Tides of Taste: Personalized Seafood Restaurant Insights **Phase 2**

Boyuan Chen
Gaurangi Agrawal
Mengxin Zhao
Varun Kaza

1

How We Got Here

Phase 1

1. Foundation: Scraping of review data updated weekly

   Initial Scrape: Oct 7, Weekly Scrape: Every Tuesday

2. ETL pipeline to automate the flow of data

   Apify Client -> JSON files to GCP -> Aggregated data for analysis to

   Motherduck -> SQL query that creates or replaces table

3. Tools Used: Apify,Prefect,Pandas,GCP, MotherDuckDB,Superset

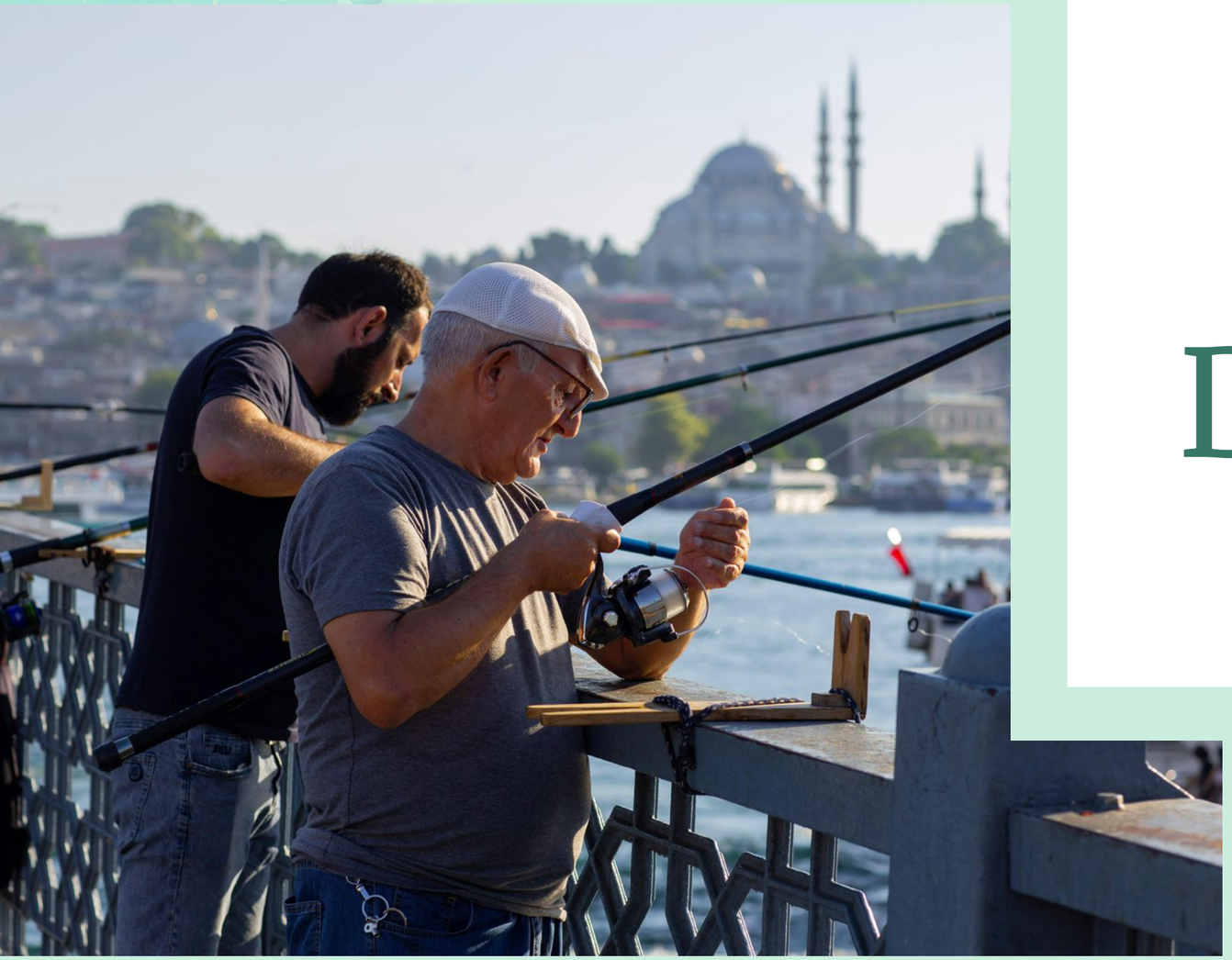
Phase 1 (Continuted)

1. **Apify Client**: Used for web scraping Google Reviews of the 100 restaurants.
2. **Prefect**: Employed for workflow orchestration and monitoring of the pipeline tasks.
3. **Pandas**: Utilized for data manipulation and transformation in the data processing steps.
4. **Google Cloud Storage**: our data lake, the storage solution for raw and processed data files.
5. **MotherDuckDB**: Acts as the cloud-hosted data warehouse where the final data is stored for analysis.
6. **Superset**: Dynamic dashboarding & analytics

Phase 2

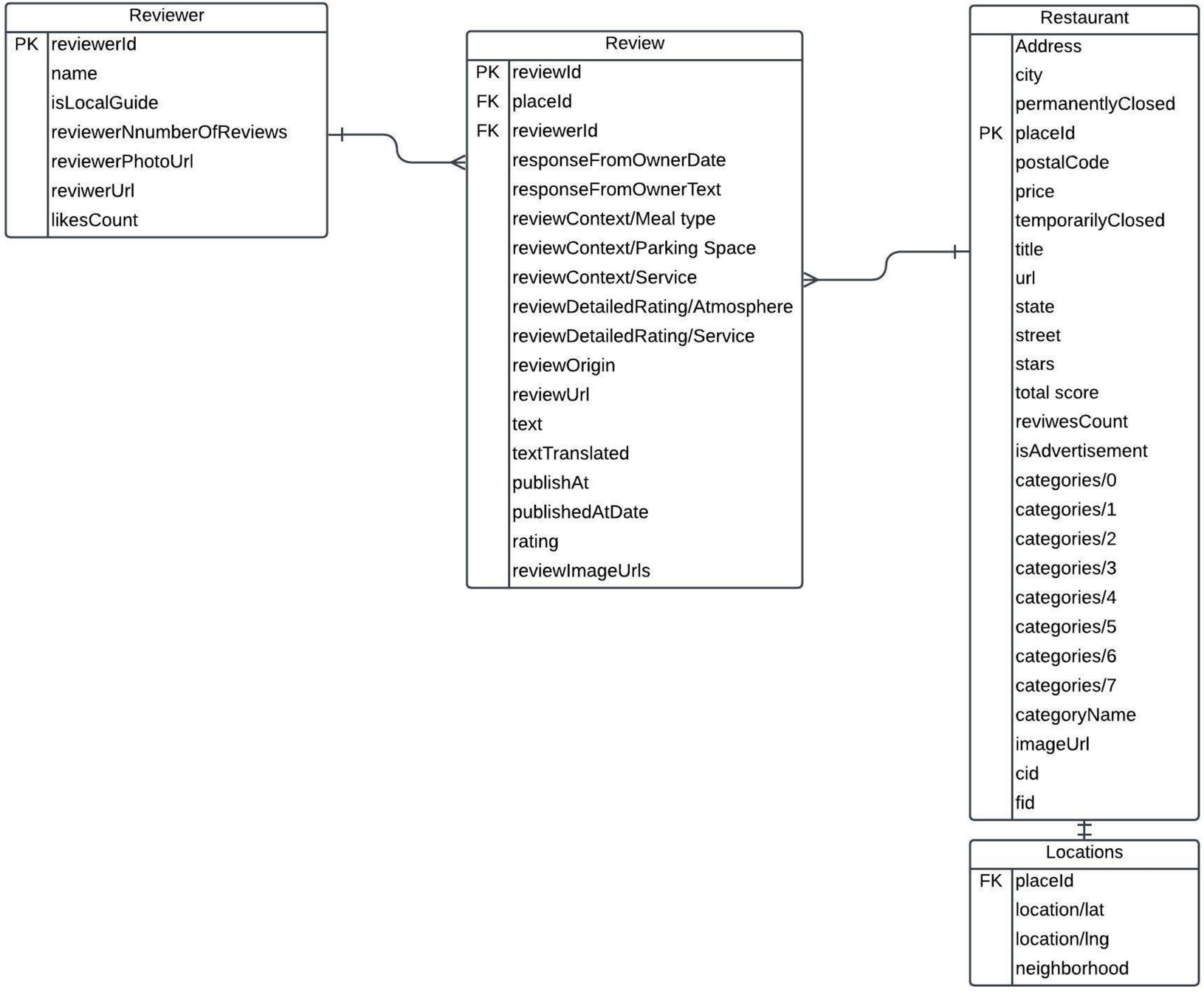
1. Topic Modelling, Sentiment Analysis
2. Further Preprocessing of Data:

- Dropping of Nulls
- Filtering for Restaurants Open All Year
- Preprocessing of Text

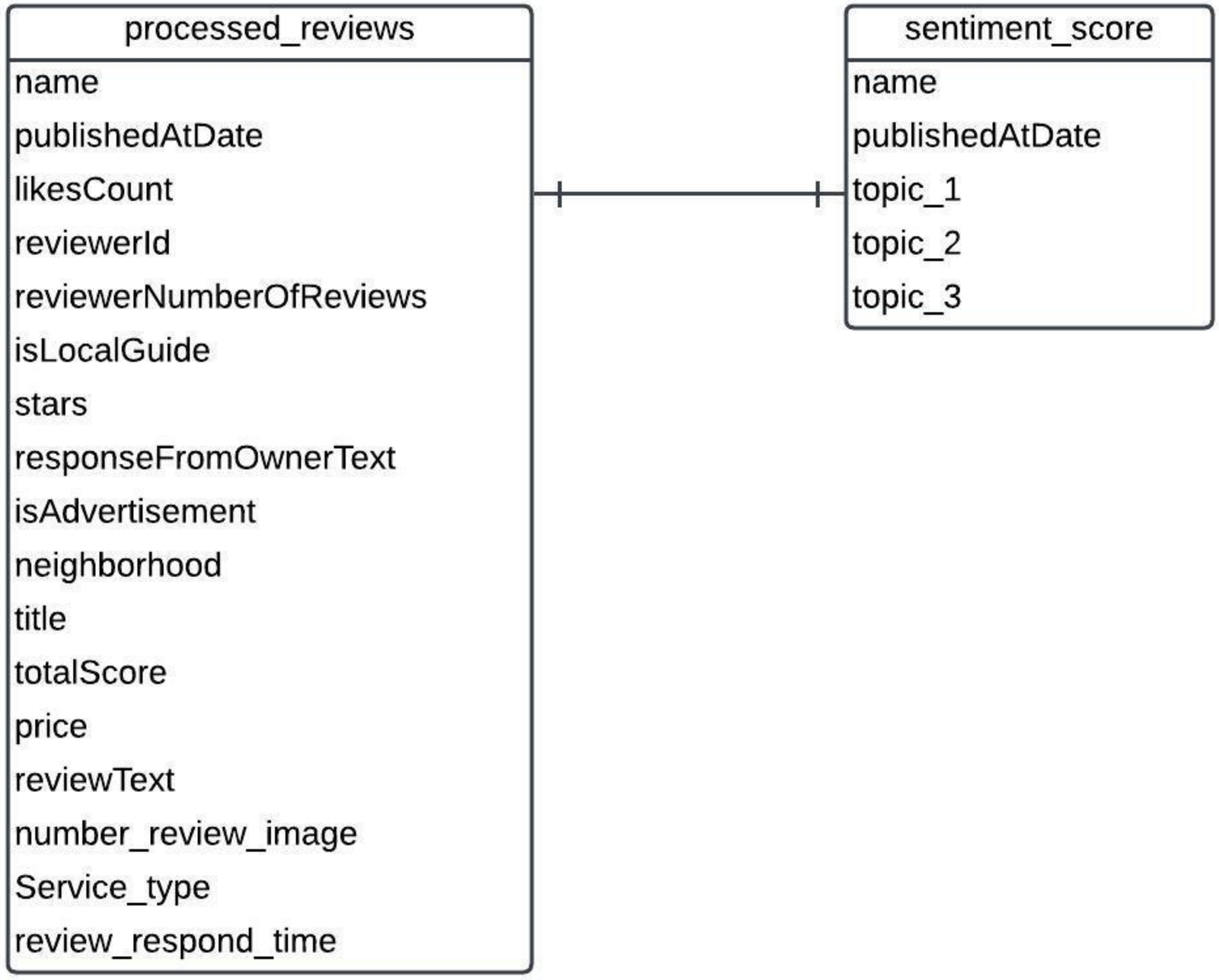


2

Data Structure



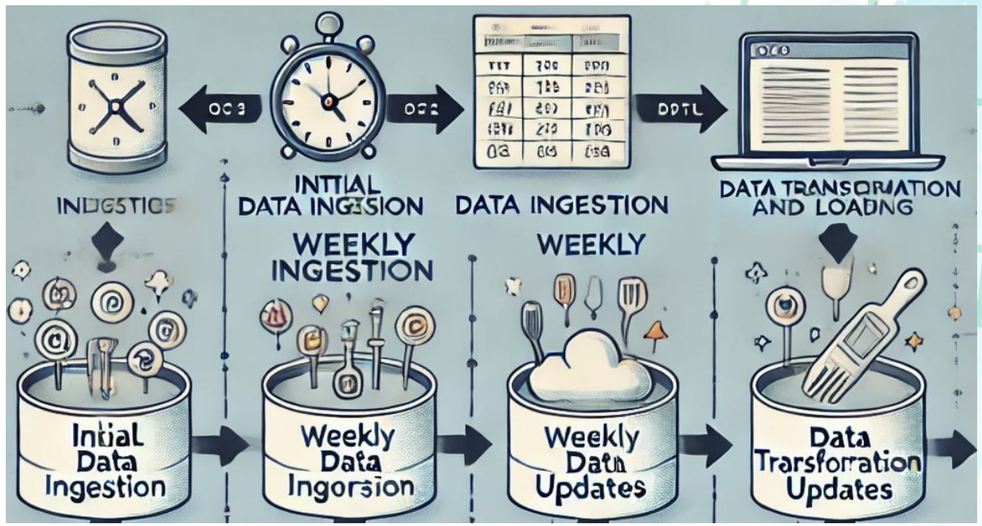
Entity Relationship Diagram



Entity Relationship Diagram

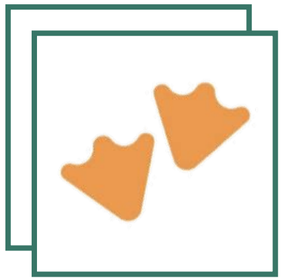
3

MLOPs Pipeline Structure



An MLOps pipeline automates the data flow to topic modeling/ABSA/predictive analysis

Tools we used before reporting



MotherDuck DB

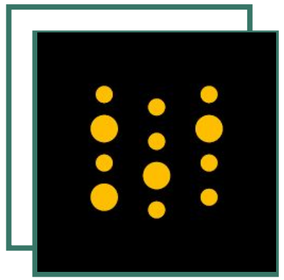
Data Warehouse

Prefect

Pipeline Orchestration.

Light GBM

Modeling



Weights & Biases

Hyperparameter Tuning

Predictive Analysis

Step 1: Non-negative Matrix Factorization

What could be the representation of satisfaction level?

1. Numeric/categorical information (price, service type….)
2. **Review Text**

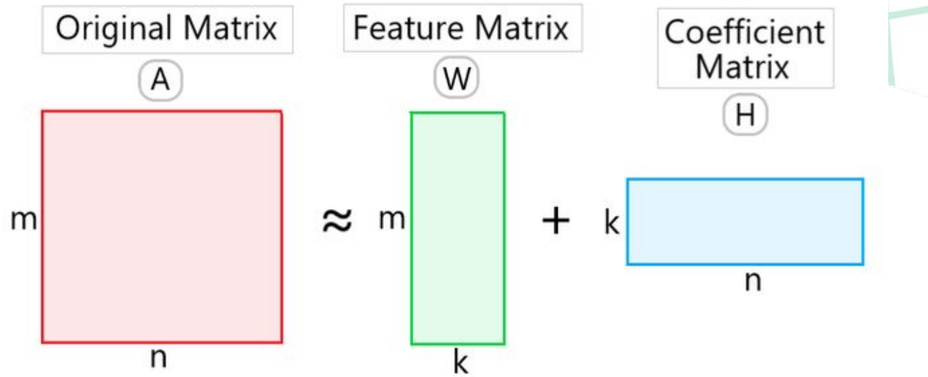
What kind of information we really want from the text?

1. Its context (What is it talking about?)
2. Its sentimental value (Attitude?)

<u>How do we transfer text information to numbers that could be integrated into a ML model?</u> (Text is complicated.)

Topic extraction: Non-negative matrix factorization, to retrieve k topics for each of the topic

Each topic could store at most *m* words and could be used as a boost of sentiment analysis



m*n: word * document m*k: word * topic k*n: topic * document

NMF explanation

topic_name	word1	word2	…..	wordN
topic 1	tasty	menu	…..	cuisine
topic 2	pricey	decoration	…..	good
topic 3	friendly	table	…..	serve

Topic-Word Table

Predictive Analysis

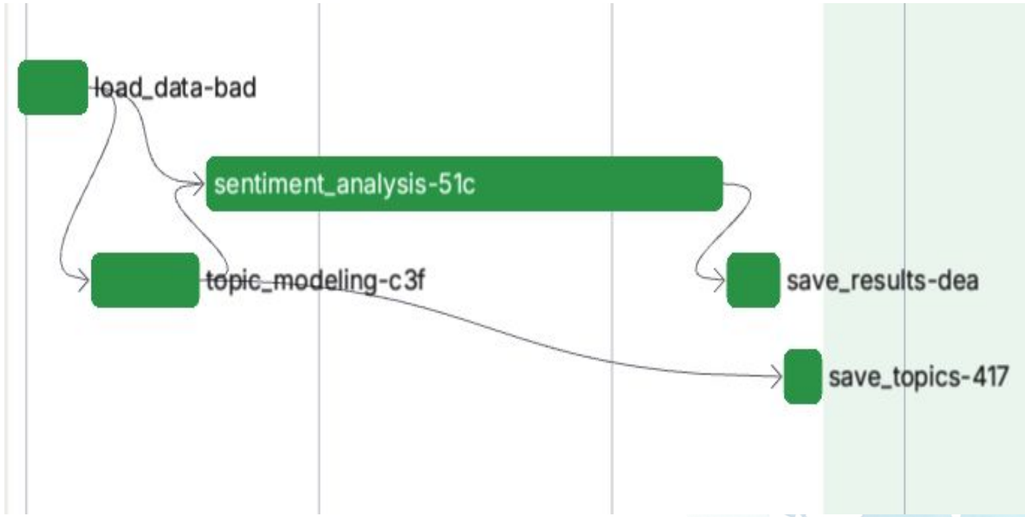
Step 2: Granting Sentiment Score

We use
1. keyword matching to identify the topic of each sentences in each review
2. Calculate the sentiment score for each topic, and multiply them with the coefficient from topic-document association matrix

After executing the script, we have:
1. Score table
2. Table with numerical/categorical information

Merge them based on the reviewerID & publish date

For doc i:
Score for each topic =
topic coefficient for topic j * sentiment score on topic j



NMF & ABSA flow

Predictive Analysis

Step 3: Modeling

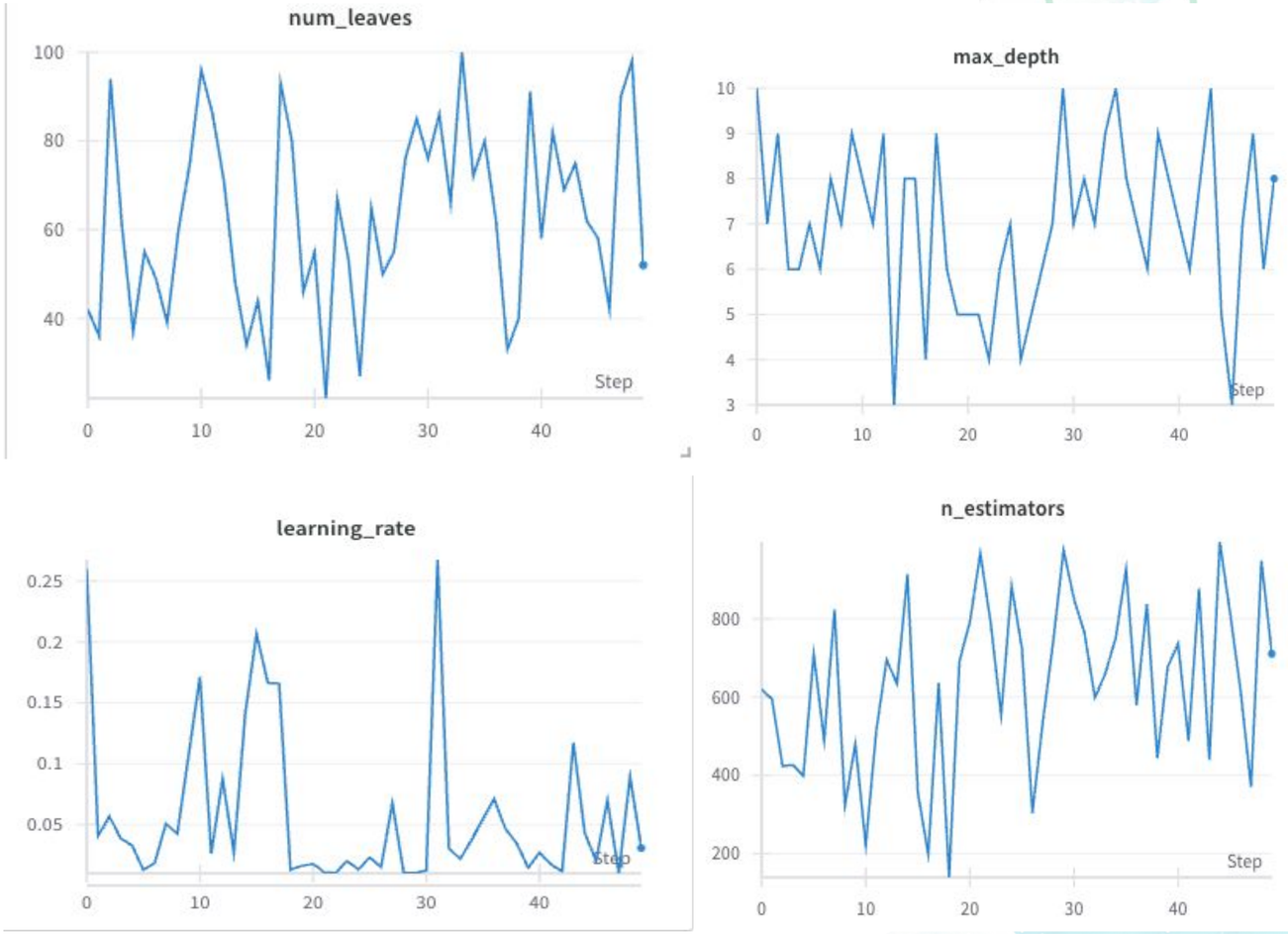
We used:

- stars as our Y
- numerical/categorical features like sentiment score for each topic, review response time, price…. as Xs
- Weight: number of images & number of review posted by the reviewer before
- and run a Light GBM regression.
- Target: validation MSE

Hyperparameters:

1. num_leaves (maximum no. of leaves in one tree)
2. max_depth (maximum no. of a tree)
3. learning_rate
4. n_estimator (boosting round)

Tuning method: random search



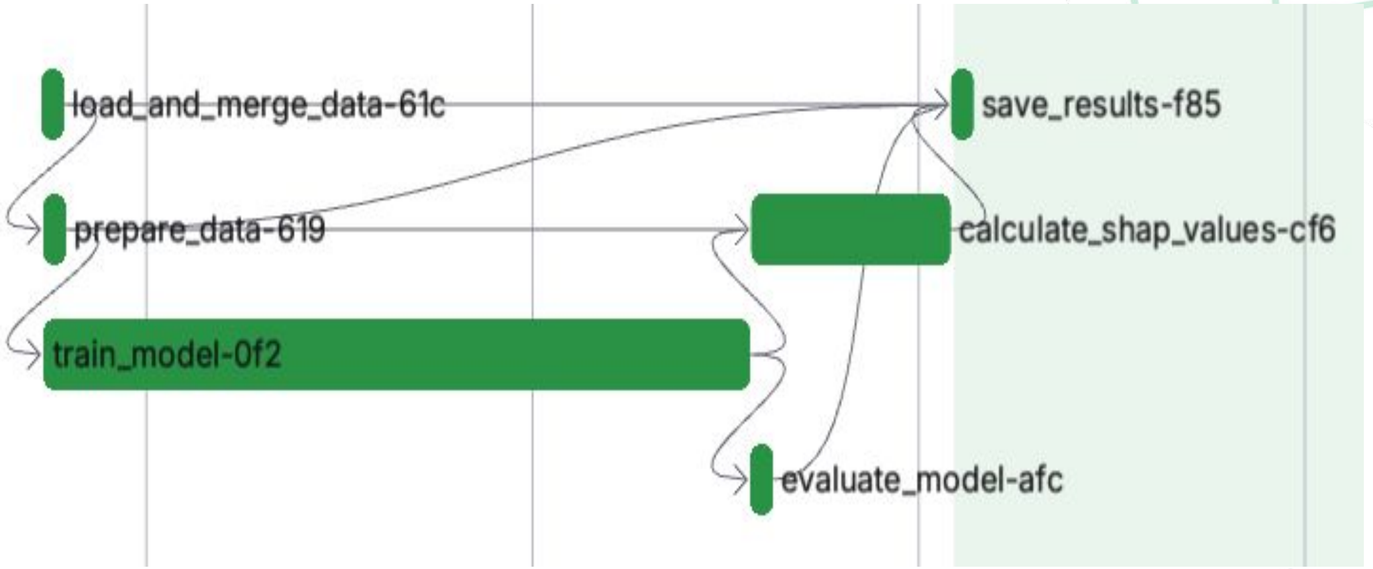
Hyperparameter Tuning Results from W&B

Predictive Analysis

Step 3: Modeling

After the modeling:

- SHAP value is calculated to inspect the impact of each feature on the target variable
- Model performance are saved to inspect the model quality



Modeling flow

4

Dashboard Insights & Plans

Dashboard on Superset & Potential improvements on the current model

Superset Dashboard

Link: http://34.172.134.179:8088/superset/dashboard/p/v1y5K6A3LbY/

Next Step:

- Develop AI-Driven, User-Friendly Interfaces (UX/UI Design using Streamlit): Leverage AI and LLMs to create intuitive interfaces that allow customers to input preferences and receive personalized restaurant recommendations, with clear explanations for each suggestion. Achieve the functionality of helping customers make informed choices while enabling restaurant owners to gain actionable feedback to improve their services.

- Enhance Model Accuracy: Refine the current model by trying different hyperparameter tuning methods to improve prediction accuracy.

Thanks For Listening!