Alternate Sprint 0 Idea Expanded: Analyzing Hotel Reviews

The Problem area:
My area of interest falls in the hotel industry. Hotel's may not be staffed to have a regular analysis of guest reviews for their property. Therefore, when a hotel decides to look at reviews on their property, whether it be on google maps, trip advisor, or social media posts, it may be overwhelming to read through hundreds of reviews to identify any useful insights. Are guests enjoying their experience at a newly renovated property? Was their decision to put only 1 floor of rooms out of order enough to prevent disturbing the surrounding guests due to their restaurant renovation? It would take a lot of time skimming and filtering through review websites to identify common themes. Even then, it is very likely that user error and user bias may mislead interpretation of the reviews found.

The User:
Hotel general managers, hotel owners, hotel companies, or other stakeholders experience the problem of the reputation of their property(s). They would benefit from an automated process for analyzing, categorizing, and prioritizing hotel reviews so that actions can be taken quickly to correct issues that may damage their hotel's reputation.

The Big Idea:
By being able to decode customer emotions in reviews, and then categorize and prioritize them, root causes can be identified and presented to the operations team for quick corrections. On the positive review side, we can identify positive review trends to help with future promotions that highlight a hotel's strengths or create new packages to further draw in customers to their property. In my initial dataset, I want to predict emotions in reviews, categorize them, and identify a root problem in each review. Some Machine Learning techniques available that may help may include:

1. Text Preprocessing
    a. Break down text into individual words and reduce them to base forms
    b. Remove data errors
    c. Ie: Tokenization and Lemmatization/Stemming
2. Sentiment analysis
    a. Identify and classify reviews as positive, negative, and neutral
    b. Ie: Naive Bayes, SVM, Logistic Regression, Long Short-Term Memory
3. Aspect-Based Sentiment Analysis
    a. Identifying themes from the aspects within the reviews
    b. Ie: Attention Mechanisms
4. Topic Modeling
    a. Identify underlying topics in reviews
    b. Ie: Latent Dirichlet Allocation, Non-Negative Matrix Factorization
5. Feature Extraction:
    a. Identify important words in reviews

    b. Ie: Term Frequency-Inverse Document Frequency, Word Embeddings
6. Review Summarization
    a. Provide a summary for reviews
    b. Ie: TextRank, Bidirectional Encoder Representations from Transformers
7. Natural Language Processing (NLP)
    a. Understanding and generating human language
    b. Ie: BERT, GPT-3

Doing some research, there are Customer Review Oriented Decision Support Systems (CRDSS) that help businesses make informed decisions by analyzing customer reviews and ratings using machine learning techniques and potentially predict future ratings based on historical review data and trends. Some of the key components overlap with some techniques above, such as Sentiment Analysis, feature extraction, and review summarization.

The Impact:

In the hotel industry, like many other service industries, brand reputation is important. It is what drives customers or guests to choose your product vs. a competitor. A damaged reputation is costly to fix and if problems go unaddressed, may ultimately lead to losses to a businesses bottom line. In the hotel industry, brand reputation is monitored using Guest Satisfaction Scores(GSS).

Another insight to consider are franchises. Hotel companies like Marriott international are asset-light, meaning they only own a small number of hotels. The majority of the hotels are franchised and Marriott generates revenue by collecting franchise fees from hotels that want to use their already established brand. GSS scores play a big role in keeping brand standards. The franchisee must meet certain GSS criteria in order to keep their hotel operating under that brand or flag or else they may have their flag revoked. This will provide significant costs to the hotel owner and operator to have to rebrand their hotel.

By being able to quickly assess customer reviews to improve their product, services, and marketing strategies, hotels can better understand customer feedback, enhance customer satisfaction by addressing concerns, build customer loyalty, and better help understand their target audience and provide services or products that tailor to their guests.

The Data:

I have identified multiple datasets with hotel reviews, however the most beneficial so far is the “Hotel Reviews” dataset from Kaggle. This dataset provide data with over 19 columns, include multiple hotels, their location/address, review dates, review title, review text (with over 34,000 unique values), the reviewer name, and reviewer location.

- A sample dataset can be found here: https://www.kaggle.com/datasets/datafiniti/hotel-reviews

Alternatively, I found other datasets that have less columns available, but has the key variables needed such as the review summary, review text, and the rating. Only 1 of these datasets have the hotel name available.

- Contains multiple hotel names, review title, review text, rating

    - https://www.kaggle.com/datasets/mexwell/hilton-hotel-london-reviews
- Contains data for multiple hotels, however no hotel identified available. Only review title, review text, and rating
    - https://www.kaggle.com/datasets/thedevastator/tripadvisor-hotel-reviews
    - https://huggingface.co/datasets/Aditya1010/17k-hotel-reviews-dataset
- Contains data for a single hotel, review title, review text, rating
    - https://zenodo.org/records/1219899