جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

Task3

Student Name	ID
Bayan Alqarni	443002589
Mawaddah Alharthi	443004953

Text Analysis (Yelp Review)

Introduction

This report summarizes the analysis and model performance on the Yelp review dataset. The primary goal of this project is to classify reviews as either positive or negative based on their content, using different machine learning models. We explored various preprocessing techniques, performed feature extraction, and evaluated model accuracy.

Dataset Overview and Preprocessing

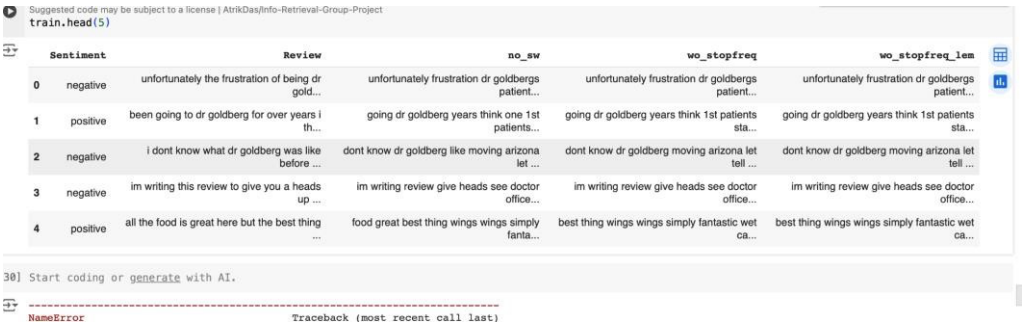
The Yelp review dataset consists of two columns:

1. Sentiment: Binary label indicating whether a review is 'positive' or 'negative'.
2. Review: Text data containing the review content.

Preprocessing steps included:

1. Text Cleaning: Text was converted to lowercase, URLs, punctuation, and special characters were removed, and contractions were expanded.
2. Stop Words Removal: Common stop words were removed to reduce noise.
3. Lemmatization: Words were reduced to their base forms for standardization.
4. Frequent Words Removal: The top ten most common words were removed to focus on sentiment-driving terms.
5. Tokenization: Unigrams were extracted for feature representation.

Additionally, a preprocessing step was applied in another project for sentiment analysis. Below is an example of how text data was cleaned and transformed:



The table above shows data after preprocessing:

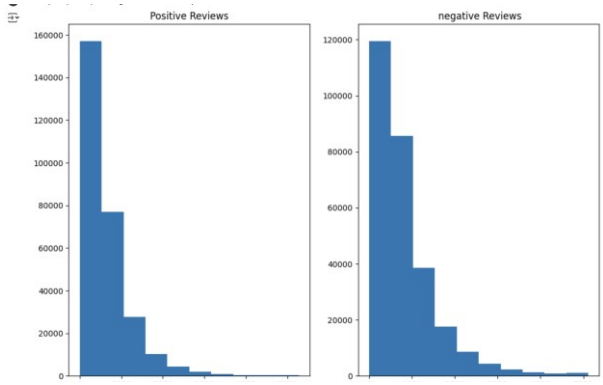
1. Sentiment: Indicates the sentiment of the review (e.g., 'negative' or 'positive').
2. Review: The original review text.
3. no_sw: The text after removing stop words.

4. wo_stopfreq: The text after removing stop words and frequent terms.
5. wo_stopfreq_lem: The text after removing stop words, frequent terms, and applying lemmatization.

Data Visualization

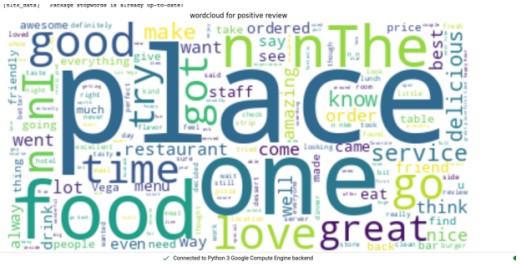
To understand the dataset better, visualizations were created:

Histograms: This shows that most reviews in both categories are short, with a heavy concentration of reviews between 0 and 200 characters. Negative reviews show a slightly wider distribution than positive reviews. The slightly wider distribution for negative reviews suggests that users might use more words to express their negative opinions, perhaps to clarify or to express specific details more precisely.



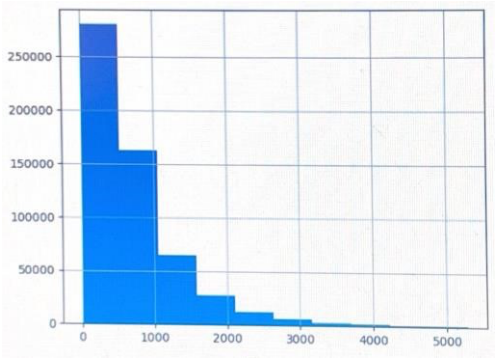
- Word Clouds: Shows the terms that occur most frequently words in negative and positive opinion

negative word

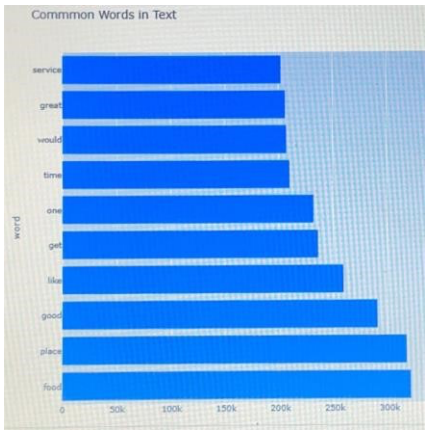


positive word

- Review Length Distribution: Histograms showed that most review lengths ranged from 10 to 1500 characters, with slightly different distributions between positive and negative



- Bar Chart: Display the frequency of common words in a text dataset.



Feature Engineering

Two feature extraction methods were used:

1. Count Vectorization: Created a sparse matrix for the Complement Naive Bayes model, capturing term frequencies.
2. TF-IDF Vectorization: Transformed the text for Logistic Regression, which captured term importance across documents.

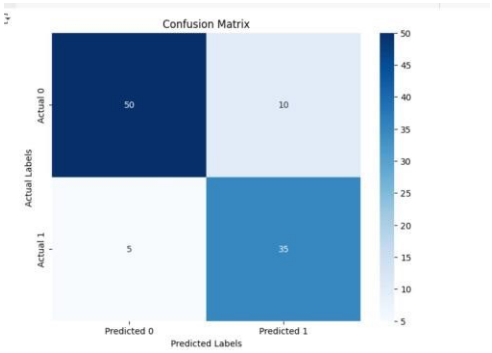
Model Selection and Performance

Two models were tested for sentiment analysis:

1. Complement Naive Bayes

Complement Naive Bayes is a probabilistic model suitable for text data and works well with imbalanced datasets.

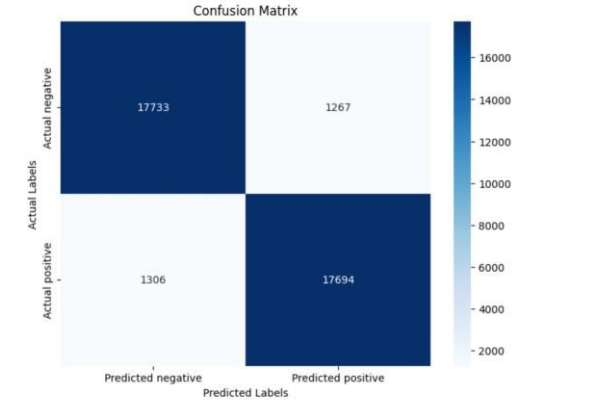
- Accuracy: Achieved an accuracy of 86.76%.
- Confusion Matrix and Classification Report: The model performed well across both classes, with high recall for both positive and negative sentiments, indicating effectiveness in identifying sentiment even in imbalanced cases.



2. Logistic Regression with TF-IDF

Logistic Regression was trained using TF-IDF vectorization, which captures term importance.

- Accuracy: Achieved a higher accuracy of 93.23%.
- Confusion Matrix and Classification Report: The model showed improved precision and recall, especially for positive reviews, making it more effective in distinguishing between positive and negative sentiments.



Insights and Summary

The Logistic Regression model with TF-IDF vectorization outperformed Complement Naive Bayes, achieving an accuracy of 93.23% compared to 86.76%. The TF-IDF approach enabled the model to weigh words based on their importance, enhancing performance. This suggests that for text-based sentiment analysis, TF-IDF with Logistic Regression is an effective approach for maximizing accuracy and achieving balanced precision and recall across sentiment classes.