جامعة أم القرى

UMM AL-QURA UNIVERSITY

Data Analysis Project

" Text Analysis REPORT"

NAME:	ID:
Wed Alshehri	444015020
Ftoon Alshmaimri	444003233

Dr. Omima Fallatah

1. Text Preprocessing

The text data underwent extensive preprocessing to ensure optimal model performance and interpretability. This process included:

1. **Removing Missing and Duplicate Rows**: Initial data cleaning included removing rows with missing values and duplicates in critical columns (Score and Text). This step reduced noise and redundancy in the dataset, leaving us with a cleaner set of unique reviews.
2. **Tokenization and Stopword Removal**: Tokenization split the reviews into individual words, enabling further processing steps. Stop words (common, non-informative words) were removed, while negative stop words like "not" were retained to preserve review sentiment.
3. **Stemming and Lemmatization**: Using NLTK's `PorterStemmer`, we stemmed words, reducing them to their base forms (e.g., "running" to "run"). This normalization improved model generalization by grouping word variants together.
4. **Data Balancing**: With an imbalance across review classes (Positive, Neutral, Negative), subsampling was performed on the Positive and Negative categories to match the count of Neutral reviews, creating a balanced dataset.
5. **Word Cloud Visualization**: Word clouds were generated for each sentiment category (Positive, Neutral, Negative) to visualize the most frequent words. These word clouds helped identify common themes in each review type, such as frequently used positive and negative adjectives.

2. Model Performance

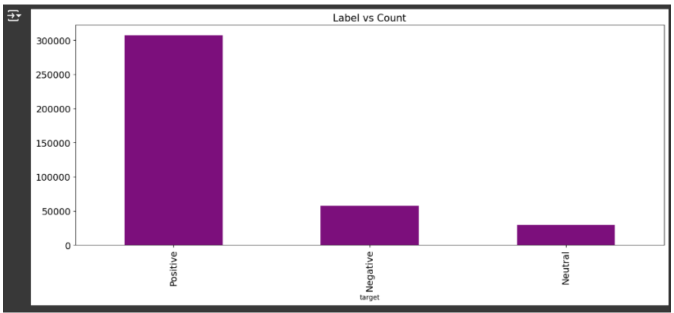
Two models were trained to classify review sentiments: **Naive Bayes** and **Logistic Regression**, both utilizing a Bag-of-Words (BoW) approach.

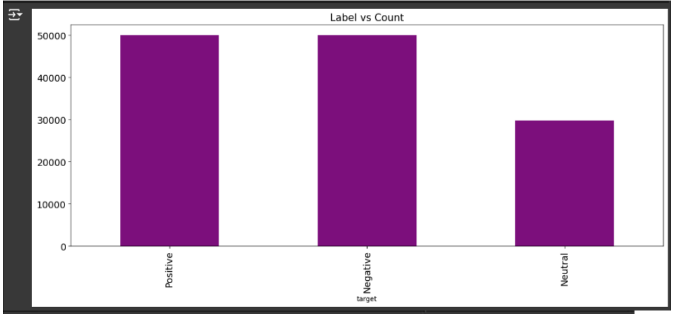
1. **Naive Bayes Classifier**: This model achieved satisfactory accuracy, leveraging word frequency features to predict sentiments. Its performance is reliable in text classification but may sometimes struggle with more nuanced or complex language.
2. **Logistic Regression**: Logistic Regression outperformed Naive Bayes, likely due to its ability to capture linear relationships in the data. We achieved high accuracy on the test data, as Logistic Regression handled BoW representations effectively. The model's confusion matrix revealed that most errors were between Neutral and Negative categories, suggesting some overlap in language among these reviews.

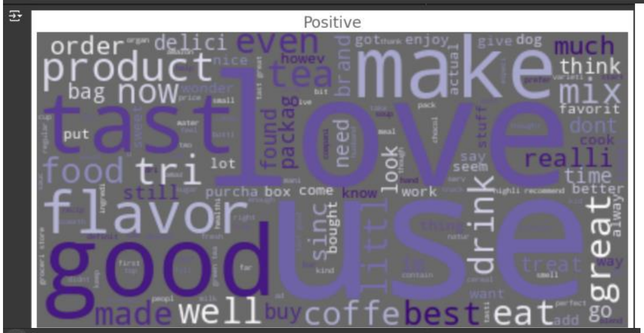
3. Insights from the Analysis

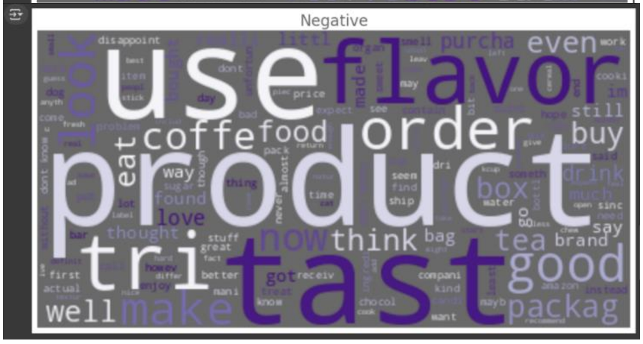
1. **Sentiment Distributions**: Positive reviews are more prevalent, a common occurrence in product review datasets, while Neutral reviews are less frequent. The balanced dataset allows us to compare each class fairly and achieve reliable predictions across categories.
2. **Common Themes in Reviews**: Word clouds demonstrated that Positive reviews frequently included terms like "love," "great," and "recommend," while Negative reviews had terms like "disappointed," "bad," and "waste." This suggests distinct sentiment language patterns, which can help organizations identify key factors driving customer satisfaction or dissatisfaction.
3. **Model Recommendations**: Based on accuracy and interpretability, Logistic Regression is recommended for production deployment, as it offers a balance of performance and computational efficiency. However, refining the model with additional NLP techniques, such as TF-IDF or word embeddings, may enhance performance further.

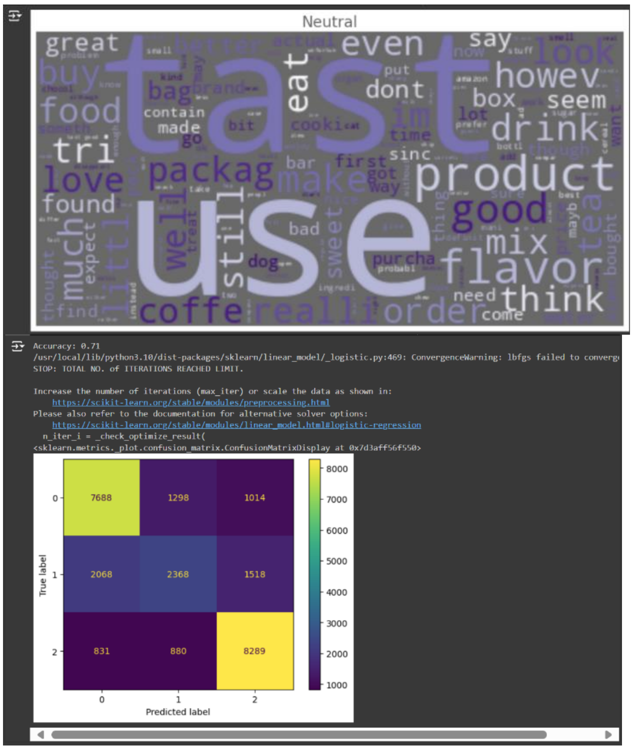
4. Visualization











```
Accuracy: 0.71
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:469: ConvergenceWarning: lbfgs failed to converg
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7d3aff56f550>
```

5. Conclusion

In conclusion, this analysis provided insights into customer sentiment and model capabilities, showing how specific text preprocessing techniques enhance model accuracy in sentiment classification.

جامعة أم القرى

UMM AL-QURA UNIVERSITY