Sentiment Analysis of IMDB Movie Review Dataset

Project Description: This project focuses on analysing the IMDB Movie Review Dataset to understand the sentiments or emotions expressed in movie reviews and build a classification model that accurately categorizes the sentiment as either positive or negative. Sentiment analysis, a natural language processing task, has applications in various domains such as sentiment classification, customer feedback analysis, social media monitoring, market trend research, and more.

Project Tasks:

Data Analysis: The project begins with an initial analysis of the dataset to gain insights into its content and structure. Key observations include the dataset's size, balance between positive and negative sentiments, absence of missing data, and identification of data noise.

Data Cleaning and Structuring: To prepare the data for analysis, several pre-processing steps are performed, including:

- Converting sentiments and reviews to lowercase for consistency.
- Using regular expressions (Regex) to remove special characters and punctuation.
- Eliminating numbers, HTML tags, and irrelevant characters.
- Removing stop words using NLTK (Natural Language Toolkit).
- Tokenizing the text and lemmatizing words to reduce them to their base forms.
- Eliminating occurrences of the word "br."

Exploratory Data Analysis: A function is used to identify the most common and least common words in the cleaned text data, providing insights into the dataset's vocabulary.

Modelling - Naive Bayes: The project employs a Multinomial Naive Bayes classifier to build the initial sentiment analysis model. The following steps are carried out:

- Data is split into training and testing sets.
- TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is applied to convert text data into numerical features.
- The Multinomial Naive Bayes model is trained and evaluated on the testing data.
- Accuracy, precision, and recall are computed and reported.

Testing with New Data: The trained Naive Bayes model is tested with new movie reviews to predict their sentiments.

Modelling - Support Vector Machine (SVM): A Support Vector Machine model with a linear kernel is implemented and evaluated, providing an alternative approach to sentiment analysis.

Libraries Used:

- Pandas
- String
- Seaborn

- Plotly
- NLTK (Natural Language Toolkit)
- Re (Regular Expressions)
- Sklearn (Scikit-Learn)
- WordNetLemmatizer
- Counter
- Svc (Support Vector Machine)
- Multinomial Nb (Multinomial Naive Bayes)

Results:

The Naive Bayes model achieved an accuracy of 87% on the test dataset, indicating its effectiveness in sentiment classification.

Precision and recall values are also reported in the classification report.

The SVM model's performance is evaluated, and accuracy and a classification report are provided.

Usage:

The project can be used as a template for performing sentiment analysis on text data.

Users can replace the dataset with their own text data for sentiment analysis.

The code can serve as a reference for pre-processing text data and training classification models.

Future Enhancements:

- Hyperparameter tuning to improve model performance.
- Integration of more advanced NLP techniques such as word embeddings (Word2Vec, Glove) or deep learning models (LSTM, BERT) for potentially higher accuracy.
- Deployment of the model for real-time sentiment analysis.

Dataset source: IMDB Movie Review Dataset