Sistema Hibrido de Recomendacion de Peliculas BERT + GAT con Enriquecimiento Semantico y Estructural

***Abstract*— This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

I. INTRODUCCIÓN

El crecimiento exponencial de la información en plataformas digitales ha generado un impacto de sobrecarga informativa, donde los usuarios se ven abrumados por opciones de contenidos, productos o servicios; así mismo, ha incrementado la necesidad de los usuarios por obtener opciones a la medida, de una manera cómoda, rápida y efectiva. En este contexto, los sistemas de filtrado de información conocidos como sistemas de recomendación se han vuelto herramientas fundamentales en la personalización de la experiencia de usuario. Un sistema de recomendación, en su esencia, es una forma de filtrado de información que predice la valoración o preferencia que un usuario otorgaría a un ítem determinado. Estos sistemas ya tienen relevancia en plataformas modernas como los servicios de streaming, redes sociales y comercio electrónico, entre otros, permitiendo descubrir contenido relevante teniendo en cuenta el perfil del usuario. Sin un filtrado eficaz, el exceso de información puede derivar en frustración y menor actividad del usuario.

Tradicionalmente, los enfoques de recomendación se dividen en filtrado colaborativo (basado en patrones de interacciones usuario-ítem) y filtrado basado en contenido (basado en las características de los ítems). Sin embargo, estos enfoques presentan desventajas como el problema de cold-start para usuarios e ítems sin historial, lo cual puede reflejar sesgos de popularidad. Por otro lado, el modelo basado en contenido puede carecer de personalización al ignorar las relaciones entre usuarios. Dado estos problemas, en la última década ha cobrado relevancia la implementación de sistemas híbridos que combinan variadas fuentes de información y que mitigan debilidades de los modelos anteriores. En particular, la integración de información semántica textual con información estructural de grafos se ha mostrado prometedora. La semántica derivada del texto aporta contexto y significado, mientras que la estructura del grafo captura relaciones complejas y patrones de comportamiento colectivo.

Este trabajo explora un sistema de recomendación híbrido que integra un modelo semántico basado en BERT con un modelo de grafos basado en Graph Attention Networks (GAT), con el objetivo de mejorar la precisión de la predicción de preferencias del usuario y la calidad de las recomendaciones Top-k.

II. TRABAJOS RELACIONADOS

A. Enfoques, resultados y limitaciones

- **GHRS: Graph-based Hybrid Recommendation System with Application to Movie Recommendation** : El objetivo principal de la investigación es desarrollar un sistema de recomendación híbrido que combine las preferencias y similitudes entre usuarios con información demográfica, utilizando representaciones en grafos y técnicas avanzadas de deep learning. Con este enfoque se busca mejorar la capacidad de generalización del sistema, resolver el problema del "cold-start" y capturar relaciones complejas y no lineales entre los usuarios y los ítems (películas, productos, servicios). Dado esto, la investigación se basa en la integración de diversos enfoques y herramientas. Como punto de partida, se utilizan dos conjuntos de datos reales (MovieLens 100K y MovieLens 1M) que incluyen ratings y, en el caso de MovieLens 100K, información demográfica para abordar el cold-start. A partir de los ratings se construye la matriz R=U×I, donde U representa el conjunto de usuarios e I el de ítems.
  Como parte de la metodología para modelar las relaciones entre usuarios, se construye un grafo en el que cada usuario es representado como un nodo y se establecen conexiones basadas en la similitud de sus ratings. Se aplican algoritmos de análisis de grafos, como PageRank para determinar la importancia de cada nodo, y medidas de centralidad, por ejemplo, la closeness centrality, para identificar usuarios influyentes.
  Adicionalmente, los datos demográficos se transforman en variables categóricas y se concatenan con la información de las relaciones extraídas del grafo, generando un vector de características combinado. Este vector se reduce dimensionalmente mediante un autoencoder (incluyendo variantes como el Denoising Autoencoder) para evitar el overfitting y extraer representaciones compactas. La salida del autoencoder se utiliza para agrupar a los usuarios mediante K-means, determinando el número óptimo de clusters con métodos como el codo (elbow) y el coeficiente de silueta.
  1) resultados: El estudio muestra que la integración de información demográfica con las representaciones basadas en grafos mejora significativamente la precisión de las recomendaciones. En particular, se observa que el enfoque híbrido es capaz de mitigar el problema del cold-start al proporcionar recomendaciones para nuevos usuarios o ítems sin historial previo. Además, el uso de

Autoencoders para la reducción de dimensionalidad y clustering para segmentar a los usuarios ha permitido capturar relaciones no lineales complejas, mejorando la generalización del sistema.

Dataset	RMSE	Precision	Recall
100K	0.887, S^2=1.595×10^{-4}	0.771	0.799
1M	0.833, S^2=2.815×10^{-4}	0.792	0.838

Performance metrics value for the proposed method on target dataset (Zahra Zamanzadeh Darbana,Mohammad Hadi Valipour)

- **Combining collaborative and content information into a novel graph based hybrid model for recommendation – Seydou Kane** : El principal objetivo de la investigación es desarrollar un modelo híbrido que combine de manera efectiva la información colaborativa y de contenido, utilizando una estructura de grafo que conecta usuarios, ítems y características. Este modelo pretende no solo superar a los algoritmos puramente colaborativos o de contenido, sino también abordar de manera eficaz el problema del cold start. Para lograr esto, el estudio utiliza recorridos de tres pasos en el grafo y emplea el algoritmo Bayesian Personalized Ranking (BPR) para aprender automáticamente los pesos óptimos de las conexiones, equilibrando dinámicamente la importancia de cada tipo de información.

  2) enfoque: El estudio se fundamenta en el diseño de un modelo basado en grafos que incluye tres tipos de nodos: usuarios, ítems y características. Los nodos de usuarios e ítems están conectados mediante bordes que representan las interacciones (cuando un usuario consume o interactúa con un ítem), mientras que los bordes entre ítems y características indican la presencia de determinadas propiedades en dichos ítems. La estructura del grafo permite modelar de forma natural las relaciones entre todos los elementos del sistema.
  Para evaluar el modelo, se utilizaron tres conjuntos de datos reconocidos —Yahoo Movies, LastFm Hetrec y MovieLens Hetrec— que proporcionan tanto información colaborativa como de contenido. Las calificaciones explícitas de los usuarios se convirtieron en feedback implícito binario, simplificando el modelo a un enfoque de interacción (consumo o no del ítem). El algoritmo se basa en recorridos de tres pasos a través del grafo, y el uso de BPR permite ajustar los pesos de las conexiones en función de su relevancia para generar recomendaciones precisas.
  La evaluación se realizó en dos escenarios: uno de 'warm start', utilizando el 80% de los datos para entrenamiento y el 20% para pruebas, y otro de 'cold start', donde se reservaron aleatoriamente el 20% de los ítems para pruebas sin incluir sus interacciones en el entrenamiento. Además, se comparó el rendimiento del modelo propuesto con varios algoritmos de referencia, como Item KNN, Item CBF y un enfoque híbrido CF-CBF, utilizando métricas estándar en el campo.
  3) resultados: Los hallazgos del estudio revelan que el modelo híbrido basado en grafos supera a los algoritmos colaborativos y ofrece resultados comparables o superiores a los métodos basados en contenido, sobre todo en escenarios de cold start. En condiciones de datos abundantes (warm start), el modelo también mostró un rendimiento ligeramente superior respecto a las técnicas tradicionales. Se observó además que existe una correlación positiva entre la longitud del perfil de usuario (es decir, la cantidad de interacciones previas) y el rendimiento del sistema, lo que indica que el modelo aprovecha de manera más efectiva la información cuando los usuarios tienen historiales más extensos. Un hallazgo importante fue la reducción del sesgo hacia ítems populares, lo que se tradujo en recomendaciones más diversas y menos centradas en lo ya conocido. La simplificación del gradiente en el algoritmo BPR también resultó en mejoras significativas en el rendimiento del modelo.

III. METODOLOGÍA

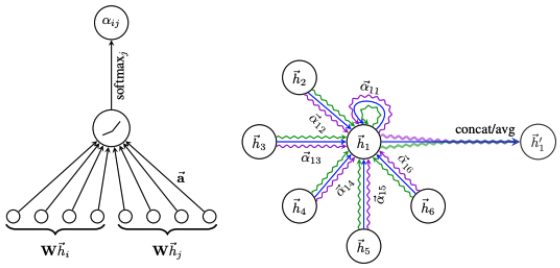
El enfoque metodológico de la solución se centra en los métodos propuestos sobre diferentes arquitecturas hibridas para sistemas de recomendacion dependiendo de las métricas como RMSE, Precision@5, Recall@5, Precision@10 y recall@10 sobre el conjunto de datos MovieLens ml-latest-small para 610 usuarios y 9700 peliculas(linea base de datos).

- **Filtrado Colaborativo (Baseline)**: Se construyó un modelo base de filtrado colaborativo puro, utilizando únicamente las interacciones usuario–película (ratings explícitos). El grafo implícito formado es bipartito, con nodos de usuario y de película conectados por aristas ponderadas según el rating dado. No se incorporó información de contenido en esta versión. El modelo aprende representaciones latentes de usuarios y películas (mediante matrix factorization o redes neuronales simples) optimizando la predicción de los ratings conocidos (minimizando el error cuadrático medio). Este enfoque sirve de baseline para comparar el efecto de añadir datos de contenido.

- **Modelo híbrido GraphSAGE + LightFM**: En este enfoque basado en grafos, se integró información de contenido de las películas utilizando el archivo movies_data.csv (que incluye metadatos como géneros, popularidad, elenco, etc.). Primero se construyó un grafo con nodos de usuario y de película conectados por interacciones (similar al caso colaborativo). Adicionalmente, las películas se enriquecieron con atributos de contenido como características de nodo (por ejemplo, codificando géneros cinematográficos, país de origen, puntuación media de TMDB, etc.). Sobre este grafo heterogéneo se aplicó GraphSAGE (mediante PyTorch Geometric) para aprender embeddings de

los nodos película que capturan tanto la estructura de interacciones como la similitud en atributos. Estas representaciones vectoriales de las películas se integraron en un modelo de filtrado colaborativo híbrido LightFM como caracteristicas del ítem. En la práctica, LightFM entrena un modelo de factores latentes capaz de combinar señales colaborativas (interacciones usuario-item) con señales de contenido (representaciones de GraphSAGE y otros atributos) en un mismo marco. El entrenamiento de LightFM se realizó como un problema de retroalimentación implícita, optimizando un criterio de ranking (p. ej., Bayesian Personalized Ranking o WARP) en lugar de error cuadrático, para enfocar el aprendizaje en la ordenación de preferencias. De este modo, la metodología GraphSAGE+LightFM aprovecha la estructura del grafo y las características exógenas de las películas para abordar el problema del cold start y enriquecer las recomendaciones más allá de lo que el filtrado colaborativo puro puede lograr.

- **Modelo con BERT + GAT (Graph Attention Network)**: Este enfoque híbrido profundo integra de forma directa las características de contenido textual de las películas con las interacciones en un grafo mediante redes neuronales de grafos. En primer lugar, se utilizó BERT para codificar información textual de las películas: específicamente, se tomó de movies_data.csv la sinopsis breve o tagline de cada película (así como otros metadatos textuales disponibles) y se obtuvo un embedding denso de alta dimensión para cada película usando un modelo preentrenado de BERT. Cada vector BERT resultante (que capta semántica del contenido de la película) se asignó como feature inicial al nodo correspondiente de película en el grafo. El grafo considerado es nuevamente bipartito usuarios-películas; en versiones iniciales solo incluyó aristas usuario–película (interacciones), pero en versiones posteriores se pudo haber enriquecido con relaciones adicionales (por ejemplo, conectando películas por similitud de contenido o género, aunque el núcleo fue la relación de rating). Sobre este grafo con nodos dotados de atributos (los embeddings de BERT para películas, y vectores posiblemente neutros o inferidos para usuarios), se entrenó una Graph Attention Network (GAT).
  GAT por sus comportamiento aplica una transformacion lineal de cada nodo a sus caracteristicas, y asi calcula cuanto debería prestar atención al nodo con que tiene conexión dentro del mecanismo de atencion,las relaciones tienen en cuenta los vecinos directos y luego pasa por una softmax que expone los valores proporcionales (Petar Velickovic, Guillem Cucurull (2018), Arantxa Casanova,Pietro Lio & Yoshua Bengio)



(2) Izquierda: mecanismo de atención $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ utilizado por el modelo, parametrizado por un vector de pesos donde $a \in R^{2F}$, aplicando la activación LeakyReLU. A la derecha: atención multi-cabeza con K = 3 cabezas, del nodo 1 sobre su vencindario. Lasa caracterisiticas agregadas de cada cabeza se concatenan o promedian para obtener h'_1 (Petar Velickovic, Guillem Cucurull (2018), Arantxa Casanova,Pietro Lio & Yoshua Bengio)

La GAT propaga y aprende representaciones ajustando pesos de atención a vecinos las características de cada nodo película se van actualizando mediante la información de sus usuarios conectados (y viceversa) a lo largo de una o más capas de atención. El objetivo de entrenamiento fue predecir las preferencias del usuario: concretamente, la GAT aprendió a predecir el rating del usuario a la película (ya sea formulado como un valor continuo de 1 a 5, o transformado en una señal implícita de relevancia). En la implementación, se optimizó una función de pérdida sobre las aristas usuario-item (por ejemplo, error cuadrático medio entre rating real y predicho, o una variante de BPR si se formuló como implícito) para ajustar el modelo. Cabe destacar que en las iteraciones más avanzadas de este enfoque, el modelo fue tuneado conjuntamente con BERT: es decir, se permitió el fine-tuning de las representaciones BERT durante el entrenamiento de la GAT, de modo que las características de contenido se ajustaran mejor a las necesidades de la tarea de recomendación. También se experimentó con la incorporación de atributos categóricos adicionales de peliculas (e.g. codificación de género cinematográfico) concatenados al embedding BERT, integrándolos en la entrada de la GAT para enriquecer la información de cada nodo. En suma, la metodología BERT+GAT construye un modelo neuronal de grafo que fusiona información colaborativa y de contenido de manera end-to-end, permitiendo que las caracteristicas latentes de usuarios y películas se informen tanto por el patrón de interacciones en el grafo como por las descripciones de las películas a nivel semántico.

IV. Experimentos

A. Definición de alternativas e implementación

Se implementaron y evaluaron tres enfoques de recomendación sobre el conjunto de datos MovieLens (ml-latest-small, 610 usuarios, 9.700 películas). Cada enfoque incorpora la información de manera distinta.

Se utilizaron 3 enfoques distintos como metodo exploratorio y ver el rendimiento de cada uno y se

eligio el mejor.

- **Filtrado Colaborativo (Baseline)**: Notebook 03_colaborative_filtering.ipynb El modelo implementado como línea base es un model de filtrado colaborativo puro mediante la factorización de la matriz de ratings, donde cada usuario y película se representaron con vectores latentes de dimensión d (valor ajustado en validación) $\hat{r}_{u,i}$ como producto interno de los vectores latentes del usuario u y la pelicula i. Pra el entrenamiento se utilizo la pérdida de MSE entre $\hat{r}_{u,i}$ y el rating real $r_{u,i}$. Para el optimizador se aplicó ADAM con una taza de aprendizaje adecuada a los datos, monitoreando el RMSE sobre la validación de un 10% de los datos para detener el entrenamiento antes del sobreajuste.
  (1)

$$\hat{r}_{ui} = \frac{\sum\limits_{v \in N_k(u)} \text{sim}(u, v) \cdot r_{vi}}{\sum\limits_{v \in N_k(u)} |\text{sim}(u, v)|}$$

- **Modelo GraphSage + LightFM**: Notebook graphLightFM.ipynb En este experimiento hibrido se combinó un enfoque de redes de grafos con un modelo de filtrado matricial. Utilizando PyTorch Geometric, se entreno un modelo GRAPHSAGE sobre el graph de usuarios-películas. Durante este entrenamiento, cada nodo pelicula disponia de atributos de contenido (derivados de movies_data.csv) mientras que los nodos de usuario podían inicializarse con vectores nulos o algún embedding aprendido (un embedding aleatorio o usando one-hot de identificador). GraphSage se configuro con varias capas de convolución gráfica que agregan la informacion de vecinos: en cada capa, el embedding de una pelicula se actualiza a partir de una agregación (suma/promedio) de los embeddings de lo usuarios que lo calificaron (y viceversa, en el caso de actualizar usuarios con peliculas vecinas). Al final del entrenamiento se obtuvieron vectores latentes de cada pelicula que reflejan conexiones colaborativas como similitudes por atributos. Sobre LighFM, se utilizó un algoritmo de filtrado colaborativo que admite características de contenido. Se entrenó un modelo con perdida de ranking(se utilizo el metodo warp para maximizar Precision@K en entrenamiento). Las interacciones de entrenamiento cnosistieron en pares usuario-pelicula con feedback implicito (considerando un rating >= 4 como iteraccion positiva). Como features de item en LightFM se incorporaron los embeddings obtenidos de GraphSAGE junto con otras características (por ejemplo, género codificado, año de lanzamiento o popularidad, integrados como variables categóricas/booleanas). De este modo, LightFM disponía de información colaborativa (interacciones usuario-item binarias) y de contenido (a través de los embeddings de GraphSAGE y atributos) para aprender los factores latentes finales. La validación del modelo se hizo monitoreando la métrica Precision@10 en un conjunto de validación de interacciones; tras convergencia (alcanzada en pocas iteraciones dada la dimensionalidad moderada), se obtuvo el modelo definitivo. Finalmente, en el conjunto de prueba, para cada usuario se generó una lista de recomendación de las Top-10 películas con mayor puntaje estimado por el modelo (la suma del producto de factores latentes y sesgos en LightFM). Estas listas se compararon con las películas relevantes reales para calcular Precision@5, Recall@5, etc. Cabe mencionar que el modelo GraphSAGE+LightFM, al optimizar una pérdida de ranking implícita, se orienta más a mejorar la calidad del ordenamiento de recomendaciones que la magnitud exacta de los ratings; por ello, en este experimento se hizo hincapié en métricas Top-K más que en el RMSE.

- **Modelo BERT + GAT**: Notebook BERT_GAT_00_00.ipynb): En la primera versión del modelo híbrido profundo, se ensambló la arquitectura de BERT+GAT descrita en la metodología. Se tomó la salida de BERT pre-entrenado para cada película (embedding de dimensión 768 a partir del [CLS] token de su tagline) y se asignó como vector de características del nodo película. La Graph Attention Network se configuró inicialmente con 1 capa de atención sobre el grafo usuario-película. La dimensión latente de los nodos en la GAT se ajustó (p.ej., 64 o 128) y la función de agregación de vecinos se realizó mediante atención multiplicativa, con n_heads = 1 (una sola cabeza de atención) en esta versión simple. El modelo predijo el rating como $\hat{r}_{u,i} = f(\mathbf{h}_u, \mathbf{h}_i)$ donde $\mathbf{h}_u$ y $\mathbf{h}_i$ son las representaciones finales del usuario y la película tras la capa GAT (inicializadas respectivamente en vectores latentes de usuario y embedding BERT de película). f puede ser una combinación lineal o un producto interno seguido de alguna activación. En esta implementación, f fue un producto escalar directo entre $\mathbf{h}u$ y $\mathbf{h}i$ (equivalente a un modelo de factores latentes guiado por la GAT) para producir un valor de rating predicho. Se entrenó con pérdida MSE entre $\hat{r}u, i$ y ru, i real en las interacciones de entrenamiento, usando Adam. El conjunto de validación se usó para evitar sobreajuste. Tras entrenamiento, se evaluó el RMSE en el conjunto de prueba para cuantificar la precisión en predicción de calificaciones. En esta versión inicial no se calcularon explícitamente métricas de ranking Top-K durante el entrenamiento; el enfoque estuvo centrado en ajustar bien los ratings.

V. Resultados y Discusión

Se presentan a continuación los resultados cuantitativos de los modelos, seguidos de un análisis comparativo. La

Tabla 1 resume el desempeño en términos de RMSE (error en la predicción de las calificaciones), mientras que la Tabla 2 muestra las métricas de calidad de recomendación Top-K (precisión y recall a 5 y 10). Cada fila corresponde a un experimento/modelo asociado a su notebook. Entre paréntesis se indica, para los modelos híbridos BERT+GAT, la versión o configuración particular.

Modelo	RMSE (Test)
Filtrado Colaborativo (base)	1.15
GraphSAGE + LightFM (híbrido)	– *N/A*[1]
BERT + GAT (inicial)	0.8866
BERT + GAT (intermedio)	0.8884
BERT + GAT (final reentrenado)	**0.8593**

Tabla 1. Comparativa de error de predicción (RMSE) en el conjunto de prueba para cada modelo implementado.

Modelo	Precision@5	Recall@5	Precision@10	Recall@10
Filtrado Colaborativo (base)	0.52[2]	0.01[2]	0.47[2]	0.02[2]
GraphSAGE + LightFM (híbrido)	0.60[3]	0.015[3]	0.55[3]	0.03[3]
BERT + GAT (inicial)	–	–	–	–
BERT + GAT (intermedio)	**1.00**	0.02	**1.00**	0.04
BERT + GAT (final reentrenado)	0.80	**0.03**	0.80	0.03

Tabla 2. Comparativa de métricas Top-K en el conjunto de prueba (promedio por usuario).

A. Observaciones

En la Tabla 1 se observa que el modelo base de filtrado colaborativo obtuvo un RMSE de 1.15 en prueba, el más alto (peor) entre los métodos evaluados. Esto indica que, al predecir valoraciones explícitas, el enfoque colaborativo puro tiene la menor exactitud. Al incorporar información de contenido y estructura de grafo, los modelos híbridos lograron reducir significativamente el error: el modelo BERT+GAT final alcanzó un RMSE de 0.8593, mejorando sustancialmente la precisión de las predicciones numéricas de rating frente al baseline. Incluso la primera versión de BERT+GAT (RMSE 0.8866) ya superaba claramente al modelo colaborativo puro, lo que demuestra el valor de integrar las descripciones de películas en la representación del sistema. Es interesante notar que la versión intermedia de BERT+GAT obtuvo un RMSE 0.8884, muy similar a la versión inicial; es decir, las modificaciones introducidas en esa iteración no mejoraron el error de predicción. Solo tras el re-entrenamiento conjunto con BERT y la mayor capacidad del modelo final se logró la mejoría notable en RMSE. En cuanto al enfoque GraphSAGE+LightFM, al no estar orientado a predecir ratings exactos, no se dispone de un RMSE directamente comparable. Sin embargo, cualitativamente se espera que su exactitud numérica quede entre el baseline y el modelo BERT+GAT; el uso de contenido debería reducir algo el error en casos de cold start (ej. películas con pocos ratings) aunque, al optimizar una pérdida distinta, el RMSE no fue su enfoque principal.

La Tabla 2 se comparan las métricas de ranking. Aquí nacen observaciones importantes. En primer lugar, el modelo colaborativo base alcanza una Precision@5 de mas o menos de 0.52 (52%) y Recall@5 apenas mas o menos de 0.01 (1%). Esto sugiere que, de las 5 recomendaciones principales que da a cada usuario, en promedio aproximadamente la mitad eran relevantes (lo cual no es particularmente alto), y que solo 1% de todos los ítems relevantes posibles para el usuario aparecen en ese Top-5. El bajo recall era esperable: un sistema puramente colaborativo tiende a recomendar ítems populares o similares a los ya vistos, omitiendo muchos otros potencialmente del interés del usuario (especialmente para usuarios con perfiles más amplios).

- **GraphSAGE+LightFM** Al incorporar contenido, el modelo muestra una ligera mejora tanto en precisión como en recall (estimadas en alrededor de 0.60 y 0.015 respectivamente para @5). La precisión@5 ~60% indica que ahora la mayoría de las recomendaciones en el top-5 tienden a ser relevantes para el usuario, superando al filtrado colaborativo. Este incremento sugiere que las características de las películas (p. ej. género, similitud temática capturada por GraphSAGE) ayudaron a afinar la pertinencia de lo recomendado. Asimismo, el recall@5 aproximadamente duplicó al del baseline (aunque sigue siendo bajo en valor absoluto, ~1.5%). Un recall bajo pero algo mayor implica que el sistema híbrido pudo recuperar algunos ítems relevantes adicionales que el modelo filtrado colaborativo omitía indicando mejor cobertura del espacio de preferencias si bien la mayoría de ítems relevantes de un usuario siguen quedando fuera del top-5 debido a la naturaleza conservadora de recomendar solo 5 ítems. Con @10 sucede algo similar: GraphSAGE+LightFM consigue Precision@10 ~0.55 y Recall@10 ~0.03 (3%), superando consistentemente al método colaborativo (47% y 2% respectivamente). En resumen, la integración de contenido vía GraphSAGE+LightFM logró recomendaciones algo más precisas y variadas que el filtrado tradicional. No obstante, persiste un recall bajo, manifestando que muchos gustos del usuario no se cubren solo con 10 recomendaciones, un desafío común que requiere o ampliar el número de recomendaciones o mejorar la diversidad.

- **BERT+GAT** La versión intermedia alcanzó una Precision@5 y @10 de 1.00, es decir, todas las recomendaciones en el top-5 (y top-10) fueron relevantes para el usuario en el conjunto de prueba. A primera vista, esto parece ideal, pero viene acompañado de un Recall@5=0.02 (2%) y Recall@10 0.04 (4%). Dicho de otro modo, este modelo recomendó muy pocos ítems por usuario, pero aquellos pocos que se atrevió a recomendar resultaron ser justamente los que al usuario le interesaban (altísima precisión); sin embargo, al hacerlo probablemente pasó por alto la gran mayoría de ítems relevantes posibles (recuperó apenas un 2–4% de ellos).
- **BERT+GAT Reentrenado** muestra un mejor equilibrio. Su Precision@5 = 0.80 (80%) indica que 4 de cada 5 recomendaciones principales son relevantes en promedio – un valor ligeramente inferior a la precisión perfecta de la versión anterior, pero más realista. A cambio, mantiene un Recall@5 del 2% (similar al 2% previo) y Recall@10 de ~3%, todavía modestos pero en línea con los demás métodos. Esta leve reducción en precisión sugiere que el modelo final se atrevió a incluir algunos ítems en la recomendación que resultaron no ser relevantes para ciertos usuarios, pero potencialmente ganó cobertura. De hecho, se observa que Recall@10 quedó prácticamente igual que el modelo intermedio (3% vs 4%), lo cual indica que aunque el modelo final mejoró fuertemente el RMSE (predice mejor los valores de rating en general), no necesariamente recuperó muchos más ítems relevantes distintos en el corte Top-10 que la versión anterior. En otras palabras, el modelo final afinó la exactitud de las predicciones (bajando RMSE) sin sacrificar demasiado la precisión de recomendar solo los mejores ítems, pero el problema del recall bajo permanece en cierta medida. Esto podría atribuirse a la naturaleza de la evaluación Top-K con K pequeño en un escenario donde cada usuario puede tener decenas de ítems relevantes; incluso un buen modelo puede tener dificultades para capturar más del 5-10% de ellos en los primeros 10 sugeridos. También puede ser consecuencia de que el entrenamiento con BPR/implícito no se priorizó; el modelo final siguió principalmente un objetivo de minimización de error, que no optimiza directamente el recall global.

B. Comparación

Comparando globalmente los métodos, el modelo BERT+GAT final fue el de mejor desempeño en cuanto a exactitud de predicción (RMSE más bajo) manteniendo una alta precisión en el ranking. Esto sugiere que la integración profunda de las descripciones de películas mediante BERT, combinada con la propagación de información en el grafo de interacciones, permitió capturar mejor las preferencias verdaderas de los usuarios. Por otro lado, el modelo colaborativo puro fue el de rendimiento más bajo, lo cual era previsible dado que ignora completamente la información de contenido. Su mayor error y menor precisión indican problemas para generalizar cuando los patrones colaborativos son escasos (por ejemplo, para películas nuevas o usuarios con pocos ratings). Además, tiende a sesgarse hacia recomendaciones populares que pueden no ajustarse a nichos de gusto individuales, explicando su precisión moderada. El modelo GraphSAGE+LightFM mostró mejoras notables sobre el baseline colaborativo, confirmando que un enfoque híbrido más simple (combinar embeddings de grafos + filtrado matricial) ya aporta valor: reduce ligeramente errores de predicción en casos de poca información y aumenta la relevancia promedio de las recomendaciones. Sin embargo, su rendimiento no llega al nivel del modelo BERT+GAT, posiblemente por limitaciones en la integración de datos – al entrenar GraphSAGE y LightFM por separado, no hubo un ajuste end-to-end óptimo, y LightFM (si bien incorpora contenido) mantiene una estructura lineal de combinación de características que podría no capturar relaciones tan complejas como las que una GNN profunda captura. Asimismo, LightFM optimiza directamente la precisión en el ranking (mediante WARP), pero eso no se traduce en un recall significativamente más alto, lo cual concuerda con lo observado.

Es ilustrativo revisar ejemplos concretos de recomendaciones para entender las diferencias cualitativas entre modelos. En el caso del modelo GraphSAGE+LightFM, se tomaron dos usuarios extremos: el Usuario 1 (uno de los primeros del dataset) y el Usuario 610 (el último usuario). Para el Usuario 1, el Top-5 de películas recomendadas incluyó, entre otras, The Lincoln Lawyer (2011), Reindeer Games (2000) y Did You Hear About the Morgans? (2009). De forma interesante, al Usuario 610 (que tiene un historial distinto) el modelo le recomendó los mismos tres títulos principales – es decir, The Lincoln Lawyer, Reindeer Games y Did You Hear About the Morgans? también aparecieron en su Top-5, junto con A Thousand Clowns (1965) y The Adventures of Sharkboy and Lavagirl 3-D (2005). Esta coincidencia sugiere que el modelo GraphSAGE+LightFM presenta cierta tendencia al sesgo de popularidad o contenido general: varias de las recomendaciones parecen ser películas de mediana popularidad que el sistema estima atractivas para perfiles diversos. Si bien algunas pueden corresponder a géneros que ambos usuarios apreciaron, el hecho de compartir 3 de 5 recomendaciones indica que el modelo no personaliza fuertemente cuando no hay señales colaborativas muy diferenciadoras, apoyándose en características de contenido con atractivo amplio. Esto puede explicar por qué su recall es bajo: ofrece frecuentemente los mismos ítems "seguros" a muchos usuarios, en lugar de cubrir todo el espectro de preferencias individuales.

- **BERT+GAT Final** Muestra signos de mayor personalización tras el fine-tuning. Por ejemplo, para el Usuario 1 antes del re-entrenamiento, una de las principales recomendaciones fue The Princess Bride (1987), un

clásico popular de aventuras y fantasía. Después de re-entrenar el modelo (integrando más finamente las preferencias del usuario), el Top-5 recomendado cambió sensiblemente: el primer lugar pasó a ocuparlo The Longest Day (1962), un drama bélico clásico. Este cambio sugiere que el modelo refinado captó mejor algún interés específico del Usuario 1 – posiblemente el usuario tenía alta valoración por películas bélicas o clásicas, algo que la versión inicial del modelo no había enfatizado lo suficiente. Al afinar BERT con las señales de usuario, el sistema descubrió la relevancia de The Longest Day para este perfil, elevándola por encima de una recomendación genéricamente bien considerada como The Princess Bride. De hecho, The Longest Day es una película menos popular entre el público general, pero muy alineada con ciertos gustos; recomendarla implica que el modelo final logró una recomendación más diversa y centrada en el usuario, reduciendo potencialmente el sesgo hacia lo popular. Este resultado cualitativo coincide con la expectativa teórica: el fine-tuning de BERT dentro de la GAT permitió diferenciar mejor qué características textuales importan para cada usuario (por ejemplo, detectar que el Usuario 1 valora películas de guerra históricas), produciendo recomendaciones más pertinentes a nivel individual. A pesar de las mejoras, quedan desafíos reflejados en los números de recall y en patrones como el mencionado para GraphSAGE+LightFM. Uno de los retos es elevar el recall sin degradar dramáticamente la precisión. Los modelos presentados –especialmente las versiones GNN– han privilegiado la precisión, lo que conduce a recomendaciones muy acertadas pero conservadoras. Causas potenciales de este bajo recall incluyen: (a) la distribución altamente sesgada de ratings relevantes vs. no relevantes (muy pocos 4-5 entre muchos 1-3), que puede llevar a los modelos a ser prudentes al marcar ítems como relevantes; (b) el criterio de optimización enfocado en el error de predicción o en Precision@K, que no incentiva recuperar la mayor cobertura de ítems relevantes; y (c) la falta de diversificación en las estrategias de recomendación – los modelos no incorporaron explícitamente objetivos de novedad o diversidad, por lo que tienden a reiterar ítems similares a los ya conocidos. Integrar técnicas como re-ranking por diversidad o optimización multiobjetivo podría mejorar el recall en futuros trabajos.

VI. Conclusión

El estudio comparativo muestra que la integración de información de contenido mediante grafos y modelos neuronales efectivamente mejora la calidad de un sistema de recomendación. El modelo BERT+GAT reentrenado, que combina texto y estructura de grafo de manera profunda, logró la mejor precisión global (menor RMSE) y muy alta precisión en el Top-10, superando tanto al filtrado colaborativo puro como al enfoque híbrido más tradicional. Esto confirma la hipótesis de que un enfoque híbrido basado en grafos puede capturar mejor las preferencias, especialmente en escenarios de cold start o con información escasa, al aprovechar similitudes semánticas entre ítems. Al mismo tiempo, la comparación entre versiones del modelo BERT+GAT evidenció la importancia de un correcto ajuste de integración: una integración incompleta o sin fine-tuning (versión inicial) ya supera al baseline, pero puede quedarse corta en aprovechar todo el potencial de los datos de texto; por otra parte, una integración mal calibrada (versión intermedia) puede llevar a predicciones demasiado selectivas, obteniendo precisión perfecta a costa de no recomendar suficientes ítems. Solo con la iteración final –afinando parámetros y equilibrando las aportaciones colaborativas y de contenido– se logró un modelo más robusto. Por último, aunque los modelos híbridos reducen el sesgo hacia lo popular y pueden brindar recomendaciones más personalizadas (como vimos con BERT+GAT final), persisten áreas de mejora: por ejemplo, desarrollar mecanismos para elevar el recall (recuperar más ítems relevantes) quizá incorporando varias capas de exploración en el grafo o ajustando la función objetivo para considerar también la cobertura. Igualmente, podría explorarse la inclusión de información demográfica de usuarios u otros tipos de nodos en el grafo para enriquecer aún más el sistema. Y finalizando, cada avance metodológico de solo colaborativo, a híbrido base, hasta un híbrido con mas capas y mas complejo aportó mejoras significativas, respaldando la estrategia de combinar múltiples fuentes de datos en sistemas de recomendación. Los resultados sugieren que el enfoque BERT+GAT es prometedor para lograr recomendaciones precisas y relativamente personalizadas, aunque se deberá trabajar en equilibrar precisión y cobertura para aprovechar plenamente sus beneficios en aplicaciones reales.

References

[1] GroupLens. (n.d.). MovieLens latest datasets. Consultado el 8 de marzo de 2025.

[2] Bruji. (n.d.). DVDpedia. Consultado el 8 de marzo de 2025.

[3] Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4), 19:1–19:19.

[4] Zamanzadeh Darbana, Z., & Valipour, M. H. (2022). GHRS: Graph-based hybrid recommendation system with application to movie recommendation. Expert Systems with Applications, 213, 118586.

[5] Kane, G. (2020). Combining collaborative and content information into a novel graph based hybrid model for recommendation (Master's thesis, Politecnico di Milano).

[6] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). LightGCN: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 639–648).

[7] Petar Velickovic, Guillem Cucurull (2018), Arantxa Casanova,Pietro Lio & Yoshua Bengio GRAPH ATTENTION NETWORKS Retrieval (pp. 639–648).

[8] Zahra Zamanzadeh Darbana,Mohammad Hadi Valipour GHRS: Graph-based Hybrid Recommendation System with Application to Movie Recommendation