Installed and imported required packages

Python version used – 3.7.13

1. Dataset Preparation

Read the data after downloading and extracting the rar file.

Read only the star_rating and review_body columns

Shuffled the dataframe

Removed rows containing null values

Changed the datatype of star_rating column to integer and review_body column to string

Randomly selected 20000 reviews randomly from each rating class

Shuffled the dataframe again

2. Data Cleaning

Converted to lowercase

Removed html tags using beautifulSoup

Removed urls using regex

Performed contractions using the contractions library fix function

Removed non-alphabetic characters using regex

Removed extra white spaces using regex

Removed rows containing null values (if any)

3. Preprocessing

Removed English stopwords using nltk package

Removed rows containing null values (if any)

Performed lemmatization on the remaining words

4. Feature Extraction

Used TfidfVectorizer to extract features from the cleaned and preprocessed data

Split the data into train:test as 80:20

Imported metrics for evaluation

5. Perceptron

Used the linear perceptron model from sklearn to train and test

Printed the precision, recall, f1score and their average

6. SVM

Used the linear SCV model from sklearn to train and test

Printed the precision, recall, f1score and their average

7. Logistic Regression

Used the linear Logistic Regression model from sklearn to train and test

Printed the precision, recall, f1score and their average

8. Multinomial Naive Bayes

Used the linear Multinomial Naive Bayes model from sklearn to train and test

Printed the precision, recall, f1score and their average