

Name : Ajay Tanala

Intern Task Report

- 1) Clearly, To classify this kind of dataset we must use the Naive Bayes Classifier which is based on class conditional probability. I solved this problem from scratch which means I didn't use the python inbuilt functions of NB classifier.
- 2) Accuracy of my model is 65.2%.
- 3) Limitations of NB classifier is it assumes all features are independent which is very rare in real world use cases. This algorithm faces the zero frequency problem where it assigns zero probability to a categorical variable whose category in the test data is not available. Smoothing techniques need to be done.

Given data has empty cells so need to remove those data points from datasets. Accuracy of my model is comparatively low because I divided my dataset randomly. There may be a chance of fewer data points of a particular class in training data. I had to divide train and test data separately for a particular class and after that we need to combine every train data of each class and every test data of each class. Then we can use NB classifier to get more accuracy.

**** Stop words and stemmer techniques need to be done on data before training.**