جامعة أم القرى
UMM AL-QURA UNIVERSITY

project of Data Analysis 2

Name	ID
Mawadh Saleem Aloufi	444005971
Reham Hameed Alqurashi	444002142

جامعة أم القرى
UMM AL-QURA UNIVERSITY

Text Analysis

Using Logistic Regression to Classify Yelp Reviews

Objective:

The goal of this task is to build a text classification model to categorize Yelp reviews as positive or negative. This model helps automate sentiment analysis and derive insights from customer feedback.

- **Text Preprocessing**

The following steps were performed to preprocess the raw text data:

• **Lowercasing and Punctuation Removal:** All reviews were converted to lowercase, and punctuation was removed to standardize the text.

• **Tokenization and Stop Words Removal:** Reviews were split into individual words (tokens), and common stop words (e.g., " the"," is", " and ") were removed to focus on meaningful words.

• **TF-IDF Vectorization:** The text data was transformed into numerical form using TF-IDF (Term Frequency-Inverse Document Frequency) to capture the importance of each word.The top 5000most frequent words were used for modeling.

- **Model Performance**

The Logistic Regression model was trained to classify the reviews as positive or negative. Below are the performance metrics of the model.

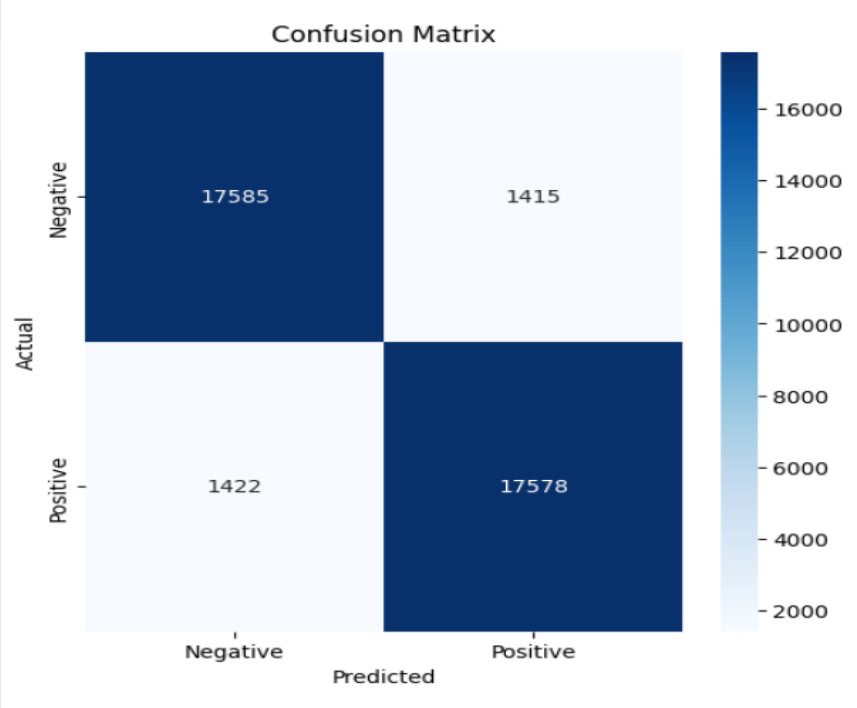
1 Accuracy

The model achieved an accuracy of 92.53% on the test set, indicating high effectiveness in classifying customer reviews.

2 Confusion Matrix

The confusion matrix illustrates the number of correctly and incorrectly classified reviews:

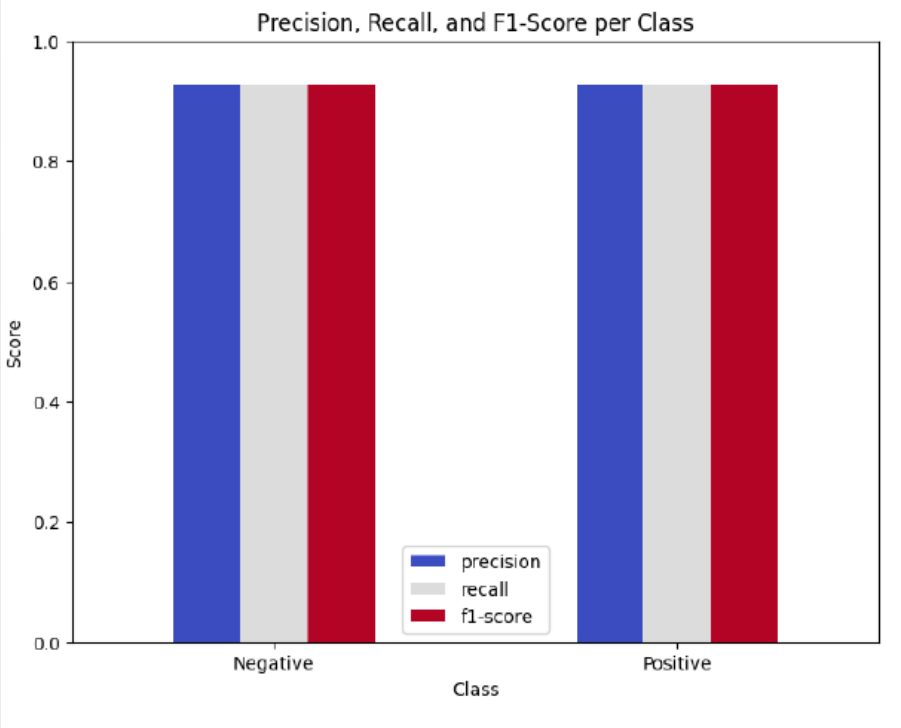
- 17,585 negative reviews were correctly classified.
- 17,578 positive reviews were correctly classified.
- 1,415 negative reviews were misclassified as positive.
- 1,422 positive reviews were misclassified as negative.



- ## Precision, Recall, and F1-Score

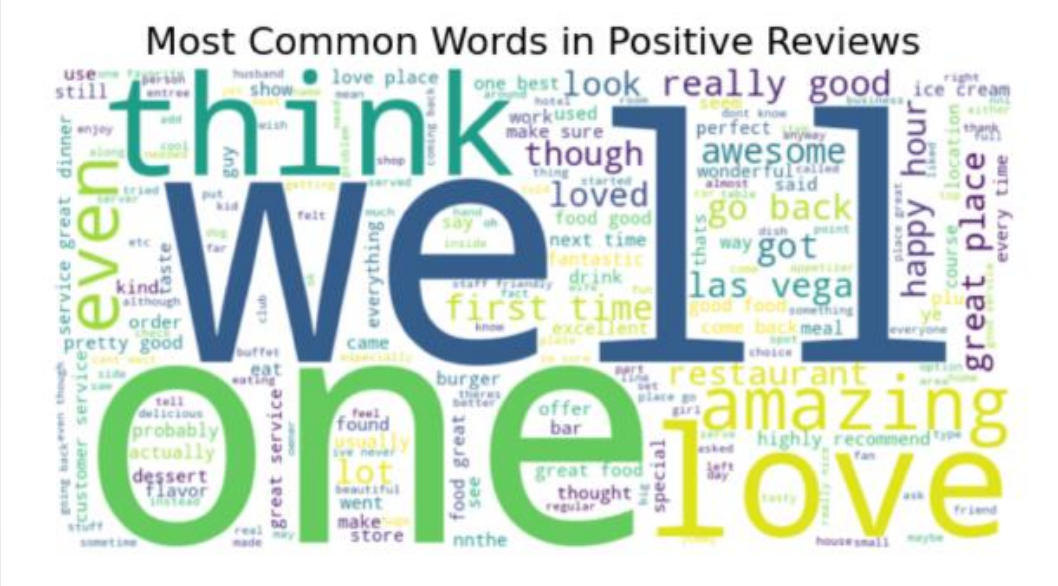
The model is performance across various metrics for both positive and negative reviews is shown below:

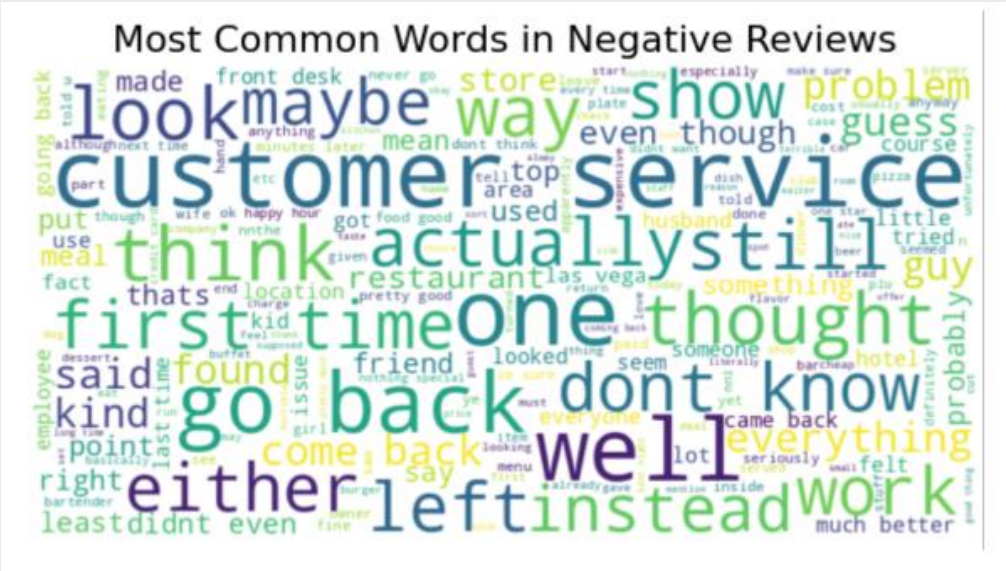
- **Precision:** 0.93 for both positive and negative reviews.
- **Recall:** 0.93 for both positive and negative reviews.
- **F1-Score:** 0.93 for both positive and negative reviews.



- **Common Words Visualization**

The most frequent words from the reviews were visualized to provide insights into the language used in positive and negative reviews.





- **Testing with Real Reviews**

The model was tested with real-world reviews to assess its ability to generalize beyond the test set:

- Review 1: "The service was terrible, and the food was cold. I will never come back! "

Predicted Sentiment: Negative

- Review 2: "Absolutely loved the place! The staff was friendly, and the food was amazing! " Predicted Sentiment: Positive

جامعة أم القرى
UMM AL-QURA UNIVERSITY

- **Conclusion**

This text classification model demonstrates its effectiveness in categorizing Yelp reviews, with strong performance across key metrics. The insights derived from this model can help businesses automate sentiment analysis and make informed decisions based on customer feedback.