Assignment 3 Report

Damanpreet Singh – 000741359

# **Data Description:**

There are two columns, Text and Emotions. Quite self-explanatory right. The Emotions column has various categories ranging from happiness to sadness to love and fear. Have fun building models which can identify what words denote what emotion.

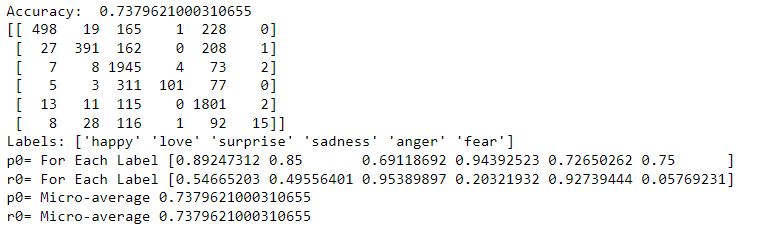
**Description of the Text Representations:**

**Test and training data was split into 30/70**

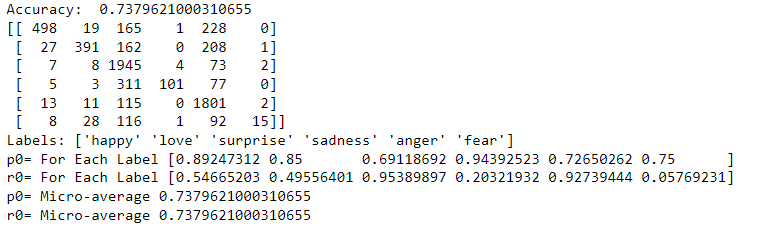
1. Using ComplementNB and no changes were made to Count Vector
2. Using MultinomialNB and default countVector
3. Using MultinomialNB and countVector where n-bag of words = 2
4. Using MultinomialNB and where n-bag of words = 2 and using stop words
5. Using ComplementNB and default count vectorcountVector with stop\_words.
6. Using MultinomialNB and countVector where n-bag of words = 3
7. Using MultinomialNB and with TfidVectroizer that uses stop words and removes words with high and low frequency to avoid skewness.
8. Using ComplementNB and with TfidVectroizer that uses stop words and removes words with high and low frequency to avoid skewness.

**Results:**

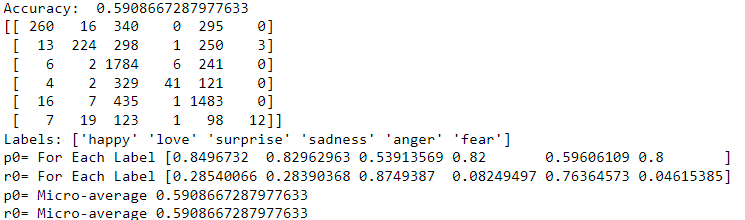
**1.** Using ComplementNB and no changes were made to Count Vector



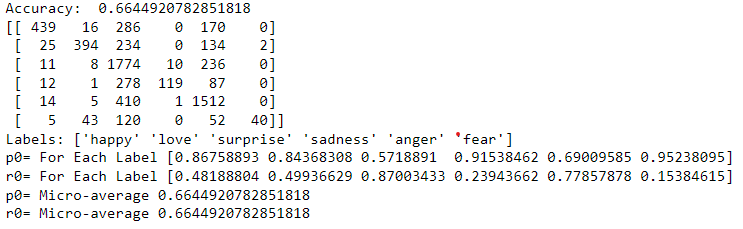
**2.** Using MultinomialNB and default countVector



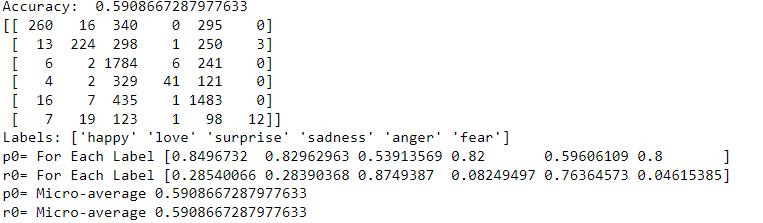
**3.** Using MultinomialNB and countVector where n-bag of words = 2



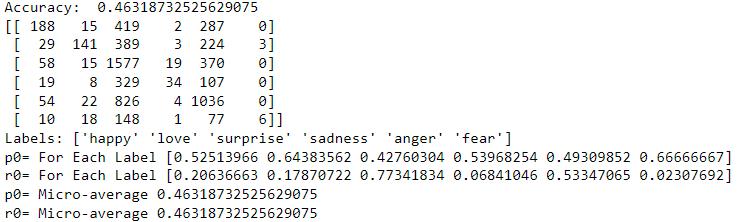
**4.** Using MultinomialNB and where n-bag of words = 2 and using stop words.



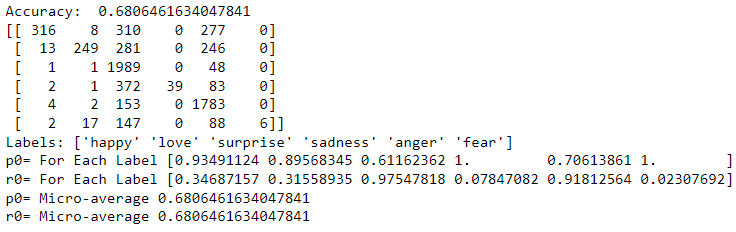
**5.** Using ComplementNB and default count vectorcountVector with stop\_words.



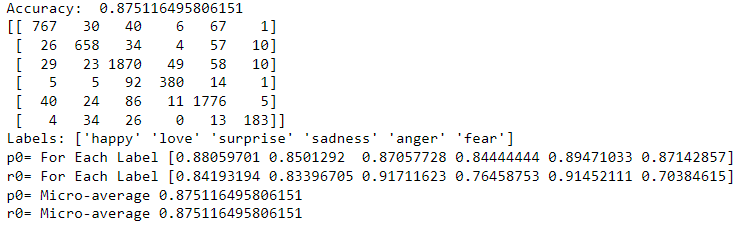
**6.** Using MultinomialNB and countVector where n-bag of words = 3



**7.** Using MultinomialNB and with TfidVectroizer that uses stop words and removes words with high and low frequency to avoid skewness.



**8** Using ComplementNB and with TfidVectroizer that uses stop words and removes words with high and low frequency to avoid skewness.



**Discussion:**

Using ComplementNB has yielded great accuracy and precision for this specific data set. Multinomial naïve bayes tests gave and average test around 60 % and once test accuracy dipped below 50 when n-bad if words were increased to set of Three. Removing Stop words, the words with most frequency and lowest frequency gave a great accuracy of over 87% as the data became more normalized, I tried some more test, but Test 5 and 8 accuracy was the highest I could achieve. Which is a good fitting and can be relied upon.

**Future Work:**

In this report you can see the screenshots that also show precision and recall score for each label that can help me find what other changes I can make to get better results. If you look and the results of test five each label has an incredibly good precision score and did not require too much data cleaning. I would like to do more research on why other tests had more skewed results for some labels.