Sweet Sentiment: Sentiment Analysis of Dessert Restaurant Reviews

April 24, 2025

DATA641 Applied Natural Language

Yen Jo (Sally) Lee, Dongni Li,

Chen Hsu, Pin Tzu Tseng

Research Statement

- Measure semantic similarity between BLIP2 captions and user reviews using BERT Score
- Evaluate text usefulness through classification tasks using DistilBERT embeddings
- Compare performance across three inputs:
  - Original text
  - Generated text
  - Combined text
- Explore the potential of AI-generated content in review understanding and recommendation systems

Research Questions

1. How do different classifiers perform in binary sentiment classification of dessert reviews, and which model offers the best balance of accuracy and fairness?
2. Can AI-generated text from food images accurately reflect the meaning and sentiment of human-written restaurant reviews?
3. How semantically similar are the BLIP2-generated captions to original user reviews?
4. Does BLIP2-generated text alone perform well in sentiment or relatedness prediction?
5. What are the implications of integrating AI-generated content into review interpretation or recommendation systems?

Methodology Overview

- Data preprocessing
- Image-to-Text Generation
- BLIP-2, Flan T5-xxl
- Cosine Similarity
- BertScore
- Evaluation
- Recommedation & Limitation

Dataset Overview

- 49 dessert restaurants Yelp reviews on Kaggle, from the year **2021 and beyond**.
- Each review paired with a manually collected food image
- BLIP2-generated captions from each image

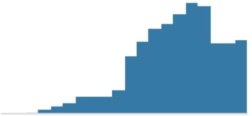
Jul 11, 2022

5 photos

First time here. Great service. Order came through quickly. Clean. The front desk guy explained the ordering system very clearly.

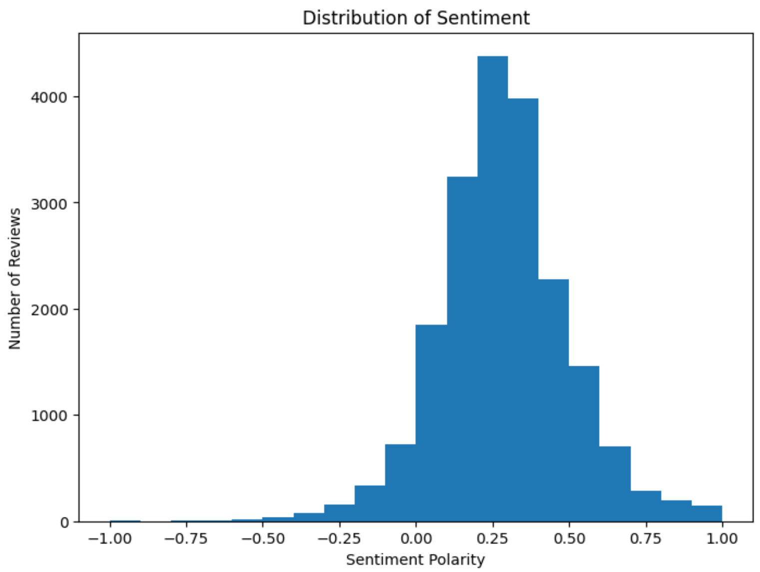
It was really hot so I didn't get the waffle fish, just Ebi ice cream swirl in a cup. The gal preparing my order asked if I wanted toppings which is included in the price. I chose the toasted coconut and a fresh strawberry.

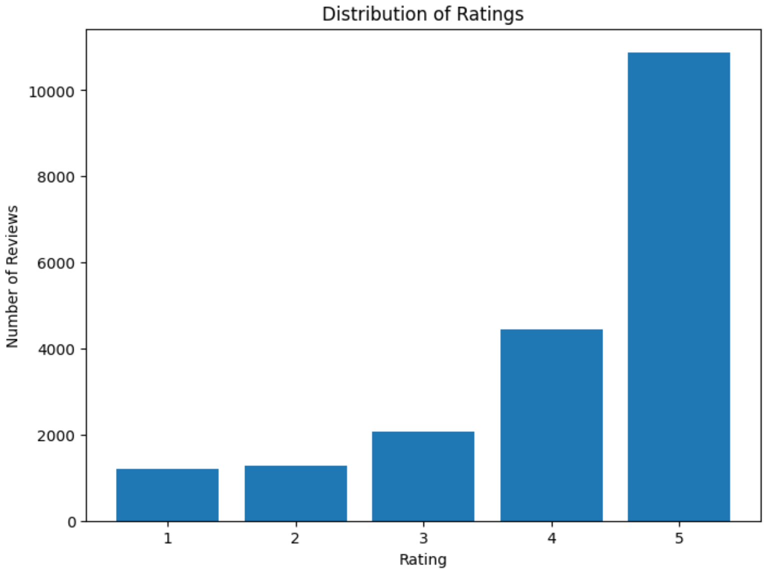
Loved the cold creaminess and the coconut and strawberry provided a nice contrast.

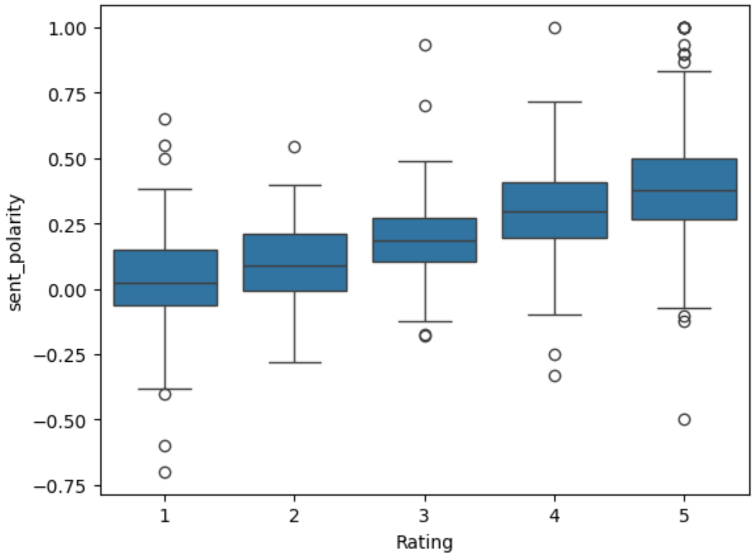
Yelp URL Restaurant store URL	Rating Rating between 1 and 5	Date Date of the review posted	Review Text Text of the review
https://www.yelp.... 10% https://www.yelp.... 7% Other (16468) 83%	 1 5	 2005-06-23 2022-08-01	**19895** unique values
https://www.yelp.com/biz/sidney-dairy-barn-sidney	5	1/22/2022	All I can say is they have very good ice cream I would for sure recommend their cookies and creme ic...
https://www.yelp.com/biz/sidney-dairy-barn-sidney	4	6/26/2022	Nice little local place for ice cream.My favorite is their pumpkin shake (Fall season special).(My...

Exploratory Data Analysis (Original Reviews)

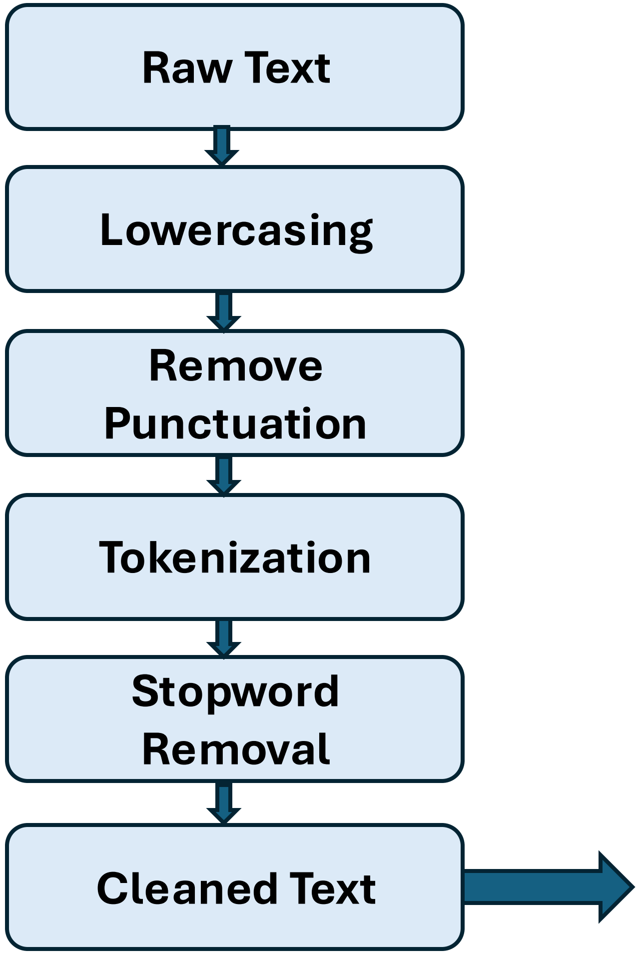
- Most reviews have **4 or 5 stars**, indicating a strong positive skew in user sentiment
- **Negative reviews are underrepresented**, which presents a challenge for model learning and fairness
- Calculated **TextBlob sentiment polarity** for each review
- Found a **moderate correlation (0.49)** between sentiment polarity and actual star rating







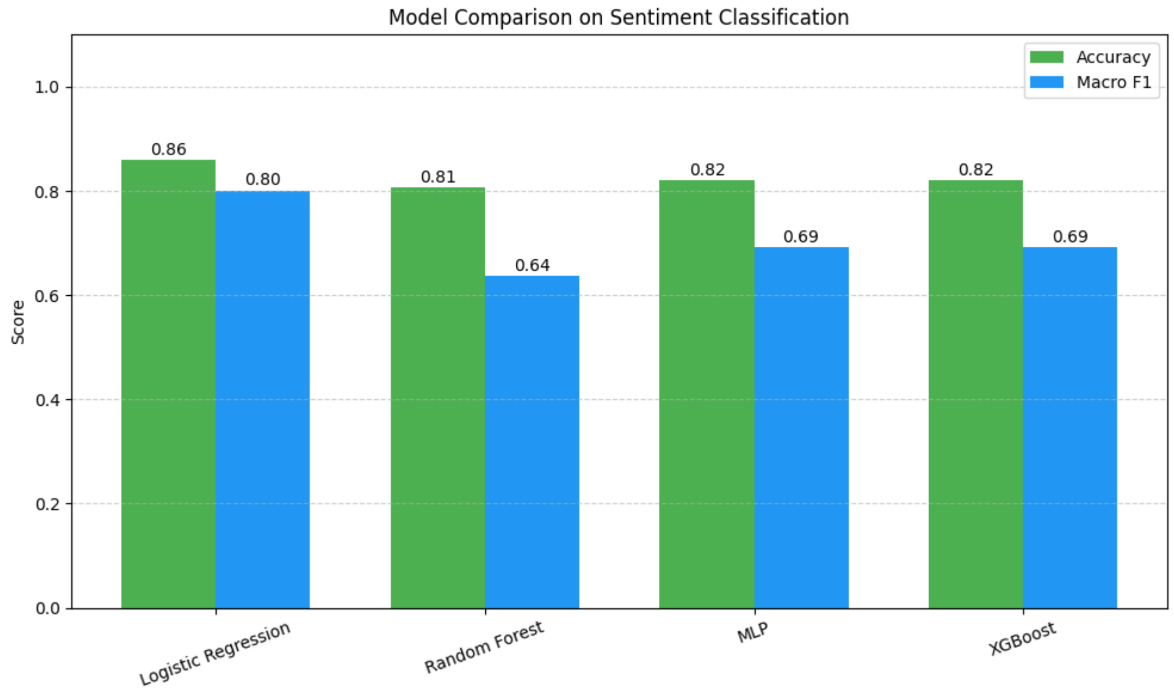
Text Preprocessing Overview



- Converted all text to lowercase for consistency
- Removed punctuation, digits, and common English stopwords
- Tokenized text into individual words
- Retained only sentiment-rich and meaningful words
- Frequent words like "ice-cream," "place," or "flavor" would indicate these are key aspects reviewers focus on.

Sentiment Classification Model Comparison

- Classified Yelp reviews as **positive (rating ≥ 4)** or **negative (rating ≤ 3)**
- Cleaned and embedded review text using **DistilBERT**
- Trained and evaluated four classifiers:
    - Logistic Regression
    - Random Forest
    - Multi-Layer Perceptron
    - XGBoost
- Handled **class imbalance** using `class_weight='balanced'`
- Used **Macro F1-score** to fairly evaluate both classes



Logistic Regression performed best overall, offering strong accuracy and balanced treatment of both sentiment classes — especially important given the positive skew in review data.

Image-to-Text Generation-BLIP2 Image Captioning

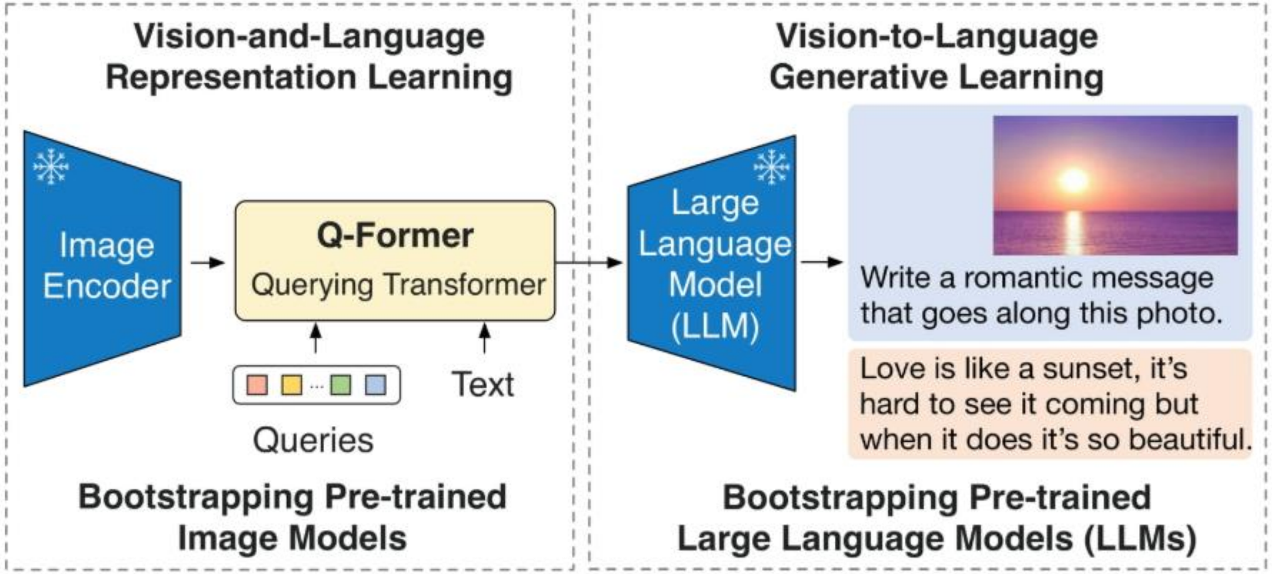


Figure retrieved from https://huggingface.co/Salesforce/blip2-flan-t5-xxl

- The system is divided into two stages:
  - Vision-and-Language Representation Learning
  - Vision-to-Language Generative Learning
- **Image Encoder**: Extracts features from the input image using a pre-trained visual model.
- **Q-Former**: A "querying transformer" that translates image features into a format suitable for language models.
- **Large Language Model (LLM)**: Generates natural language descriptions based on the translated image representation.

Example of **BLIP2-Flan-T5-XXL**

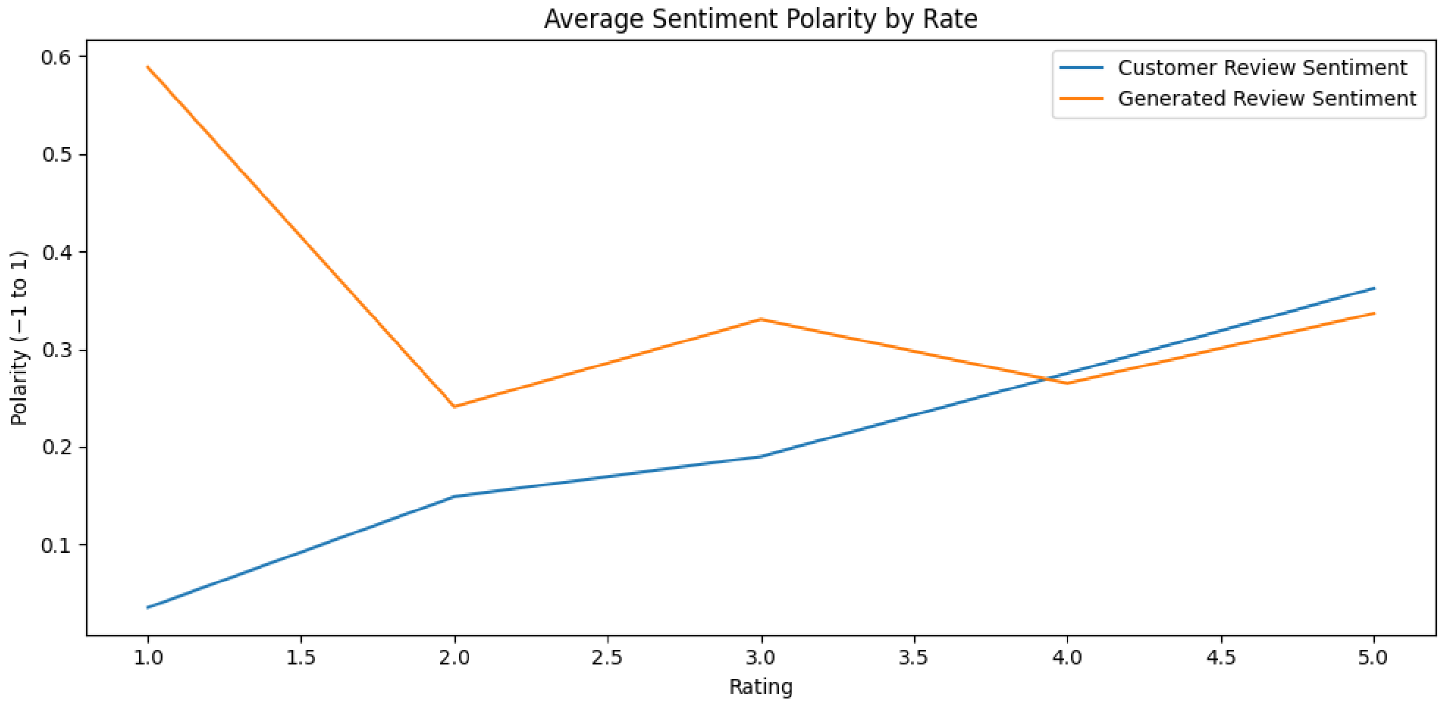
📝 Generated Review Text:
the rainbow cake was delicious and the staff was very friendly

Input Image

📝 Generated Review Text:
i love the croissants here

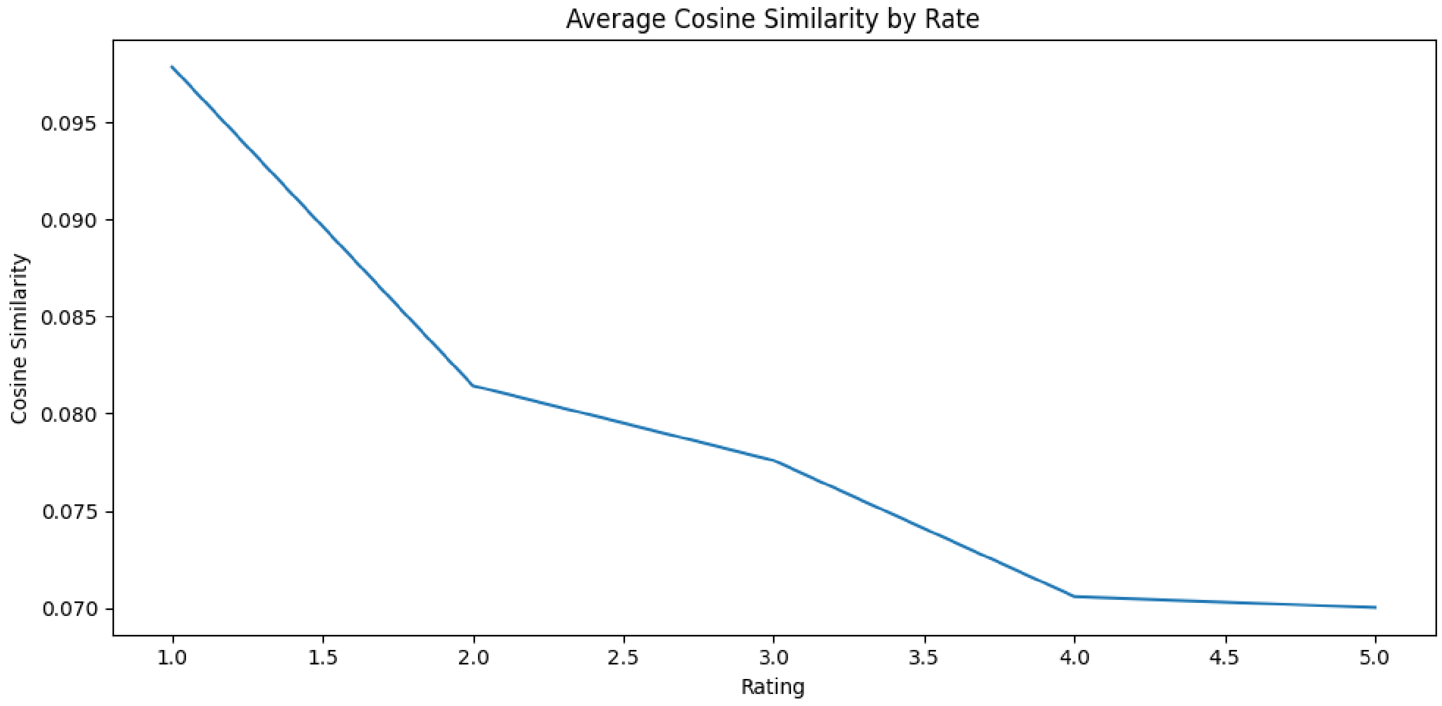
Average Sentiment Polarity by Rating

- **Blue line (Customer Reviews)**
  - Shows a positive trend as the rating increase, which is expected.
- **Orange line (Generated Reviews)**
  - Unusually high sentiment (around 0.6) at rating of 1, which should normally be negative.
- **Possible reasons**
  - The model struggles to capture negative tone, or it has a bias toward generating positive statements.



Average Cosine Similarity by Rating

- The higher the rating, the less similar the generated review is to the original.
- At a rating of 1, generated reviews are the most similar to the original ones.



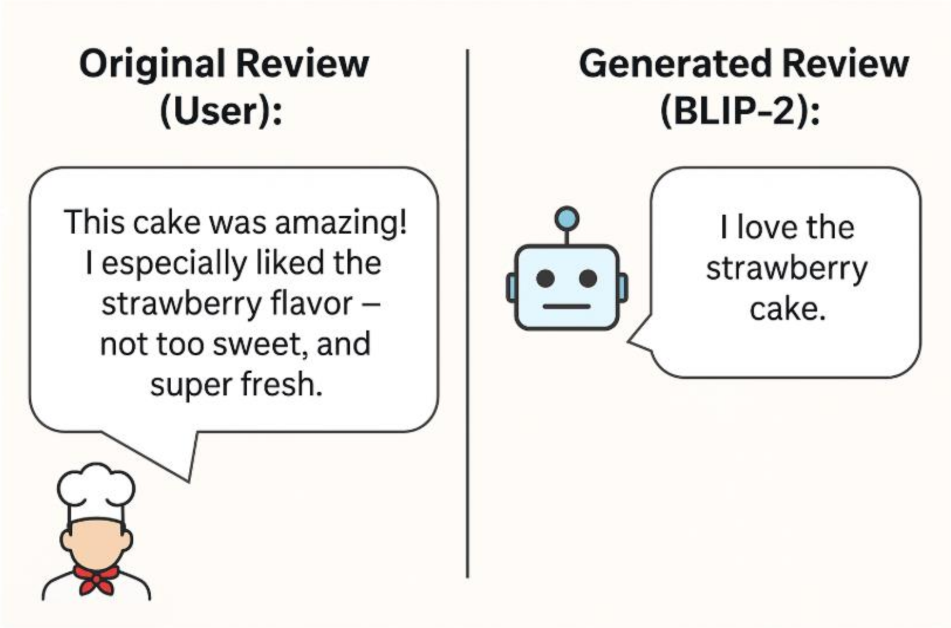
BERT Score: Evaluating Semantic Quality of Generated Reviews

Through TF-IDF:

There is a significant difference in the **lexical level** between the generated text and user comments (Cosine similarity: 0.035).

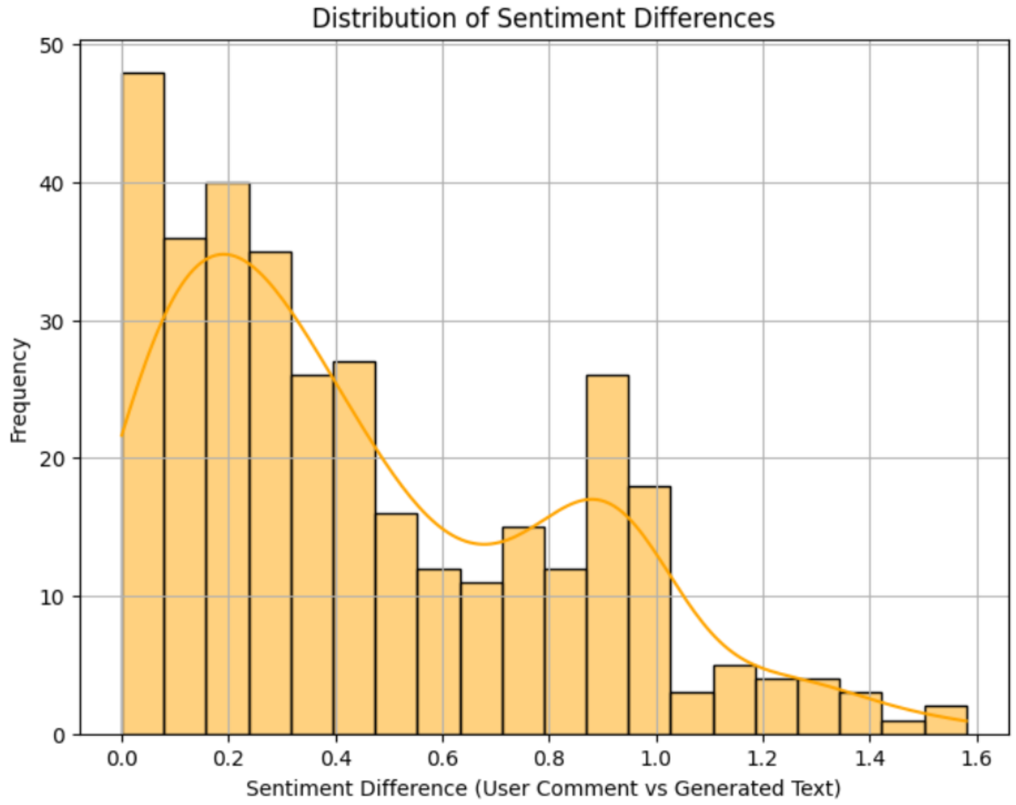
Using BERTScore:

- **Precision**: 0.812 → Generated content is mostly relevant
- **Recall**: 0.779 → Captures most of the reference meaning
- **F1 Score**: 0.795 → Strong overall semantic alignment



VADER Sentiment Analysis

- Average Sentiment Difference: 0.456
- The sentiment difference is centered around 0, indicating that the generated text is similar to the user comments in terms of emotional expression



Evaluation Metrics

- Accuracy and Macro F1 scores
- Logistic Regression, Random Forest, MLP and XGBoost
- Sentiment classification is applicable to generated text, demonstrating that the model effectively captures sentiment, even when the generated text differs in syntax and wording from user reviews.

Limitation & Future Recommendation

- Combine **Yelp/Google Reviews** with **image-based AI generation**, allowing users to generate customized review texts based on their preferences and their uploaded photos.
- Expand Dataset Diversity
- Enhancing the emotional intensity of generated content, making it more vivid and emotionally layered.
- Emotion Strength Control

Conclusion

- The **generated texts and user reviews** demonstrate a **high degree of semantic alignment**.
- The **emotional expression in generated texts** tends to be **simplified**, while **user-written reviews** are more **personalized** and exhibit **greater emotional variability**.
- The **best-performing sentiment classifier (Logistic Regression)** can be effectively applied to the generated texts.

Thank you for listening