Text Sentiment Analysis of IMDB Movie Reviews

1. Exploration and Data Loading:

- Importing the required data analysis and visualization tools is the first step in the project.

- A structured table format is used to load the movie review dataset.

- We look at the features of the dataset, such as the types of data, whether any information is missing, and whether there are any duplicate entries.

- The dataset's distribution of positive and negative sentiments is understood through visualizations.

- In order to spot any trends, we also examine the movie reviews' lengths.

2. Preprocessing Data:

- To guarantee data quality, duplicate entries are eliminated from the dataset.

- To protect the original data, a copy of the dataset is made.

- To be used in machine learning models, sentiment labels are converted into numerical representations.

- The reviews' text is processed to eliminate superfluous words and return words to their most basic forms.

- This processed text is used to create a new representation of the reviews.

3. **Visualization of Word Clouds:**

- The most common words that appear in both positive and negative reviews are highlighted using visual representations known as "word clouds."

- This aids in comprehending the important words connected to every sentiment.

4. **Feature Extraction:**

- We transform the text data into a numerical format suitable for machine learning algorithms.

- This is done by calculating the importance of each word within a review and across the entire dataset.

- A limit is set on the number of words considered to manage the complexity of the analysis.

5. Training and Assessing Models:

- The dataset is split into two sections: one for model training and another for performance evaluation.

❖ The Logistic Regression Model:

- To identify patterns in the training data, a statistical model known as logistic regression is used.

- To make sure the model converges with a solution, steps are taken.

- The test data's review sentiment is predicted by the model.

- A few metrics, including accuracy, precision, and recall, are used to evaluate the model's performance.

❖ **Multinomial Naive Bayes:**

- This statistical model is another one that is used to analyze the data.

- Methods are employed to deal with situations in which specific words may be absent from the training set.

- The test data's sentiment is predicted by the model.

- The same metrics are used to assess this model's performance.

- To determine the best method for sentiment analysis, the outcomes of the two models are contrasted.

6. Conclusion:

- This task shows how Natural Language Processing and Machine Learning can be applied to analyze movie reviews and determine their sentiment.

- The process involves systematic data preparation, feature extraction, and model training.

- The evaluation results highlight the model's ability to accurately classify sentiments.