Problem Description

Small and medium-sized businesses receive large volumes of customer feedback across channels such as Yelp, surveys, and social media, but often lack effective tools to extract structured, actionable insights. Although this feedback contains valuable information about customer preferences and pain points, it is largely unstructured, making it difficult to identify recurring themes and understand how customer concerns change over time.

This project designs and implements an NLP system that analyzes Yelp reviews to identify recurring customer experience themes for a business and track how those themes evolve over time.

Implementation Overview

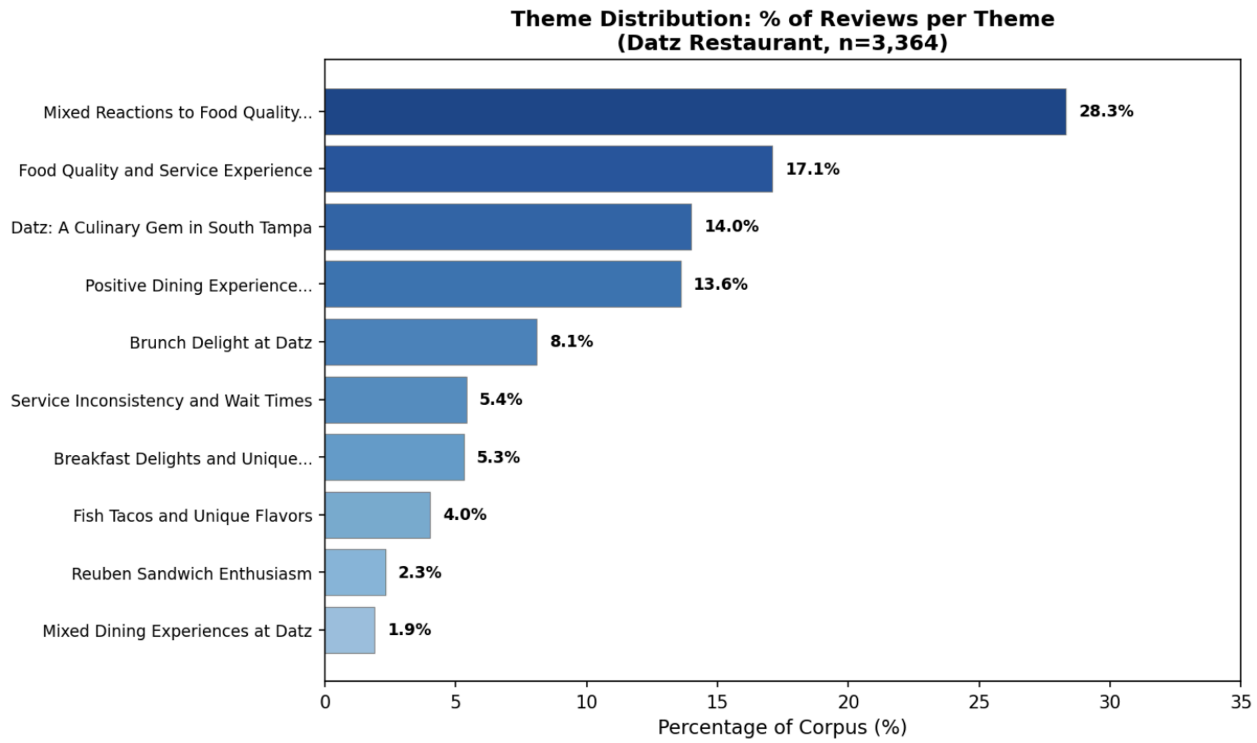
The system is applied to Yelp reviews for Datz Restaurant in Tampa, Florida, using approximately 3,300 reviews collected over a 13-year period from 2009 to 2022. Reviews are loaded efficiently using PyArrow pushdown predicates, and all available reviews are used in Version 1 of the project.

Review text is deduplicated and cleaned through normalization of HTML and URLs, casing, Unicode characters, whitespace, and contractions. Short, low-information texts are removed to produce analysis-ready input for downstream modeling.

Cleaned reviews are embedded using a SentenceTransformer model (all-MiniLM-L6-v2), generating 384-dimensional semantic vectors. Unlike TF-IDF, these embeddings capture semantic meaning, allowing conceptually similar reviews to be close in embedding space even when they share little or no word overlap.

Theme discovery is performed in two stages. First, hierarchical clustering is used to separate the main body of reviews from noise, including off-topic content, foreign-language reviews, and outliers. KMeans clustering is then applied to the remaining reviews to group them into semantic themes. The number of clusters is selected using a k-selection evaluation that balances intra-cluster cohesion with inter-centroid separation.

The figure below shows the distribution of review counts across discovered themes, highlighting which customer experience topics dominate overall discourse.



Each theme is labeled using a combination of TF-IDF keywords unique to the cluster, representative reviews closest to the cluster centroid, and extracted noun phrases that capture common multi-word customer expressions. These signals are combined using a large language model to generate concise, human-readable theme summaries.

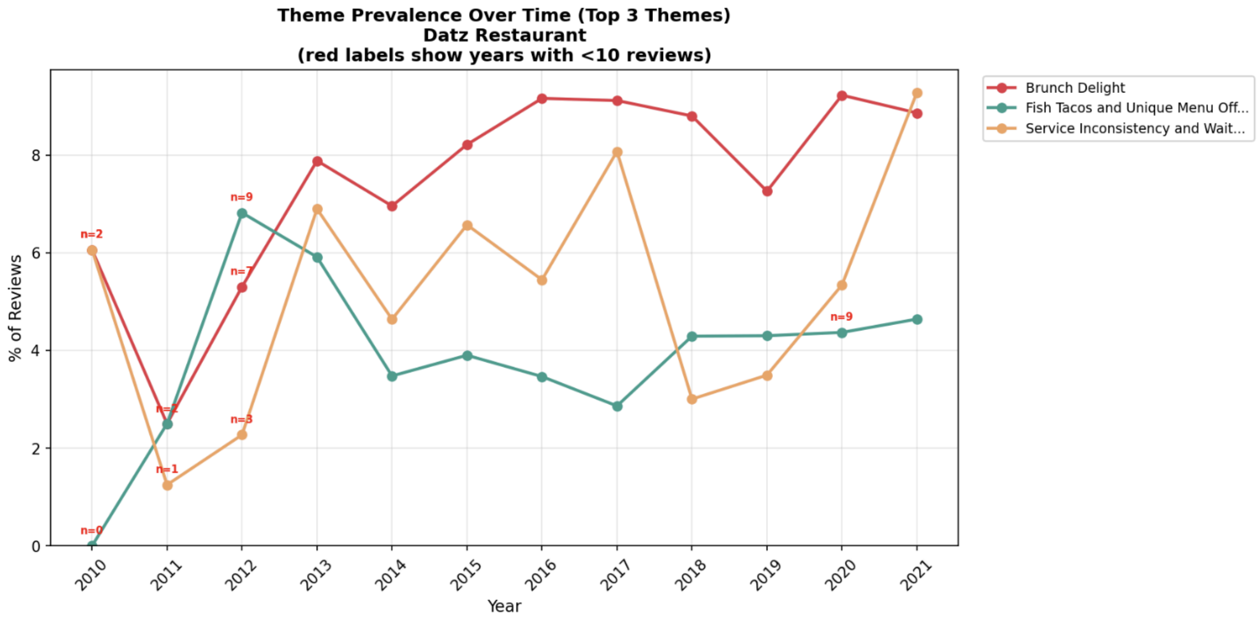
Example Discovered Themes

One prominent theme captures the brunch experience at Datz, characterized by high review volume and frequent discussion of popularity, wait times, and signature menu items. A second theme focuses on food quality and unique offerings, emphasizing distinctive dishes and ingredients. A third theme highlights inconsistent service, reflecting mixed customer experiences with wait times and service reliability over time.

Business Use Cases

Theme frequency and supporting reviews allow operational managers to prioritize customer experience issues based on their prevalence and share of total discourse. Tracking theme prevalence across time windows enables identification of rising, declining, or stable issues, serving as an early-warning signal for emerging operational risks.

The figure below illustrates the prevalence of selected customer experience themes over time, with annotations indicating years with fewer than 10 reviews where estimates are less reliable.



Representative reviews within each theme also support targeted qualitative follow-up, enabling domain experts to contextualize issues and guide internal investigations, staff interviews, and operational audits.

Model Evaluation

Because this project uses unsupervised semantic clustering, no ground-truth labels are available for traditional train–test evaluation. Model quality is therefore assessed using internal validation metrics rather than accuracy. Specifically, inter-centroid cosine separation is used to evaluate theme distinctiveness, reaching 0.61 at the selected number of clusters, with cluster count determined via the elbow method. LLM-generated theme labels are additionally reviewed through human evaluation for interpretability, actionability, and semantic distinctiveness.

Key Observation and Future Work

Individual customer reviews often contain multiple themes, which can dilute clustering performance when reviews are treated as the unit of analysis. While review-level clustering is effective for discovering broad thematic structure, future iterations of the system will adopt sentence-level modeling to improve thematic precision.

Results

The system successfully identifies coherent, semantically distinct, and actionable customer experience themes from 13 years of Yelp reviews. These results support operational prioritization, temporal monitoring of customer concerns, and targeted qualitative follow-up, while clearly motivating sentence-level analysis in future iterations.