Title: Yelp Reviews Sentiment Analysis
We Came, We Saw, We Modeled
Names: Kaitlyn Chou (group leader), Jensen Harvey, and Emily Friedman
DS 4002
8 February 2026

Hypothesis: Positive sentiment language (and shorter review length) will be associated with higher star ratings (4-5 stars), while negative sentiment language will be associated with lower star ratings (1-2 stars).

Research Question: To what extent can dining-related language in Yelp restaurant reviews be used to predict customer star ratings, and which attributes of the dining experience contribute most to these predictions?

Model Approach: Our modeling strategy will be to create a sentiment classification pipeline that maps the textual features of the review to the associated star rating. Once the text is preprocessed, we will convert the review into a numerical format using techniques such as TF-IDF or word embedding techniques. We will then be able to experiment with using both interpretable models, such as regression analysis, to determine what words or phrases are the most significant indicators of customer sentiment, as well as using more complex models, such as random forests, to determine the nonlinear relationships between the customer review and the associated rating. By doing so, we will be able to balance the effectiveness of the model with the level of interpretation, allowing us to determine what factors of the dining experience are most important to the customer, such as the food, the service, or the ambiance, etc.

Executive Summary: This project applies sentiment analysis and machine learning to predict Yelp restaurant star ratings with 90% accuracy while identifying key drivers of customer satisfaction. Exploratory analysis revealed that ratings skew toward 4-5 stars, higher-rated reviews tend to be shorter, and distinct linguistic patterns emerge with positive words like "amazing" and "delicious" associated with high ratings and negative words like "bad" and "disappointed" associated with low ratings. Our three-phase approach includes text preprocessing with TF-IDF vectorization, training both interpretable and ensemble models, and rigorous performance evaluation to deliver actionable insights about which dining aspects most influence customer reviews

Data set establishment details:

- Goal: The goal of this project is to use Yelp restaurant review content and length to predict customer star ratings with 90% accuracy, while identifying which dining-related aspects (such as food quality, service, ambiance, and value) most strongly influence overall customer satisfaction.
- Summary of Dataset: The dataset consists of Yelp restaurant reviews from Kaggle (mainly from dessert locations), containing customer-written review text paired with star ratings on a 1-5 scale. The data exhibits class imbalance skewed toward higher ratings

(4-5 stars), with lower-rated reviews tending to be longer and containing negative sentiment words like "bad" and "disappointed," while higher-rated reviews are shorter and feature positive words like "amazing" and "delicious."

Data Dictionary: The table below provides a data dictionary describing each variable we will use for this project, its type, and its role in the analysis:

Variable	Type	Description	Role
rating	numeric (1–5)	Yelp star rating assigned by reviewer	Target variable
text	string	Full written Yelp review	Primary text feature
review_length	integer	Number of characters in review text	Derived feature / control

Questions Explored in EDA: Several exploratory questions were investigated to better understand the structure and limitations of the dataset and to assess its suitability for the project goal. First, we examined the distribution of Yelp star ratings across the dataset to understand the balance of customer satisfaction levels. This analysis revealed that ratings are skewed toward higher values, with a large proportion of reviews receiving four or five stars.

Second, we explored which words tend to be most frequently associated with each star rating by examining word frequencies segmented by rating level. This analysis helped identify clear linguistic differences between low-rated and high-rated reviews.

We then plotted the average length of Yelp reviews, measured in characters, by star rating. Reviews with lower and mid-range ratings (1–3 stars) tend to be longer on average, with 2- and 3-star reviews having the greatest length. In contrast, 5-star reviews are noticeably shorter than reviews at other rating levels. This suggests that dissatisfied customers may write longer, more detailed reviews to explain their experiences, while highly satisfied customers tend to leave shorter, more concise feedback.

We narrowed our analysis to sentiment-bearing words by removing dessert-related nouns such as “cream,” “ice,” and “chocolate.” After filtering these terms, clear patterns emerged: negatively connoted words like “time” and “disappointed” appeared more frequently in 1- and 2-star reviews, while positively connoted words such as “delicious” and “love” were more common in 4- and 5-star reviews.

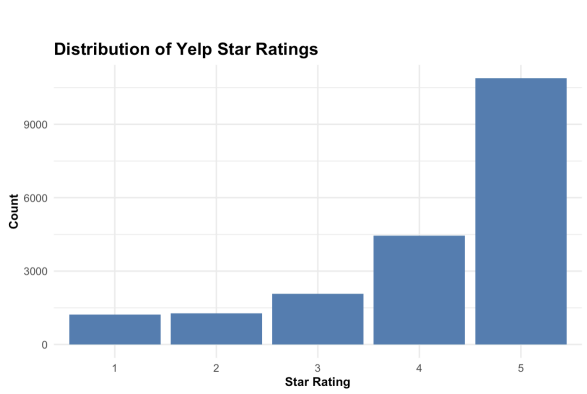
Current Unknowns: One key issue is the dominance of food-related nouns, particularly dessert-related terms such as chocolate, cream, and cake, among the most frequent words in the dataset. While these terms reflect the content of the reviews, they may obscure sentiment-driven language related to service quality, ambiance, or overall satisfaction. To

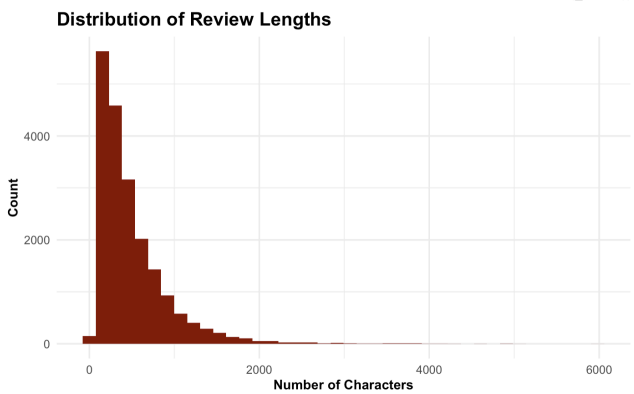
mitigate this issue, the analysis plan includes filtering out common food nouns and focusing on opinionated and sentiment-bearing words.

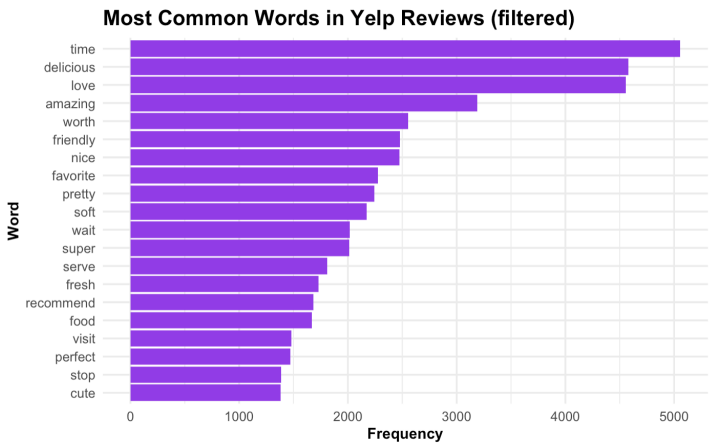
Another uncertainty involves the presence of reviews that may be incomplete, extremely short, or not meaningfully related to the dining experience. Such reviews could introduce noise into the analysis. This limitation will be addressed by incorporating preprocessing steps that focus on sentiment-bearing tokens and, if necessary, applying minimum review-length thresholds to ensure that reviews contain sufficient information for modeling.

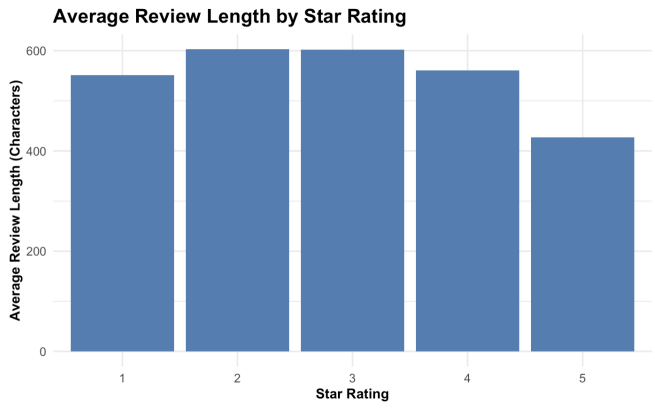
Refinement of Model Plan: The exploratory data analysis led to several refinements in the project's analytical focus and modeling approach. While the original goal emphasized predicting star ratings using Yelp review text, the EDA revealed that raw word frequencies are heavily influenced by food-specific nouns, particularly dessert-related terms such as chocolate, cream, and cake. As a result, the focus of the analysis was refined to emphasize opinionated and sentiment-bearing language, rather than descriptive food nouns, in order to better capture customer evaluations of the dining experience.

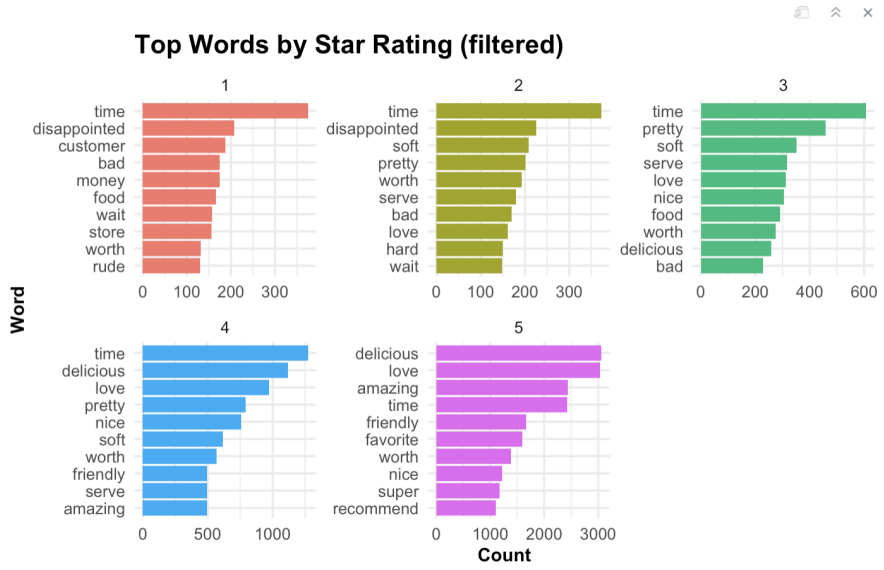
Exploratory Plots:





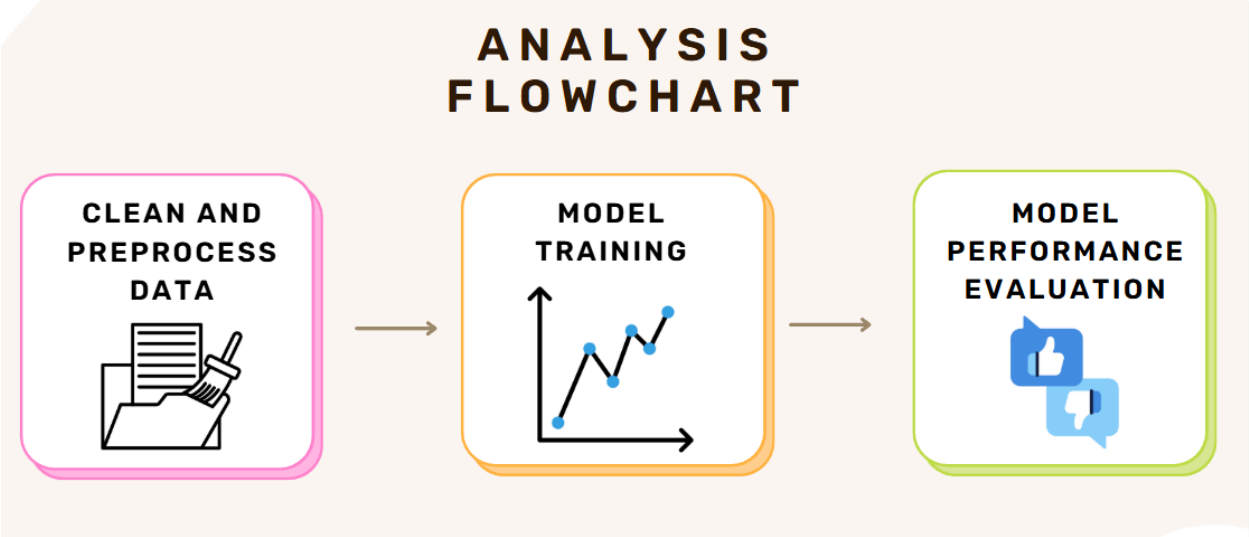






Analysis Plan:

Graphic:



Preprocessing: In order to clean and preprocess the Yelp review text, we will first apply tokenization and replace any potentially sensitive information (e.g., names or addresses) with randomly generated tokens. We will then perform stop word removal to eliminate common, low-information words, followed by lemmatization to improve consistency and overall model performance. Finally, all text will be converted into numerical features in Python using either TF-IDF or word embedding methods for modeling purposes.

Method: To analyze the data, we plan to use predictive modeling analysis. The dataset will first be split into training and test sets to ensure unbiased evaluation of model performance. We will then use our training data to train either a linear or logistic regression model, in order to identify which words most strongly predict star ratings. We will also train a random forest model to capture potential nonlinear relationships in review language. Cross-validation will be applied

during model training and hyperparameter tuning to improve robustness and reduce overfitting. Model performance will be evaluated using the test set and compared to select the approach that best balances predictive accuracy and interpretability.

Evaluation: To evaluate model performance, we will use appropriate metrics depending on the modeling approach. For regression models, performance will be assessed using metrics such as mean squared error (MSE) and R^2. For classification models, we will evaluate accuracy, precision, recall, and F1-score. Cross-validation results will be used to assess generalizability and reduce overfitting. Feature importance and model coefficients will also be examined to interpret which words or features most strongly influence star rating predictions.

Quantifiable Goal: The goal of this project is to develop a predictive model using Yelp restaurant reviews that achieves at least 90% accuracy on a held-out test set when predicting customer star ratings, while quantitatively identifying and ranking the most influential dining-related factors (e.g., food quality, service, ambiance, and value) and review lengths based on model coefficients or feature importance scores.

Goal put in another way: The goal of this project is to use Yelp restaurant reviews to predict customer star ratings, while identifying which dining-related factors (e.g., food quality, service, ambiance, and value) and review lengths most strongly influence overall customer satisfaction.

References:

Kaggle, “Yelp restaurant reviews.” [Online]. Available: https://www.kaggle.com/datasets/farukalam/yelp-restaurant-reviews [Accessed: Jan. 30, 2026.]