1. Introducción

El presente informe describe el diseño e implementación de un sistema de recuperación de información multimodal basado en embeddings vectoriales, orientado a la búsqueda de productos de Amazon utilizando texto e imágenes. El sistema integra un pipeline completo de *retrieval*, *re-ranking* implícito y generación de respuestas mediante técnicas de *Retrieval-Augmented Generation* (RAG). El objetivo principal es analizar el impacto del re-ranking sobre la calidad de los resultados recuperados y la coherencia de las respuestas generadas, así como identificar las limitaciones inherentes al enfoque propuesto y posibles líneas de mejora en escenarios reales de búsqueda multimodal.

2. Descripción del Corpus

Se utilizó el corpus **Consumer Reviews of Amazon Products**, proporcionado por Datafiniti, el cual contiene más de 34.000 reseñas de productos de Amazon como Kindle, Fire TV y Fire Tablet. El dataset incluye información estructurada y no estructurada, destacando los siguientes campos relevantes para el proyecto:

- Nombre del producto.
- Categorías.
- URLs de imágenes del producto.
- Título y texto de las reseñas.
- Indicador de recomendación del usuario.

El corpus fue obtenido desde Kaggle y posteriormente filtrado para conservar únicamente las columnas necesarias. Se realizó un proceso de limpieza que incluyó normalización del texto, eliminación de valores nulos y selección de una única imagen válida por producto, garantizando consistencia en la representación multimodal.

3. Pipeline del Sistema

El sistema implementado sigue un pipeline compuesto por tres etapas principales: recuperación inicial (*retrieval*), re-ranking y generación aumentada por recuperación (RAG).

3.1. Codificación Multimodal

Para la representación vectorial de los datos se empleó el modelo **CLIP ViT-B/32**, capaz de codificar texto e imágenes en un espacio vectorial compartido. Se generaron embeddings independientes para:

- Texto descriptivo del producto (nombre y categoría).
- Imágenes asociadas a cada producto.

Los embeddings fueron almacenados en archivos persistentes para optimizar el uso de recursos computacionales.

3.2. Indexación Vectorial

Se construyeron dos índices vectoriales utilizando **ChromaDB**:

- Un índice para descripciones textuales.
- Un índice para imágenes de productos.

Cada vector fue almacenado junto con metadatos que incluyen reseñas, categorías y atributos del producto, permitiendo enriquecer la recuperación y facilitar etapas posteriores de re-ranking.

3.3. Retrieval Multimodal

El sistema permite consultas de tipo:

- Text-to-product.
- Image-to-product.
- Text + Image.

Cuando se combinan texto e imagen, el vector de consulta se calcula como:

$$\vec{v}_{query} = \frac{0{,}7 \cdot \vec{v}_{text} + 0{,}3 \cdot \vec{v}_{image}}{2}$$

dando mayor peso al componente textual, debido a su mayor precisión semántica en búsquedas descriptivas.

3.4. Re-ranking y RAG

El re-ranking se realiza de forma implícita mediante la combinación ponderada de embeddings y el uso de metadatos como reseñas y recomendaciones. Posteriormente, los documentos recuperados se utilizan como contexto para un esquema RAG, permitiendo generar respuestas más informadas y alineadas con la evidencia disponible en el corpus.

4. Ejemplos de Consultas y Resultados

A continuación, se presentan ejemplos representativos de consultas realizadas al sistema:

4.1. Consulta Textual

Consulta: “pink table”

El sistema recuperó productos relacionados principalmente con mobiliario y dispositivos electrónicos de color similar, priorizando coincidencias semánticas en la descripción del producto y su categoría. Los resultados incluyen información básica como título, categoría e imagen.

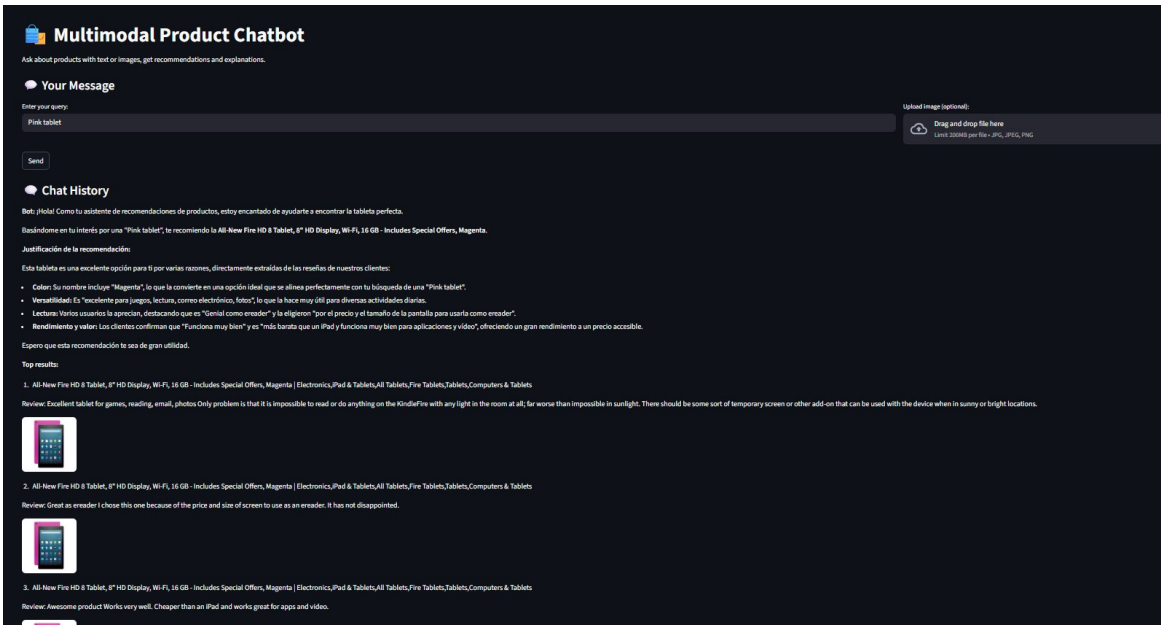

Figura 1: Resultados de la consulta textual “pink tablet”

4.2. Consulta Textual Alternativa

Consulta: “better in red”

De manera análoga, si el usuario realiza la consulta **“better in red”**, el sistema interpreta la intención semántica asociada al color y al tipo de producto, retornando como resultado principalmente *tablets de color rojo*.

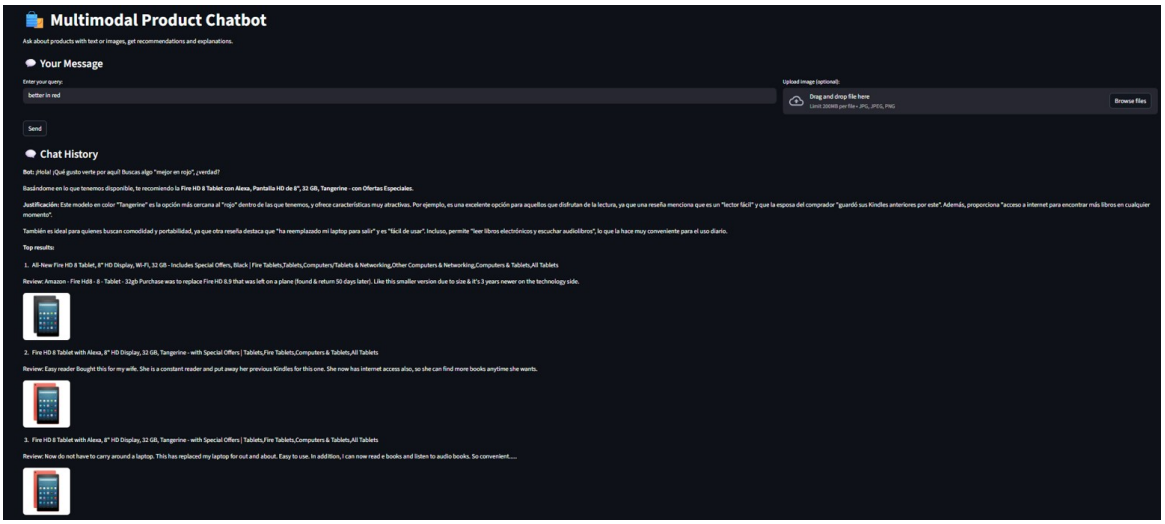

Figura 2: Resultados obtenidos para la consulta textual “better in red”

4.3. Consulta Multimodal

Al combinar texto descriptivo con una imagen de referencia, se observó una mejora en la precisión visual de los resultados, reduciendo la ambigüedad presente en consultas únicamente textuales.

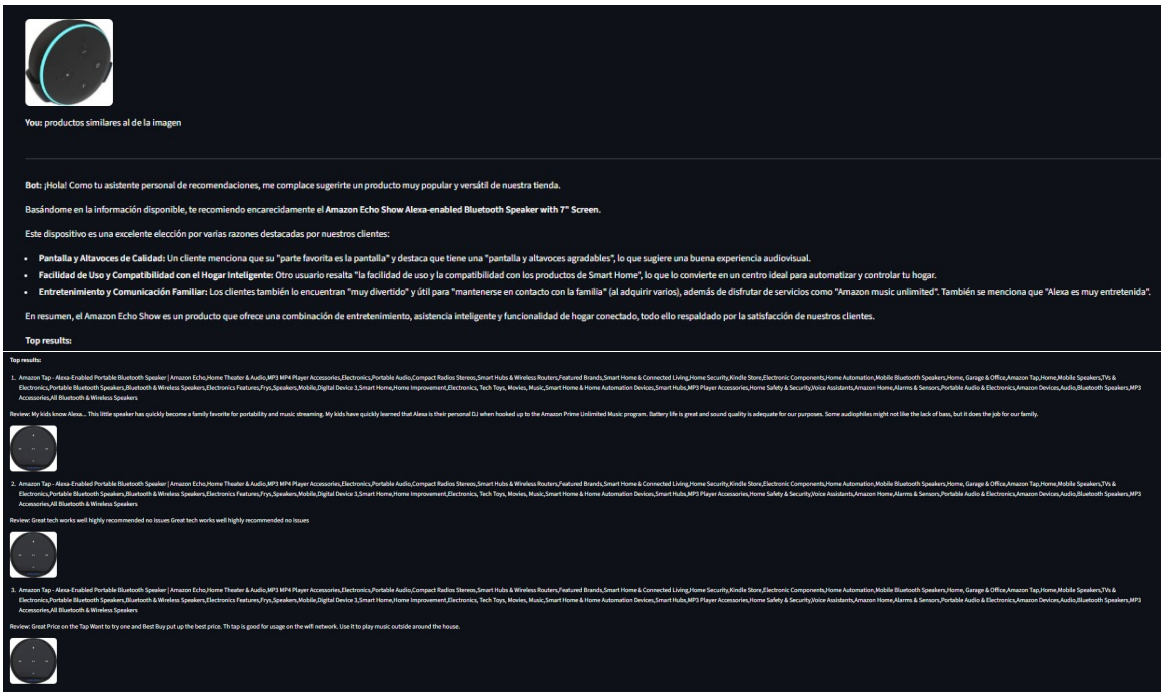

Figura 3: Resultados obtenidos para la consulta de imagen + texto

5. Análisis Cualitativo

El impacto del re-ranking se evidencia en una mayor coherencia entre la intención del usuario y los resultados finales. La ponderación diferencial entre texto e imagen permitió:

- Reducir resultados visualmente irrelevantes.
- Priorizar productos con reseñas alineadas a la consulta.
- Mejorar la calidad del contexto utilizado en RAG.

En cuanto a la generación de respuestas, el uso de información recuperada del corpus permitió producir salidas más específicas y fundamentadas, disminuyendo respuestas genéricas y mejorando la trazabilidad de la información.

6. Análisis Crítico y Limitaciones

A pesar de los resultados favorables observados, el sistema presenta diversas limitaciones que deben ser consideradas para una evaluación objetiva de su desempeño.

En primer lugar, el uso de CLIP, aunque eficiente, presenta limitaciones en la discriminación fina entre productos visualmente similares. En varios casos se observaron resultados con alta similitud visual pero escasa relevancia funcional, especialmente en categorías con diseños homogéneos como tablets y dispositivos electrónicos.

Otro aspecto crítico es que el esquema RAG depende directamente de la calidad de los documentos recuperados. Cuando el retrieval inicial presenta ruido, la generación de respuestas puede incorporar información irrelevante o ambigua, afectando la coherencia final.

7. Posibles Mejoras

Como trabajo futuro, se identifican múltiples oportunidades de mejora. En primer lugar, el re-ranking podría beneficiarse de modelos de aprendizaje supervisado o cross-encoders, capaces de evaluar la relevancia consulta-documento de manera más precisa que una combinación ponderada fija. El uso de modelos multimodales más recientes y de mayor capacidad, así como la integración de feedback del usuario, podrían mejorar tanto la calidad del retrieval como la generación de respuestas en el esquema RAG.

8. Conclusiones

El sistema implementado demuestra el potencial de integrar recuperación multimodal, re-ranking y generación aumentada por recuperación en un único pipeline funcional. Si bien los resultados cualitativos evidencian mejoras en relevancia y coherencia, el análisis crítico revela limitaciones relacionadas con la adaptabilidad del re-ranking, la evaluación cuantitativa y la dependencia del retrieval inicial.

No obstante, el proyecto constituye una base sólida para el desarrollo de sistemas avanzados de búsqueda multimodal, y abre múltiples líneas de mejora orientadas a incrementar la precisión, robustez y escalabilidad del enfoque propuesto.