Title: Yelp Reviews Sentiment Analysis
We Came, We Saw, We Modelled
Names: Kaitlyn Chou (group leader), Jensen Harvey, and Emily Friedman
DS 4002
February 1, 2026

Goal Statement: The goal of this project is to use Yelp restaurant reviews 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.

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?

Narrative Paragraphs:
Dining out is a universal experience, as nearly everyone has eaten at a restaurant before. The restaurant industry also represents a significant portion of the overall business economy [3]. For both customers and restaurant owners, online reviews play an important role: they help customers decide where to eat [4] and drive traffic and revenue to restaurants [1]. As a result, understanding which specific aspects of the dining experience matter most is essential for interpreting reviews and improving customer satisfaction.

Many restaurants use AI-driven sentiment analysis to identify patterns in customer reviews and guide operational improvements [1]. This project contributes to that line of work by applying sentiment analysis to identify the key words and recurring themes in reviews that are most strongly associated with specific restaurant ratings, helping clarify how customer language translates into overall evaluations.

Modeling Strategy:
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 [2]. 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.

References:

[1] M. J. Baker and B. Hashimoto, “Expression of Customer (Dis)satisfaction in Online Restaurant Reviews: The Relationship Between Adversative Connective Constructions and Star Ratings,” International Journal of Business Communication, vol. 61, no. 1, pp. 148–180, Jan. 2024, doi: 10.1177/23294884231200245. [Accessed: Jan. 30, 2026.]

[2] GeeksforGeeks, “Understanding TF-IDF (Term Frequency–Inverse Document Frequency),” Jan. 20, 2021. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/ [Accessed: Jan. 30, 2026.]

[3] National Restaurant Association, “National statistics.” [Online]. Available: https://restaurant.org/research-and-media/research/industry-statistics/national-statistics/ [Accessed: Jan. 30, 2026.]

[4] J. Pitman, “Local consumer review survey 2022,” BrightLocal, 2022. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey/ [Accessed: Jan. 30, 2026.]

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