Document-Based Question Answering System with Retrieval-Augmented Generation, Translation, and Sentiment Analysis

Objective:
This document-based question answering (QA) system leverages **Retrieval-Augmented Generation (RAG)**, **document processing**, **translation**, and **sentiment analysis** to provide users with a comprehensive solution for extracting relevant information from documents. The system allows users to upload documents in multiple formats (PDF, DOCX, TXT), split and process the content into manageable chunks, and generate contextual answers to user queries. It also supports translation of the answers into several languages (French, Spanish, or German) and includes sentiment analysis to gauge the tone of the responses.

How the Problem is Solved:

The first challenge in this system is the processing of documents in various formats (PDF, DOCX, TXT). To address this, the system provides a file upload mechanism where users can submit documents in any of these formats. The text is then extracted using appropriate libraries: PdfReader for PDFs, python-docx for DOCX files, and direct reading for TXT files. Following text extraction, the content is split into smaller, manageable chunks using the RecursiveCharacterTextSplitter from LangChain. This ensures the text is in a format suitable for embedding and retrieval, with a chunk size of 500 characters and an overlap of 100 characters to prevent truncation of context.

After text extraction and splitting, the next step involves converting the chunks into embeddings. The system uses the SentenceTransformer (paraphrase-mpnet-base-v2) model to generate vector embeddings for the text, which are then indexed using FAISS (Facebook AI Similarity Search). FAISS enables efficient similarity search and retrieval of relevant document chunks when users query the system. The embeddings are stored in LangChain's FAISS vector store, making the system highly scalable and capable of quick retrieval for real-time applications.

To generate accurate answers, the system employs the Retrieval-Augmented Generation (RAG) model. When a user asks a question, the system retrieves the most relevant document chunks using the RetrievalQA chain. The retrieved context is then passed to the Google Gemini Generative AI model (gemini-pro), which produces the final response. The answer generation process ensures that the response is contextually relevant by utilizing the specific information retrieved from the document.

Given that not all users speak English, the system also provides multi-language support. Using the Helsinki-NLP MarianMT model, the system translates the generated answers into French, Spanish, or German based on the user's selection. Translation is performed only when the user requests a language other than English, ensuring accessibility for a broader audience.

In addition to the core features, the system includes sentiment analysis to assess the tone of the generated responses. A pre-trained sentiment analysis model is integrated to determine whether the answer is positive, neutral, or negative. This allows users to gain insights not only into the content but also into the emotional undertone of the response, enhancing the overall experience.

Finally, the user interface is built using Streamlit, ensuring a smooth and intuitive experience for non-technical users. The interface allows users to upload documents, ask questions, select the desired language for the answer, and view the sentiment analysis result. It also includes display areas for both the original and translated answers, with clear instructions and visual indicators to enhance user interaction.

Why This Approach Works:

This approach addresses several key challenges effectively. First, document handling is optimized through the splitting of large documents into smaller chunks, preventing overload and ensuring that only relevant sections are retrieved based on user queries. Embedding and indexing using FAISS ensures that the system can perform fast, efficient similarity searches, enabling real-time responses to user questions. The use of the RAG model guarantees that answers are accurate and contextually relevant, making the system more effective than traditional models that rely on generic answers.

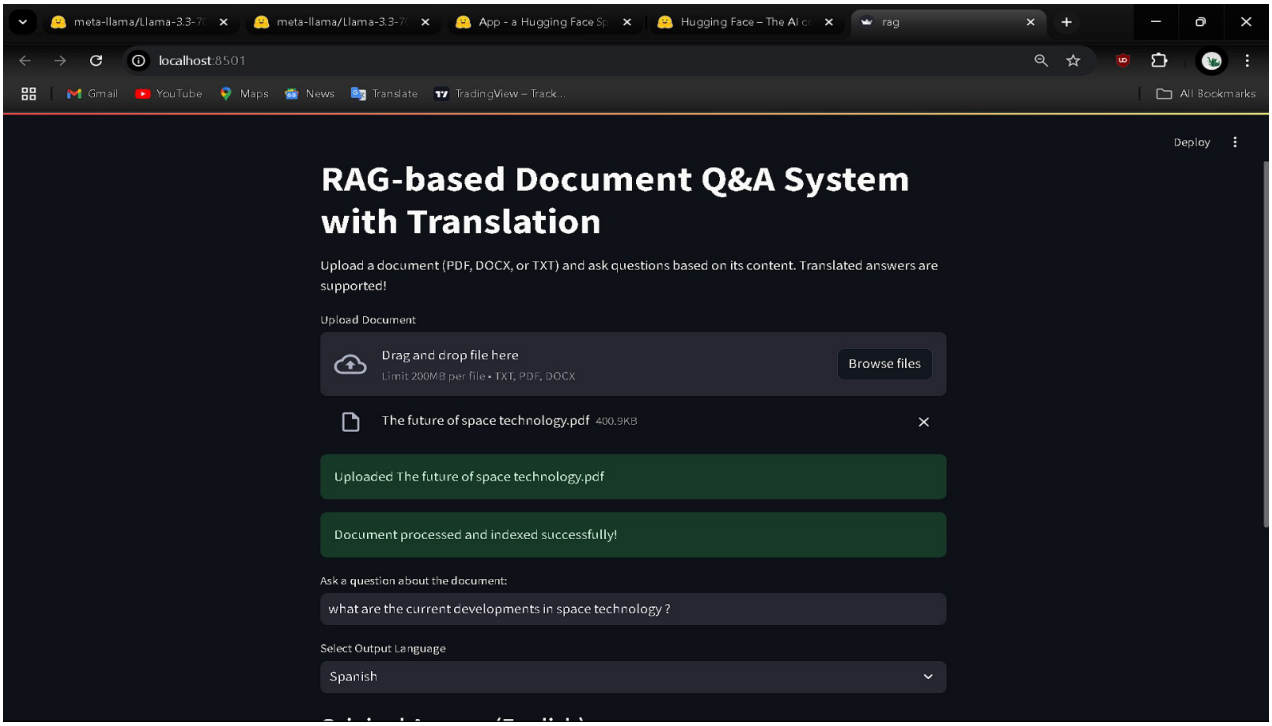
Moreover, the system's flexibility allows users to upload documents in multiple formats without worrying about conversion, and it supports multiple languages to cater to a global audience. The integration of sentiment analysis further enhances the user experience by providing insights into the emotional tone of the generated answers. This feature is particularly useful for applications where understanding sentiment is important, such as customer support or social media analysis.

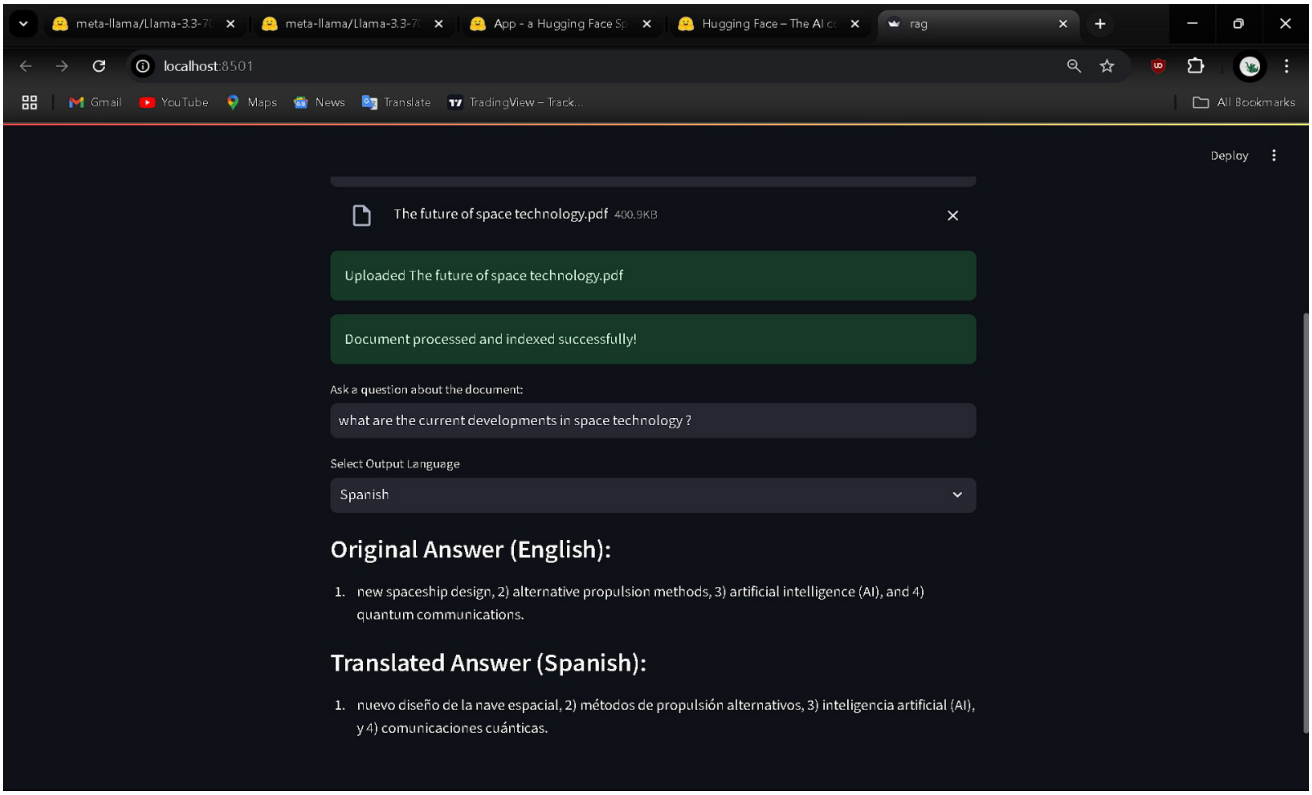
The user interface, built with Streamlit, provides a seamless, interactive experience, making it easy for users to interact with the system without requiring technical knowledge. The combination of RAG, document processing, translation, and sentiment analysis offers a comprehensive solution that delivers accurate, contextual, and translated responses, while also providing valuable sentiment insights.

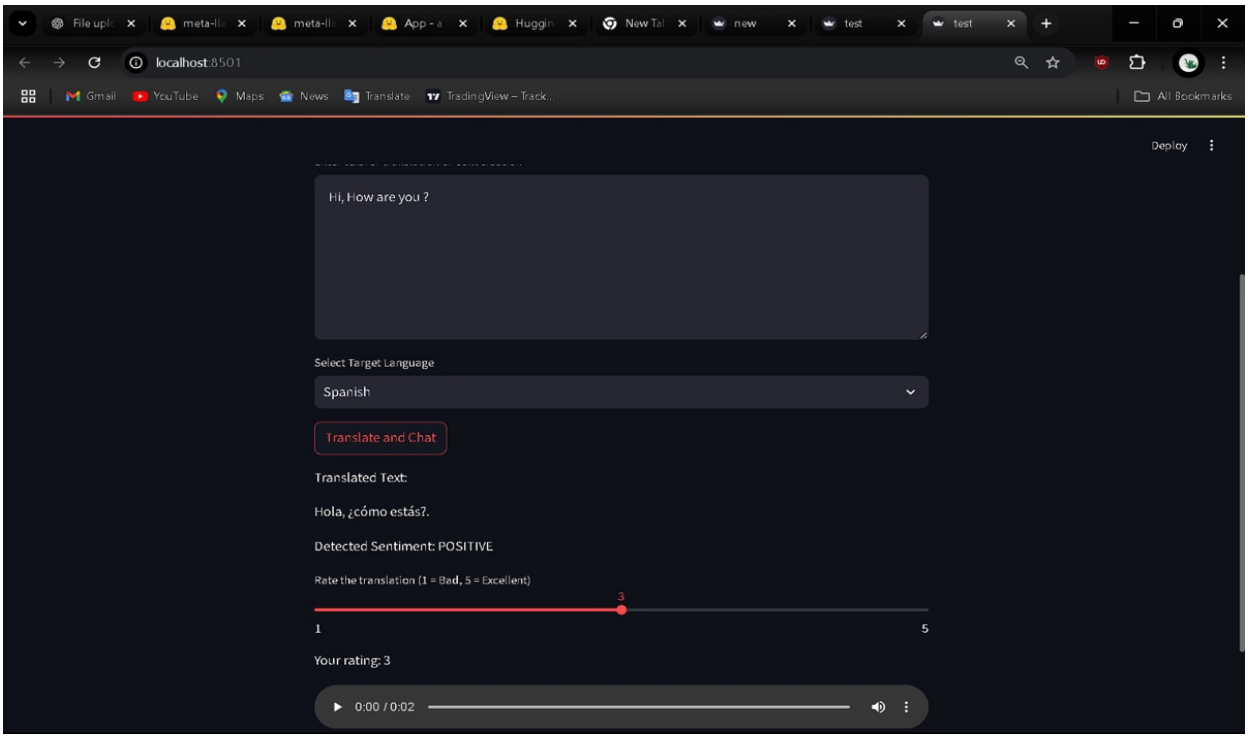
Conclusion:

This document-based question answering system solves the problem of efficiently processing large documents, retrieving relevant information, and providing accurate and contextual answers. With the added capability of sentiment analysis and multi-language support, the system is well-suited for a wide range of use cases. By combining RAG, document processing, embeddings, translation, and sentiment analysis, it offers a scalable, flexible, and user-friendly solution that enhances both the accuracy and accessibility of answers. This approach is ideal for users in diverse, multilingual environments seeking detailed insights from complex documents.

Output Screenshots:






localhost:8501
Deploy
Hi, How are you ?
Select Target Language
Spanish
Translate and Chat
Translated Text:
Hola, ¿cómo estás?.
Detected Sentiment: POSITIVE
Rate the translation (1 = Bad, 5 = Excellent)
3
1
5
Your rating: 3
0:00 / 0:02