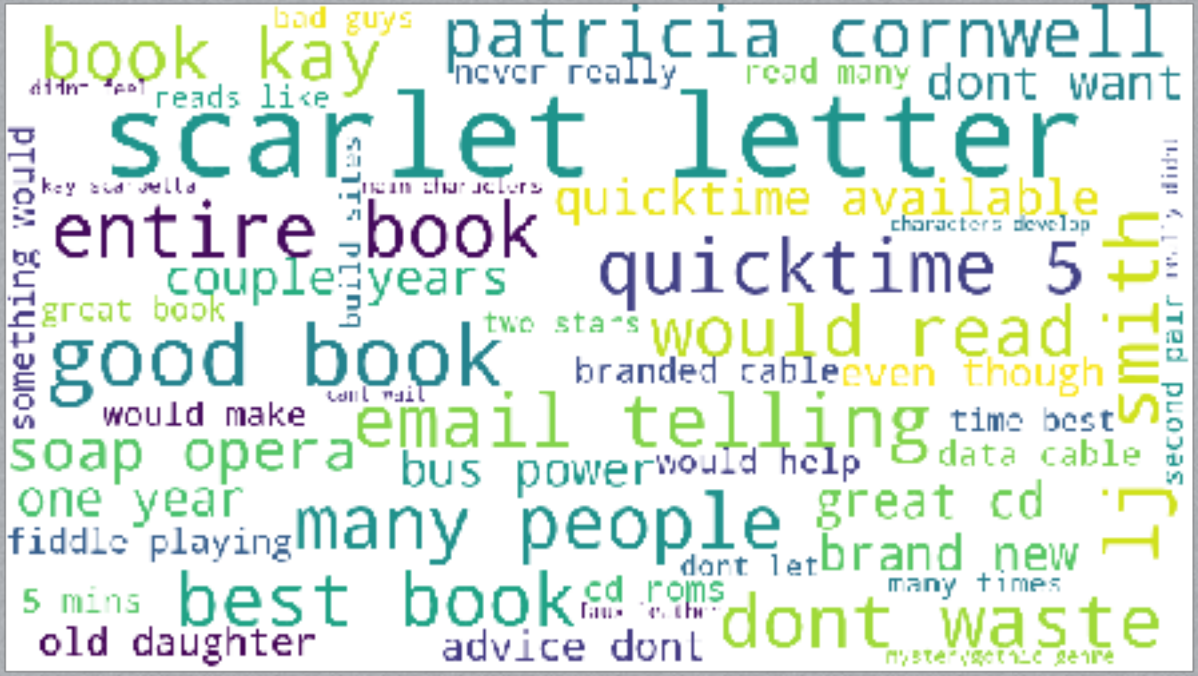
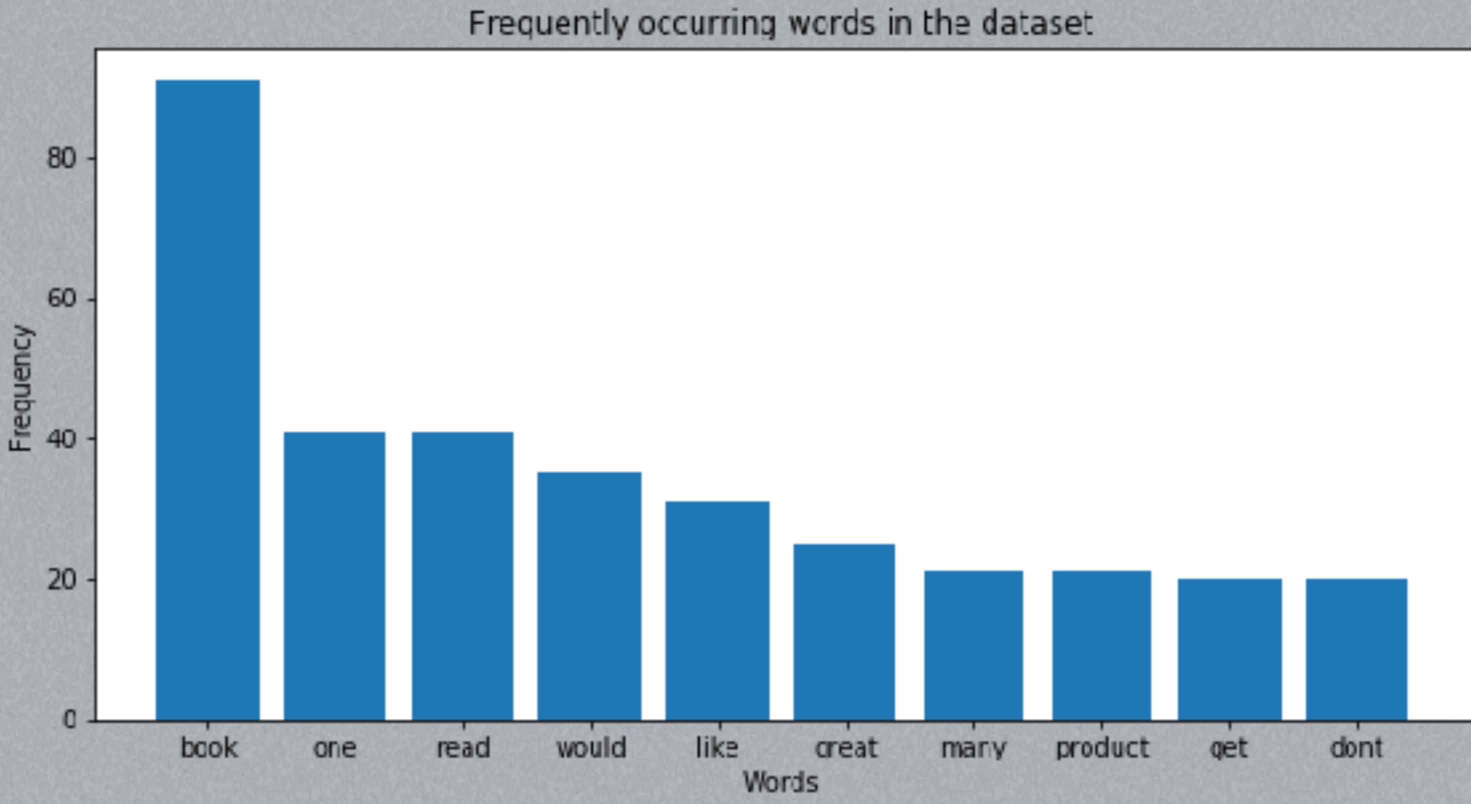
Sentiment Analysis on Amazon Reviews

Most frequent Bigrams





Test Data : Every fifth review from the given data set. (79987 reviews)
Training Data : All other rows excluding Test data. (319952 reviews)

Data Preprocessing:
Removal of all characters which are not words or spaces.
Conversion to lowercase.
Removal of stop words.

Feature Extraction : Tfidf vectorizer to tokenise sentences and convert words to vectors

Algorithms

Neural Network:

- MLP Classifier with one hidden layer of 50 nodes. Solver used: adam, Activation: relu, early stopping=True
- Works well on very large data sets.
- Learns many features and stores a lot of information.
- Gives a lot of flexibility.

Decision Tree:

- DecisionTreeClassifier with min_sample_split : 2.
- Simple and fast algorithm.
- Does not do well as it is difficult to choose key tokens to perform the split.
- It is a greedy algorithm so it searches only some of the possibilities.

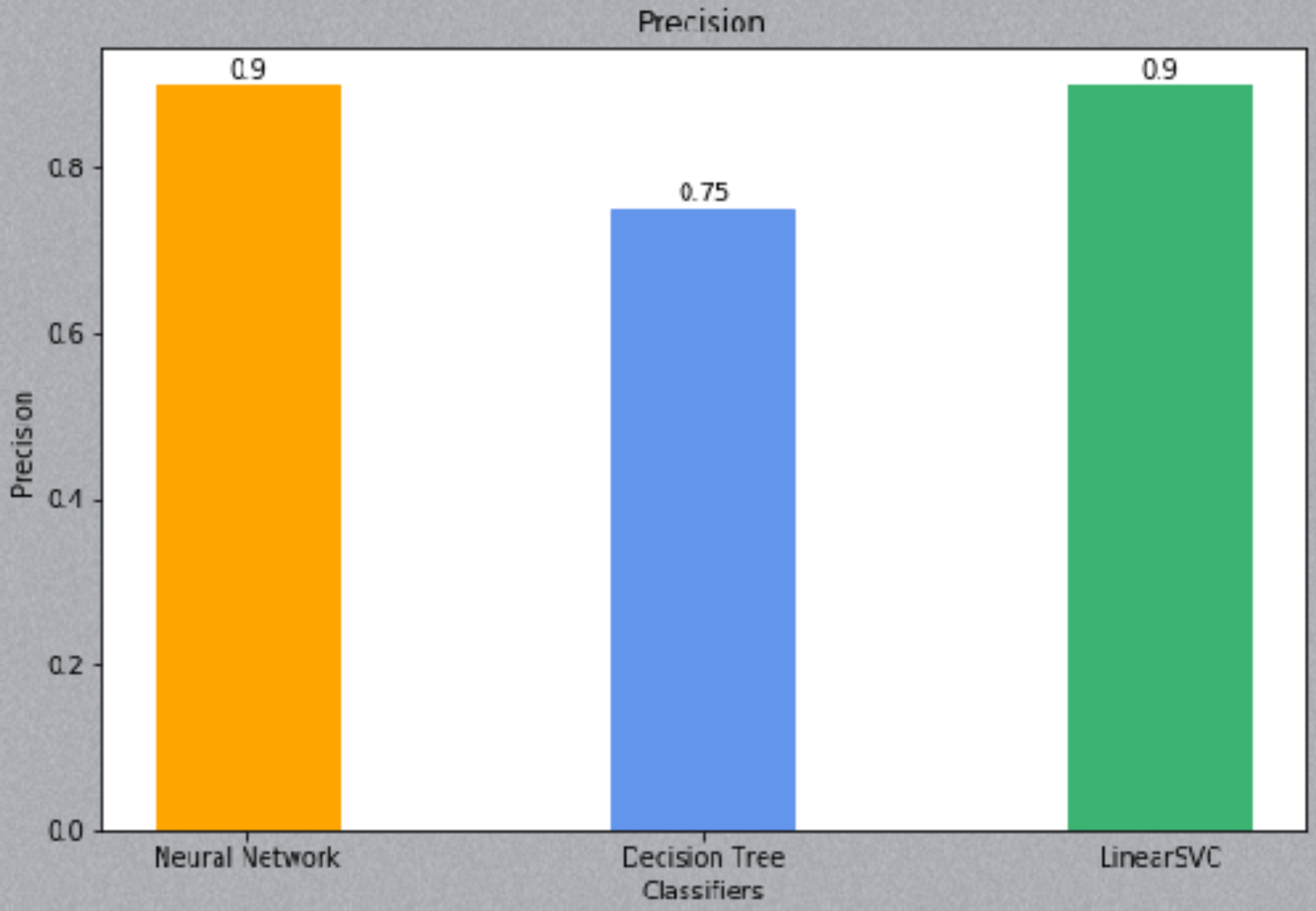
Reasons for choosing Linear SVC :

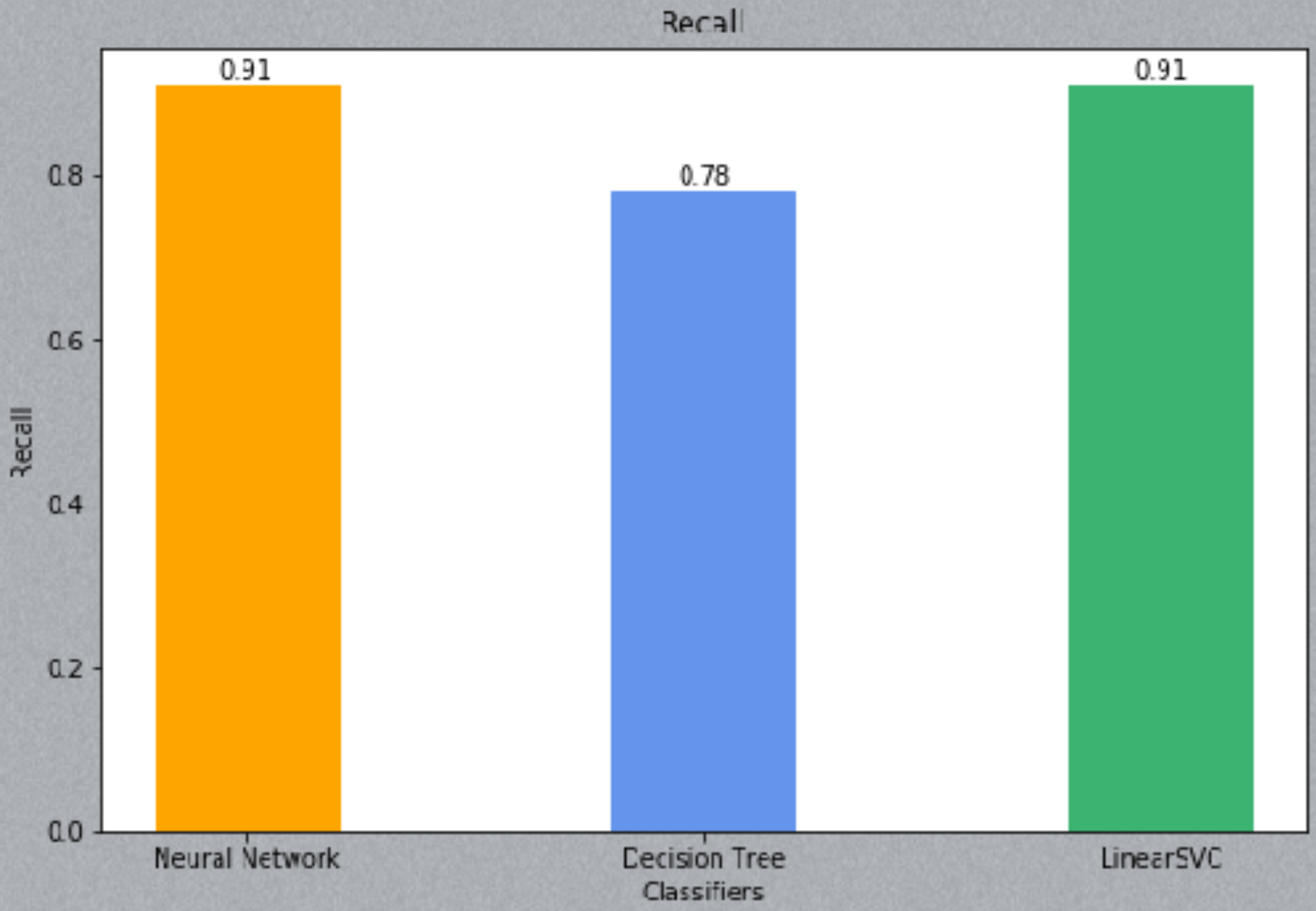
- Returns “best-fit” hyperplane that classifies data. C: 0.15 , tolerance:1e-6 ROC metric was used to evaluate classifier output quality. AUC was maximised with Linear SVC.
- Simpler and faster algorithm as compared to Neural Network and Decision Tree.
- Less prone to overfitting.
- Handles sparse matrices well.
- Less memory intensive than other algorithms.

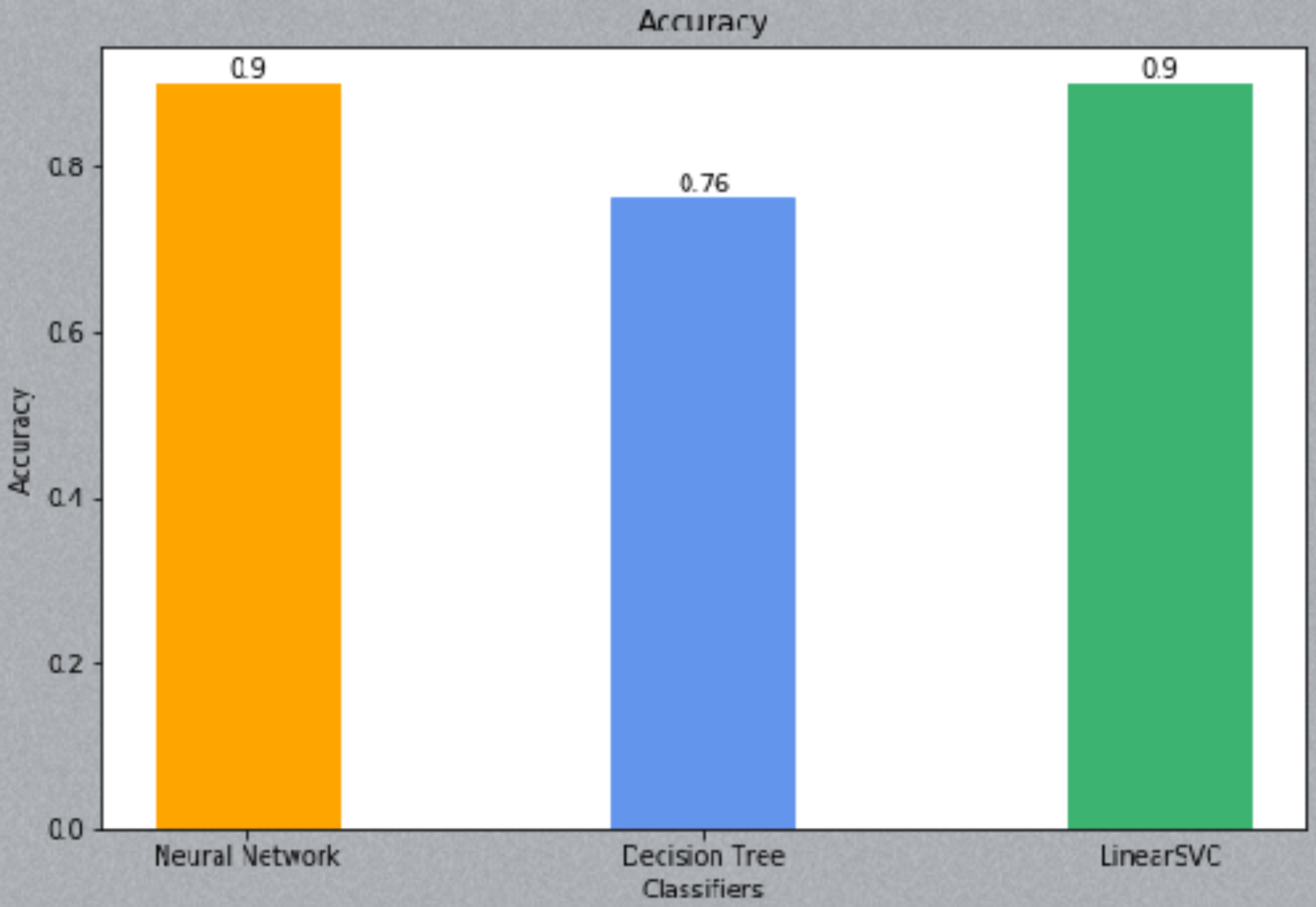
Conclusions:

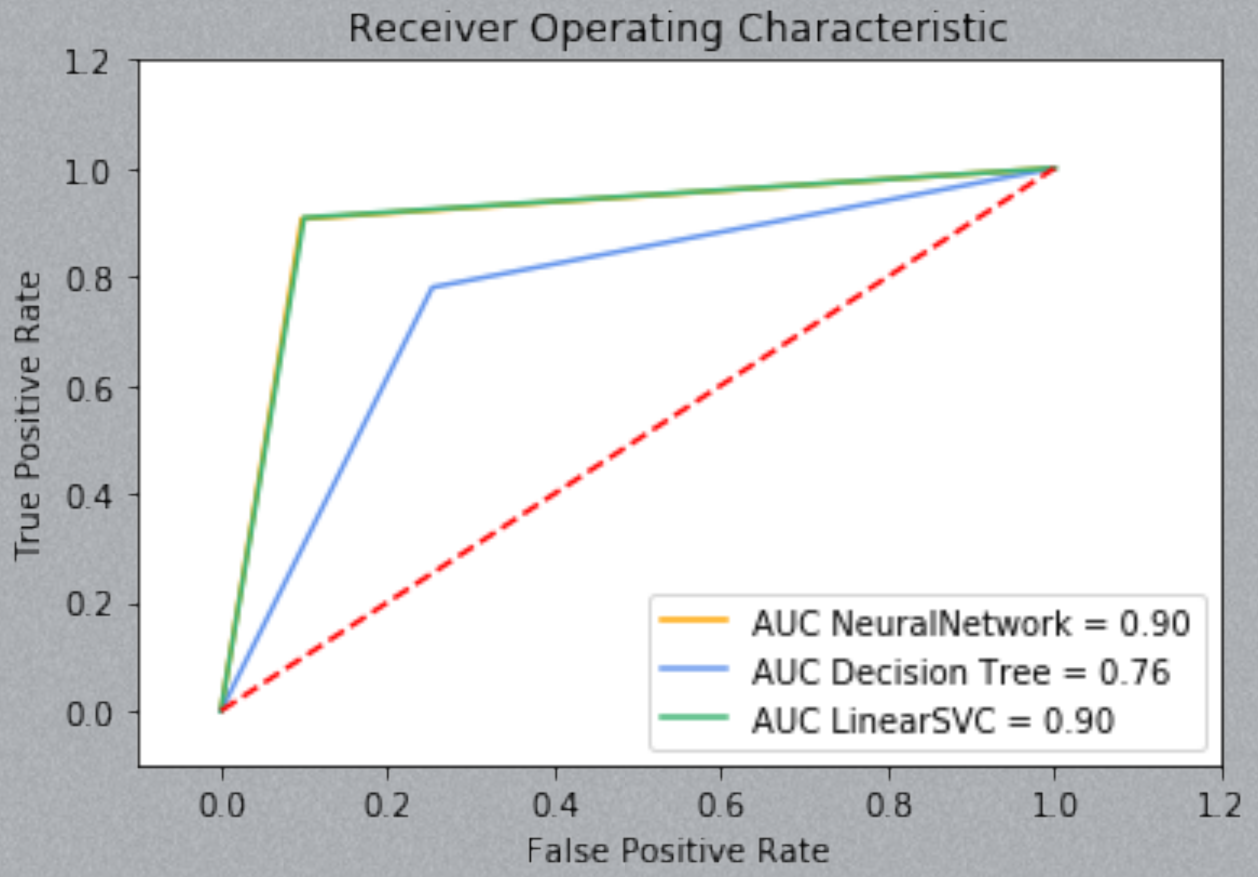
- The most sophisticated algorithms are not necessarily the best.
- For text datasets, preprocessing data helps in improving accuracy.
- ROC curves are reliable indicators of performance.

Visualisations of Results


Precision
Precision
Neural Network
Decision Tree Classifiers
LinearSVC


Recall
Recall
Neural Network
Decision Tree Classifiers
LinearSVC


Accuracy
Accuracy
Neural Network
Decision Tree Classifiers
LinearSVC


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
AUC NeuralNetwork = 0.90
AUC Decision Tree = 0.76
AUC LinearSVC = 0.90