**Project Document**

**Libraries for Project:**

* Numpy
* Pandas
* Re
* Sk-learn
* Nltk
* Os

**Data Pre-processing:**

* **Text Cleaning** : Remove any irrelevant information such as special characters, numbers, and punctuation. Convert the text to lowercase to ensure uniformity.
* **Tokenization:** Split the text into individual words or tokens. Tokenization is necessary to convert the text into a format that can be processed by machine learning algorithms.
* **Lowercase:** Convert the text to lowercase to ensure uniformity.

**Loading Processed Data as:**

**Function:** Defining function for cleaning, tokenization and to load

**Splitting Data:** Train reviews, Test reviews, Train labels, Test labels

**Data Directory:** Specifying Data directories and their sub-directories

**BoW Feature Extraction:**

**Training Phase:**

* I use the BoW representation of the training data to train your sentiment analysis model.
* The model learns the patterns and relationships between the BoW features and the corresponding sentiment labels.

**Testing Phase:**

* I use the BoW representation of the test data as input to your trained model.
* The model applies the learned patterns and relationships to predict the sentiment labels for the test data.

**Model Selection:**

The choice of the best model can vary depending on the dataset, the specific problem you are tackling.

**Logistic Regression:** Despite its simple nature, logistic regression can perform surprisingly well on text classification tasks, especially when combined with effective feature extraction techniques like Bag-of-Words or TF-IDF.

**Random Forest:** Random forests are ensemble methods that combine multiple decision trees to make predictions. They can handle non-linear relationships in the data and often provide good generalization.

**Ensemble Methods**: Combining predictions from multiple models can often lead to improved performance. Methods like stacking or blending can be used to combine the outputs of different models.

**Model Training:**

Train my machine learning model using the preprocessed and vectorized data. Fit the model to the training data and use the validation data for hyper-parameter tuning if needed.

**Model Evaluation:**

Evaluates my model's performance on the test data using metrics like accuracy, precision, recall, F1-score, etc.