ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENERIA DE SISTEMAS

INGENERIA EN CIENCIAS DE LA COMPUTACIÓN

RECUPERACIÓN DE LA INFORMACIÓN

Sistema de Recuperación Multimodal de Información

PROYECTO 2DO BIMESTRE

INTEGRANTES:

DARLIN ANACICHA

MICHAEL PERUGACHI

1. Introducción

En el presente proyecto se desarrollo un sistema de recuperación multimodal de información, utilizando el corpus Consumer Reviews of Amazon Products. Este sistema integra datos heterogéneos, procesando tanto descripciones textuales como representaciones visuales de los productos para habilitar una experiencia de búsqueda híbrida.

La arquitectura propuesta soporta dos modalidades de interacción: la búsqueda visual (Image-to-Product), donde se recuperan artículos semánticamente similares a una imagen de referencia, y la búsqueda textual (Text-to-Product), basada en consultas en lenguaje natural.

Para lograrlo, se combina un recuperador denso basado en CLIP para la indexación multimodal, seguido de una etapa de refinamiento mediante Re-ranking (Cross-Encoder) para maximizar la relevancia. Finalmente, se integra el modelo generativo Gemini para sintetizar respuestas fundamentadas en la evidencia recuperada, todo ello orquestado bajo un sistema de memoria de sesión que permite el refinamiento conversacional de las consultas.

2. Descripción del corpus utilizado

Como se menciono anteriormente, para el desarrollo del presente proyecto se utilizó un conjunto de datos extraído de la plataforma Kaggle, específicamente el dataset titulado “Consumer Reviews of Amazon Products” (versión Datafiniti May 2019). Este corpus proviene del sector del comercio electrónico general, con una fuerte predominancia en dispositivos electrónicos, lo que lo convierte en una fuente ideal para el análisis y desarrollo de sistemas de recuperación de información multimodal en escenarios de ventas reales.
El dataset se estructura en un archivo consolidado (CSV) que incluye:

- **Referencias visuales:** En lugar de archivos locales, el corpus proporciona enlaces web (URLs) en la columna imageURLs, las cuales son procesadas dinámicamente para obtener las representaciones visuales de los productos.

- **Metadatos categorizados:** Cada registro cuenta con información estructurada del producto, destacando columnas como name (título del producto) y primaryCategories (categoría principal), que permiten una organización jerárquica y temática del catálogo.
- **Información textual:** Una dimensión semántica profunda aportada por la columna reviews.text, la cual contiene reseñas y opiniones de usuarios reales. Estos textos no solo describen el producto, sino que aportan contexto de uso y valoraciones cualitativas, enriqueciendo la capacidad del sistema para interpretar la intención de búsqueda más allá de la descripción técnica.

3. Explicación del pipeline y métodos utilizados

Para la ejecución del proyecto se opto por un pipeline que abarca desde el preprocesamiento de datos multimodales hasta la generación de respuestas con IA, se utilizaron los siguientes métodos:

CLIP (Contrastive Language–Image Pretraining) Se utilizó el modelo pre-entrenado clip-ViT-B-32 a través de la librería SentenceTransformers. Su función crítica es proyectar tanto las imágenes de los productos (obtenidas vía URL) como las consultas de texto y reseñas en un espacio vectorial compartido de 512 dimensiones.

- **Bi-Encoder:** Actúa como un codificador dual que permite calcular la similitud semántica entre una consulta (sea texto o imagen) y el inventario de productos mediante operaciones vectoriales, habilitando la búsqueda agnóstica a la modalidad.

Indexación y Recuperación (Retrieval) En lugar de una búsqueda lineal ineficiente, se implementó un sistema de recuperación densa basado en operaciones tensoriales optimizadas (PyTorch).

- **Espacio Vectorial:** Se generó un índice en memoria (corpus_embeddings) que contiene las representaciones vectoriales de todos los productos (Nombre + Categoría + Imagen).
- **Similitud del Coseno:** Se utilizó la función util.cos_sim para calcular la distancia angular entre el vector de la consulta y los vectores del corpus, permitiendo recuperar instantáneamente los Top-K (50) candidatos más relevantes.

Re-ranking Semántico (Cross-Encoder) Para elevar la precisión del sistema (criterio de excelencia), se integró una etapa de reordenamiento explícito utilizando el modelo cross-encoder/ms-marco-MiniLM-L-6-v2.

- **Diferencia con CLIP:** A diferencia del Bi-Encoder (que es rápido pero menos preciso en matices), el Cross-Encoder procesa la consulta y el documento simultáneamente. Esto permite evaluar la relevancia profunda y reordenar los 50 candidatos iniciales para seleccionar el Top-5 definitivo, mejorando drásticamente la calidad de la recomendación final.

RAG (Retrieval-Augmented Generation) con Gemini Es el componente generativo encargado de la interfaz final con el usuario. Se utilizó el modelo Google Gemini (Pro/Flash) vía API.

- **Grounding:** Se diseñó una estrategia de prompt engineering que inyecta el contexto de los 5 productos seleccionados (título, categoría y fragmentos de reseñas) y obliga al modelo a responder "basado exclusivamente" en esa información, mitigando alucinaciones.

Interfaz Gráfica (Gradio) La interacción usuario-sistema se desplegó utilizando la librería Gradio, diseñada para soportar el flujo multimodal. La interfaz incluye:

- **Entradas Híbridas:** Cajas de texto para consultas naturales y un componente de carga de imágenes (Image Uploader).
- **Panel de Análisis Técnico:** Una sección dedicada a la visualización de métricas y depuración, mostrando la comparación entre los resultados antes y después del re-ranking.
- **Galería Visual:** Renderizado HTML dinámico para mostrar las imágenes reales de los productos recomendados.

4. Ejemplos de consultas y resultados obtenidos

Para la evaluación del sistema se realizaron múltiples ejecuciones y pruebas con consultas ingresadas por el usuario, primero se realizo una consulta de tipo text-to-product, esto se puede visualizar en la figura a continuación.

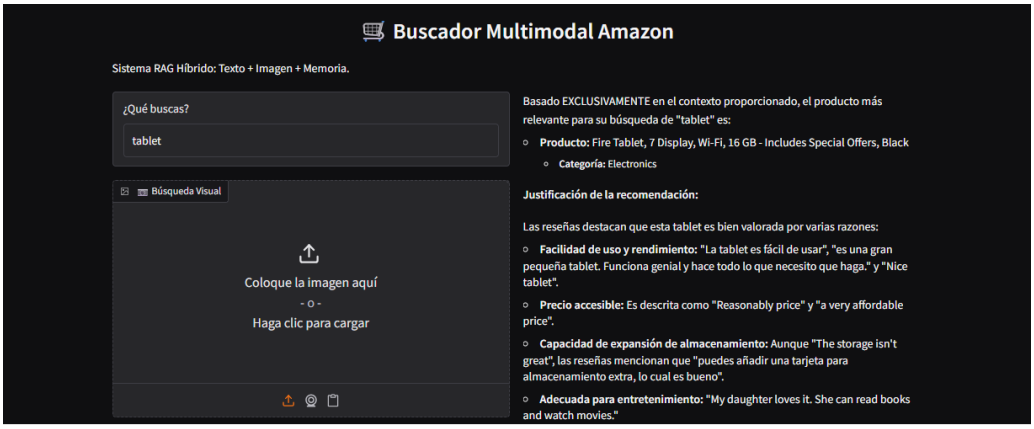

Figura 1. Consulta 1 “text-to-product”

Como se puede observar en la figura 1, se ingresó una búsqueda por texto del producto “tablet”, primero se observa la capacidad del modulo RAG para recuperar el producto más relevante (Fire Tablet 7) y generar una justificación basada exclusivamente en las reseñas

recuperadas, citando textualmente opiniones de usuarios sobre "facilidad de uso" y "precio accesible" para validar la recomendación.

Para complementar lo dicho anteriormente, en la figura 2 que se encuentra a continuación se puede observar el apartado '3. Información no disponible en el contexto', el modelo no intenta inventar especificaciones técnicas que no existen en las reseñas. En su lugar, reporta transparentemente las limitaciones de la información recuperada.

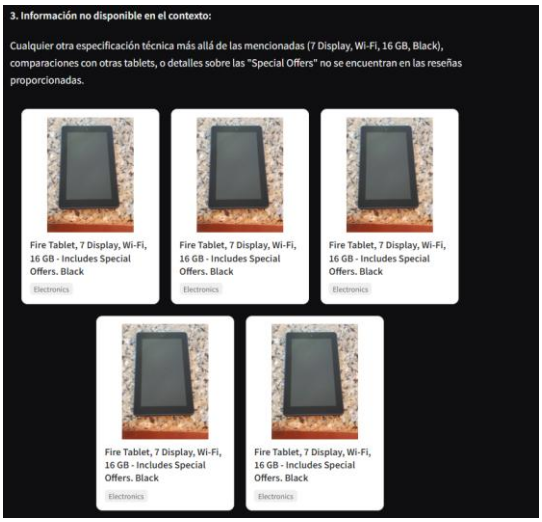

Figura 2. Productos recuperados.

Adicionalmente se presentan la visualización de los candidatos recuperados, confirmando el módulo de recuperación multimodal CLIP.

Ahora realizaremos una consulta de tipo Image-to-Product, para lo cual realizamos una consulta por imagen, como se observa en la figura a continuación se utiliza la imagen de una Tablet.

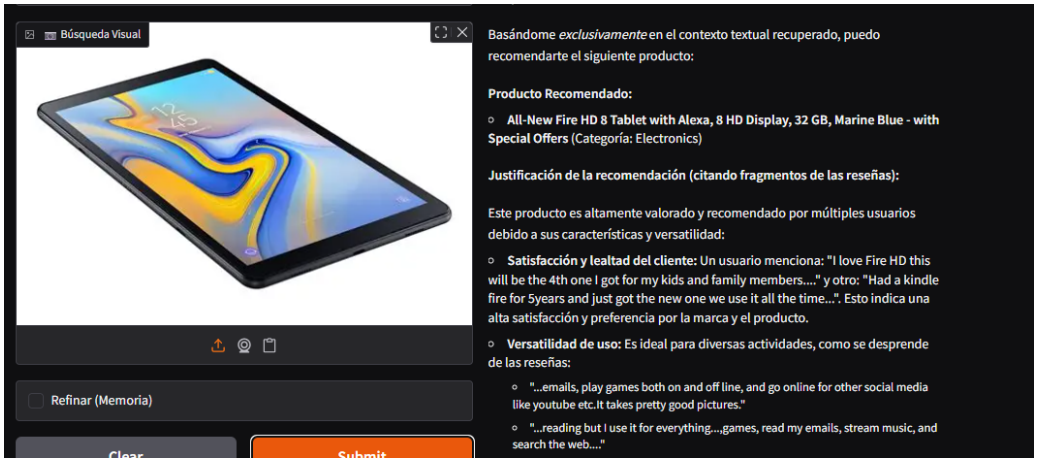

Figura 3. Consulta 1 “image-to-product”.

Como se observa en la figura 3 en el lado derecho se muestra la respuesta RAG generada, donde se justifica la recomendación citando textualmente reseñas de usuarios sobre la durabilidad y versatilidad del dispositivo.

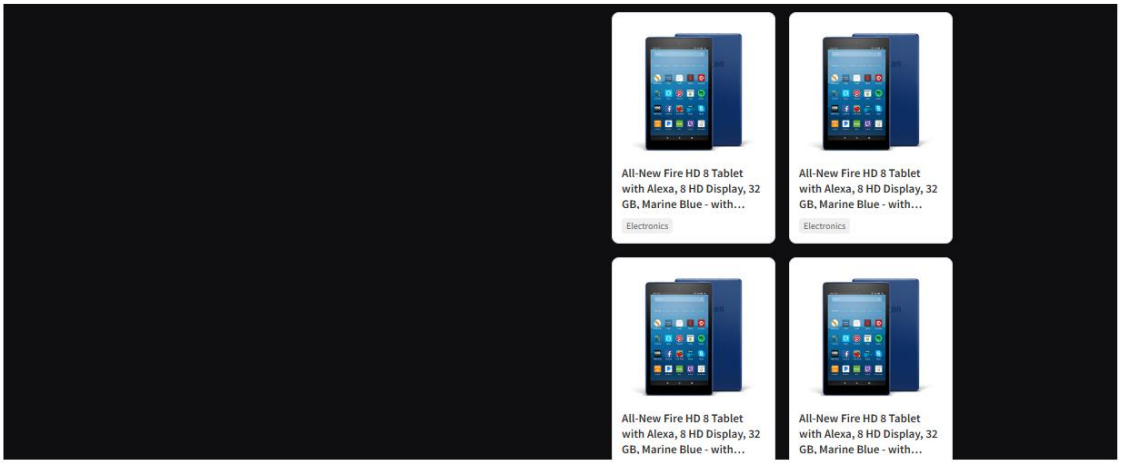

Figura 4. Productos recuperados.

También podemos observar los resultados visuales obtenidos para la búsqueda por imagen, el sistema logra recuperar múltiples productos Fire HD 8 Tablet, demostrando que le motor de búsqueda identifico correctamente el objeto de la imegn de entrada, una Tablet.

5. Análisis cualitativo de los resultados

Después de ejecutar el sistema en diversos escenarios y con múltiples consultas, se observo una alta dependencia de la calidad de la imagen, es decir, cuando la imagen tiene fondos muy recargados y poca iluminación el modelo tiende a recuperar productos no tan cercanos al deseado. Por otro lado, el sistema logra una gran eficiencia al capturar los productos basándose en la figura geométrica, un claro ejemplo son las tablets.

Adicionalmente al ingresar una consulta muy larga el sistema llega a tener dificultades, sin embargo, a pesar de algunas dificultades, el sistema demuestra una gran capacidad para recuperar resultados bastante similares a los deseados, adicionalmente el modelo de IA, es muy bueno para generar las respuestas en base a las descripciones y reseñas de los productos.

6. Conclusiones

- Se logro implementar exitosamente un sistema de Recuperación Aumentada por Generación (RAG) Multimodal, integrando un flujo híbrido que procesa tanto imágenes como texto.
- La utilización del modelo CLIP fue fundamental para crear un espacio vectorial compartido, permitiendo la funcionalidad de búsqueda cruzada (Image-to-Product y Text-to-Product).

- Finalmente, la integración del modelo generativo Gemini con técnicas de Prompt Engineering orientadas al anclaje permitió sintetizar respuestas justificadas. El sistema no solo recuperó productos, sino que utilizó las reseñas de usuarios reales como evidencia para explicar las recomendaciones

7. Video demostrativo

Link video:
https://drive.google.com/file/d/1VCv6MdfuxQWhAL14Q_uHPxPXuQf2Gps6/view?usp=drivesdk