

**Artificial Intelligence**

**Project Documentation**

**SECTION: “B”**

**SUBMITTED TO:**

MISS HIRA KHALID

**SUBMITTED BY:**

SHAHZAD BABAR 22-SE-087

**DEPARTMENT OF COMPUTER SCIENCE HITEC UNIVERSITY, TAXILA**

**Sentiment Analysis on IMDB Movie Reviews**

**Project Overview**

The primary goal of this project is to perform sentiment analysis on a dataset of IMDB movie reviews. Sentiment analysis involves determining whether a piece of text, in this case, a movie review, is positive or negative. This can help in understanding the general reception of movies and can be valuable for various stakeholders in the film industry.

**Methodology:**

**Data Collection**

The dataset used for this analysis is the IMDB Dataset, which contains 50,000 movie reviews labeled as either positive or negative.

**Data Preprocessing**

Data preprocessing is a crucial step in preparing the dataset for machine learning models. The following preprocessing steps were performed:

1. **Reading the Dataset:** The dataset was read into a Pandas DataFrame.
2. **Exploratory Data Analysis (EDA):**

* Visualized the distribution of sentiments.
* Displayed a few sample reviews and their corresponding sentiments.

**Text Cleaning:**

* Lowercased the text.
* Removed HTML tags, URLs, punctuation, and special characters.
* Tokenized the text.
* Removed stop words.
* Performed stemming using the Porter Stemmer.

1. **Handling Duplicates:** Removed duplicate reviews from the dataset.
2. **Feature Engineering**: Added a feature for the word count of each review.

**Data Visualization;**

To gain insights from the data, various visualizations were created:

1. **Sentiment Distribution:** Count plot showing the distribution of positive and negative sentiments.
2. **Word Count Distribution:** Histograms showing the distribution of word counts for positive and negative reviews.
3. **Review Length Distribution:** Histograms showing the length of reviews for positive and negative sentiments.
4. **Word Clouds:** Word clouds depicting the most frequent words in positive and negative reviews.
5. **Bar Plots:** Bar plots showing the most common words in positive and negative reviews.

**Model Training and Evaluation:**

Three different machine learning models were trained and evaluated for sentiment analysis:

* **Logistic Regression**
* **Multinomial Naive Bayes**
* **Linear Support Vector Classifier (SVC)**

The dataset was split into training and test sets (70-30 split). The TF-IDF Vectorizer was used to convert text data into numerical vectors suitable for model training.

**Hyper-parameter Tuning;**

Hyper-parameter tuning was performed for the Linear SVC using Grid-Search-CV to find the best parameters.

**implementation details:**

**Data Loading and Initial Exploration:**



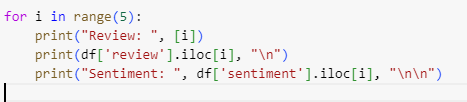
**Libraries:** Imports essential libraries for data manipulation, visualization, and machine learning.

**Data Loading:** Reads the IMDB dataset into a Data-Frame and displays the first few rows.

**Shape and Info:** Outputs the shape and summary information about the Data-Frame.

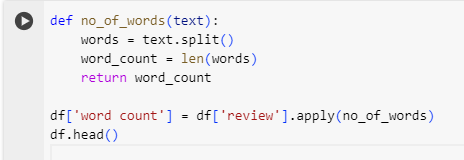
**Sentiment Distribution:** Visualizes the distribution of sentiments using a count plot.

**Data Exploration:**

****

This loop prints out the first five reviews along with their corresponding sentiments, giving a sense of the dataset's contents.

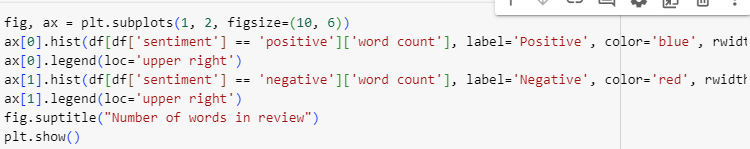
**Feature Engineering:**

****

**Function:** Defines a function to count the number of words in each review.

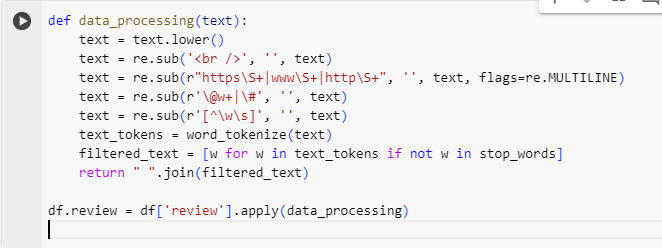
**Apply:** Adds a new column to the Data-Frame for word counts.

**Visualization of Word Count:**

****

This section visualizes the word count distribution for positive and negative reviews separately.

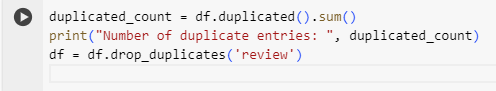
**Text Preprocessing:**

****

**Data Cleaning:** A function that lowercases text, removes HTML tags, URLs, special characters, and stop words.

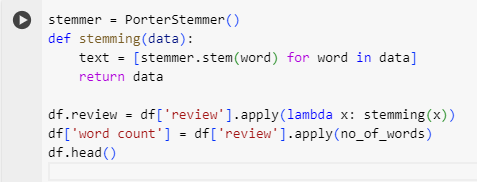
**Apply:** Cleans the review texts.

**Handling Duplicates:**

****

Checks for and removes duplicate reviews

**Stemming:**

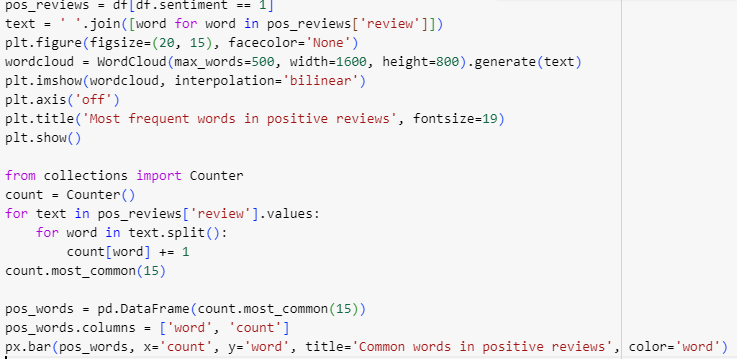
****

**Stemming Function:** Applies the Porter Stemmer to reduce words to their base forms.

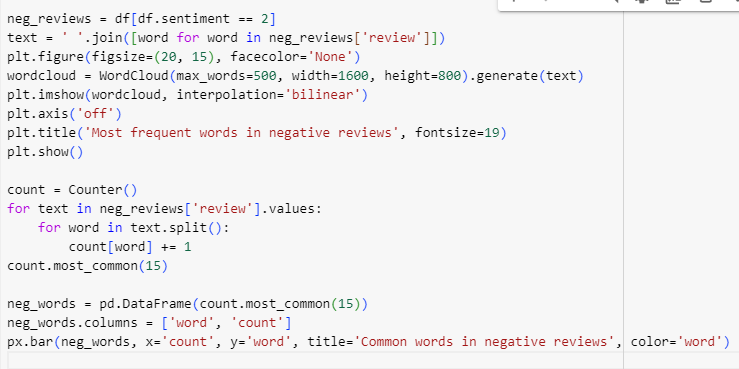
**Apply:** Adds stemming to the review texts.

**Word Cloud and Common Words:**

**Positive Reviews.**

****

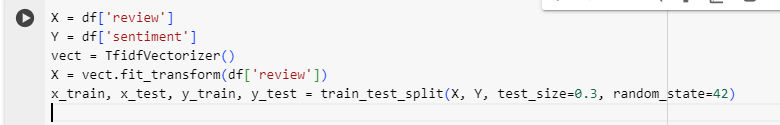
**Negative Reviews:**

****

**Word Cloud:** Generates and displays word clouds for the most frequent words in positive and negative reviews.

**Common Words:** Finds and plots the most common words in positive and negative reviews.

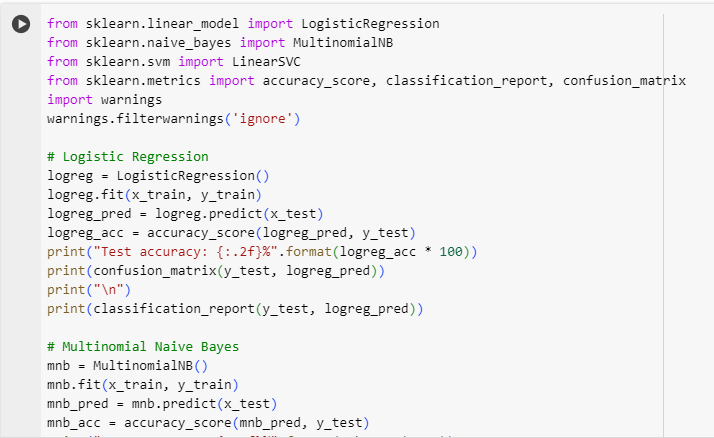
**Model Training and Evaluation:**

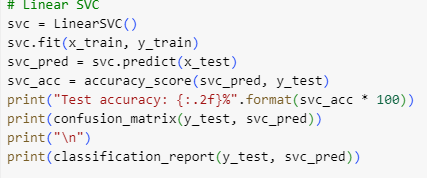
****

**Feature Extraction:** Uses TF-IDF vectorization for the review texts.

**Train-Test Split:** Splits the data into training and testing sets.

**Model Training and Evaluation:**

****

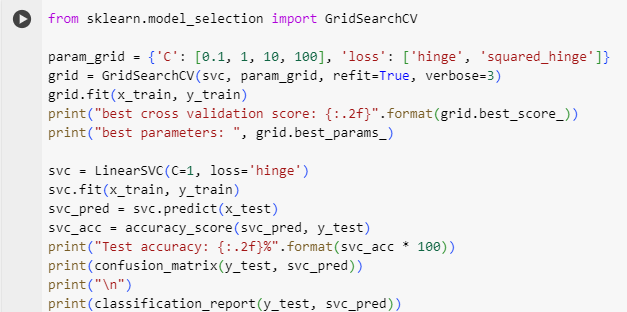
****

**Logistic Regression:** Trains and evaluates a logistic regression model.

**Multinomial Naive Bayes:** Trains and evaluates a multinomial Naive Bayes model.

**Linear SVC:** Trains and evaluates a Linear Support Vector Classifier.

**Hyper-parameter Tuning:**

****

Performs hyperparameter tuning for the SVC model using grid

### N

**THANK YOU**