Project Proposal

Progress made thus far

- Access to Data: The Disneyland Reviews dataset, publicly available on Kaggle (https://www.kaggle.com/datasets/arushchillar/disneyland-reviews/data), was successfully loaded for analysis.
- Exploratory Data Analysis: Generated a bar chart to visualize Rating Distribution as a Percentage, providing insights into the distribution of customer satisfaction levels.
- Text Preprocessing: Lowercasing; Removal of Unwanted Patterns; Stemming and Lemmatization; Stop Word Removal; Generated Word Cloud.
- Data Preprocessing: we preprocessed the Disneyland review dataset by extracting sentiment features, adjusting scores for low ratings, and combining these with numerical features. One-hot encoding was applied to the branch column, and TF-IDF vectorization was used for text. All features were combined and data was split into training and testing sets.
- Model Training: Built multiple Classifier and fine-tuning models but results in low performance on evaluation (need more feature engineering)

Remaining tasks

- More feature selection and feature enhancements
- Fine-tuning on models
- Save and load model for prediction

Any challenges/issues being faced

- Word Cloud Analysis Limitation: Frequent words across all ratings are too similar, making it challenging to identify unique terms that indicate satisfaction or dissatisfaction, potentially impacting model accuracy.
- Imbalance dataset (much less low-rating reviews than high-rating reviews)
- Hard to detect features for medium-rating reviews
- Some rating are unmatched with the review context (positive review context but with low rate)
- Still need larger amount of dataset to get an accurate rating prediction