ML Challenge Report

Siqi Liu Rafay Usman Ryan Shiels Albert Li

I. Data

The multiple-choice questions (Q1, Q3, Q7, Q8) were split into individual one-hot features, and Q3 Q7, which both allowed multiple answers, also had each combination represented as a one-hot feature (if they appear in the training data at least 5 times) and a feature that represented the number of selected answers.

Q2 and Q4 were converted into numbers using the following logic:

- If there is a range (two numbers with the word “to”, “and” or “-” between), then take the average
- Otherwise, take the first number that appears (it can handle numeric, and also written numbers up to 20)
- Otherwise, if it’s the ingredients question, take the number of commas or line breaks, plus one
- Otherwise, return the average seen in the training

Q5 and Q6 were more complicated free-form text answers. To generalize as broadly to any input, we used a fuzzing library to cluster similar responses into singular features, assuming they appeared frequently enough in the training set. Then, for movie answers we only allowed a singular matched movie (or if nothing matched, “other”), but for drink answers multiple drinks could be listed.

Another approach we adopted was using a word2vec model on the free-form questionnaire responses. Word2vec learns vector representations for words, assigning similar embeddings to words with related meanings. This captures semantic relationships in the text and allows distances between embeddings to quantify word similarity. For each free-form answer and label (pizza, sushi, shawarma), we calculated the average word vector. Each question had three features, one per food type, that recorded the cosine similarity between the free-form answer’s averaged vector and the averaged vector for each label.

This was all condensed into a function that took in a CSV file, split the data as described, and output a dataframe of the flattened vectors for every feature.

We visualized the cleaned data using histograms and boxplots to explore the data. Some questions such as Q3: ”In what setting would you expect this food to be served?” have multiple choice as inputs, and others such as Q5: ”What movie do you think of when thinking of this food item?” allow participants to type their responses. For those questions, we counted the number of occurrence for each input, and only show the top 5 responses in the graph so it is readable while capturing the most important information.

Figure **??** shows a histogram of responses for Q1: From a scale 1 to 5, how complex is it to make this food?. For Sushi, the difficulty is generally higher with more inputs of 4 and 5. For Pizza, we see medium complexity(3) being the highest choice. For Shawarma, we see more votes in the range of 4 and 5.

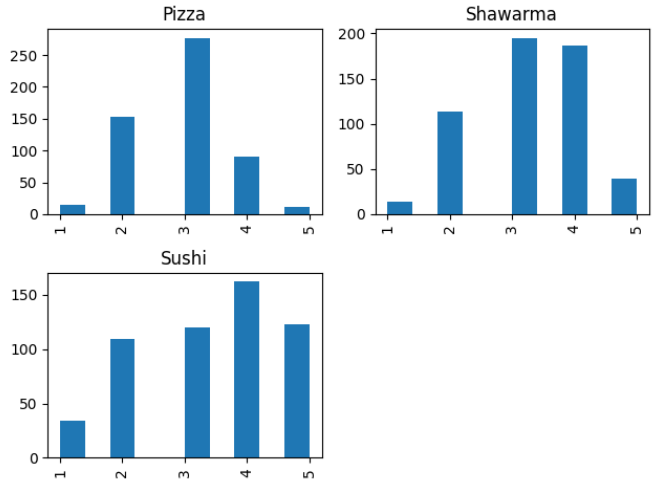

Fig. 1: Histogram for Q1: How complex is it to make the food

In figure **??**, the boxplot illustrates the distribution of expected ingredient counts for each food class. We see the participants generally expect the fewest ingredients in Sushi (median 4), followed by Pizza (median 5), and then Shawarma (median 7). Shawarma shows the greatest variability in expected ingredient counts, while Pizza and Sushi show less variability. All three food types have outliers, suggesting some individuals expect significantly more ingredients than the majority. We see in **??** that the histograms shows a right-skewed distribution for all three food items, indicating most respondents expect a low number of ingredients. However, the skewed pattern suggests a minority of participants anticipate a considerably higher ingredient count.

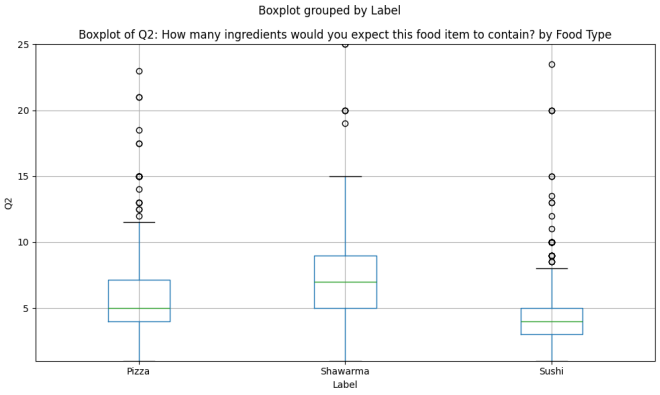

Fig. 2: Boxplot for Q2: How many ingredients would you expect the food to have

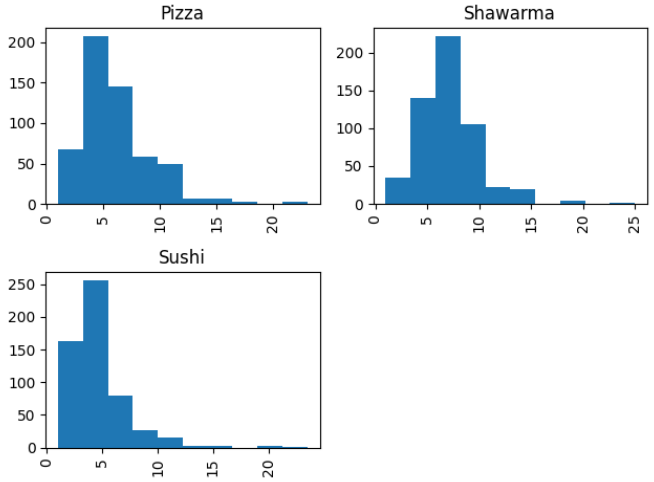

Fig. 3: Histogram for Q2: How many ingredients would you expect the food to have

Question 3 is ”In what setting would you expect this food to be served?”. For Q3: “In what setting would you expect this food to be served?”, we see a trend that Pizza is more appropriate for most situations, and shawarma and sushi are more specific in when they are expected to be served.

In figure **??**, we see a majority of participants expecting to pay a lower price for the food. Interestingly we see much higher outliers for sushi, with a peak of someone willing to pay 100 dollars for one serving of sushi. We also see different distributions for the three food items - pizza has a peak at 5 dollars, with a tail towards higher prices; shawarma has a normal distribution with mean at 10 dollars.

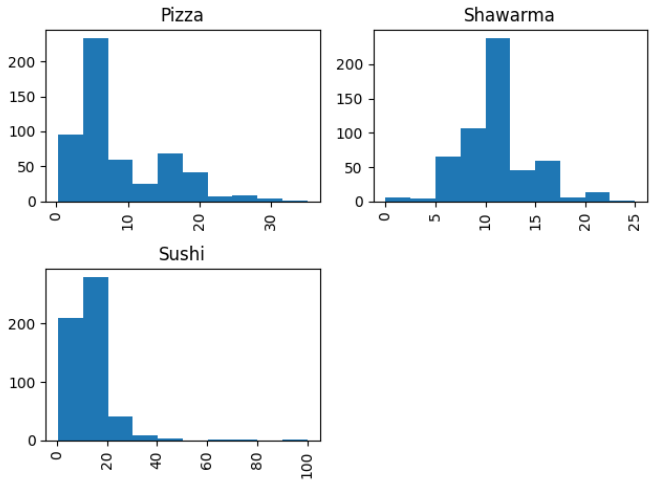

Fig. 4: Histogram for Q4: How much would you expect to pay for one serving of this food item?

Figure **??**, **??**, **??** shows the top responses for each class for Q5. For both Pizza and Sushi, we saw “none” as the most popular input. However for Shawarma we saw “Avengers” to be by far the most popular response for Shawarma, suggesting a correlation between ”Avengers” and ”Shawarma”, making Q5 a good indicator.

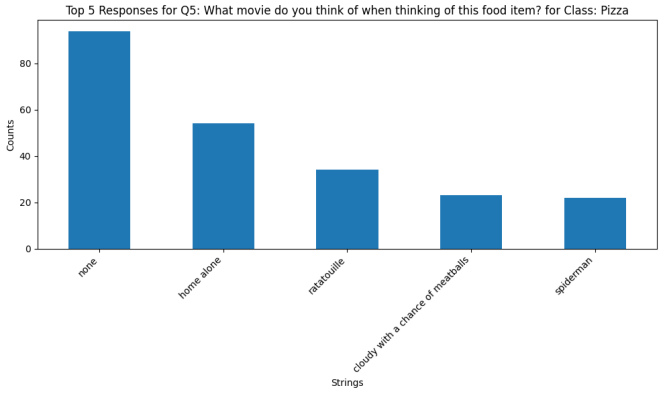

Fig. 5: Counts for Pizza for Q5: What movie do you think of when thinking of this food item?

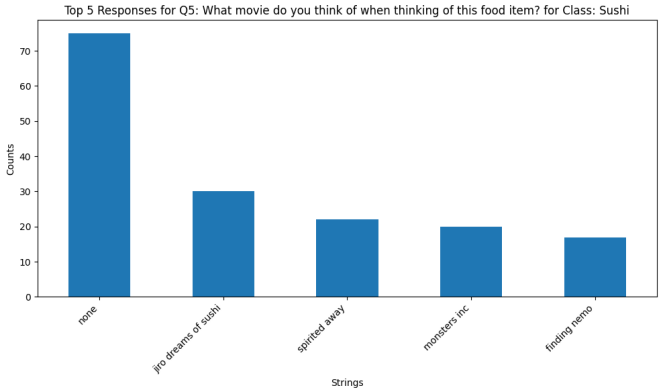

Fig. 6: Counts for Sushi for Q5: What movie do you think of when thinking of this food item?

For Q6: ”What drink would you pair with this food item?” we see a substainal difference between the answers for the three foods. With sushi participants preffered water, tea or sake. With Piazza there was a clear prefference for Cock-cola and other fizzy drinks and Shawarma had a tie between water and coca cola for the most frequent response with a siginficant minority preffering juice or nothing at all.

With regards to Q7: “when you think about this food item, who does this remind you of?”, there are no clear indicating responses between the food classes. For example, the most popular response accross the three foods for Q7 was “Friends”, with the remaining responses being, similar in size and frequency across the three food items. Therefore we decided to remove this Q7 from the parameters.

We split the dataset into 3 sets: 60% training, 20% validation and 20% test. This allowed us sufficient data to train the model as well as data for testing that the models generalizes to unseen data.

With regards to Q8, ”How much hot sauce would you add to this food item?”, the responses for Piazza and Sushi were extremely similar with the responses being exactly the same, but there was a very high preference for a moderate amount of hot sauce when considering Shawarama. As such this was a good feature to include.

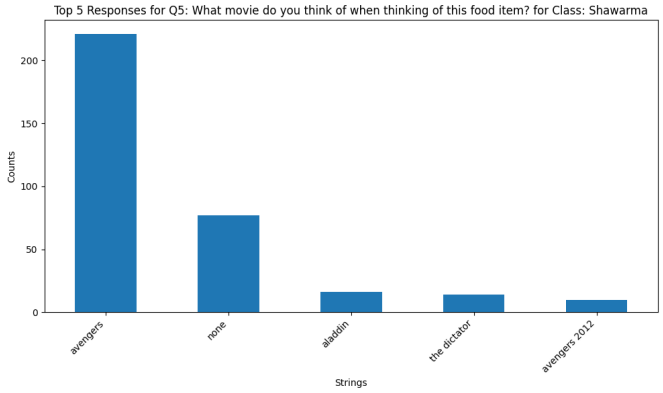

Fig. 7: Counts for Shawarma for Q5: What movie do you think of when thinking of this food item?

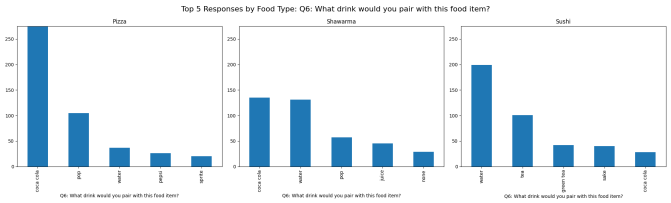

Fig. 8: Top 5 Responses for Q6

II. Model

A. Logistic Regression

B. Neural Network

C. Decision Trees

I tested a regular decision tree, an ensemble of decision trees, and a Random Forest model, using the built-in RandomForestClassifier from sklearn. I found the decision tree to perform the worst, and it was easier to manipulate the RandomForestClassifier's parameters, so I decided to explore that further.

I decided to include all of the possible features (even initially 'id' accidentally), as the Random Forest was very robust regardless of the feature choice or hyperparameters. To allow the text-based features to generalize as well as possible, I used a library to cluster text by Levenshtein distance, and to limit overfitting, I only included the categories with enough representatives. This did not seem to have too large an impact on validation accuracy, but would allow it to correctly categorize otherwise unseen data with unique typos or spelling choices. I tested many of the hyperparameter options, but found little change in accuracy aside from increasing the number of estimators to around 250, and limiting the minimum samples split to 10. Some randomness was introduced when generating text clusters, but I found it to still be very consistently accurate regardless of what data it trained on.

It consistently achieved an 87% validation accuracy. The classification report initially showed it was particularly effective at identifying Pizza and Shawarma, but had some difficulty with Sushi, which I would expect if respondents have less familiarity with Sushi. However, after splitting the training

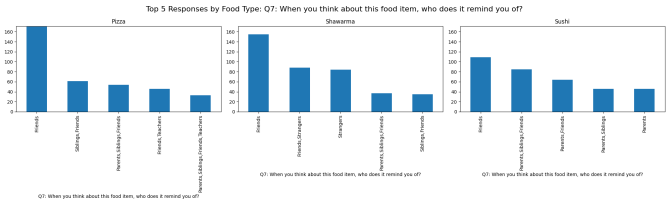

Fig. 9: Top 5 Responses for Q7

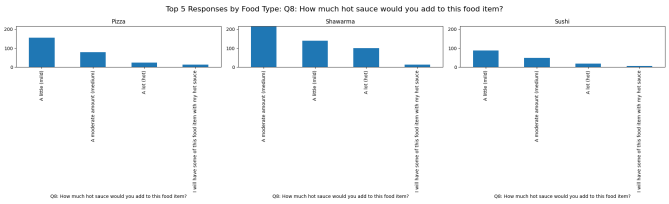

Fig. 10: Top 5 Responses for Q8

data such that there is the same amount of each in the training and validation portions, this was no longer the case.

Category	Precision	Recall	F1-score
Pizza	0.85	0.93	0.89
Shawarma	0.88	0.84	0.86
Sushi	0.90	0.86	0.88
Accuracy	0.88 (263 samples)		
Macro avg	0.88	0.88	0.88
Weighted avg	0.88	0.88	0.88

TABLE I: Classification Report

After comparing the loss, bias, variance decomposition to our other models, I found the expected higher bias, but lower variance that Random Forests are known to have:

Loss	Bias	Variance
0.1380	0.1255	0.0410

III. Model Choice and Hyperparameters

IV. Prediction

V. Workload Distribution

Ryan Shiels worked on the data cleaning and input functions, and explored Decision Trees/Random Forest models.
Siqi Liu: data visualization and implementing the MLP Classifier
Rafay Usman: data cleaning and implementing/tuning the MLP Classifier
Albert Li