

## **American University DATA 641 Applied Natural Language Project**

**Project Name:**

**Sentiment Analysis of Dessert Restaurant Reviews with AI-Generated Image Captions**

**Team Members: Yen Jo Lee, Dongni Li, Chen Hsu, Pin Tzu Tseng**

**Last updated: April 28, 2025**

## **1. Introduction**

This project explores sentiment analysis in the context of dessert restaurant reviews, focusing on how AI-generated text, specifically BLIP-2 captions from food images, aligns with human-written reviews. We begin by exploring the sentiment of user reviews, followed by evaluating the performance of sentiment classification models applied to these reviews. Next, we introduce image-generated text and perform initial analysis to understand its relation to user-generated content.

The analysis progresses through the following stages:

- Examining the lexical overlap and differences between generated text and user reviews
- Assessing semantic similarity using BERTScore
- Comparing emotional expression through sentiment polarity analysis
- Evaluating the performance of sentiment classification models (Logistic Regression, Random Forest, MLP, and XGBoost) on both human and AI-generated reviews

This approach allows us to thoroughly examine how well AI-generated content mirrors human-written reviews across multiple dimensions, and assess the potential of AI-generated reviews for tasks such as recommendation systems.

## **2. Research Questions:**

- How well do different classifiers perform in predicting sentiment?
- Can BLIP-2 captions truly reflect the meaning and sentiment of human-written reviews?
- How semantically aligned are the captions with user reviews?
- Does AI-generated text perform well in sentiment prediction?

- What are the implications of integrating AI-generated content into recommendation systems?

### **3. Methodology**

#### **Data Preprocessing:**

The initial dataset is [sourced from Kaggle](#), consisting of 49 Yelp reviews from dessert restaurants, spanning a period of more than 15 years, including all ~20,000 reviews. We focused on the dates from 2021 onwards. Because the original dataset only included text reviews and did not provide accompanying images, we manually collected relevant food images for each restaurant to enable multimodal analysis. To generate descriptive captions for these images, we employed the BLIP-2 model with the Flan-T5-XXL decoder, leveraging its vision-to-language generation capabilities.

For the text data, we performed standard natural language preprocessing to clean and structure the reviews for downstream tasks. Each review was converted to lowercase to maintain uniformity, and all punctuation, digits, and common English stopwords were removed to eliminate noise. The text was then tokenized into individual words to facilitate embedding and analysis. To gain an initial understanding of the linguistic focus of the reviews, we generated a word cloud, which highlighted prominent terms such as “ice cream,” “flavor,” and “place.” These preprocessing steps ensured that the review texts were consistent, semantically meaningful, and ready for embedding using DistilBERT for later classification tasks (Patil et al., 2024).

### **Sentiment Classification Model Comparison(Original Text):**

To classify the sentiment of original Yelp reviews as either positive or negative, we developed several machine learning models. Given the class imbalance in our dataset, with a majority of reviews being positive, we addressed this issue by applying `'class_weight='balanced'` during model training. To transform the textual data into a format suitable for classification, each review was embedded using DistilBERT, producing a 768-dimensional vector representation that captures the semantic meaning of the text (Bagui & Li, 2021).

We evaluated four classifiers: Logistic Regression, Random Forest, Multi-Layer Perceptron (MLP), and XGBoost. MLP, a type of feed-forward neural network, was selected for its ability to model complex nonlinear relationships. To ensure a fair evaluation despite the imbalance, we relied on the Macro F1-score, which gives equal importance to both classes. Among all the models, Logistic Regression outperformed the others, achieving an accuracy of 86.0% and a Macro F1-score of 0.801. In comparison, Random Forest, MLP, and XGBoost achieved slightly lower Macro F1-scores of 0.637, 0.692, and 0.692, respectively. These results demonstrate that even a relatively simple model like Logistic Regression, when combined with high-quality embeddings, can effectively and fairly capture sentiment patterns, especially important given the strong positive skew present in the review data.

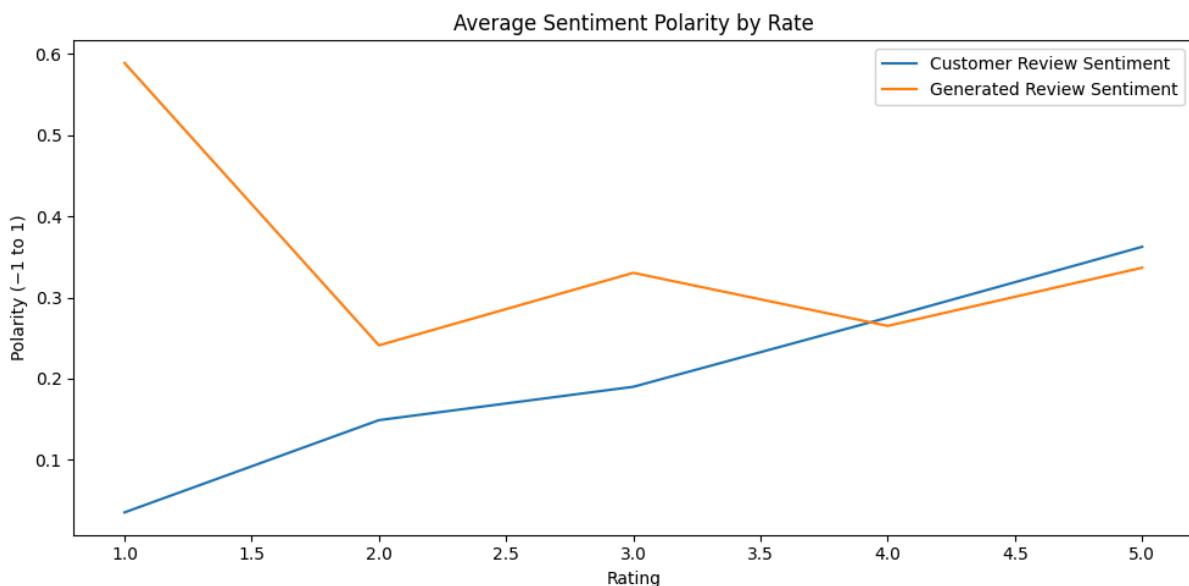
Logistic Regression	Accuracy: 0.860	Macro F1: 0.801
Random Forest	Accuracy: 0.807	Macro F1: 0.637
MLP	Accuracy: 0.820	Macro F1: 0.692
XGBoost	Accuracy: 0.820	Macro F1: 0.692

## **Image-to-Text Generation:**

We used BLIP-2 for image captioning, which operates in two key stages. First, during the vision-and-language representation learning phase, the image encoder extracts features from the input image. Then, in the vision-to-language generative learning phase, a transformer generates natural language descriptions based on the extracted features(Li et al., 2023).

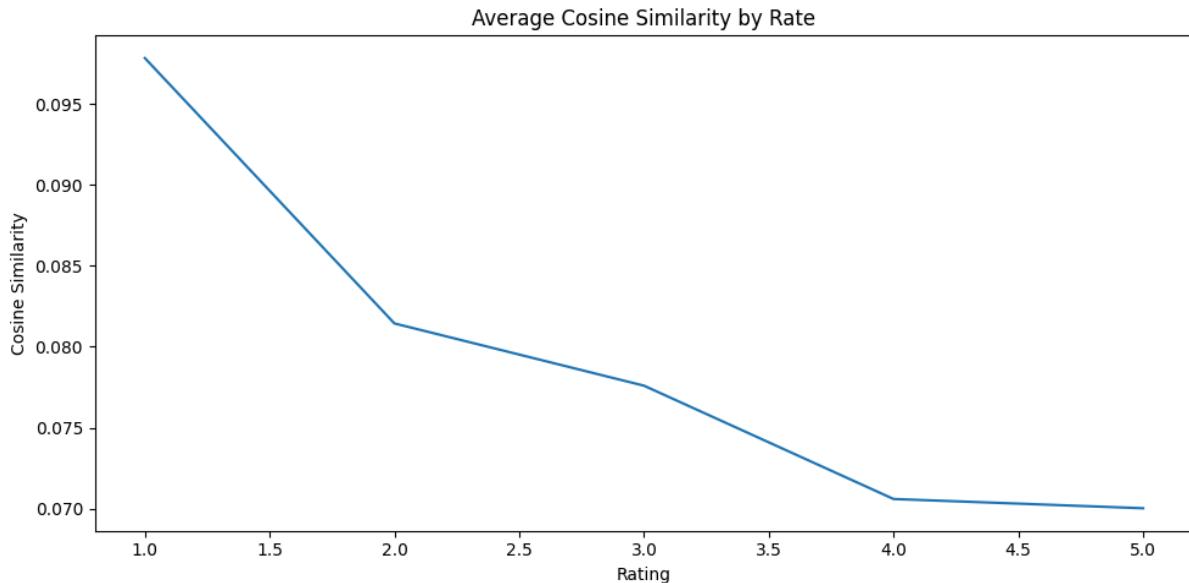
## **Average Sentiment Polarity by Rating**

The blue line, representing customer reviews, shows a positive trend as the rating increases, which is expected. In contrast, the orange line, representing generated reviews, exhibits unusually high sentiment scores (around 0.6) even at a rating of 1, where the sentiment would typically be negative. This discrepancy may be due to the model's difficulty in capturing negative tones or a potential bias toward generating positive statements.



## Average Cosine Similarity by Rating

The similarity between generated and original reviews decreases as the rating increases, with the highest similarity observed at a rating of 1.



From the sentiment polarity perspective, the generative model performs more accurately on high-rated reviews, but tends to be overly positive for low-rated ones indicating a bias in tone generation. On the other hand, cosine similarity reveals that generated reviews are most similar to original ones at low rating, likely due to more consistent tone and language, while high-rated generated reviews show more variability.

## Sentiment Classification Model Comparison:

### Lexical Similarity (TF-IDF)

Using TF-IDF cosine similarity, we found that the lexical similarity between generated text and user reviews was low (0.035), showing that the models produce different wording but maintain semantic meaning.

### BERTScore Evaluation: Semantic Quality of Generated Reviews

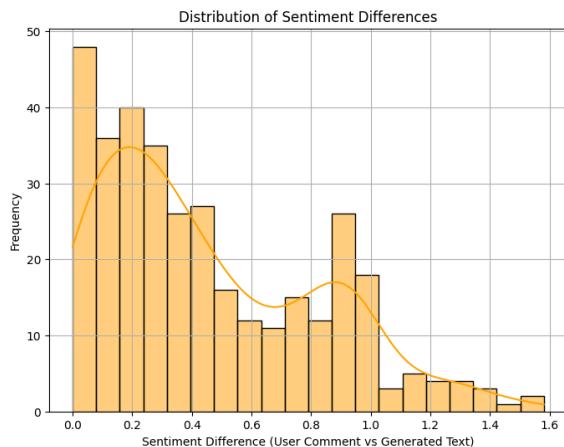
We used BERTScore to measure the semantic similarity between generated text and user reviews:

- Precision: 0.812
- Recall: 0.779
- F1-score: 0.795

These results indicate that while the vocabulary differs, the semantic alignment between generated and human-written reviews is strong.

## Sentiment Comparison Using VADER

We calculated the average sentiment difference between the generated reviews and user reviews, which was 0.456. This suggests that despite the differences in presentation, both convey similar emotional messages.



The histogram shows that most sentiment differences between user comments and generated texts are centered around 0, indicating similar sentiment. However, there are still noticeable differences at higher values, suggesting some generated texts diverge in sentiment. The smooth curve (KDE) reveals that sentiment differences are more frequent at lower values, with fewer differences as they increase, indicating moderate consistency in sentiment.

We trained and evaluated four classifiers to predict sentiment:

Logistic Regression	Accuracy: 0.884	Macro F1: 0.859
Random Forest	Accuracy: 0.812	Macro F1: 0.730
MLP	Accuracy: 0.884	Macro F1: 0.855
XGBoost	Accuracy: 0.855	Macro F1: 0.812

The chart compares four classifiers—Logistic Regression, Random Forest, MLP, and XGBoost—on sentiment classification of generated text using Accuracy and Macro F1 Score. Logistic Regression performs best with 0.88 accuracy and 0.86 Macro F1, followed by XGBoost and MLP. Random Forest has the lowest scores (0.81 accuracy, 0.73 F1), but still performs reasonably well. Overall, Logistic Regression and MLP capture sentiment effectively.

#### **4. Limitation & Future Recommendation**

While our project demonstrates the potential of AI-generated text to mirror human-written restaurant reviews, several limitations remain. First, the emotional expression in the BLIP-2-generated captions was often overly simplified, lacking the nuance and depth typically found in authentic user reviews. Additionally, our dataset focused solely on dessert restaurants from Yelp, which limits the generalizability of our findings to other restaurant types or review domains.

For future work, we recommend expanding the dataset by incorporating reviews from multiple platforms such as Yelp and Google Reviews to provide a broader range of styles, emotions, and customer perspectives. We also suggest enhancing the emotional complexity of generated texts by fine-tuning language generation models specifically for sentiment-rich contexts. With these improvements, future models could achieve more expressive, human-like review generation and stronger performance in nuanced sentiment analysis tasks.

## 5. Conclusion

Our results show that BLIP-2-generated captions are semantically well-aligned with user-written reviews, even though there are noticeable differences in style and emotional depth. The generated text captures the general meaning and sentiment fairly well but tends to simplify emotions, missing some of the richness and personal touches that real user reviews often have.

When it came to sentiment classification, Logistic Regression performed surprisingly well, maintaining high accuracy even on AI-generated text. This suggests that while AI-generated captions might not fully match the complexity of human expression, they can still be very useful for structured tasks like classification. Overall, our findings highlight that AI-generated reviews have strong potential to support review interpretation and recommendation systems, especially if future models can generate more expressive and emotionally varied content.

## **Appendix: Files and Code Submitted**

As part of this project, we have submitted the following files to document and support our analyses:

### **Code Files:**

1. **Original Text Analysis:** Preprocessing, embedding, and sentiment classification based on user-written Yelp reviews.
2. **Image-to-Text Generation:** Caption generation using the BLIP-2 model and Flan-T5-XXL decoder.
3. **Combined Analysis: Cosine Similarity:** Semantic similarity evaluation between original and generated reviews.
4. **Combined Analysis: BERTScore, VADER, and Sentiment Comparison:** Deeper semantic and emotional alignment analysis between the two text sources.

### **Datasets:**

1. **Kaggle Original Dataset:** *Yelp Restaurant Reviews.csv* - original user-written reviews collected from Kaggle.
2. **Merged Dataset:** *merged\_output.csv* - a combined dataset containing both the original reviews and the BLIP-2-generated captions for further analysis.

## Reference

Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-020-00390-x>

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2301.12597>

Patil, R. N., Singh, Y. P., Rawandale, S. A., & Singh, S. (2024). Improving sentiment classification on restaurant reviews using deep learning models. *Procedia Computer Science*, 235, 3246–3256. <https://doi.org/10.1016/j.procs.2024.04.307>

Md Faruk Alam. (August 15th, 2017). Yelp Restaurant Reviews CSV version. Retrieved August 15th, 2017 from  
<https://www.kaggle.com/datasets/farukalam/yelp-restaurant-reviews>