EMOTIONS CLASSIFICATION IN HINDI TEXT

Steps followed to built the model:

1. Data has been read into lists and have been finally converted into a Pandas DataFrame.
2. Input to the model (X) and output (y) has been separated.
3. Using CountVectorizer text has been converted into numerics as model needs numbers not String or tokens. Here we are giving it some **stop words** so that they can be removed as they do not contribute to the model building. Also a tokenizer defined manually has been passed as the built-in tokenizer will split everything into characters which we don't want so the **manually built tokenizer (my_tokenizer)** splits about white spaces only. Also we are using **n_grams** to preserve some local meanings.
4. A **logistic regression model** has been trained as Logistic Regression works well for the **Sparse Matrices**.
5. Train Accuracy and Test Accuracy has been evaluated.
6. After that I have trained the model on the **whole data Set** and have shown **cross validation score** in each case as well as the overall cross validation score.

Obtained Results:

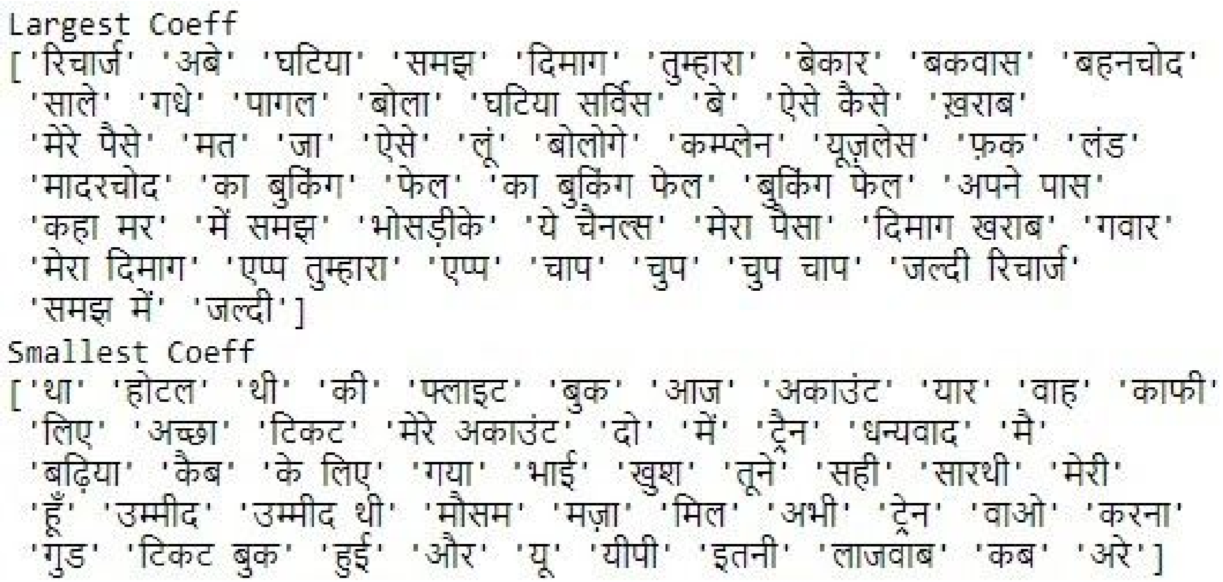
Train Accuracy: 0.76

Test Accuracy: 0.65

Cross Validation Accuracy: 0.61

Analysis of the results:

1. Clearly the model is a bit **overfitting** as there is a small difference between Train and Test Accuracy.
2. Also since we have **very little data** (considering the Hindi aspect where we have too many characters and other things) it is difficult to increase the accuracy. However the trained model has been able to catch some important words for the purpose of classification which can be seen in this diagram taken from the Jupyter Notebook.



```
Largest Coeff
['रिचार्ज' 'अबे' 'घटिया' 'समझ' 'दिमाग' 'तुम्हारा' 'बेकार' 'बकवास' 'बहनचोद'
 'साले' 'गधे' 'पागल' 'बोला' 'घटिया सर्विस' 'बे' 'ऐसे कैसे' 'ख़राब'
 'मेरे पैसे' 'मत' 'जा' 'ऐसे' 'लूं' 'बोलोगे' 'कम्प्लेन' 'यूज़लेस' 'फ़क' 'लंड'
 'मादरचोद' 'का बुकिंग' 'फेल' 'का बुकिंग फेल' 'बुकिंग फेल' 'अपने पास'
 'कहा मर' 'में समझ' 'भोसड़ीके' 'ये चैनल्स' 'मेरा पैसा' 'दिमाग खराब' 'गवार'
 'मेरा दिमाग' 'एप्प तुम्हारा' 'एप्प' 'चाप' 'चुप' 'चुप चाप' 'जल्दी रिचार्ज'
 'समझ में' 'जल्दी']
Smallest Coeff
['था' 'होटल' 'थी' 'की' 'फ्लाइट' 'बुक' 'आज' 'अकाउंट' 'यार' 'वाह' 'काफी'
 'लिए' 'अच्छा' 'टिकट' 'मेरे अकाउंट' 'दो' 'में' 'ट्रैन' 'धन्यवाद' 'मै'
 'बढ़िया' 'कैब' 'के लिए' 'गया' 'भाई' 'खुश' 'तूने' 'सही' 'सारथी' 'मेरी'
 'हूँ' 'उम्मीद' 'उम्मीद थी' 'मौसम' 'मज़ा' 'मिल' 'अभी' 'ट्रेन' 'वाओ' 'करना'
 'गुड' 'टिकट बुक' 'हुई' 'और' 'यू' 'यीपी' 'इतनी' 'लाजवाब' 'कब' 'अरे']
```

3. Stop words are only available for a few languages. For **Hindi we don't have inbuilt stop words**. So we need to provide stop words on our own.