**Problem overview:** The problem is to build a multi-label sentiment classifier that analyzes restaurant reviews and assigns sentiment scores not only to the overall review but also to specific aspects like service, ambiance, food, price, and context. This goes beyond simple positive/negative classification by providing a more granular understanding of customer opinions. The goal is to provide restaurant owners with a rapid classification of all restaurant reviews on a frequent basis in order to provide timely feedback as part of a continuous improvement program. This is an interesting problem as the restaurant business is highly competitive and reviews by food critics and restaurant reviewers can strikingly influence the future success of restaurants. Our classifier will also provide breakdown by critical categories as outlined above. Moreover, restaurants that are committed to receiving feedback and attempting to incorporate such feedback to facilitate continuous improvement are more likely to proceed. An appropriate classifier that is reliably constructed can provide near real-time feedback integrating data from disparate sources. Other interesting aspects:

1. **Granularity:** Analyzing sentiment across multiple aspects provides deeper insights than a single overall score. This can help restaurants pinpoint specific areas for improvement.
2. **Contextual Understanding:** Incorporating context (e.g., "for a quick lunch," "for a special occasion") can significantly enhance sentiment analysis accuracy.
3. **Practical Application:** The results can be directly used by restaurants for customer feedback analysis, marketing, and operational adjustments.
4. **Natural Language Complexity:** Restaurant reviews often contain nuanced language, sarcasm, and implicit opinions, making it a challenging and interesting NLP task.

**Data:** We selected the Yelp Open Dataset of businesses and reviews (<https://business.yelp.com/data/resources/open-dataset/>), specifically the yelp\_academic\_dataset\_review.json and yelp\_academic\_dataset\_business.json files. We downloaded the datasets and chose to focus on the New Orleans Greater Metropolitan Area including target cities of New Orleans, Metairie, Kenner, Gretna, Harahan, Westwego, Chalmette and Slidell. We then searched for businesses in these cities and specifically selected those businesses related to “restaurant/food”. We created a list of target business ID numbers and then searched from 01/01/2019 to 12/31/2024 pulling all reviews of these target businesses. This resulted in 4099 unique restaurant business IDs. We then processed the data collecting the review ID, text, stars and date of review. The dataset for analysis comprised 20,377 reviews. We then cleaned and preprocessed the text data. We calculated overall sentiment scores and aspect sentiment scores (food sentiment, service sentiment, ambiance sentiment, price sentiment and context sentiment).

### **Method:** We approached our problem using the following methods:

* **Aspect Definition & Keyword Spotting:**
  + Five primary aspects were defined: 'food', 'service', 'ambiance', 'price', and 'context'.
  + Keyword lists were created for each aspect.
  + Sentences within each review were classified into the aspect with the highest keyword count, following the methodology described in related literature. Sentences without keywords were labeled 'other'.
* **Sentiment Analysis (Initial):**
  + NLTK's VADER (Valence Aware Dictionary and sEntiment Reasoner) was used to calculate a compound sentiment score (-1 to +1) for each sentence.
  + Weighted sentiment scores for each aspect per review were calculated based on the proportion of sentences assigned to that aspect and their summed sentiment scores, aiming to replicate a methodology described in related work.
* **Text Preprocessing:**
  + Standard text preprocessing techniques were applied to the review text: lowercasing, removing URLs and punctuation, tokenization using NLTK, stop word removal (NLTK English stopwords), and lemmatization using WordNetLemmatizer.
* **Target Variable Creation (Sentiment Discretization):**
  + The calculated continuous sentiment scores per aspect were discretized into three categories (negative, neutral, positive) using quantile-based binning (splitting into thirds). These discrete labels (0, 1, 2) served as the target variables for the classification models.
* **Modeling:**
  + **Baseline Model:**
    - **Feature Extraction:** TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was applied to the preprocessed text, limited to the top 5000 features.
    - **Classifier:** A Logistic Regression model was used as the base estimator, wrapped in Scikit-learn's MultiOutputClassifier to handle the prediction for all five aspects simultaneously.
  + **Advanced Model (Transformer):**
    - **Model:** A pre-trained DistilBERT model (distilbert-base-uncased) from the Hugging Face Transformers library was fine-tuned for the multi-aspect classification task.
    - **Architecture Modification:** The standard classifier head was replaced with a linear layer outputting scores for each class within each aspect (total output size: num\_aspects \* num\_classes\_per\_aspect, e.g., 5\*3=15).
    - **Training:** The model was trained using PyTorch, with the AdamW optimizer, a linear learning rate scheduler, and CrossEntropyLoss (calculated by reshaping logits and labels appropriately for the multi-output task).
* **Evaluation:**
  + Both models were evaluated on a held-out test set (20% of the data).
  + Metrics included precision, recall, F1-score, and accuracy, calculated per aspect.
  + Overall performance was assessed using micro and macro averages across all aspects.
  + Scikit-learn and PyTorch libraries were used for implementation and evaluation.

### **Intermediate/Preliminary Experiments and Results:** We performed the training and evaluation of both the baseline (Logistic Regression with TF-IDF) and the fine-tuned Transformer (DistilBERT) model.

| **Baseline Model** (Logistic Regression) | **Transformer Model** (Fine-tuned DistilBERT) |
| --- | --- |
| **Food:** Weighted F1-score: 0.61, Accuracy: 0.65  **Service:** Weighted F1-score: 0.74, Accuracy: 0.77  **Ambiance:** Weighted F1-score: 0.88, Accuracy: 0.90  **Price:** Weighted F1-score: 0.92, Accuracy: 0.94  **Context:** Weighted F1-score: 0.82, Accuracy: 0.83  **Overall:** Micro F1: 0.78, Macro F1: 0.68 | **Food:** Weighted F1-score: 0.73, Accuracy: 0.73  **Service:** Weighted F1-score: 0.82, Accuracy: 0.83  **Ambiance:** Weighted F1-score: 0.94, Accuracy: 0.94  **Price:** Weighted F1-score: 0.96, Accuracy: 0.96  **Context:** Weighted F1-score: 0.87, Accuracy: 0.87  **Overall:** Micro F1: 0.87, Macro F1: 0.79 |

**Comparison:** The fine-tuned DistilBERT model consistently outperformed the TF-IDF + Logistic Regression baseline across nearly all aspects, demonstrating significantly higher F1-scores and accuracy, particularly in overall micro and macro averages. This highlights the effectiveness of transformer models in capturing complex patterns and nuances in the review text for this task. Example predictions show the model assigning sentiment labels ('positive', 'neutral', 'negative') to each aspect for given review texts.

Here are some examples from our experiments:

**Review:** "The food was good, the waiter was fast!" **Predicted Sentiments:** Food: positive, Service: positive, Ambiance: negative, Price: negative, Context: negative

**Review:** "The food was bad, the waiter was slow, the building was dirty, and the prices were expensive!" **Predicted Sentiments:** Food: negative, Service: negative, Ambiance: negative, Price: negative, Context: negative

When we tried numerous examples, our transformer model is able to distinguish well between positive and negative sentiment regarding food and service, but struggles with ambiance and price. This appears to be due to the training data having fewer examples in these categories. For the final iteration, more input for these categories will likely be necessary combined with manual tuning of score cutoffs.

### **Related Work**

1. **Duan (2018):** Aspect-based opinion mining with yelp restaurant reviews. This work focuses on identifying common topics using LDA and then extracting aspects and sentiments specifically for restaurants. It discusses preprocessing, aspect extraction, categorization, and sentiment analysis, providing a pipeline similar in goal to this project but potentially using different techniques (like LDA for topic modeling).
2. **Alamoudi and Alghamdi (2021):** Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. This paper analyzes Yelp restaurant reviews using machine learning, deep learning (DL), and transfer learning models for binary and ternary sentiment classification. It proposes an unsupervised approach for aspect-level sentiment analysis (food, service, ambiance, price) based on semantic similarity using pre-trained embeddings like GloVe, achieving high accuracy with models like ALBERT.
3. **Peng (2020):** Aspect based rating prediction for yelp customer review**.** This work predicts Yelp review ratings based on review text and other features. It creates dictionaries to identify <aspect-noun, sentiment word> pairs and uses these as features. The focus is rating prediction rather than just aspect sentiment classification, but the aspect identification step is relevant.
4. **Akhtar et al. (2024):** Explainable Aspect-Based Sentiment Analysis Using Transformer Models. This article examines the performance of various transformer models on ABSA and utilizes explainability techniques (LIME, SHAP, Attention Weights, etc.) to understand their decision-making. While not specific to Yelp, it validates the use of transformers for ABSA, similar to this project's approach, and adds the dimension of explainability.
5. **Musa et al. (2024):** HauBert: A transformer model for aspect-based sentiment analysis of hausa-language movie reviews. This paper presents HauBert, a fine-tuned mBERT model for ABSA on Hausa movie reviews. Although the domain (movies) and language (Hausa) differ, it demonstrates the successful application of fine-tuning transformer models (like BERT) for ABSA tasks, achieving high accuracy in aspect extraction and polarity classification, which aligns with the methodology used in this project with DistilBERT.
6. **Gan Q et al. (2017):** A text mining and multidimensional sentiment analysis of online restaurant reviews. The authors attempt to delineate the structure of online restaurant reviews and investigate the influence of review attributes and sentiments on restaurant star ratings. The authors utilized the Yelp Dataset Challenge consisting of restaurants in Phoenix with 7,508 restaurants and 268,442 reviews identified for inclusion. The authors used a naive Bayes algorithm. The authors demonstrate that five attributes, food, service, ambience, price and context, explain the differences in star ratings with the greatest influence attributed to food, service and context. The inclusion of context as an attribute was an incremental contribution of this work.

**Comparison:** This project aligns with existing research by employing aspect-based sentiment analysis on Yelp data and utilizing transformer models (specifically DistilBERT) for classification but also incrementally builds on this foundation. It differs from some related work by focusing on a specific geographic location (New Orleans) and implementing a multi-output classification approach directly with the transformer model (as opposed to other papers using Naive Bayes, for example), rather than separate models per aspect or solely relying on keyword spotting/lexicons for the final classification step. The discretization of sentiment scores using quantiles is also a specific methodological choice.

**Division of Labor:** The labor was divided equally. Aaron and Jacob both conceived the program components and design, jointly developed and edited the code. Aaron conducted the initial literature review and Jacob added additional papers and input. Jacob conducted the model refinement based upon testing and training data.

### **Timeline:** Over the weekend we will refine our aspect keyword list. Over the next 7 days we will conduct model improvement and hyperparameter with fine tuning and add data for training. Over the next 10 days we will prepare our final report and prepare our presentation for April 22 or 29th.

### **References**

1. Akhtar, Z., Ahmad, T., Khan, M. A., & Ul Abidin, S. Z. (2024). Explainable Aspect-Based Sentiment Analysis Using Transformer Models. *Applied Sciences*, *8*(11), 141.<https://www.mdpi.com/2504-2289/8/11/141>
2. Al-Twairesh, N., & Al-Khalifa, H. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, *30*(2-3), 226-252.<https://www.tandfonline.com/doi/abs/10.1080/12460125.2020.1864106>
3. Duan, T. (2018). *Aspect-based Opinion Mining with Yelp Restaurant Reviews*. Duke University Statistical Science Theses.<https://dukestatsci.github.io/thesis-sp18-duan-sentiment/>
4. Gan, Q., Ferns, B.H., Yu, Y., Jin, L (2016). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality and Tourism*.<https://www.tandfonline.com/doi/abs/10.1080/1528008X.2016.1250243?casa_token=3343pqRoCCwAAAAA:nULDsc5QgFROIP2shCOffBtXiwK_DoOktl2cm0OvD5CdNiU9nCmnH_mOKjlzP-yrIPv5Cfo4SSoumw>
5. Musa, A., Adam, F. M., Ibrahim, U., & Zandam, A. Y. (2024). HauBert: A transformer Model for Aspect-based Sentiment Analysis of Hausa-Language movie reviews. *Proceedings of the 5th International Electronic Conference on Applied Sciences*.<https://sciforum.net/paper/view/20914>
6. Peng, G. (2020). *Aspect Based Rating Prediction For Yelp Customer Review*. Master's Paper, University of North Carolina at Chapel Hill.<https://cdr.lib.unc.edu/downloads/4f16c8090?locale=en>