Nicolás Casella, Uriel Arias

Predicción de géneros de películas/documentales basados en sus sinopsis y título

Resumen

Se busca predecir todos los géneros a los que pertenece una película considerando una lista predefinida de ellos y entrenando con diversas sinópsis escritas en inglés. Se trata entonces de una tarea de clasificación de clases que no son mutuamente excluyentes. De esta manera este clasificador podría ser útil para catalogar películas que estén por salir al público, así como también como un parte de un sistema de recomendación.

Se trabajará con un dataset que deberá ser preprocesado para eliminar campos, valores en null, filas repetidas y en el caso de las sinopsis eliminar stopwords y signos de puntuación que no aporten información.

El análisis que se realizará buscará encontrar correlaciones entre palabras que aparecen en la sinopsis de la película y los géneros a los que pertenece. A su vez, se intentará comparar los resultados obtenidos al analizar solo el título de la película o agregarlo a la sinopsis como parte de los datos de entrenamiento.

Datos

Los datos serán seleccionados del siguiente dataset obtenido de HuggingFace: https://huggingface.co/datasets/pkchwy/letterboxd-all-movie-data . En el mismo se cuenta con más de 800.000 filas sin filtrar por lo que luego de hacer una limpieza y división en datasets de *training*, *validation* y *test* se espera manejar una cantidad cercana a 100.000 películas en total, en donde se prioriza que exista una buena distribución para que el entrenamiento y la posterior evaluación.

Los datos del dataset elegido se brindan como archivos json con los siguientes campos: "url", "title", "year", "directors", "genres", "cast", "synopsis", "rating", "poster_url", "reviews": {"username", "review_text", "likes"}.Sin embargo, dada la tarea definida, se mantendrán sólo los campos de “title”, "synopsis" y "genres".

Por otro lado, se realizará una limpieza de los campos con la sinopsis y el título en donde se eliminarán signos de puntuación, stopwords y cualquier otro símbolo que no aporte información.

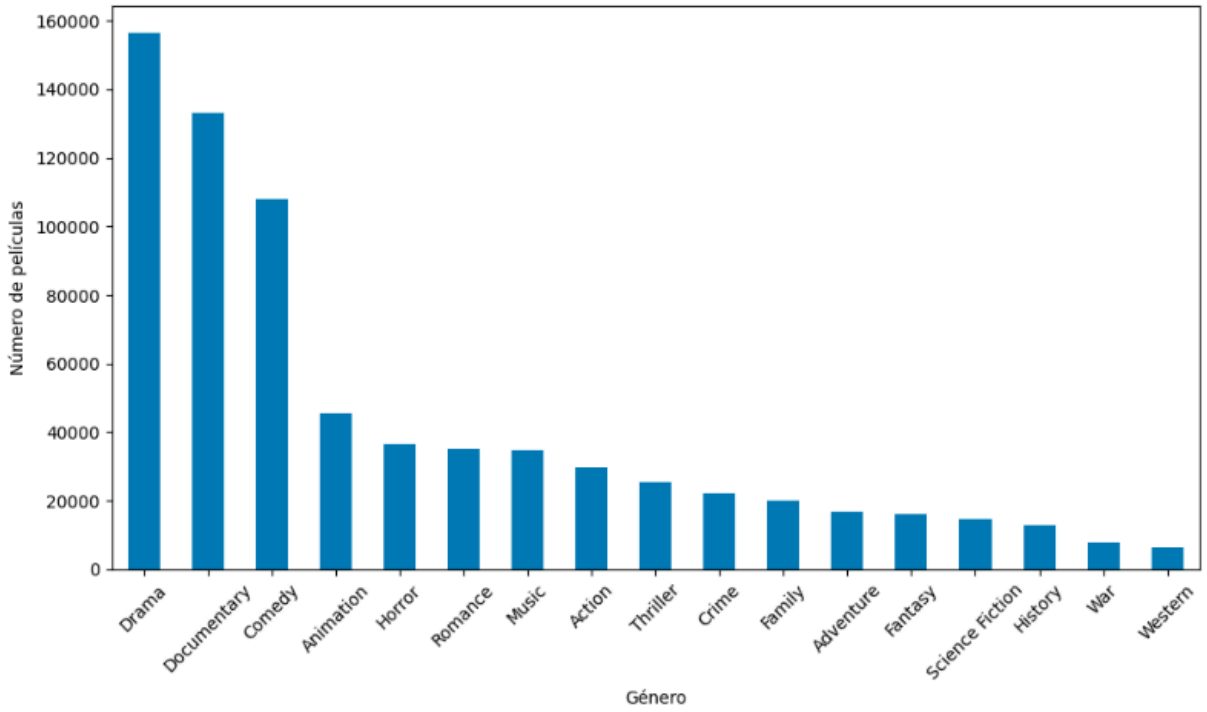
Análisis exploratorio

A continuación se presentan algunos resultados que muestran la distribución de géneros en el dataset. Se realizó una limpieza que consistió en eliminar repetidos, así como los casos en donde las sinopsis estaban vacías. Luego, se seleccionaron los 18 géneros más frecuentes para usar en los experimentos y sobre los cuales se hizo el siguiente análisis.

Nicolás Casella, Uriel Arias

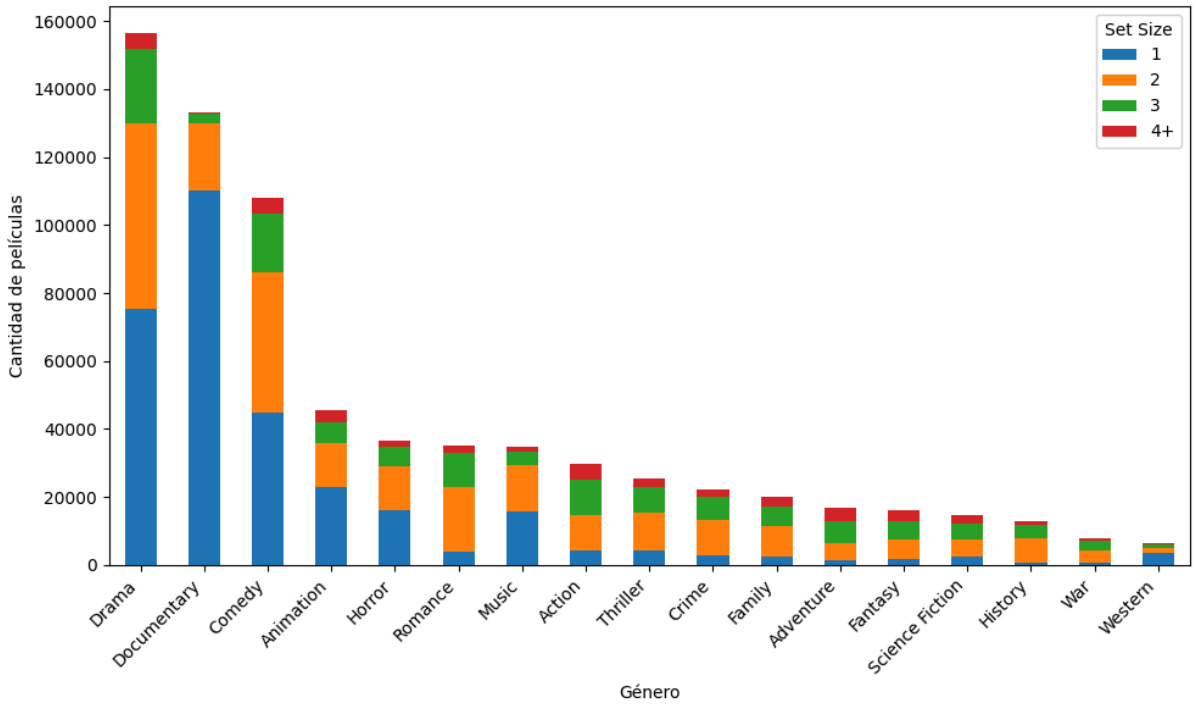
Los géneros elegidos considerando un orden descendente de frecuencias son: "Drama", "Documentary", "Comedy", "Animation", "Horror", "Romance", "Thriller", "Music", "Action", "Crime", "Family", "Adventure","Fantasy", "Science Fiction", "Mistery", "History", "War", "Western".

Análisis de Número de películas vs Género:



Análisis de Número de películas vs Número de géneros por película:

Análisis de Cantidad de películas agrupadas por el número de géneros asignados vs Género:



Es importante remarcar en este último gráfico que la cantidad de películas totales dado el número de géneros asignados a cada una se distorsiona al hacer el análisis separado por cada género en particular. Por ejemplo, para una película con los géneros “Animation” y “Family” la misma película se cuenta dos veces, una vez en cada barra al analizar géneros por separado.

Propuesta de análisis

El objetivo que planteamos es poder analizar e identificar palabras clave que puedan dar indicios sobre qué género/s está categorizada la película.

También planteamos hacer un análisis del título de la película para poder estudiar qué tanto puede influenciar hacia un género. Por ejemplo, la película “Nightmare on Elm Street” naturalmente puede catalogarse como una película de terror con solo el título, mientras que la película “It” no da mucha información sobre a qué género pertenece, sin tener contexto previo de su sinopsis. Por otro lado, es posible que si bien la sinopsis sea descriptiva no sea suficiente para detectar todos los géneros asignados como target, por lo que se podría evaluar si agregando el título como parte del entrenamiento permite mejorar esta clasificación.

Experimentos

- **¿Qué importancia tiene la sinopsis de una película a la hora de clasificarla por género? ¿Cuánto cambia el resultado al agregar el título de la misma?** e.g. que el título en películas de terror de indicios de que en efecto entra en la calificación de terror, mientras que otros títulos no den indicios a que género/s puede pertenecer.

  El experimento se hará a partir de realizar fine-tuning de un modelo de la familia BERT como bert-base-cased (https://huggingface.co/google-bert/bert-base-cased)

usando un conjunto de datos de entrenamiento separados usando un split 80-10-10, en donde para cada género se separe en ese porcentaje para *training*, *dev* y *test* respectivamente. Si bien no es el mejor modelo disponible actualmente (como deBERTaV3 o ModernBERT) sigue teniendo resultados robustos que esperamos que se traduzcan en resultados aceptables para realizar esta tarea de clasificación.

Se evaluará el rendimiento del modelo usando *Macro F1 score* para comparar el rendimiento del modelo en la clasificación de la película incluyendo o no el título.

Esta métrica está definida como el promedio de aplicar la métrica de F1 score entre diferentes clases que en este caso son representadas por los géneros definidos. La misma es útil para evaluar el grado de precisión al representar un promedio ponderado entre las métricas de *precision* (cantidad de predicciones correctas respecto del total de elementos de esa clase) y *recall* (cantidad de predicciones correctas respecto del total de elementos que se clasificaron para esa clase).

Formalmente queda definida como:

$$F1 - Score = \frac{2 * TP}{2 * TP + FP + FN} = \frac{2 * precision * recall}{precision + recall}$$

donde:

TP = *True positives* (predicciones correctas sobre pertenencia a una clase)

FP = *False positives* (predicciones incorrectas sobre pertenencia a una clase)

FN = *False negatives* (predicciones incorrectas sobre la no pertenencia a una clase)

A su vez, junto con esta métrica se podrá mostrar una matriz de confusión con todas las predicciones y la entropía cruzada del modelo final finetuneado.

- **¿Existen entidades que suelen estar más relacionadas con ciertos géneros?** e.g ubicaciones/profesiones (espías con acción, payasos con terror, casa abandonada en terror, pueblo chico en thriller, bar con comedia).

  Se podría detectar entidades como una forma de mejorar los sistemas de recomendación, por ejemplo, si se quiere recomendar a una persona películas que estén ambientadas en Nueva York o si se quiere detectar películas que están vinculadas por nombres de personajes.

  Se podría abordar esta pregunta haciendo uso de NER (Named Entity Recognition) haciendo inferencia sobre un modelo preentrenado como *bert-base-ner* (https://huggingface.co/dslim/bert-base-NER). De esta manera, luego de aplicarlo sobre cada conjunto de películas separadas por género, sobre el conjunto de entidades detectadas se hará un conteo de las entidades y palabras más importantes para cada uno de los géneros.

- **¿Cuáles son los adjetivos/adverbios que están asociados a cada género?** e.g. que en terror se usen ciertos adjetivos con una connotación negativa. La idea de esto es poder evaluar cómo se usan los adjetivos, dependiendo del género de la película.

  Una opción es usar POS Tagging por medio de librerías como spaCy (https://spacy.io/usage/linguistic-features#pos-tagging) en el que se distinguen adjetivos y adverbios. Nuevamente sobre cada conjunto de películas separado por género se destacarán las que más se repiten.

- **¿Existe una correlación entre la longitud de la sinopsis y el género?** e.g. que en los documentales se haga una pequeña introducción sobre el tema y su importancia respecto de lo que podría resumirse de la sinopsis de una película.

  Se podría ilustrar las métricas de precisión elegidas en función de la longitud de las sinopsis usadas para cada género. Es decir, podríamos ver si existen géneros que con sinopsis más cortas ya son detectadas rápidamente a diferencia de otras.

  Para esto se podrían separar las películas de cada género separadas por bines y luego mostrar un promedio de f1-score para cada bin.

Anexo

En este caso no es necesario.