Sentiment Analysis or Opinion mining

CSCI 561 – Foundations of Artificial Intelligence

Summer 2018

Laxmipooja Anegundi
USC_ID:2896786817
Email_id:anegundi@usc.edu

Sentiment Analysis And Data preprocessing

Introduction to our data set and Problem Statement:

- Each review is labelled as positive and negative in the 400k amazon review dataset.
- Using a Machine learning technique to train the model and predict any new given review to either positive or negative review.

Data Preprocessing:

- Cross Verified the Dataset to find whether there exist any additional labels apart from positive and negative for example "Positive,piositive,Negative,neagtive" using **Google Refine(Data Cleaning Tool)**
- Using the CountVectorizer for the tokenizing, removal of English stop words, convert to lowercase and punctuation from the dataset and then create a vocabulary of unique tokens(words)
- The given labeled dataset is balanced.

Machine Learning Algorithms

Decision Tree Algorithms

- **Random Forest** is an ensemble algorithm which uses multiple decision trees for making the right classification and prevents overfitting.
- Random Forest is also considered as a very handy and easy to use algorithm because of its simplicity.
- Accuracy : 79

Neural Networks MLP

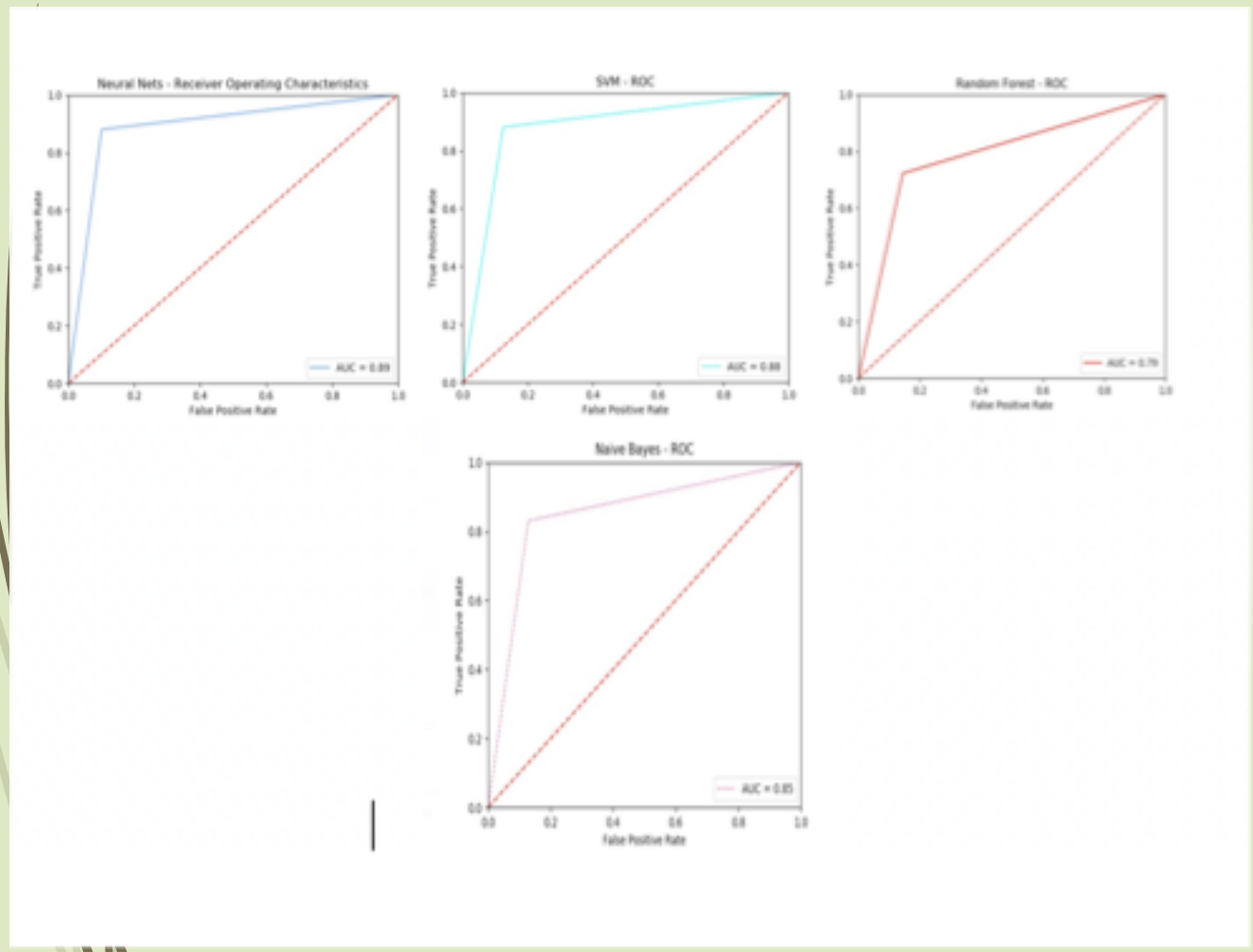
- **MLP** (Multi Layered Perceptron)is used to solve complex tasks.
- Can use Back Propagation, to adjust weights, which is used to increase the accuracy of the model
- Accuracy: 89

SVM(Third Algorithm chosen)

- **SVM** creates hyperplane that have the largest margin in Dimensional space
- Avoids overfitting
- It is best for the dataset given 2-Classes works best if the number Of positive training is same as the number of the negative training set

Accuracy: 88

Measures of Classifier



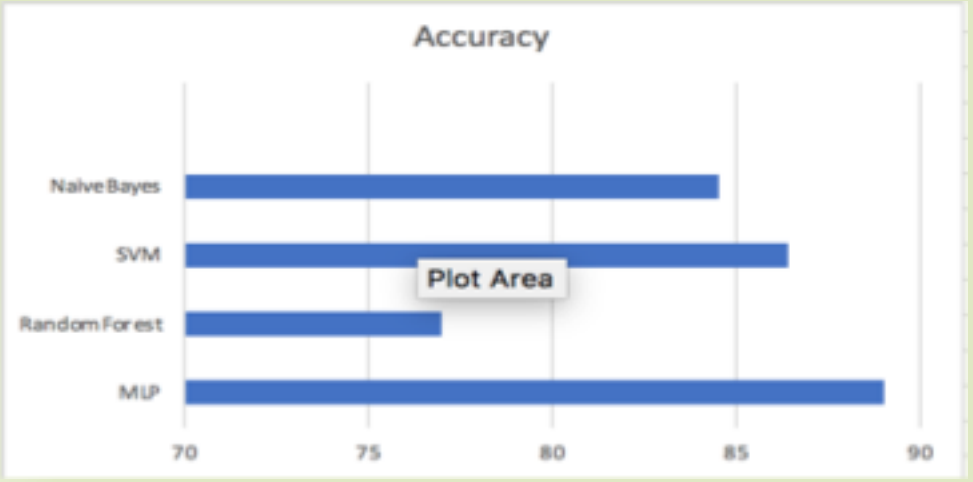


Chart Title