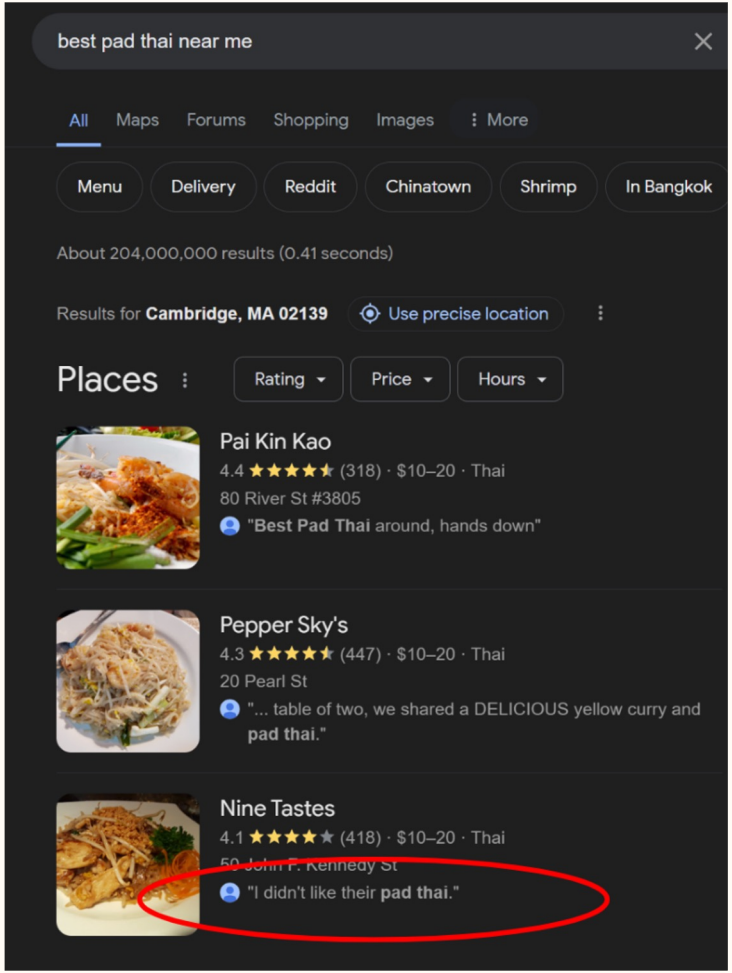
# AC209b Final Project

*A restaurant recommendation system that uses sentiment analysis, summarization, and user metadata to deliver personalized, dish-specific suggestions*

Ethan Tan, Dries Rooryck, Susannah Su, Janice Nam, Dhati Oommen

# 01 - Motivation

While current restaurant recommendation platforms such as Google Maps are effective at providing general ratings and feedback, they often overlook the nuances of individual dish quality. This oversight can frustrate users who are craving certain dishes especially rather than merely visiting well-rated restaurants. Our project seeks to bridge this gap by developing a search engine that leverages sentiment analysis and language modeling techniques to meticulously analyze text reviews for positive experiences with specific dishes at restaurants.



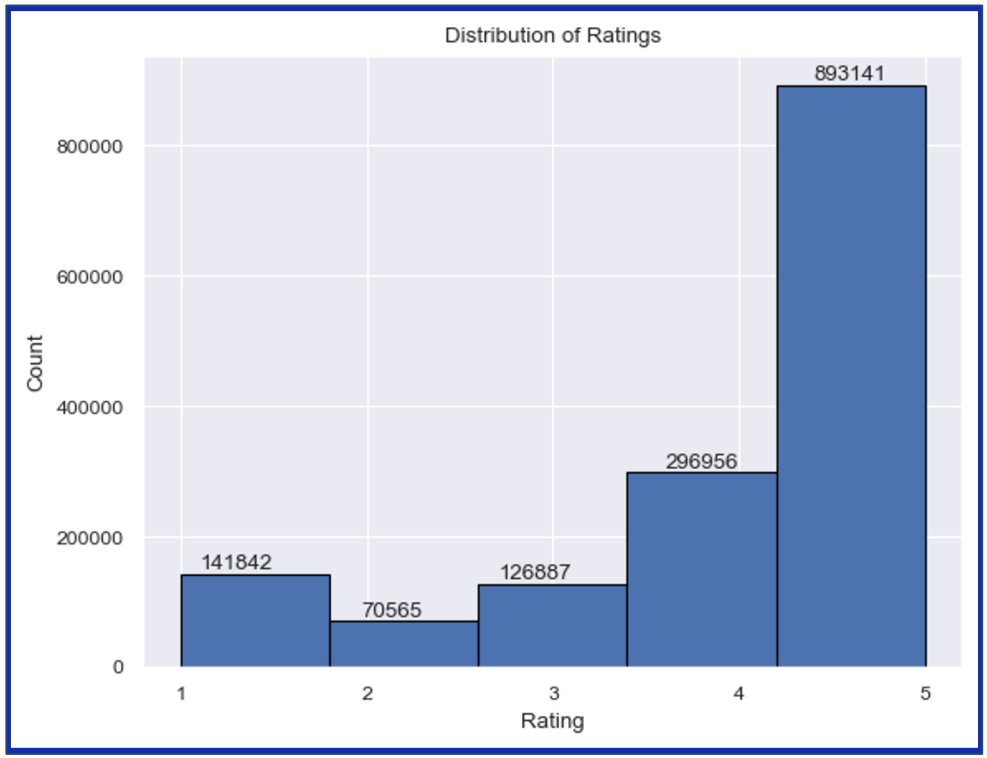
# 02 - Working Question

Given that current restaurant recommendation platforms like Google Maps predominantly focus on overall restaurant evaluations rather than specific dish quality, how can we utilize advanced sentiment analysis and language modeling techniques on textual reviews to develop a search engine that provides more accurate and personalized restaurant recommendations based on specific dishes? This system would aim to enhance the dining decision-making process by pinpointing the best establishments for a particular dish, such as the best pasta Alfredo in town, based on user-generated content.

# 03 - Data Cleaning & Pre-processing

- **Data description:** Our primary dataset contains 100,000 reviews from Massachusetts, including ratings, texts, and business IDs. The secondary dataset provides business metadata like name, description, and category, helping us focus on restaurants.
- **Data Integration:** We merged the review dataset with the business metadata, allowing us to associate each review with the corresponding restaurant, providing a comprehensive view of the data.
- **Filtering Relevant Data:** We refined the metadata to include only restaurants by filtering the 'category' field. This step ensured that our analysis would be specific to restaurants.
- **Dropping Unnecessary Columns:** We removed columns that were not crucial for our analysis, such as the reviewer's name, time of review, and pictures included in reviews. This reduction simplified our dataset, getting us on the order of 1 GB of data.
- **Handling Duplicates and Missing Values:** We identified and eliminated duplicate entries to ensure unique reviews. Additionally, we dropped rows where essential fields like were missing, or locations with latitude and longitude of '0'.
- **Expanding and Encoding Attributes:** We expanded the 'MISC' column (service, dining, etc.) into distinct columns and converted the 'price' column to numeric values for easier analysis.
- **Final Refinements:** We refined the dataset by removing rows with missing critical values and encoded missing 'Atmosphere' and 'Crowd' data for integrity without losing too much information.

# 04 - Visualization and EDA




Distribution of Ratings

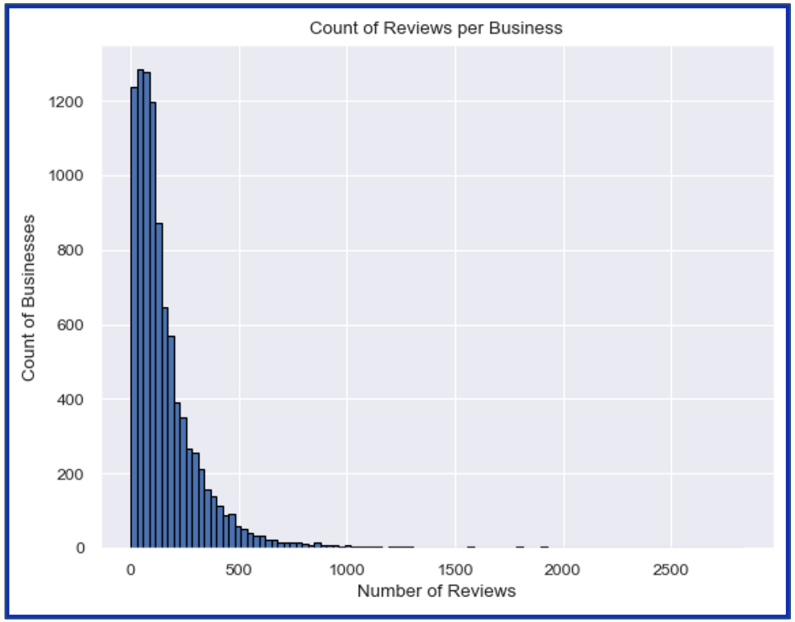
The graph illustrates a marked preference for higher ratings, with 5-stars dominating, suggesting users predominantly post reviews after positive experiences.

This skewed distribution is advantageous for a recommendation system focused on top-rated restaurants but also highlights the need for detailed analysis of negative reviews to pinpoint areas for improvement.

This insight will inform our sentiment analysis model, ensuring it effectively captures the full spectrum of user experiences to enhance our recommendation system.


06

# 04 - Visualization and EDA



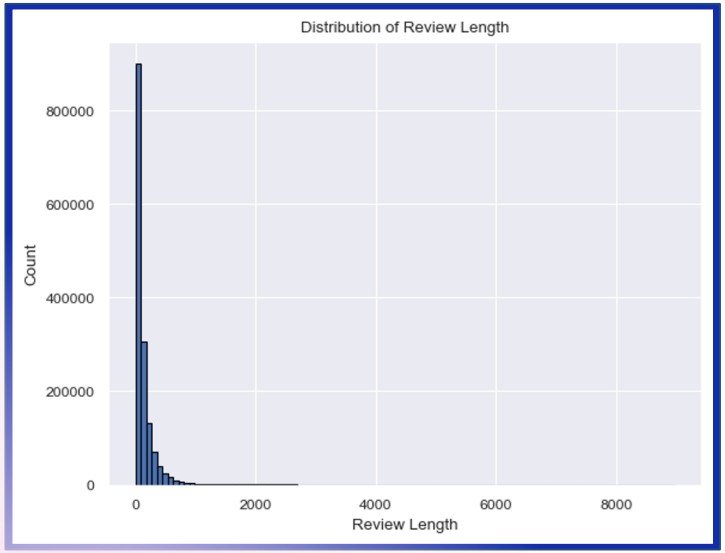

Count of Reviews per Business

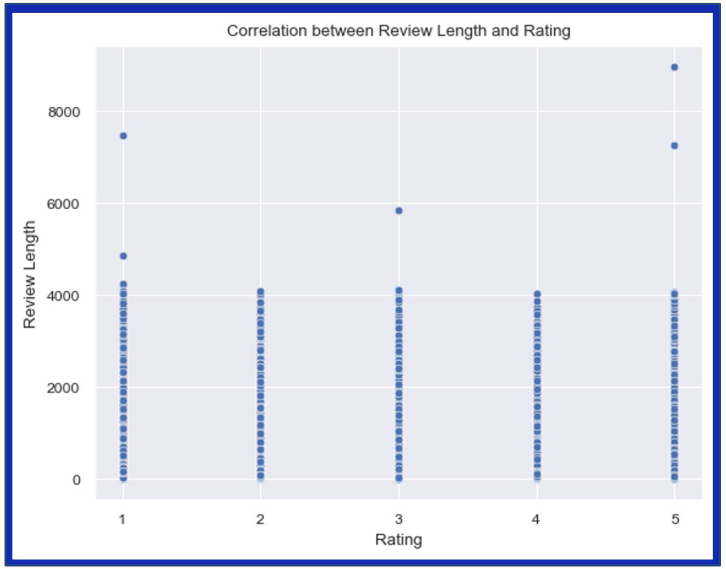
```
Summary statistics for review length:
count    1.529379e+06
mean     1.320390e+02
std      1.923557e+02
min      1.000000e+00
25%      3.200000e+01
50%      6.900000e+01
75%      1.550000e+02
max      8.973000e+03
```

The graph shows a sharp decrease in the number of businesses as the count of reviews increases, with most businesses accumulating between 0 and 200 reviews. This trend indicates a concentration of review volume among a small number of highly-reviewed establishments, some receiving as many as 2831 reviews. This skewed distribution suggests potential bias in our dataset towards popular venues, which could affect the balance and accuracy of our recommendation system. Addressing this in our analysis is crucial to ensure that our recommendations are not disproportionately influenced by these outliers.

# 04 - Visualization and EDA

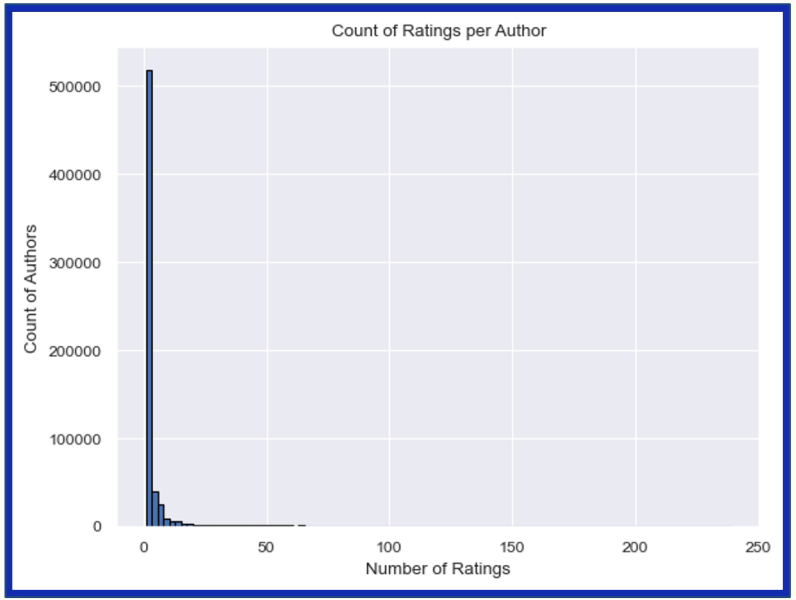
Review Length & Rating




The analysis of review lengths shows that the majority of reviews are concise, mostly under 200 characters, reflecting a preference for brief feedback. Additionally, a weak negative correlation of -0.25 between review length and rating indicates that while longer reviews tend to have slightly lower ratings, the relationship isn't strong. This suggests that, although longer reviews might occasionally express more detailed negative experiences, this is not a consistent trend across the dataset.

# 04 - Visualization and EDA

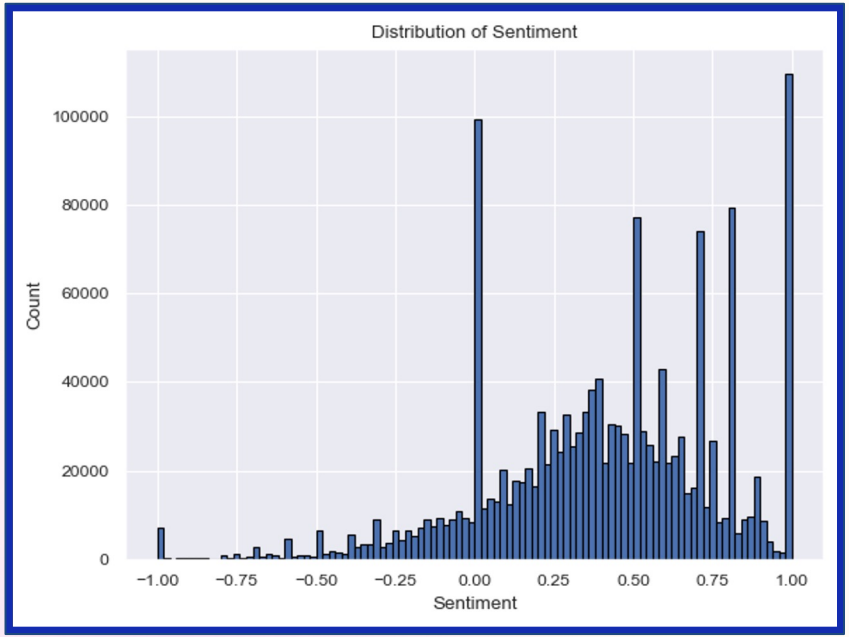



Count of Ratings per Author

```
Number of unique authors:  612146
Summary statistics for number of ratings per author:
count    612146.000000
mean          2.498389
std           4.292764
min           1.000000
25%           1.000000
50%           1.000000
75%           2.000000
max         239.000000
```

The vast majority of authors contribute between 1 and 5 reviews, with a mean of approximately 2.5 reviews per author. The distribution shows a sharp decrease in the number of authors as the number of reviews per author increases, confirming that a small group of prolific reviewers exists, but they are exceptions rather than the norm. Specifically, over 75% of authors have left only two reviews or fewer, highlighting that most users contribute sparingly. This pattern is significant for understanding user behavior on the platform.

# 05 - Simple Sentiment Analysis - Baseline Model




Sentiment Analysis

```
Summary statistics for sentiment:
count    1.529379e+06
mean     4.036130e-01
std      3.707949e-01
min     -1.000000e+00
25%      1.666667e-01
50%      4.367187e-01
75%      6.916667e-01
max      1.000000e+00
```

```
Correlation matrix:
                rating  sentiment
rating        1.000000   0.600953
sentiment     0.600953   1.000000
```

In our baseline sentiment analysis using TextBlob, we removed textless reviews to focus on substantive data. We analyzed sentiment polarity, ranging from -1 (most negative) to 1 (most positive), and found an average sentiment of 0.36, aligning with the prevalence of higher ratings. The sentiment distribution is bell-shaped with spikes at neutrality and positive extremes. There is a moderate positive correlation of about 0.60 between sentiment scores and ratings, confirming that higher-rated reviews generally have more positive sentiments. This correlation indicates our sentiment analysis captures relevant sentiments effectively.

# 06 - Final Review Sentiment Model

*Relevant reviews are ranked by the probability of having positive sentiment using RoBERTa (Robustly Optimized BERT Pretraining Approach), (Liu et. al., 2019).*

- **Motivation:** reviews and tweets are similar in terms of jargon and sparsity.
- The pretrained **RoBERTa-based model** `cardiffnlp/twitter-roberta-base-sentiment-latest` is appropriate.
- Trained on **~124M tweets** from January 2018 to December 2021.
- Finetuned for sentiment analysis with the ***TweetEval* benchmark.**

Example Usage:

"I love AC209B!" → **+ positive**: 0.980
**/ neutral :** 0.017
**- negative:** 0.004

# 07 - Review Summarization

*For each recommended restaurant, we summarize the corpus of reviews using a BART-based model.*

- We use the **"summarization" pipeline** from the transformers library.
- The 'facebook/bart-large-cnn model' **BART-based** (Bidirectional and Auto-Regressive Transformers)
- This pre-trained model is trained on 300 unique articles from the **CNN and Daily Mail.**

Example Usage:

"The food was amazing. The service was great. I would definitely recommend this place to my friends. The atmosphere was cozy. The staff was friendly. The prices were reasonable. The location was convenient. And the food was delicious, especially the pasta alfredo, which was the best I've ever had! I can't wait to go back! The only downside was the parking, but it was worth it. Overall, a great experience! I highly recommend it. But maybe make a reservation, as it can get busy."

→

The food was amazing. The service was great. The only downside was the parking, but it was worth it. I highly recommend it. But maybe make a reservation, as it can get busy.

# 08 - Dish Recommendation - Full Pipeline

- ➔ User Location and Timezone Determination: Retrieve the user's geographic coordinates and local timezone via an API to filter search results geographically and temporally.
- ➔ Dish Input and Review Retrieval: Capture user input for a specific dish (e.g., "Pepperoni Pizza"). Use regex to extract reviews mentioning the dish, ensuring the search captures variations (e.g., we capture both "Pasta Alfredo" and "Alfredo Pasta").
- ➔ Data Filtering: Filter reviews for restaurants that are currently open, using operational hours from the dataset. Ensure reviews are within a user-defined radius from their location.
- ➔ Sentiment Analysis: Conduct sentiment analysis on relevant reviews using the RoBERTa model, designed to understand context within short texts like tweets. Classify sentiment into categories (positive, neutral, negative) and calculate a probability score for each category.
- ➔ Ranking Algorithm: Apply a weighted index that combines sentiment analysis results with average restaurant ratings. Rank restaurants based on the highest probability of positive sentiment, adjusted by average ratings to balance review positivity with overall quality perception.
- ➔ Output Generation: Display the top three restaurants based on the weighted sentiment and rating scores. Enhance the user interface by using an agent that retrieves the website name, and incorporates “evidence” into its recommendations by using a BART-based model summarizing relevant reviews, as well as average rating.

# 09 - Sample Output

**Input: top 3 results for ‘pasta’ within a maximum distance of 50 miles. In code:**

```
print(restaurant_recommender(df, 'pasta', max_distance=50, top_n=3))
```

**Output:**

Based on your craving for pasta, we recommend:

1. Great Road Kitchen 4.14 / 5 ⭐
Summary of 349 reviews: Great Road Kitchen is a gem in the area- food is prepared with care and the staff and service are 5 star! Will return regularly The Point has some of the best services and restaurants!! GRK never disappoints!
Website: https://www.greatroadkitchen.com/

2. Marcello Sandwich Shop 3.69 / 5 ⭐
Summary of 216 reviews: Marcello's is the epitome of mouth-watering, small town pizza that leaves me devastated when I see the slices diminish one-by-one from the box. The Roast beef sandwich is one of the best around. The best coleslaw on the East Coast.
Website: https://www.marcellossubsandpizza.com/

3. Londi's of Salem 3.91 / 5 ⭐
Summary of 200 reviews: The menu at this place is huge to match the portion sizes. We had a Steak Bomb sub and the name fits - it was the bomb. Small meat lovers calzone and our 22 yo couldn't finish it. The same with my wives Roast Beef Sub took half home.
Website: https://www.londisofsalemmenu.com/

# 10 - Future Work/Scope of Improvement

- ➔ Web or Mobile Application Development: Launch a user-friendly application to enhance accessibility and incorporate user feedback for dynamic updates.
- ➔ Advanced Search Capabilities: Upgrade search mechanisms using NLP technologies and machine learning models like BERT or GPT to enhance query understanding and context awareness.
- ➔ Real-Time Data Integration: Partner with data providers such as Google Maps to access current information on restaurant status and reviews, ensuring our database is up-to-date.
- ➔ Diversify Data Sources: Include additional platforms like Yelp, TripAdvisor, and social media to enrich the dataset and capture a wider range of user sentiments.
- ➔ Sentiment Analysis Enhancements: Refine sentiment analysis algorithms to accurately identify subtle emotional cues and context, enabling specific aspect-based sentiment assessment.
- ➔ System Scalability and Adaptability: Ensure the system can handle growing user numbers and data volumes efficiently while remaining responsive to evolving consumer trends.
- ➔ Automated Testing and User Feedback: Implement automated testing to maintain recommendation accuracy and establish a user feedback loop for continuous system improvement.

# Thank you!