SENTIMENT ANALYSIS

AMAZON REVIEW

Data Set and Problem

Problem:
To perform sentiment analysis on Amazon reviews to determine whether a review is positive or negative.

Data Set :
•Amazon Reviews Kaggle competition. 400000 rows, each row in the format {sentiment(text), review(text)}
•Test set is every 5th sample from the dataset. Remaining goes to Train set.

Input Analysis

Stage 1 - Preprocessing the Text:
Techniques used to preprocess the data for Sentiment Analysis.

- ✓ Converted to Lower Case
- ✓ Remove punctuations
- ✓ Negation Rule: Replace all negative words like don't with not
- ✓ Deleted the stop words[excluding the negative words]
- ✓ Taken adjectives, Opinion words : Conjunction Rule

Input Analysis

Stage 2 – Feature Extraction:
Different Feature Extraction methods(Unigram, Bigram and N- gram methods)

- Standardized vector
- TF-IDF(It is defined by multiplying value of frequency of word in review (TF) and frequency of word in whole corpus (IDF). term frequency-inverse document frequency vector
- Dimension Reduction Vector

Chosen **TF-IDF** as it is better over other feature extraction methods for Sentiment Analysis.

Input Analysis

Stage 3 – Classification:

- Classify method is used to train by giving the training set, method and feature extract function
- Classifier function is applied on test data to get output.
- Methods used for Classification:
  - Random Forest
  - Neural Networks
  - Support Vector Machines

Random Forest

Decision trees are a popular method for various machine learning tasks.
The training algorithm for random forests applies the general technique of bootstrap aggregating,
or bagging, to tree learners.
Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:
For b = 1, ..., B:
Sample, with replacement, B training examples from X, Y; call these Xb, Yb.
Train a decision or regression tree fb on Xb, Yb.

Neural Networks

Depending on the nature of the application and the strength of the internal data patterns you can generally expect a network to train quite well. This applies to problems where the relationships may be quite dynamic or non-linear. ANNs provide an analytical alternative to conventional techniques which are often limited by strict assumptions of normality, linearity, variable independence etc. Because an ANN can capture many kinds of relationships it allows the user to quickly and relatively easily model phenomena which otherwise may have been very difficult or impossible to explain otherwise.

Support Vector Machine

Support vector machines are supervised learning algorithms used for classification and regression analysis

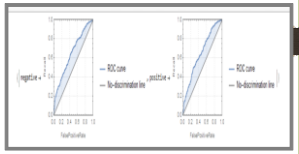
Support vector machine examines the data, identify hyper plane that classify data in to two classes with maximum margin. SVM also supports classification and regression in statistical learning.

Results

- **Accuracy**
- **Confusion Matrix**, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm,
- **Precision** also called positive predictive value, is the fraction of relevant instances among the retrieved instances
- **Recall**, defined as the number of true positives over the number of true positives plus the number of false negatives.
- **ROC curve**, tool for diagnostic test evaluation in which the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter.

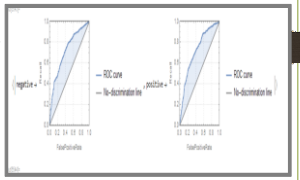
Random Forest[Decision Trees]:

- Accuracy :0.586587
- Precision: 0.603563,0.572727
- Recall: 0.535573,0.638945



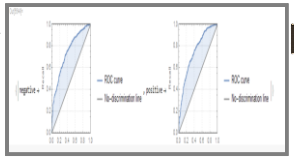
Neural Networks

- Accuracy :0.658659
- Precision: 0.666667, 0.650794
- Recall: 0.652174, 0.665314



Support Vector Machines

- Accuracy : 0.674675
- Precision: 0.696312, 0.656134
- Recall: 0.634387, 0.716024



Conclusion

Based on the Results obtained on execution of Support Vector Machines(SVM), Neural Networks(NN) and Radom Forest. **Support Vector Machines** yield better accuracy compared to others in sentiment analysis on larger inputs.