Por ejemplo, Netflix, una de las principales plataformas de streaming, utiliza modelos de aprendizaje automático para analizar las sinopsis y asignar géneros a las películas y series de su catálogo. Esto mejora la precisión de sus recomendaciones personalizadas y la experiencia general del usuario.

Además de las plataformas de streaming, los sitios web de reseñas de películas, como IMDb y Rotten Tomatoes, también se benefician de la asignación automática de géneros. Al clasificar correctamente las películas por género, estos sitios pueden proporcionar una mejor navegación y filtrado de contenido para los usuarios. Esto facilita la búsqueda de películas específicas y la exploración de nuevas opciones dentro de los géneros preferidos.

La asignación precisa de géneros también es valiosa para los estudios de cine y productoras. Al analizar los patrones de éxito de ciertos géneros, pueden tomar decisiones más informadas sobre qué tipo de películas producir y cómo comercializarlas. Además, la información de género puede ayudar a predecir el potencial de taquilla y el rendimiento en el mercado de una película antes de su lanzamiento.

En general, el mercado para soluciones de NLP en la asignación de géneros de películas está en crecimiento. A medida que la cantidad de contenido cinematográfico disponible aumenta constantemente, la necesidad de una categorización precisa y eficiente se vuelve cada vez más importante. Las empresas que puedan desarrollar e implementar modelos de aprendizaje automático precisos y escalables para esta tarea tendrán una ventaja competitiva en el mercado.

2. Entendimiento de los Datos.

El conjunto de datos de entrenamiento contiene 7.895 entradas de películas y el de prueba del modelo 3383 entradas de películas cada una con información como el título, el rating, el año, la sinopsis o descripción de la película y los géneros asociados. Es importante destacar que una película puede pertenecer a más de un género, lo que agrega complejidad al problema de clasificación. En total, el conjunto de datos abarca 24 géneros de películas.

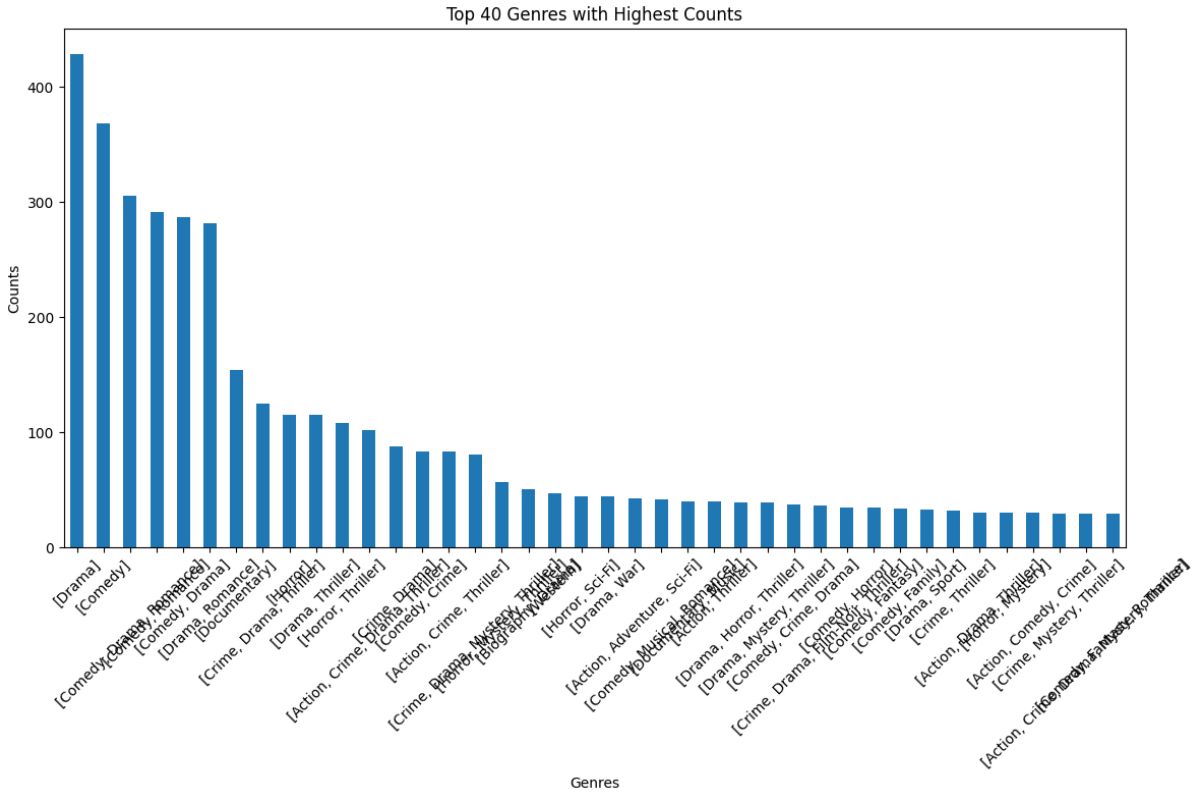

Figura 1. Top 40 Géneros de Películas. Fuente: Elaboración Propia.

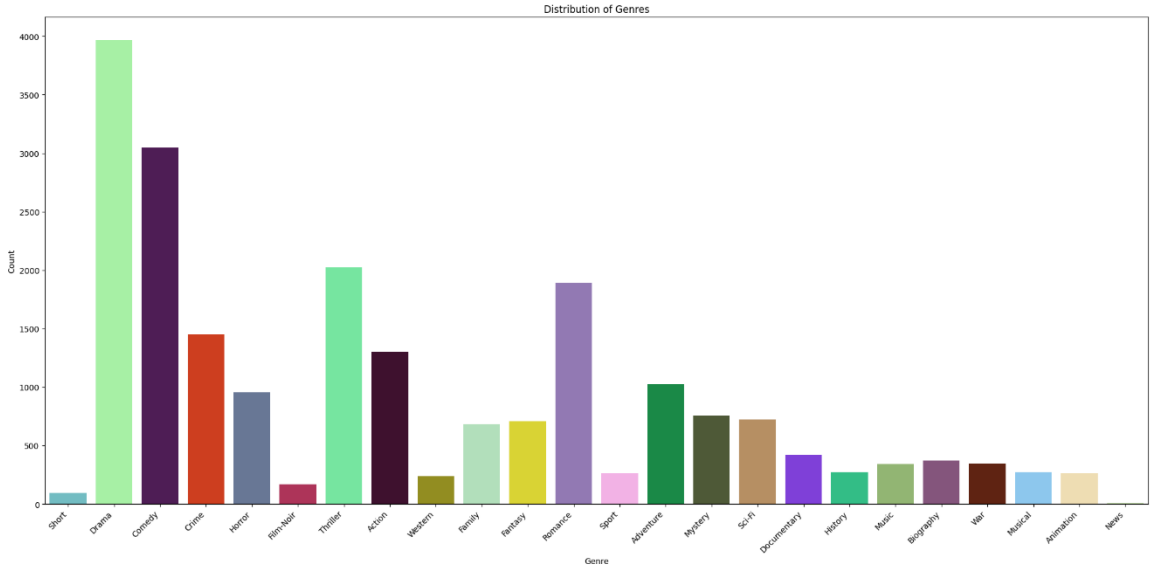

Figura 2. Géneros únicos de Películas. Fuente: Elaboración Propia.

Una característica clave del conjunto de datos es la distribución desigual de los diferentes géneros. Algunos géneros tienen una representación significativamente mayor que otros presentando así un desbalance de clases los que más documentación tienen entre 429 películas del género drama y 368 películas del género con comedy, mientras que hay películas con único género, principalmente por ser la combinación de varios esto debido a que el número máximo de géneros asignados a una sola película es de 9. Sin embargo. La mayoría de las películas tienen entre 1 y 3 géneros asociados. Esto puede sesgar los modelos de aprendizaje automático hacia las clases más representadas, lo que puede afectar el rendimiento en la predicción de géneros menos frecuentes.

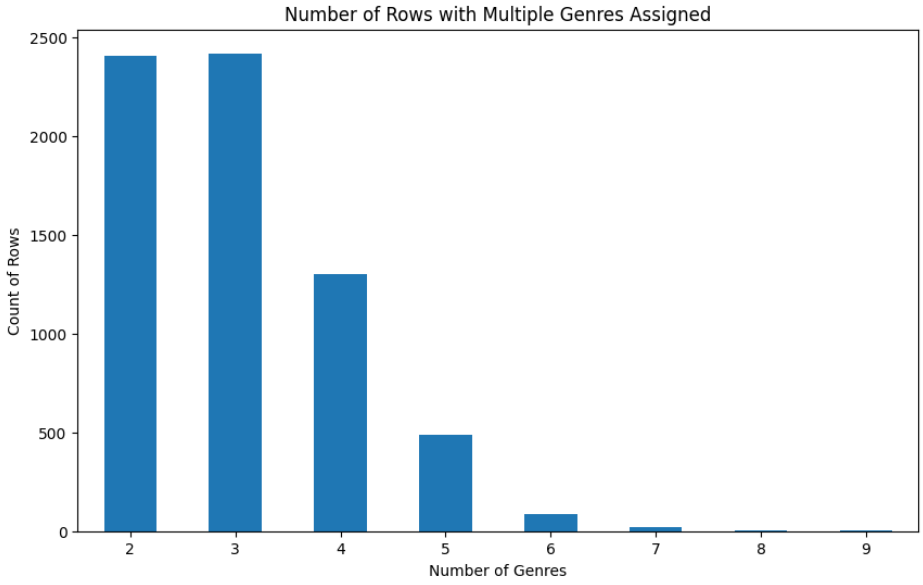

Figura 3. Número de Filas con Múltiples Géneros Asignados. Fuente: Elaboración Propia.

Además de los datos textuales, el conjunto de datos también incluye información numérica, como la cantidad de géneros asignados a cada película. Estos atributos numéricos pueden ser útiles para complementar los datos de texto y mejorar el rendimiento de los modelos de clasificación.

En general, el conjunto de datos presenta desafíos interesantes, como el desbalance de clases, la asignación múltiple de géneros y la necesidad de un preprocesamiento adecuado de los datos de texto. Sin embargo, también ofrece oportunidades para explorar diferentes técnicas de aprendizaje automático y evaluar su eficacia en la predicción de géneros de películas basada en sinopsis textuales.

3. Preparación de los Datos.

En este caso, las sinopsis de las películas se proporcionan como texto sin procesar, El primer paso consiste en tokenizar el texto (dividir las oraciones en palabras individuales). la tokenización del texto se realizó utilizando la clase CountVectorizer de la biblioteca scikit-learn. Esta clase convierte una colección de documentos de texto en una matriz de tokens, donde cada fila representa un documento y cada columna representa un token (palabra). En este caso, se especificó un parámetro max_features=1000 para limitar el número máximo de tokens a considerar.

Después de la tokenización, se aplicó la técnica TF-IDF (Frecuencia de Término-Inversa de Documento) para la limpieza y vectorización del texto. Esto se logró utilizando la clase TfidfVectorizer de scikit-learn. Se establecieron los siguientes parámetros: stop_words='english' para eliminar las palabras vacías en inglés, ngram_range=(1, 2) para considerar unigramas y bigramas, min_df=2 para eliminar términos que aparecen en menos de 2 documentos, y max_df=0.95 para eliminar términos que aparecen en más del 95% de los documentos.

En cuanto a la construcción de embeddings, se utilizaron los embeddings pre-entrenados de GloVe (Global Vectors for Word Representation). Luego, se creó una matriz de embeddings para cada sinopsis de película, donde cada palabra se representó mediante su vector de embeddings correspondiente.

Para la vectorización utilizando embeddings, se creó una función personalizada llamada create_embedding_matrix. Esta función toma como entrada una lista de sinopsis tokenizadas y los embeddings cargados, y devuelve una matriz de embeddings para cada sinopsis. La matriz se construye concatenando los vectores de embeddings de cada palabra en la sinopsis.

Además de las técnicas de vectorización mencionadas anteriormente, también se utilizó la técnica de Bag of Words mediante la clase CountVectorizer de scikit-learn. Esta técnica cuenta la frecuencia de aparición de cada palabra en un documento, sin considerar el orden de las palabras.

4. Modelamiento.

Se exploraron varios modelos de aprendizaje automático para abordar el problema de predicción de géneros de películas. Como línea base, se implementó un modelo de Random Forest utilizando la clase RandomForestClassifier de scikit-learn. Se estableció un parámetro depth_level=10 para limitar la profundidad máxima de los árboles, y se utilizó la vectorización de texto mediante CountVectorizer con max_features=1000. Obteniendo un ROC = 0.7812262183677007. se aplicó la técnica TF-IDF para la vectorización de texto con los parámetros (stop_words='english', ngram_range = (1, 2), min_df=2, max_df=0.95). Obteniendo un ROC = 0.8107723036823163

Además del modelo de Random Forest, se entrenó un modelo de regresión logística utilizando la clase LogisticRegression de scikit-learn. En este caso, se aplicó la técnica TF-IDF para la vectorización de texto, con los mismos parámetros utilizados previamente en el modelo de Random Forest (stop_words='english', ngram_range = (1, 2), min_df=2, max_df=0.95). Obteniendo un ROC = 0.875801684876896.

Se probaron otras técnicas de vectorización de texto, como la utilización de los embeddings pre-entrenados de GloVe y la técnica de Bag of Words (bolsa de palabras) mediante CountVectorizer. Sin embargo, según los resultados reportados, estas técnicas no lograron superar el rendimiento obtenido con TF-IDF en el modelo de regresión logística.

Además de los modelos lineales, se exploró un modelo de red neuronal recurrente utilizando una capa GRU (Gated Recurrent Unit). Para ello, se utilizaron los embeddings de GloVe como entrada a la red. La arquitectura de la red se determinó mediante una búsqueda de hiperparámetros utilizando Keras Tuner, lo que resultó en una capa GRU con 64 unidades, una tasa de abandono (dropout) de 0.4 y una tasa de abandono recurrente de 0.1. Obteniendo un ROC = 0.8461974367709132

También se entrenaron modelos de máquinas de vectores de soporte (SVM) y Naive Bayes utilizando las técnicas de vectorización de texto previamente mencionadas Obteniendo un ROC = 0.5270409587370896. Sin embargo, estos modelos no lograron superar el rendimiento de los modelos de regresión logística y GRU en términos de la métrica ROC.

5. Evaluación.

La evaluación de los modelos se realizó utilizando la métrica ROC (Receiver Operating Characteristic), que es una medida adecuada para problemas de clasificación multiclase. Los resultados mostraron que el modelo de regresión logística alcanzó el mejor rendimiento, con un ROC de 0.875801684876896, superando a los otros modelos explorados, como Random Forest, SVM, Naive Bayes y la red neuronal recurrente GRU.

Modelo	AUC Score
Train multi-class multi-label model	78%
Random Forest with TD-IDF	79%
One vs Rest Logistic Regression with TD-IDF	88%
One vs Rest Logistic Regression with Count Vectorizer	79%
One Vs Rest Logistic Regression with Glove embeddings	56%
One Vs Rest SVM with TD-IDF	27%
Mutlinomial Naive Bayes with TF-IDF	51%
XGBoost with TF-IDF	61%
GRU	83%

Tabla 1. Tabla comparativa de modelos.

El buen desempeño del modelo de regresión logística en esta tarea de predicción de géneros de películas tiene implicaciones relevantes para el negocio. Una predicción precisa de los géneros asociados a una película permite a las plataformas de streaming, sitios web de reseñas y bases de datos de películas organizar y categorizar su contenido de manera más efectiva, mejorando así la experiencia del usuario y facilitando la búsqueda y recomendación personalizada.

Además, la correcta asignación de géneros es fundamental para las estrategias de marketing y publicidad de las productoras cinematográficas. Al conocer los géneros predominantes en una película, pueden dirigir sus campañas promocionales a las audiencias objetivo-adecuadas, aumentando la probabilidad de éxito comercial. Esto les permitiría optimizar sus recursos de marketing y maximizar el retorno de inversión.

6. Conclusiones.

El proyecto de predicción de géneros de películas utilizando técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático demostró ser una tarea desafiante pero prometedora. Después de explorar diversos modelos, incluyendo Random Forest, regresión logística, redes neuronales recurrentes (GRU), máquinas de vectores de soporte (SVM) y Naive Bayes, se encontró que el modelo de regresión logística alcanzó el mejor rendimiento, con un ROC de 0.875801684876896. Este resultado destaca la capacidad de los modelos lineales para abordar problemas de clasificación multiclase, especialmente cuando se combinan con técnicas de vectorización de texto adecuadas, como TF-IDF.

A pesar del buen desempeño obtenido, es importante reconocer que aún hay margen para mejora. Una oportunidad interesante sería explorar técnicas de aprendizaje profundo más avanzadas, como redes neuronales convolucionales o transformers, que podrían capturar patrones más complejos en los datos de texto y mejorar la precisión de las predicciones. Además, incorporar información adicional, como reseñas de usuarios, metadatos de películas o datos de redes sociales, podría enriquecer el modelo y brindar una comprensión más profunda de los factores que influyen en los géneros de las películas.

En general, este proyecto demuestra el potencial de las soluciones de NLP y aprendizaje automático para la asignación automática de géneros a las películas. Los resultados obtenidos tienen implicaciones prácticas para diversas industrias, como plataformas de streaming, sitios web de reseñas y productoras cinematográficas. Al mejorar la organización y categorización del contenido, así como las estrategias de marketing y recomendación, estas soluciones pueden mejorar significativamente la experiencia del usuario y el éxito comercial de las películas. A medida que la cantidad de contenido cinematográfico continúa creciendo, la necesidad de soluciones precisas y escalables para la predicción de géneros se vuelve cada vez más crucial.

7. Referencias.

Zhu, H. (2023). Sentiment Analysis of Movies Based on Natural Language Processing. En *Atlantis Highlights in Computer Sciences* (pp. 1232-1240). https://doi.org/10.2991/978-94-6463-172-2_130

Alaji, Amjad. (2023). Investigate the Effect of Rotten Tomatoes and IMD b's Rating and Critic Review son Movies Publicity. 10.9756/INT-JECSE/V14I2.323.

Gachet, C. (2023, 13 febrero). Impacts of Netflix strategic AI implementation on its customer experience. *Medium*. https://medium.com/@colombagth/impacts-of-netflix-strategic-ai-implementation-on-its-customer-experience-45e7eef2d84f