Product Review Analysis Report

This report outlines the analysis of a dataset containing 1,000 customer reviews for electronic products across categories such as Smartphones, Laptops, Smart Home, Wearables, and Audio. The dataset includes features like review text, ratings, and labeled sentiments. The project addresses three tasks: developing a document processing pipeline for semantic search, applying a Large Language Model (LLM) for summarization and Q&A, and conducting sentiment analysis with visualizations.

1. Document Processing Pipeline

Reviews were loaded from a text file using LangChain's TextLoader and split into documents with CharacterTextSplitter (chunk size: 500, overlap: 50). Embeddings were generated for each chunk using OpenAIEmbeddings and stored in a Chroma vector database.The retrieve_semantic_reviews function retrieves the top_k semantically similar reviews for a given query, enabling concept-based searches. This leverages the vector database for efficient similarity matching. The retrieval of reviews is based on conceptual similarity and not, ust keyword matching.

2. LLM Application

The generate_category_summary function samples up to 10 reviews per category and uses OpenAI's GPT-3.5-turbo to generate summaries, identifying overall performance, praised features, common issues, and improvement suggestions.

The product_qa function retrieves relevant reviews via semantic search and uses the LLM to answer user questions based on those reviews. The Q&A system does the following:

1. Uses semantic retrieval to fetch top-k relevant reviews for the question.
2. Grounds gpt model with a system prompt restricting answers to supplied context
3. Returns the LLM answer strictly based on provided text.

3. Sentiment Analysis & Classification

A pre-trained model (j-hartmann/sentiment-roberta-large-english-3-classes) was used to classify review sentiments as positive, neutral, or negative.Predictions were compared to the dataset's sentiment labels, which had an Accuracy of 0.805. Generated the following plots:

1. Sentiment Distribution by Category
2. Sentiment Trends Over Time

3. Rating vs. Sentiment Heatmap
4. Sentiment by Top Features

All functions were integrated into a Streamlit app with interactive tabs:

- Overview: counts, averages, and distribution charts.
- Document Search: semantic query interface.
- LLM Insights: on-demand category summaries.
- Sentiment Analysis: real-time classifier and pre-rendered visualizations.
- Q&A System: interactive product question answering with source transparency.

Conclusion

The project successfully implemented a semantic search pipeline, LLM-driven insights, and sentiment analysis. This framework offers valuable tools for understanding customer feedback and guiding product enhancements.