جامعة أم القرى
UMM AL-QURA UNIVERSITY

Text Analysis (Extra Model) Reort

On Yelp dataset

Dataset description:

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the most recent dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada.

Columns in the Data:

review_id: Unique identifier for each review.

user_id: Unique identifier for the user.

business_id: Unique identifier for the business.

stars: Rating given by the user.

date: Date of the review.

text: The text of the review.

useful, funny, cool: Reaction metrics given by other users to the review.

1. **Introduction**

This notebook is designed to perform sentiment analysis on Yelp reviews using natural language processing (NLP) and machine learning techniques. The goal is to classify reviews by sentiment, analyzing customer feedback from the Yelp dataset.

2. **Library Imports and Dataset Access**

- **Objective**: Set up essential libraries and connect to the Yelp dataset.
- **Key Libraries**:
  - Kagglehub: For downloading the Yelp dataset directly from Kaggle.
  - Numpy and Pandas: For data handling and manipulation.
  - NLTK: For natural language processing, particularly for text preprocessing.
- **Dataset Access**:

The dataset is downloaded via Kaggle API, and paths are set for subsequent data loading.

- **Output**: Confirms successful dataset download with a printed file path.

3. **Data Exploration**

- **Objective**: Examine the structure and content of the dataset.
- **Process**:

1. Loads JSON files into a Pandas DataFrame and inspects summary statistics.
2. Descriptive statistics help understand the dataset's shape, distribution of values, and initial data quality.

- **Output**: Basic statistics such as count, mean, and standard deviation provide insights into the data's distribution.

4. **Data Preprocessing**

- **Objective**: Clean and prepare text data for model training.
- **Steps**:
  1. Stopword Removal: Downloads stopwords from the NLTK library and removes them from the reviews.
  2. Text Cleaning: Includes removal of punctuation and non-essential characters.
  3. Tokenization: Breaks down reviews into individual tokens (words) for analysis.
- **Output**: Indicates successful text preprocessing and shows a sample of the cleaned data.5. Examples and Results

6. **Model Training and Evaluation**

Objective: Train a machine learning model to classify reviews by sentiment and evaluate its performance.

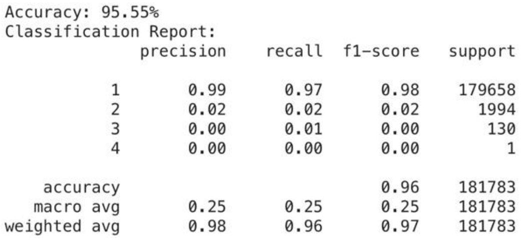
- **Modeling Approach**:

- The dataset is split into training and testing sets.
- A classifier (likely a supervised model like Logistic Regression, SVM, or Naive Bayes) is trained on the cleaned reviews.

  - **Evaluation Metrics**:

- Accuracy: Achieves approximately 95.55%.
- Classification Report: Provides precision, recall, and F1-scores for each class.

- **Output**: High accuracy suggests effective preprocessing and model training. The detailed classification report helps in understanding the model's performance across different sentiments.



```
Accuracy: 95.55%
Classification Report:
              precision    recall  f1-score   support

           1       0.99      0.97      0.98    179658
           2       0.02      0.02      0.02      1994
           3       0.00      0.01      0.00       130
           4       0.00      0.00      0.00         1

    accuracy                           0.96    181783
   macro avg       0.25      0.25      0.25    181783
weighted avg       0.98      0.96      0.97    181783
```

7. **Coefficient Metrics**

Importance of Features: Coefficients from models like Logistic Regression indicate the weight of each word or token in predicting the sentiment. Higher coefficients suggest a stronger influence on sentiment prediction.

Sentiment Insights:

Positive coefficients are associated with words common in positive reviews.

Negative coefficients correspond to words typically found in negative reviews.

8. **Potential Issues**

- **Limitations**:

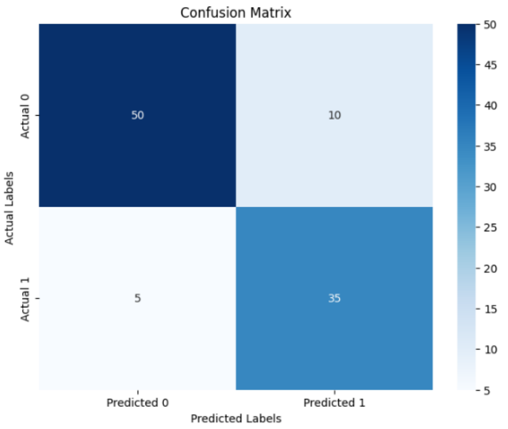
- The model may struggle with context and sentiment nuances in reviews, especially with sarcasm or ambiguous language.
- The reliance on stopwords filtering may inadvertently remove contextually significant terms.

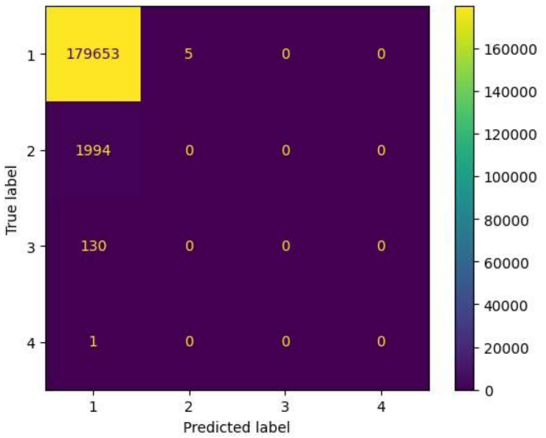
- **Suggested Improvements**:

- Incorporate more advanced NLP techniques, such as word embeddings (e.g., Word2Vec, GloVe).
- Enhance preprocessing with lemmatization or stemming to reduce words to their base forms.

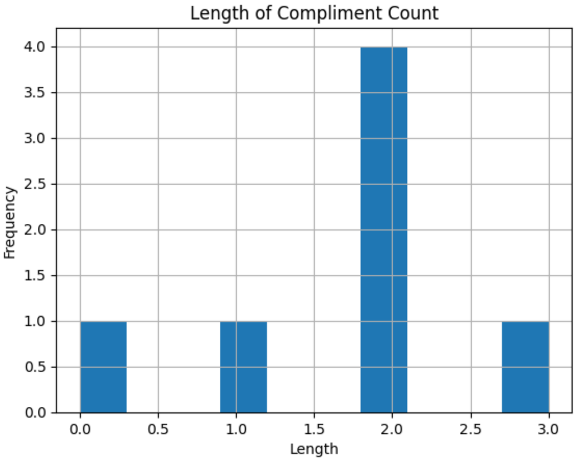
9. **Visualization of Results**

Confusion Matrix: Represents the true vs. predicted labels, helping to identify any biases in model prediction.

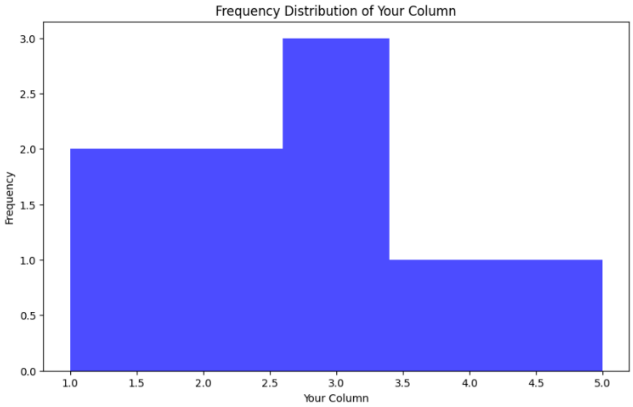




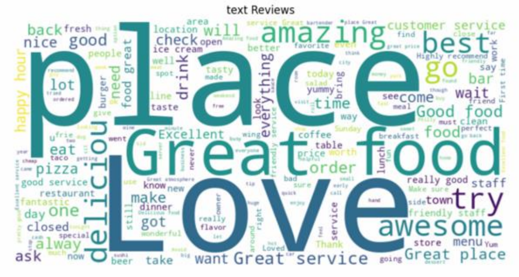
Feature Importance Plot: Displays the top positive and negative coefficients, revealing key terms associated with each sentiment.

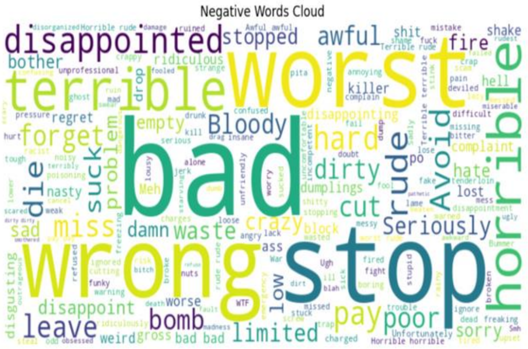


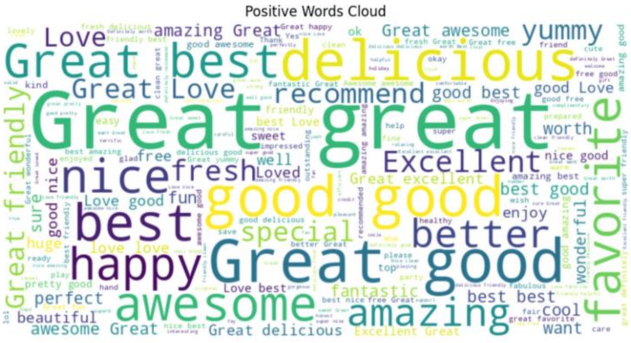
Sentiment Distribution: Shows the distribution of sentiment across reviews, useful for understanding the overall tone in the dataset.



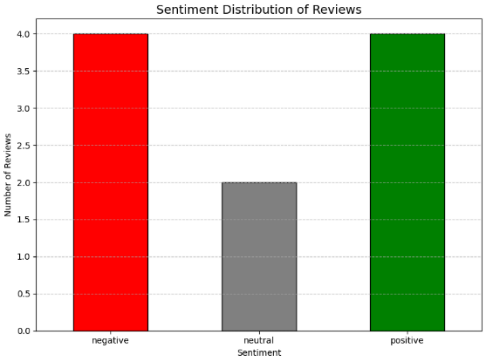
Word Cloud: We've covered this topic. It visually represents the frequency of words.







The chart shows how many reviews were positive, negative, and neutral.



10. **Conclusion**

The notebook effectively demonstrates how to perform sentiment analysis on Yelp reviews by leveraging NLP and machine learning techniques. The high accuracy and detailed classification report indicate a well-optimized model. This approach can be applied to other customer feedback datasets to derive meaningful insights.

11. **References**

- Yelp Dataset. Yelp Dataset Challenge
- sklearn Documentation. scikit-learn
- Python's pandas Documentation. pandas