Riverus Technology Solutions Pvt. Ltd.

Assignment

(2nd round of assessment process)

Submitted by:

Divyam Jain

(IIT Kharagpur)

1. **Objective:** In this report we are going to predict the sentiments of the movie reviews with the help of NLP tools and XGboost classifier.
2. **Data explanation :**
    a. Tools Used: Spyder , Notepad++
    b. Libraries Used: pandas, regex, nltk: porterstemmer, stopwords, scikit
    c. The data was pretty well balanced but it involved the anomaly of data duplication which had to be handled,
3. **Handling the Data:**
    a. From reading text files into different DataFrames such as data1 for 2.txt with reading as tab separated and no header.
    b. Also we need to consequently deal with the issue of duplicate reviews in our file. For that we drop the duplicate files while keeping the initial review.
    c. And finally List is formed while keeping the columns as Sentiment & Review. Similarly for 1.txt dataframe is formed of data2 and removal of duplicates is done.
    d. Here in data2 columns are opposite i.e review and sentiment. For further merging of data that needs to be corrected. Further re indexing is done for swapping the dataframe columns.
    e. Now finally we move onto the test data whilst following sort of similar process: reading tab separated test text file, converting pandas series into DataFrame and naming the column as review
    f. Finally we move onto the merging of dataframes. Naming the new dataframe to be data_fin with the naming of sentiment column as train_Y and removal of it.
    g. After this, we want to join both the train and test text file with resetting the index by keeping the reset.index to be True

h. Initiating the list corpus.
   In the for loop: Here moving on step by step: From i^{th} record in review we substitute all characters except all the alphabets (uppercase or lowercase), in place of the spaces in between with simultaneously lower casing the string and splitting it.
i. Further we remove the spaces present as the list element in the next step.
j. Removing of stopwords from the text and stemming of those words is done next step
k. And finally, the remaining words are again joined and appended in the corpus.

4. **Vectorisation** :
   a. In order for the data to make sense to our Machine learning models we need to convert them into numeric values through the process of vectorisation.
   b. We use tfidf Vectoriser with bigram and tuning our hyperparameters of the tfidf keeping the model hyperparameters to be constant. I set the max_features to be 500 (while i noticed to be be appropriate considering our dataset)
   c. The fit_transform function from the vectoriser modified our corpus (bag of words) dataset accordingly
   d. And, finally we sliced the “whole” dataframe again in train_new & test_new for modelling. Adding test data to the train data was important so that both training and test while modelling could be in same space.

5. Finally we run XgBoost on the processed data, while with hit and trial method on the different hyperparameters (Hyperparameter Tuning is Suggested)
6. I approached manually by changing no. of trees and seen the accuracy(from confusion matrix), while low no. of trees created bias

toward the one’s prediction, I started increasing no. of trees finally reach the conclusion for trees = 400

7. And finally we used 10-fold cross validation and confusion matrix , both of which gave the same result of around 85%.