Introduction

This report entails the findings and methodologies used on the LLM and vector database challenge The steps entail creating a data preprocessing pipeline, using a large language model to extra summaries and carry out question-answering, as well as carry out sentiment analysis and classification.

Document Processing Pipeline

- *Creating a processing pipeline that extracts key information:* The first step on this task was to clean the review text to obtain standard text for analysis. To extract phrases and better tokenization, the text was normalized by eliminating extra spaces, punctuation and case sensitivity. Ratings trend by category was then calculated by looking at sentiment distribution and average rating to identify the patterns of customer satisfaction by category. To extract useful adjective-noun combinations like poor quality, the pre-processed reviews were then tokenized with spaCy model and to finally identify the dominant themes behind the sentiments feedback. The most common compliments and complaints could be marked by dividing the reviews by sentiment so that actionable suggestions for product or service improvement could be obtained. From the analysis, positive remarks out-weighted the negative and neutral sentiments, implying that the customers are satisfied with product performances and technological innovations such as facial recognition features and wearables, as shown by the highest number of such mentions.
- *Creating embeddings:* SentenceTransformer (all-MiniLM-L6-v2) was utilized to generate dense semantic embeddings for all reviews because of its capability in encoding the sentence meaning in a lightweight way and it is fast. To avoid recalculations, the embeddings were cached.
- *Storing and retrieving reviews:* A FAISS index was then built using L2 distance to allow for efficient search of similarities even for very big datasets. This made it possible to obtain not only keyword- driven reviews that were semantically close to a user query based on meaning. One example used was “great battery life” which resulted in reviews such as long-term usage and reliability of products like PowerLaptop 15 and MobiElite 10. This gives insight into customer preferences on battery performance, even though the word battery was not in every review. By doing this, businesses are able to obtain concealed customer insights that are beyond shallow keywords.

LLM Application

To get deep insights from unstructured reviews, large language models were used in three ways:

- *Generating concise summaries of product performance:* Firstly, a category-wise summarization system was built using the facebook/bart-large-cnn model because of its ability to read and process long, messy text into clean summaries and preserve key information due to its denoising encoders.

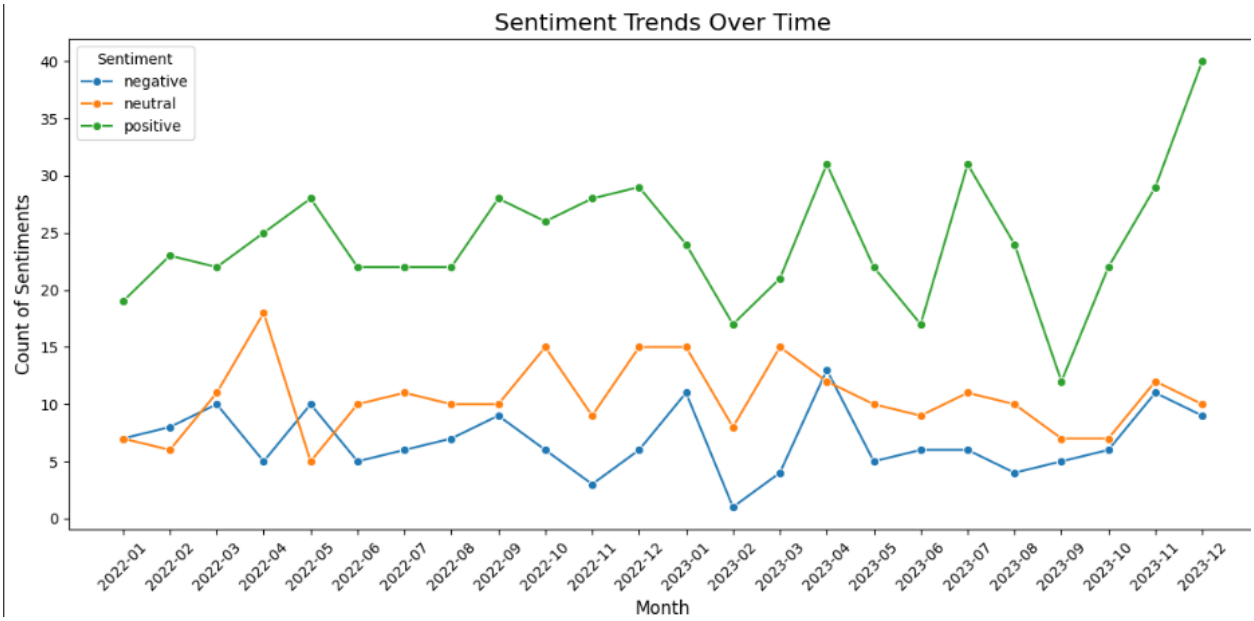
Each review was concatenated and summarized within each product category resulting to short readable summaries that highlighted the key themes per category. For example, customers showed a preference for facial recognition capabilities when shopping for smartphones while showing dissatisfaction in PixelView7 build quality. This allows for the correct and relevant enhancements.

- *Creating a Q&A system: The* second application was using the distilbert-base-uncased-distilled-squad in the Question-Answering system. It was used because of its faster speed in comparison to larger models like BERT and its ability to extract specific information from large datasets This is significant in cases where users can ask specific questions and get responses that are contextually correct without the trouble of going through the data manually.
- *Identifying common issues and praised features:* Review sentiments across the product categories were analyzed to identify common issues and praised features. The dataset was first grouped by category. To identify the top recurring issues and get actionable insights the reviews were separated based on the sentiments, identifying complaints such as connectivity in categories like smartphones. Praised features were also identified such as ease of use of products. All these issues help in improving areas customers find dissatisfying while emphasizing marketing campaigns on already loved features, ultimately improving brand loyalty and customer satisfaction.

Sentiment Analysis and Classification

Sentiment classification system: The first step was checking for the class distribution of the sentiments to identify any present imbalances. The precomputed embeddings were then loaded, followed by splitting the data into train and test splits for training and evaluating the model. Synthetic Minority Oversampling Technique (SMOTE) was used to ensure class balance, making every class more represented and improving model generalizability and fairness. cardiffnlp/twitter-roberta-base-sentiment-latest, a pre-trained RoBERTa model with high performance on text sentiment tasks was used through a HuggingFace pipeline. If not handled, the observed unbalanced class distribution would have led to biased model predictions. Context-aware predictions were made through the use of the model, critical to right perceptions of customer feedback, improving the credibility of the results from the analysis.

Sentiment trends over time: Analyzing the sentiments trend over time, there is an increase in positive sentiments towards the end of 2023. There is stability on neutral sentiments while there are relatively low negative sentiments, which increase towards the end of 2023 as well. The increasing positive sentiments suggest increase in customer satisfaction which could be as a result of better service or product improvements. Other users are neither impressed nor dissatisfied as shown by the stable neutral sentiments. However, the trend observed towards the end of 2023 could imply evolving issues such as dissatisfaction in product prices or product flaws, which need to be uncovered and addressed on time. The trends can be observed on the image below:



Recommendations

1. Increase investments on advertising products that seem to exceed customer expectations such as smartwatches because they are strong product lines.
2. Focusing on improving user experience especially on products that have installation issues such as smart home products to increase positive sentiments.
3. Opportunities for upselling and cross-selling should be created to increase overall customer expectations and satisfaction because they love products of all categories.