**NLP Project Report**

**Problem Statement :**

A restaurant wants to create a system that can sort customer reviews into two categories: positive or negative. Positive reviews show appreciation, while negative ones include criticism. The plan is to set up a customer support team to call back customers who leave negative reviews. The team will try to solve any issues and may offer discounts to encourage customers to come back. The restaurant has a dataset of reviews with labels and some new reviews that need labeling.

**Dataset:**

The dataset Name is **Barcelona\_Reviews** , this Dataset contains more than 400k Restaurant / Food Reviews on Restaurants in Barcelona.

The dataset total is 6,000 dividing it into 3,000 positive and 3,000 Negative .

Each 3,000 is distributed on the Training , Testing , and Validation. As the Training is 70% of the dataset and the Testing is 20% of the dataset and the Validation is 10% of the dataset.  
from each Positive and Negative labels :

* Training 70% = 2100
* Testing 20% = 600
* Validation 10% = 300

**Data Preprocessing :**

1. Replaced non-word characters with white spaces in each sentence
2. Word tokenization : split each sentence on white spaces to words
3. Word normalization : replaced each capital letters with small letters (no CAPS)
4. Removed Stopwords from all sentences (For Removing unimportant words )

**Feature Extraction :**

Bag of words (BOW) Algorithm is used to extract features from Sentences.

* Creating Global Vocab For all Unique Words in Train Data.
* Each sentence Has its own Vocab (Filled Global Vocab with Frequencies of words).

**Classifier :**

Naive Bayes Classifier is used (Multinomial NB() ) that uses Bayes theorem.  
Trained on each sentence Vocab from training data and its label.  
Train data is shuffled before Training to avoid overfitting .

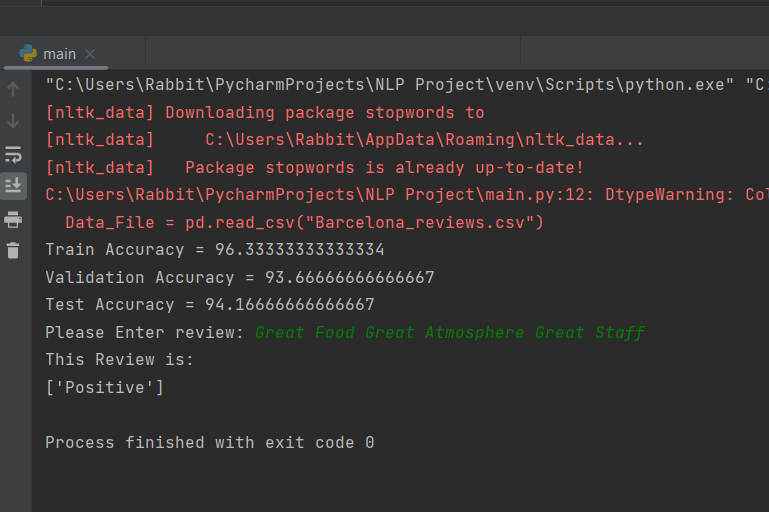
**Testing and Validation :**

Different number of sentences used in dataset and the best accuracy is reached when we used 6000 as total 70% for Training and 10% for validation and 20% for Testing each of them is split 50% positive and 50% Negative.  
in train data the data is split in equal halves (Positive , Negative) to avoid overfitting (Not to be biased to a specific label )

**Evaluation:**

1. Evaluation of the model on Train data = 96.3%
2. Evaluation of the model on Validation data = 93.7%
3. Evaluation of the model on Test data = 94.2%

The accuracy of the Training , Validation ,and Testing.  
Review checker that checks if the review is Positive or Negative.

****