ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENERIA EN
SISTEMAS INGENIERÍA EN CIENCIAS
DE LA COMPUTACIÓN

RECUPERACIÓN DE LA INFORMACIÓN

TEMA: Sistema de Recuperación Multimodal de Información

INTEGRANTES:

María Fernanda Arias

Andrés Jiménez

1. **Introducción**

   En este proyecto se diseñó e implementó un sistema RAG con enfoque multimodal, utilizando un conjunto de datos proveniente de la industria del comercio electrónico enfocado en productos de moda. El dataset incluye imágenes de productos, metadatos categorizados y descripciones detalladas en texto, lo cual permitió integrar distintas modalidades de entrada y salida.

   El sistema desarrollado opera en dos modos principales. En el primer modo, a partir de una imagen de entrada, se recuperan productos visualmente similares desde el corpus y se genera una respuesta textual que explica o narra las características comunes de los resultados. En el segundo modo, el usuario introduce una consulta en lenguaje natural y el sistema recupera imágenes relevantes junto con sus descripciones, para luego generar una respuesta textual coherente con la intención de la búsqueda.

   Para lograr esto, se combinaron técnicas de recuperación de información visual y textual mediante el uso de modelos de embeddings (CLIP) y se integró un modelo generativo para la construcción de respuestas explicativas (Gemini). De esta forma, el sistema no solo recupera información relevante, sino que también la presenta de manera interpretativa, mejorando la experiencia de búsqueda en contextos donde la moda y la estética visual son clave.

2. **Descripción del corpus utilizado.**

   Para el desarrollo del presente proyecto se utilizó un conjunto de datos extraído de la plataforma Kaggle, específicamente del dataset titulado "Fashion Product Images Dataset" publicado por Param Aggarwal. Este corpus proviene del sector del comercio electrónico y está compuesto por una amplia colección de productos de moda, lo que lo convierte en una fuente para el análisis y desarrollo de sistemas de recuperación y clasificación de información multimodal.

   El dataset incluye:

   - Imágenes de alta calidad de productos, disponibles en resoluciones estándar y en alta resolución.
   - Metadatos organizados en un archivo llamado styles.csv, donde cada producto está identificado por un ID único. Este archivo contiene diversas etiquetas categóricas como masterCategory, subCategory, gender, entre otras, que describen jerárquicamente el tipo de prenda o accesorio.
   - Descripciones textuales adicionales en archivos individuales .json, los cuales comentan sobre las características de cada producto, proporcionando así una dimensión semántica que complementa la representación visual.

3. **Explicacion de los modelos y métodos utilizados.**

   Dentro del proyecto se utilizaron diferentes modelos y herramientas que nos permiten tener un mejor control de los procesos que debemos realizar y las ejecuciones para realizar un correcto preprocesado, generación de embeddings, guardado en una base de datos vectorial y posterior recuperación de la información ya sea por ingreso de texto o por ingreso de imagen por parte de los usuarios.

   Dentro de los modelos utilizados se encuentran:

- CLIP (Contrastive Language–Image Pretraining)
  Nos permite tener una representación conjunta tanto de texto como de imagen en un mismo espacio vectorial. Su función básica dentro del proyecto es realizar las traducciones de imágenes y textos a vectores, todos ellos representados en el mismo espacio vectorial.
  Los métodos que se utilizaron dentro de CLIP fueron:
  - get_image_features: utilizado para imágenes del corpus y consultas por imagen.
  - get_text_features: se utiliza para la realización de consultas por texto y descripciones.
- FAISS (Facebook AI Similarity Search)
  Es el lugar donde se guardarán los embeddings generados después del preproceso de las imágenes y descripciones, nos permitirá encontrar los vectores más similares dentro de un índice.
  - clip_image_index.faiss: Es el método que permite contener los vectores de las imágenes en el dataset.
  - fashion_text_index.faiss: Es el método que contiene los vectores de las descripciones del dataset.
- RAG (Retrieval-Augmented Generation)
  Es el encargado de generar texto informativo usando un modelo de lenguaje, en este caso se utilizó GEMINI, y para producir las respuestas, se utilizó el contexto generado por las descripciones de los resultados relevantes recuperados para cada consulta.
- INTERFAZ WEB (GRADIO)
  La aplicación se desplego dentro del ambiente de Cola utilizando la librería de Gradio, dentro de la cual se puede realizar:
  - Entradas por imagen y texto
  - Visualización de imágenes similares para cada consulta
  - Muestra de descripciones generadas por IA

4. Ejemplos de consultas y resultados.

A continuación, se muestra un ejemplo por búsqueda por texto de un vestido y se presentan imágenes similares y una descripción de estas.

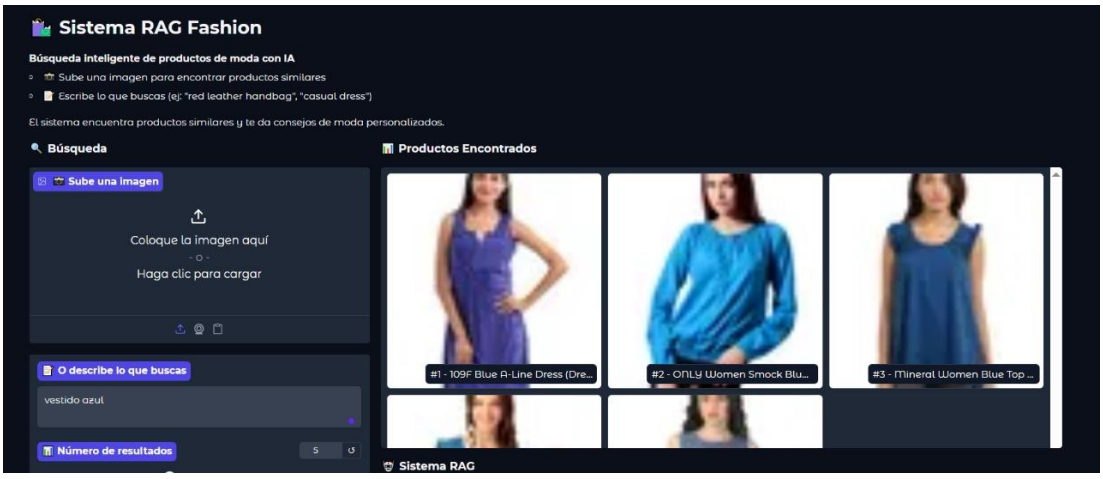

Figura 1. Búsqueda por Texto

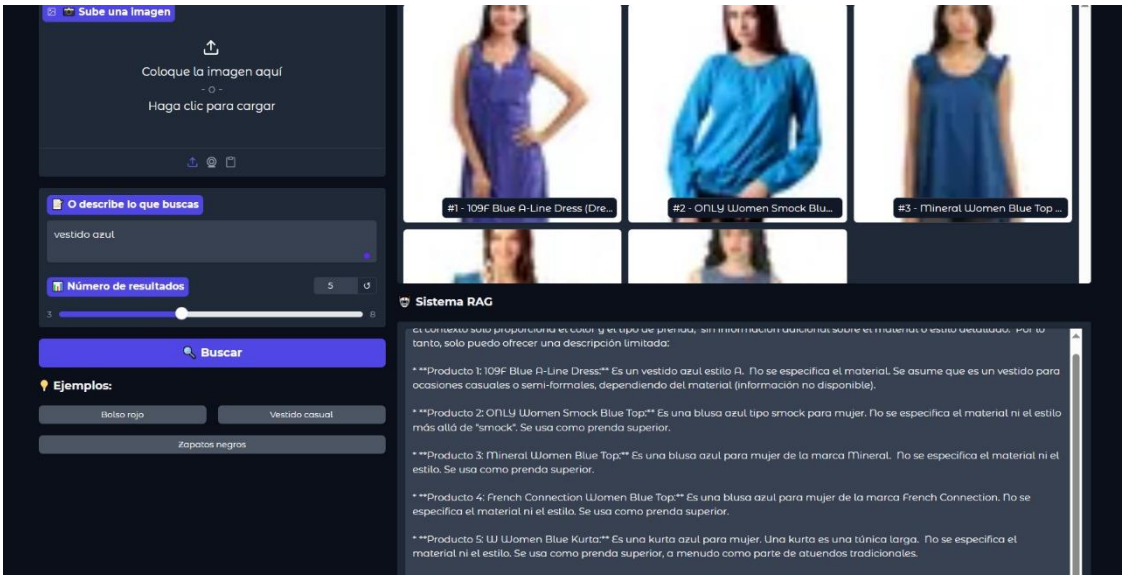

Figura 2. Descripción de la búsqueda por texto.

Como se observa en la figura 3, la búsqueda por imagen de un par de zapatos y se presentan imágenes similares y una descripción de estas.

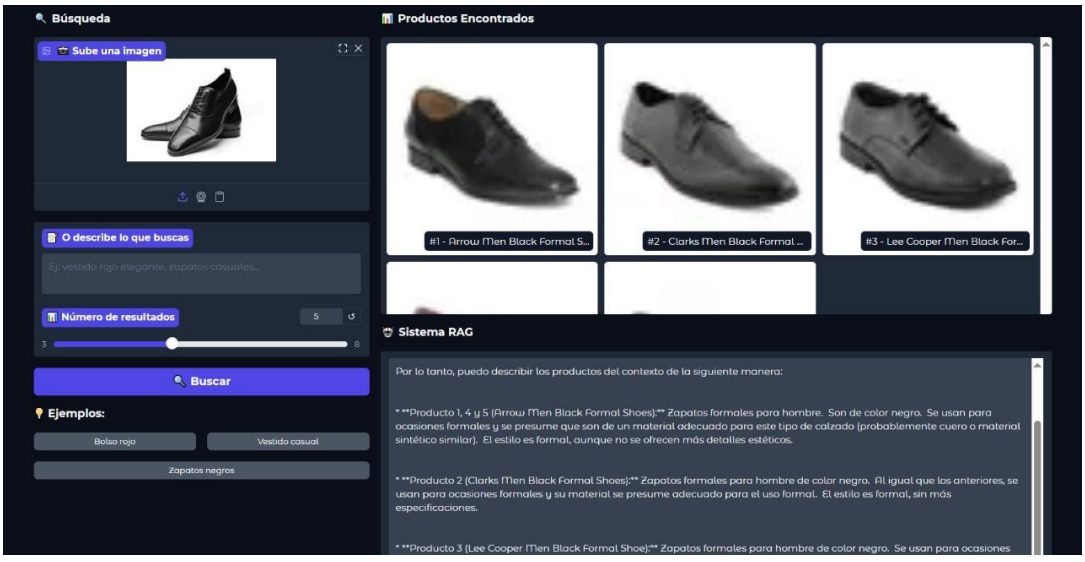

Figura 3. Búsqueda por imagen

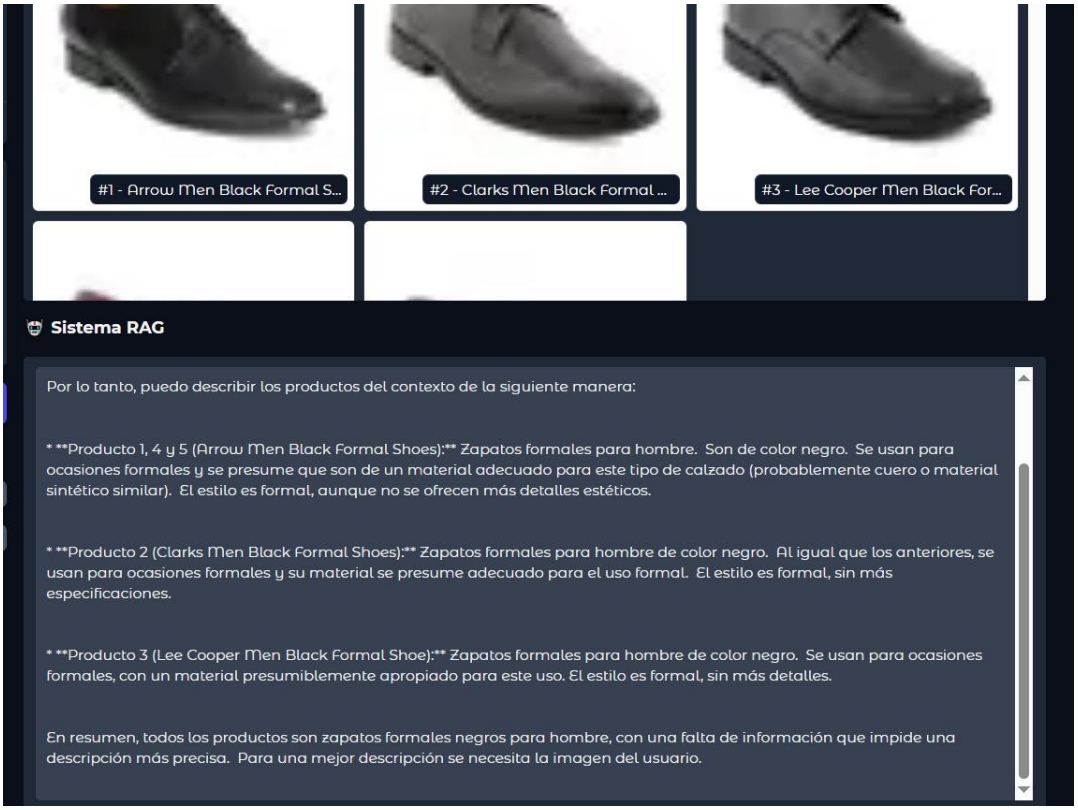

Figura 4. Descripción de la búsqueda por imagen.

5. Análisis cualitativo de los resultados.

Una vez analizado el programa y siendo ejecutado en diferentes ambientes de prueba, nos dimos cuenta de que, para consultas ingresadas por texto, el programa tiene dificultades al momento de que el usuario ingresa consultas especificas o muy largas.

También se tiene en cuenta, que cuando se ingresa consultas por imágenes, cuando la imagen tiene fondo muy obscuro o difuminado, la recuperación de imágenes tiende a demorarse en encontrar los resultados.

Los tipos de consulta que mejor genera resultados son las que especifican el tipo de ropa (zapatos, camiseta, vestido).

Finalmente, comprobamos que a pesar de las dificultades, el programa recupera resultados bastante similares a los esperados y el model de IA, Gemini, es adecuado para generar las respuestas en base a los contextos generados por las descripciones de cada imagen.

6. Conclusiones

- Se logró implementar un sistema de Recuperación Aumentada por Generación combinando imágenes procesadas con CLIP y descripciones textuales, lo cual permitió responder queries en lenguaje natural de forma precisa y contextual.
- El modelo CLIP permitió extraer representaciones semánticas robustas de imágenes de productos, mejorando la recuperación de información basada en similitud visual.

- El índice construido con FAISS facilitó una recuperación rápida y eficiente de los productos más relevantes en base a los embeddings. Al combinar los metadatos textuales recuperados con Gemini, el sistema generó respuestas coherentes y útiles, incluso con consultas ambiguas o abiertas.