FinFeed project

Objective:
Create an AI assisted resource (a conversational bot) that efficiently collates finance and economy related current news within a specified time period (eg. 3 days) from a YouTube news channel (eg. Yahoo Finance) where people can ask about finance related current news that simultaneously shows the sentiment associated with the context of the given query.
The following are the steps to the project:

1. ***Data Collection:*** Acquire relevant data to Econ and Finance. pytube library or Youtube API to download YT videos and generate audio.
2. ***Data Processing:*** once we have the audios, we do the following:
    - Whisper openAI for Audio-to-Text conversion for generating text from audio and store them in text files
    - Data Cleaning: remove irrelevant information, standardize text format
    - Tokenization: convert text into tokens so they can be processed by the model
    - Annotation(?): Label data (should we do this?)
    - Convert the text to vectors/embeddings using the OpenAi embedding model either **text-embedding-3-small** model or more powerful **text-embedding-3-large** model (we are still experimenting).
    - Store those embeddings in the **Pinecone vector database**. Pinecone is a cloud-native vector database that handles high-dimensional vector data. It is designed to store, index, and retrieve high-dimensional vectors, making it ideal for machine learning applications such as text classification
3. ***Model Selection:*** once we process the data, we do the following:
    - Choose a pre-trained LLM model, hugging face sentiment analysis model to do sentiment analysis on the context of the query (?)
    - Fine tune the model to improve performance

4. Implementing a RAG framework to retrieve context for the query (Langchain) [I'm lost on this step and where it's at]