NLP Assignment

After downloading the Amazon Reviews Dataset, the data was to be uploaded, explored, preprocessed, and used to train a Machine Learning model to act as a review classifier.

A major part of the preprocessing was NLP-based. Text cleaning: for the text to be ready to use as training and testing sets with the desired ML models.

Several Machine Learning models were used on the provided dataset to classify text into different categories depending on the text context.

The models used:

- **Multinomial Naïve Bayes (NB)**
- **Bagging Classifier**
- **Support Vector Machine (SVM)**
- **Logistic Regression**

After data exploration and preprocessing, the models were trained, validated, and tested using the prepared dataset.

After testing many models, the Logistic Regression model had the best training to test accuracies.

The model's accuracy can be improved using a series of methods, for example:

- Increasing the number of rows used to train the model, which could not be achieved due to laptop limitations
- Using the 'ngrams' method when using 'CountVectorizer' for better feature engineering
- Using a Neural Network, Deep Learning, model for a far better model training accuracy

The code can be accessed by opening the attached ".py" files and is explained in the attached "readme" file.