Ans2.) (Naïve Bayes)

The Naïve Bayes algorithm has been implemented with 5-fold cross validation. Basic preprocessing is done by using regular expression library (removing blank spaces, symbols, numbers etc.). The data set has been split into train set and test set for every fold and then accuracy and F-score is calculated for every fold. A bag of words is created comprising of all words from the train set. A bag of positive words and negative words is also created corresponding the words having positive and negative sentiments respectively.

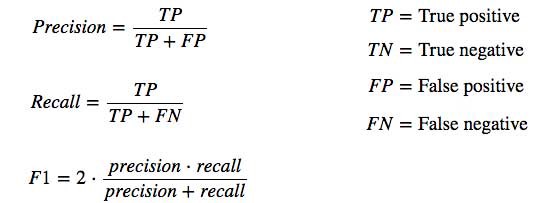
The test set is then used to calculate Naïve Bayes probability of each word of test sample using the formula –

where classes corresponding to positive and negative sentiments.

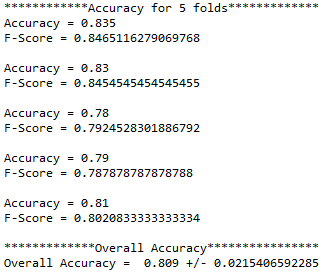
The above formula has been used to avoid getting probability as 0 for both the classes.

The probability of each word of the sentence is multiplied together for both the classes. The class with greater probability is chosen as the sentiment for the given test sample.

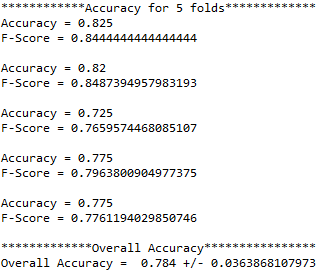
The accuracy and F-score are manually calculated by manually count true positives, false positives etc. and are manually computed by using the formulae given –



The results without removing stop words are -



A model was also made to remove stop words. The results obtained from the same are –



From the above, we observe that accuracy has decreased after removing stop words. However, significant changes have not been observed in the F-score for 3 out of 5 folds.