Sentiment Analysis Report

1-Text Preprocessing:

The preprocessing steps in this sentiment analysis project aimed to clean the textual data and prepare it for machine learning. The key steps included:

1. **Tokenization**:
    - Splitting the text into individual tokens (words).
    - For example, the sentence "I love this product!" becomes ['I', 'love', 'this', 'product'].
    - This ensures that each word can be analyzed separately.
2. **Stop Words Removal**:
    - Common words like "the", "is", "and" were removed, as they do not contribute much meaning in sentiment analysis.
    - This reduces noise and focuses on meaningful words.
3. **Lemmatization**:
    - Words were converted into their base forms (e.g., "running" becomes "run").
    - This ensures that variations of the same word are treated uniformly, reducing the dimensionality of the dataset and improving model accuracy.
4. **Word Cloud Visualization**:
    - A word cloud was generated to visualize the most frequent words, providing insight into the most discussed topics in the dataset.

2-Model Performance:

The model trained for sentiment analysis used the Yelp review dataset, where the target variable was the "stars" ratings, indicating different levels of sentiment (from negative to positive).

Key steps observed:

1. **Data Overview**:
    - The dataset was loaded, and descriptive statistics were used to understand the distribution of reviews and ratings.
2. **Model Training**:
    - A machine learning algorithm was applied to predict the sentiment based on reviews.
    - A Naive Bayes model was used to classify the texts into positive and negative texts.
3. **Evaluation Metrics**:
    - The model performance can be evaluated using standard classification metrics:
        - **Accuracy**: Ratio of correct predictions to total predictions.
        - **Precision**: True positive predictions compared to all predicted positives.
        - **Recall**: True positives compared to all actual positives.
        - **F1-score**: Harmonic mean of precision and recall.
    - These metrics help assess how well the model classifies reviews into their respective sentiment categories.

3-Insights and Analysis:

Several important insights were gained from this analysis:

1. **Sentiment Distribution**:
    - The distribution of star ratings likely reflects overall customer satisfaction. For example, a majority of 5-star reviews may indicate general customer satisfaction, while a large number of 1-star reviews might highlight specific areas of dissatisfaction.
2. **Frequent Words in Reviews**:
    - Word clouds and token frequency analysis identified the most common words in the reviews. For instance, words like "service," "food," or "price" may appear frequently if they are common topics of feedback.

4- Conclusion:

This sentiment analysis project successfully implemented text preprocessing techniques to clean and prepare the data for machine learning. The model provided valuable insights into customer sentiment through Yelp reviews, and several key findings emerged regarding frequent topics and overall sentiment trends. Future improvements could involve refining the model and preprocessing steps for enhanced accuracy and deeper insights.

5-Authors:

Name	S.ID
Eyad Mohammed Mayudh Alharthi	444000005
Mohammed saad Mohammed Alsarhani	444006145