UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

Entrega Final Proyecto

FACULTAD DE INGENIERÍA
2023

Proyecto Sistema de Recomendación de Películas

Fonsy Johan Mercado Agudelo
CC 1020472932
Ingeniería eléctrica

Orlando José Salazar Polo
CC 1152714311
Ingeniería eléctrica

Angie Dayana Rincón Mandón
CC 1091681348
Ingeniería eléctrica

Índice

1. Contexto

El crecimiento explosivo en la cantidad de información digital disponible y el número de visitantes a Internet han creado un desafío potencial de sobrecarga de información que dificulta el acceso oportuno a elementos de interés en Internet. Los sistemas de recuperación de información, como Google, DevilFinder y Altavista, han resuelto parcialmente este problema Isinkaye, Folajimi y Ojokoh [1] [1], pero la priorización y la personalización (donde un sistema asigna el contenido disponible a los intereses y preferencias del usuario) de la información estaban ausentes. Esto ha aumentado la demanda de sistemas de recomendación más que nunca.

Los sistemas de recomendación son sistemas de filtrado de información que se ocupan del problema de la sobrecarga de información al filtrar fragmentos de información vital de una gran cantidad de información generada dinámicamente de acuerdo con las preferencias, el interés o el comportamiento observado del usuario.

El sistema de recomendación tiene la capacidad de predecir si un usuario en particular preferiría un artículo o no según el perfil del usuario. Los sistemas de recomendación son beneficiosos tanto para los proveedores de servicios como para los usuarios, ya que estos reducen los costos de encontrar y seleccionar artículos en sean de interés del usuario.

2. Objeto machine learning

Ya que para recomendar un contenido según las preferencias de un usuario se debe tener en cuenta tanto las características del usuario como las características del contenido que se tiene, para ello es necesario implementar tecnicas que modelen un filtrado colaborativo. Lo que se busca es implementar un modelo que seleccione el contenido que más se ajuste a las preferencias del usuario.

3. Dataset

Para el desarrollo de este proyecto se seleccionó el dataset The Movies Dataset proporcionado en kaggle, estos archivos contienen metadatos de 45.000 películas enumeradas en el conjunto de datos completo de MovieLens. Este conjunto de datos también tiene archivos que contienen 26 millones de calificaciones de 270.000 usuarios para las 45.000 películas. Las calificaciones están en una escala de 1 a 5 y se han obtenido del sitio web oficial de GroupLens.

El dataset tiene un total de 900 MB y contiene los siguientes archivos .csv

- **movies_metadata.csv**: Archivo principal de metadatos de películas, contiene

información de 45000 películas incluyendo carteles, fondos, presupuesto, ingresos, fechas de lanzamiento, idiomas, países de producción y empresas.

- **keywords.csv**: Contiene las palabras clave de la trama de la película en forma de un objeto JSON en cadena.
- **credits.csv**: Contiene información sobre el reparto y equipo técnico de las películas en forma de objeto JSON en cadena.
- **links.csv**: Contiene los ID de TMDB e IMDB de las películas que aparecen en el conjunto de datos de MovieLens.
- **links_small.csv**: Contiene los ID de TMDB e IMDB de un subconjunto de 9.000 películas del conjunto de datos completo.
- **ratings_small.csv** El subconjunto de 100.000 calificaciones de 700 usuarios en 9.000 películas

4. Métricas de desempeño

La calidad de un algoritmo de recomendación se puede evaluar utilizando diferentes tipos de medidas, que pueden ser de accuracy o de coverage. El accuracy es la fracción de recomendaciones correctas del total de recomendaciones posibles, mientras que la coverage mide la fracción de objetos en el espacio de búsqueda para los que el sistema puede proporcionar recomendaciones. Las métricas para medir el accuracy de los sistemas de filtrado de recomendaciones se dividen en métricas de precisión estadísticas y de soporte de decisiones.

Métricas de precisión estadística: Evaluar la precisión de una técnica de filtrado comparando las calificaciones pronosticadas directamente con la calificación real del usuario.

El error absoluto medio (MAE), el error cuadrático medio (RMSE) y la correlación se utilizan normalmente como métricas de precisión estadística.

MAE es el más popular y comúnmente utilizado, es una medida de la desviación de la recomendación del valor específico del usuario. Se calcula de la siguiente manera:

$$\text{MAE } = \frac{1}{N} \sum_{u,i} |p_{u,i} - r_{u,i}| \tag{1}$$

donde $p_{u,i}$ es la calificación prevista para el usuario u en el elemento i, $r_{u,i}$ es la calificación real y N es el número total de calificaciones en el conjunto de elementos. Cuanto menor sea el MAE, con mayor precisión el motor de recomendaciones predice

las calificaciones de los usuarios. Además, el error cuadrático medio (RMSE) está dado por:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{u,i}\left(p_{u,i} - r_{u,i}\right)^2} \tag{2}$$

El error cuadrático medio (RMSE) pone más énfasis en un error absoluto más grande y cuanto más bajo es el RMSE, mejor es la precisión de la recomendación.

5. Referencias y resultados previos

Referente a los recomendadores de contenido se han tratado con diferentes enfoques y diferentes modelos, Isinkaye, Folajimi y Ojokoh [1] [1] resume las diferentes técnicas de sistemas recomendadores de contenido, como se ilustra en la Fig. Figura 1, donde la selección de una o la combinación de varias de estas técnicas depende de la aplicación que se desea obtener y los datos disponibles.

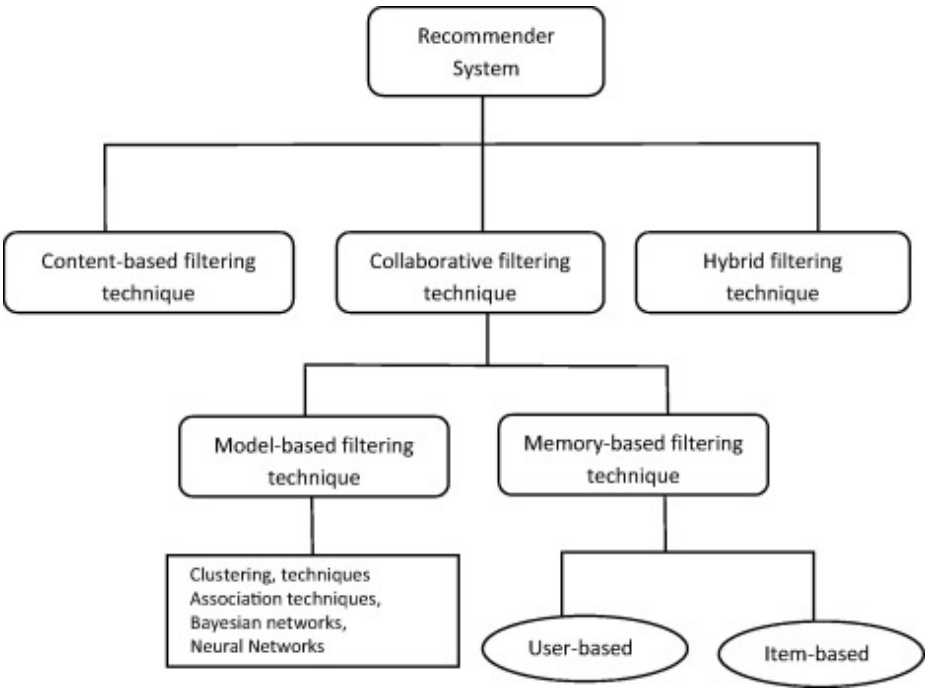

Figura 1. Técnicas recomendador de contenidos.

Zamanzadeh Darban y Valipour [2] [2] y Gan y Cui [3] [3] son otros autores que proponen diferentes modelos más enfocados en técnicas de deep learning, proponiendo

incluso un recomendador de contenido híbrido basado en grafos, comparándolo con diferentes modelos, teniendo un $RMSE$ de 0,833.

6. Desarrollo

Para el desarrollo del sistema recomendador de contenidos se proponen de dos etapas o notebooks reproducibles, los cuales se encentran en el repositorio de GitHub, en este se encuentran el notebook **01 - Exploración de datos.ipynb** donde se hace tratan los datos y se realiza una pequeña exploración sobre estos y el notebook **02 - Modelos.ipynb** donde se crean, entrenan y se evalúan los modelos propuestos para el sistema recomendador de contenidos.

6.1. Exploración de los datos

Para esta exploración solo se analiza la información de las películas, explorando el presupuesto, los ingresos, popularidad, cantidad de votos, la media de votos, géneros cinematográficos, entre otros.

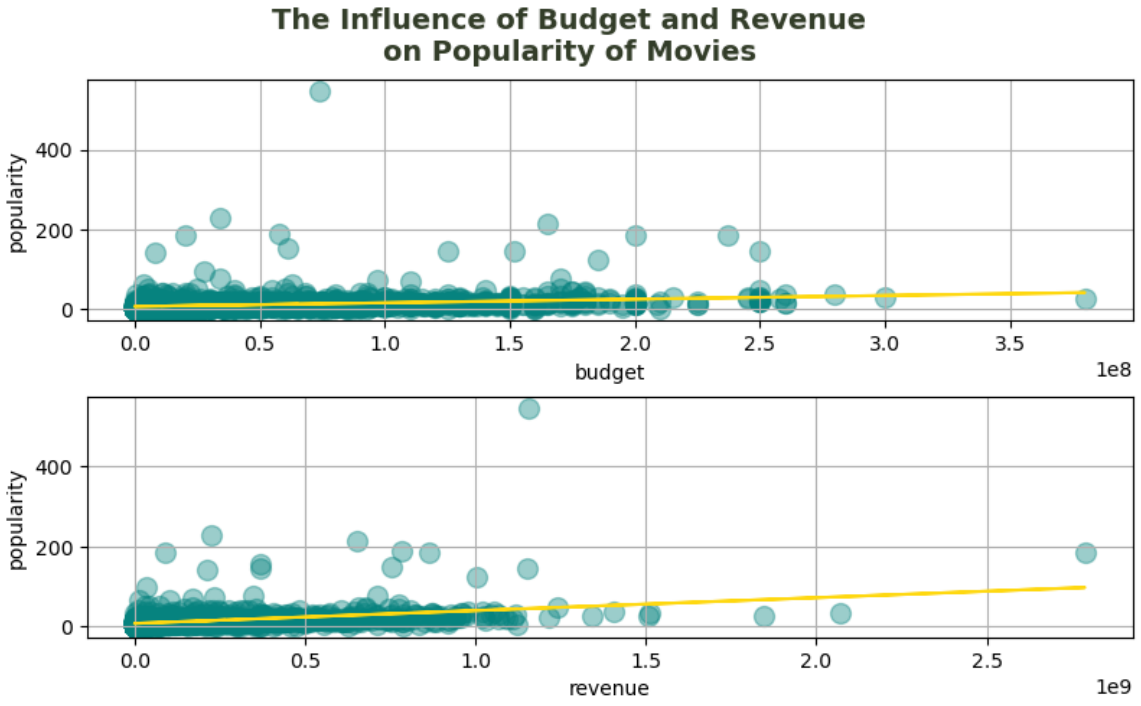

Figura 2. Influencia del presupuesto e ingresos sobre la popularidad de las películas.

De la Fig. Figura 2 se puede ver que el presupuesto que se tiene para la realicación de una película no está directamente relacionado con su popularidad, mientras que los ingresos generados por estas películas si tienen un poco más de relación con la popularidad de estas.

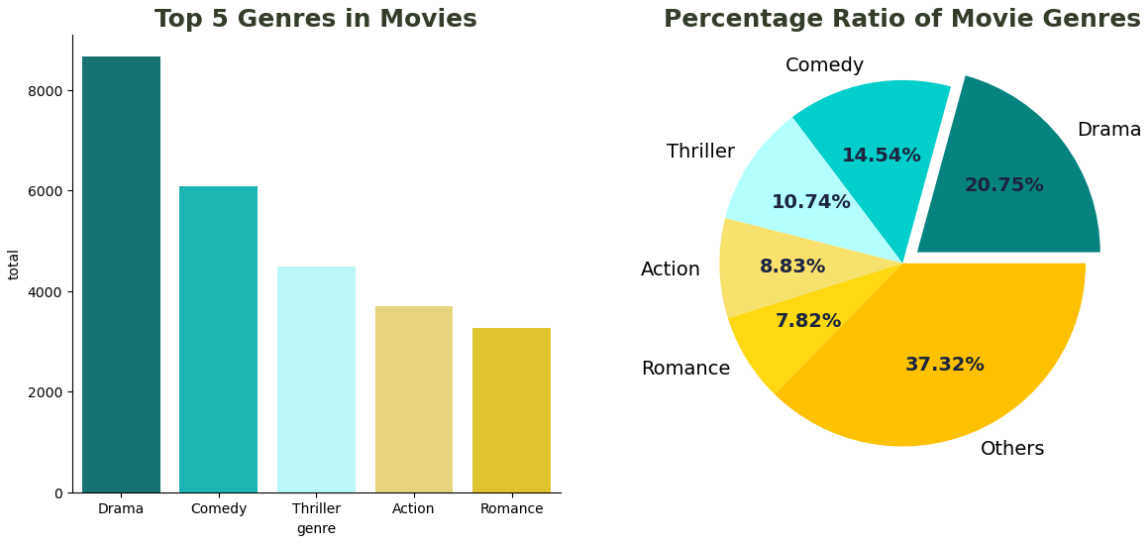

Figura 3. Top de géneros cinematográficos.

El top de géneros cinematográficos son drama, comedia, terror, acción y romance, siendo el género de drama el género más usado para la realización de las películas, como se puede ver de la Fig. Figura 3, aunque cabe resaltar que las películas suelen ser etiquetadas por más de un género. En los otros géneros se tienen etiquetas como guerra, animación, historia, documental, familia, entre otros.

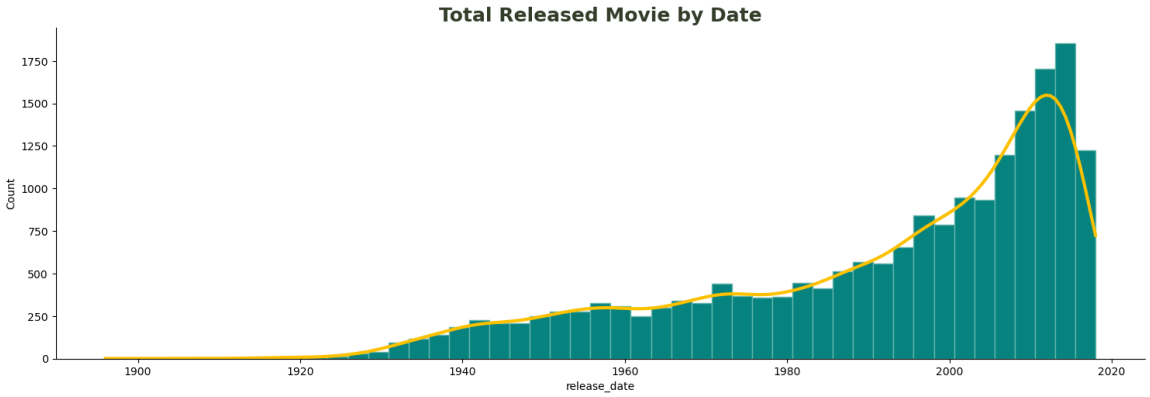

Figura 4. Cantidad de lanzamiento de películas por año.

En la Fig. Figura 4 se presenta un análisis de la cantidad de lanzamientos de películas por año. Al examinar la gráfica, es evidente que la industria cinematográfica ha experimentado un aumento significativo en la cantidad de películas lanzadas en los últimos años. En particular, se destaca una concentración de estrenos en el período que abarca desde el año 2004 hasta el 2020

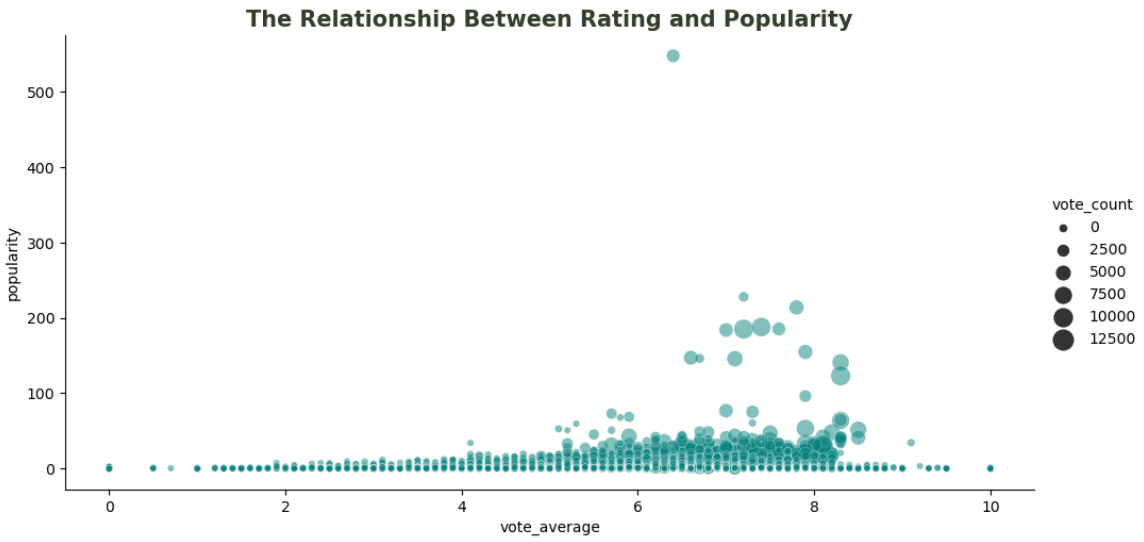

Figura 5. Relación entre rating y popularidad.

De la Fig. Figura 5 podemos ver la relación existente entre la media de los votos realizados por los usuarios y la popularidad de las películas, teniendo también una fuerte relación con la cantidad de votos que tiene una película en específico.

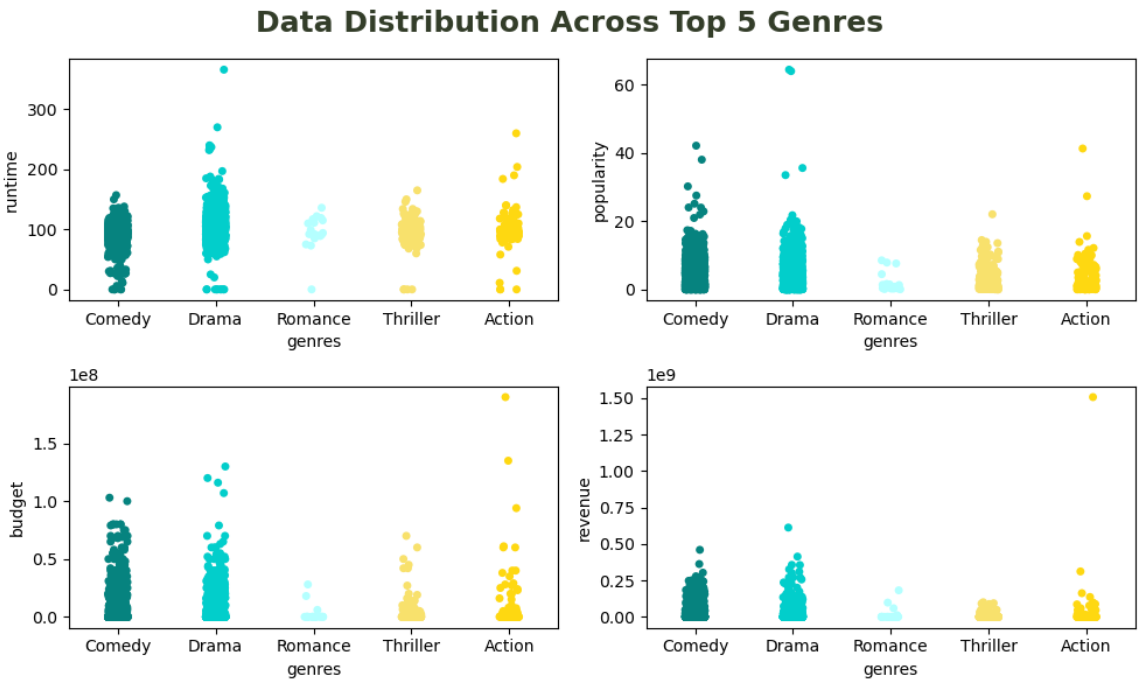

Figura 6. Distribución de los datos del top 5 de géneros cinematrográficos.

De la Fig. Figura 6 se pueden extraer varios datos de interés.

- El género cinematográfico que tiene mayor duración es el drama.
- El género menos popular en el top 5 es el romance.
- Las películas de acción gastaron más dinero que el resto de las películas.
- Una de las películas de acción obtuvo una gran ganancia en comparación con las demás.

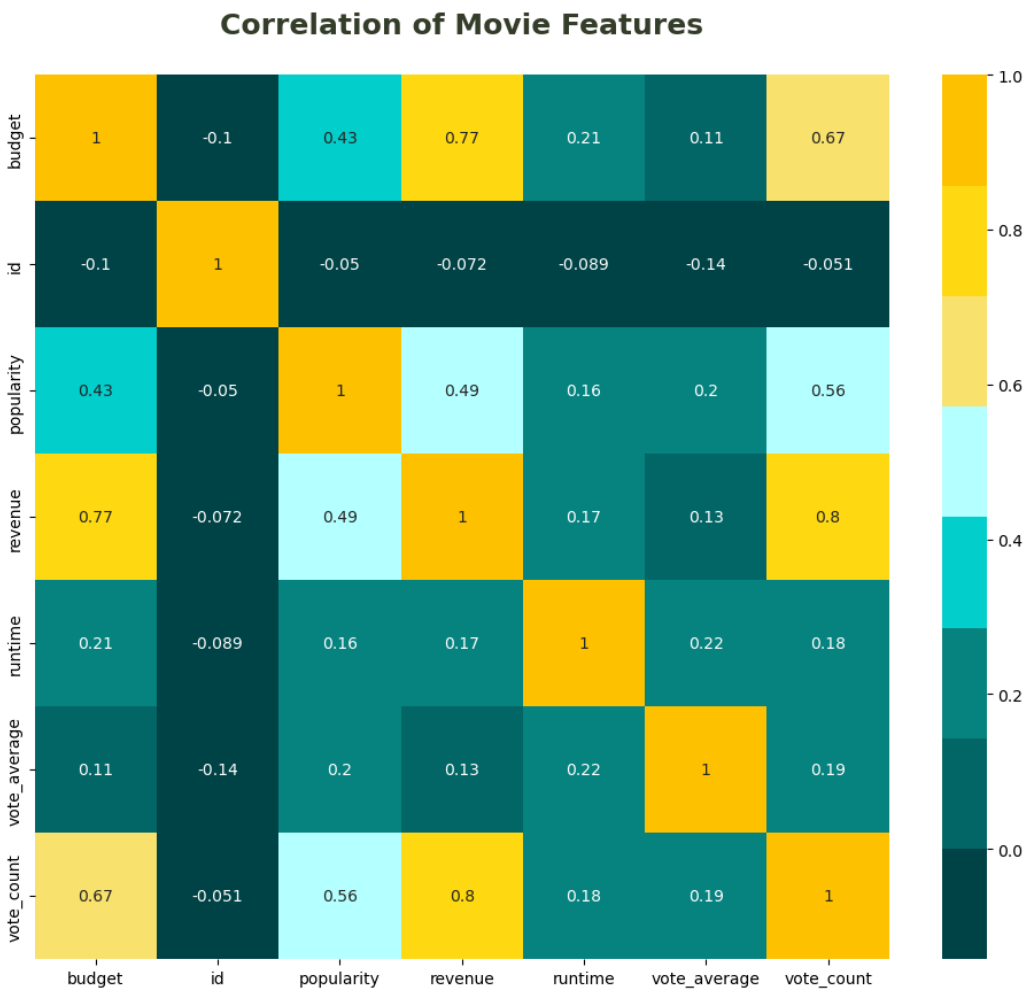

Figura 7. Correlación entre las características cinematográficas.

De la Fig. Figura 7 se puede ver una representación visual de la relación entre el costo de producción de una película, sus ingresos en la taquilla, la cantidad de votos emitidos por los usuarios y los costos de marketing asociados. A partir de esta figura, se puede notar una buena correlación entre estos indicadores

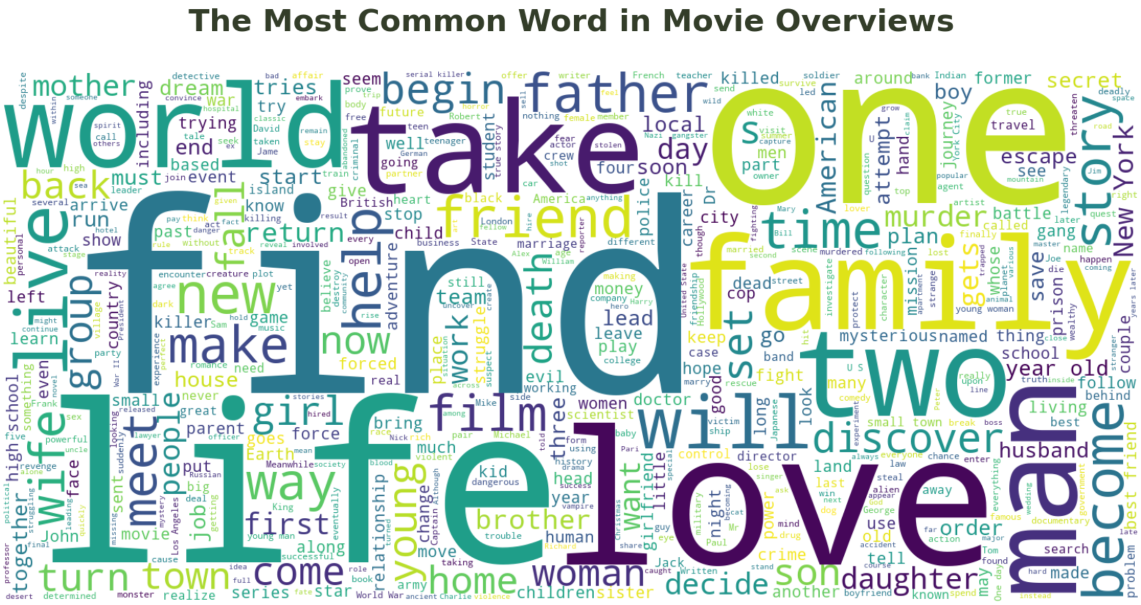

Figura 8. Nube de palabras.

Por último en la Fig. Figura 8 se puede visualizar una nube de palabras de la descripción cinematográfica, viendo las palabras más usadas para describirlas.

A partir de esta figura, se pueden identificar algunas de las palabras más utilizadas para describir las películas en la base de datos. Algunas de las palabras más frecuentes son “vida”, “amor”, “historia”, “mundo”, “familia”, “amigos”, entre otras. Estas palabras dan una idea general de los temas y motivos recurrentes en las películas que conforman la base de datos.

6.2. Modelos

Para generar un sistema recomendador de contenidos se puede categorizar por 2 tipos, por filtración colaborativa o filtración basada en contenido, como se ejemplifica en la Fig. Figura 9

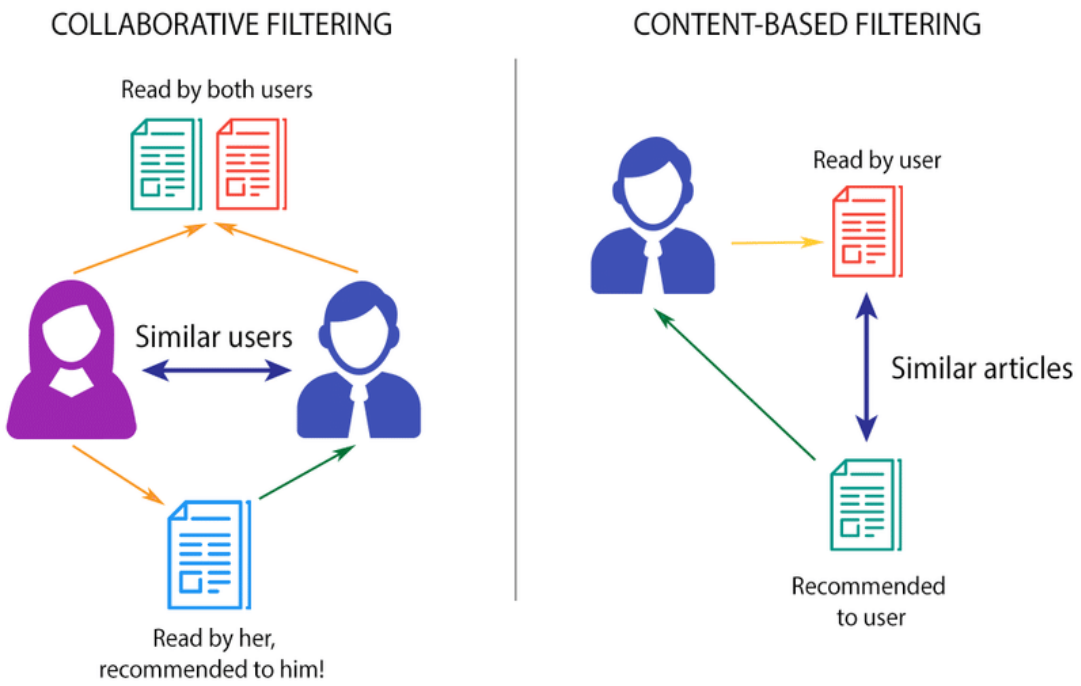

Figura 9. Tipos de recomendador de contenidos.

Para el filtrado colaborativo se propone usar *Embeddings* para representar vectorialmente los usuarios y las películas como entrada, según los ratings que cada usuario calificó cada película.

Para el filtrado basado en contenido a parte de usar *Embeddings* para representar vectorialmente los usuarios y las películas como entrada, también se propone los géneros cinematrográficos como otro parámetro de entrada, realizando un one-hot-encoding con los géneros encontrados en la base de datos.

Por último, el tercer modelo propuesto se le agrega una nueva entrada y es la representación *tf-idf* de la descripción cinematográfica, teniendo un modelo híbrido que recomiende contenido según los gustos similares de otros usuarios, los géneros y descripciones cinematográficos visualizados.

Para compilar el modelo se usó la función de pérdida *MSE* y el optimizador *adam*, para el entrenamiento se usó un *batch size* de 2048 y 20 *epochs*, las pérdidas de

entrenamiento y de pruebas por modelo se ilustra en la Fig. Figura 10 y los *RMSE* obtenidos por modelo se muestran en la Tabla Tabla 1

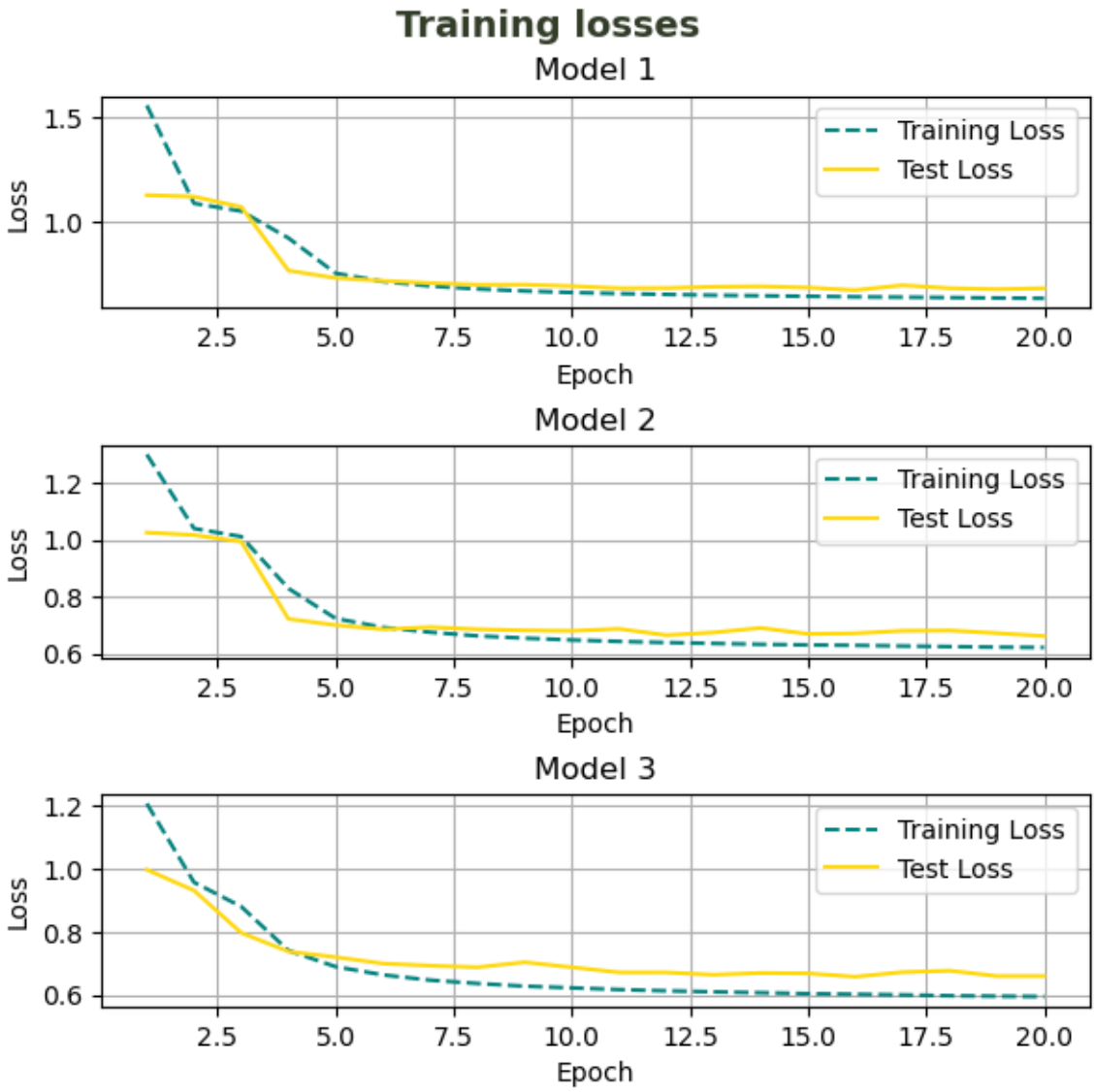

Figura 10. Pérdidas de entrenamiento.

Tabla 1. RMSE obtenido por modelo

	Model 1	Model 2	Model 3
RMSE	0.8215	0.8119	0.8116

Para concluir, hay muchas formas diferentes de configurar un sistema de recomendación de contenido y al igual que otros algoritmos de deep learning, es muy importante saber qué objetivo se debe optimizar y, por lo tanto, qué diseño se debe elegir, si se desea una configuración de filtrado colaborativo es importante recalcar el problema del inicio en frío y es cuando al sistema se ingresa un nuevo usuario o una nueva película, el cuál no se tendría suficiente información para recomendar, esto se puede solucionar teniendo en cuenta otras características, como el filtrado basado en contenido, en el cual se pueden asociar los gustos del usuario con las características del producto.

Referencias

[1] F. Isinkaye, Y. Folajimi y B. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, n.º 3, págs. 261-273, 2015, ISSN: 1110-8665. DOI: https://doi.org/10.1016/j.eij.2015.06.005. dirección: https://www.sciencedirect.com/science/article/pii/S1110866515000341.

[2] Z. Zamanzadeh Darban y M. H. Valipour, "GHRS: Graph-based hybrid recommendation system with application to movie recommendation," *Expert Systems with Applications*, vol. 200, pág. 116 850, 2022, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116850. dirección: https://www.sciencedirect.com/science/article/pii/S0957417422003025.

[3] M. Gan y H. Cui, "Exploring user movie interest space: A deep learning based dynamic recommendation model," *Expert Systems with Applications*, vol. 173, pág. 114 695, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114695. dirección: https://www.sciencedirect.com/science/article/pii/S0957417421001366.