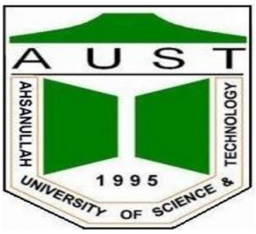

Ahsanullah University of Science and Technology

CSE 4108

Artificial Intelligence Lab

A Report on Text Classification Project

Group Id : 08

Lab Group No : B-2

Project Group Members

Name : Md Amir Faisal

Id : 14.02.04.085

Name : Tazneen Sultana

Id : 14.02.04.091

Name : A. M. N. Billah Khan

Id : 14.02.04.092

Problem Description

In Google play store there are lots of applications. Each application has some text reviews with the rating given by users. Our problem is to classify the review rating from review text. We will predict the rating from review using different models.

Dataset Description

Dataset has been created by scraping google play store website. There are 160 instances and 4 attributes. Attributes are date, rating, review text, review title. The date indicates on which day the review has been given. The rating indicates the stars of the text review, its value can be from 1 to 5. Review text indicates the actual text review given by users. Review title is the title of the text review. We have to predict the rating of a review.

Model Description

K-Nearest Neighbor :

The model will calculate the distance from the test data point which should be predicted to all other train data points. This distance can be Manhattan distance, Euclidean distance, etc. After calculating these distances the model will check the classes of nearest k number of neighbors. That means all the distances will be sorted in ascending order and model will pick the first k number of neighbors. The majority classes of that k number of nearest neighbors will be the class of the test data point.

Logistic Regression :

First, the model will be a regressor. Then this model will pass to a sigmoid function. After passing to sigmoid function the model will be converted into a classifier. Sigmoid function bounds the classification result between 0 and 1. Equation of sigmoid function is g(x) = 1/1+eˆ -x.

Decision tree classifier :

In this model, the dataset will be split according to some attribute values. Each portion will be split continuously until there is a pure subset according to an attribute. This will make a tree. In this tree, all the leaf nodes are a pure subset, where a model can take decisions for prediction. This classification is called decision tree classification.

Random forest classifier :

In this model, some data will be chosen randomly and form a subset of the main dataset. Thus many subsets will be created. These subsets can be overlapping also. From each of these subsets, we can create decision trees. Each decision tree gives us a prediction about test data. From many decision trees, the majority of the predicted class will be the final class of test data. This is the working principle of random forest classifier.

AdaBoost Classifier :

First, the model will create a classifier. In this classifier, there will be some data which are misclassified. These misclassified data has greater weight in next iteration. In next iteration, there will be another classifier which gives priority to the weighted errors. Again this classifier makes some error which will give priority to the next classifier. Thus some classifier will be created one after another and combine these weak classifiers we will get a strong classifier. This class classifier is called AdaBoost classifier.

MultinomialNB :

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Comparison Between Models

Model Name	Accuracy
MultinomialNB	0.91875
KNeighborsClassifier	0.7375
DecisionTreeClassifier	0.98
RandomForestClassifier	0.98125
LogisticRegression	0.78125
AdaBoostClassifier	0.63125

Discussion

Classification is important in data analysis. We classified here different texts by their rating with different models. In this model analysis, we learned about different models and also learned how to vectorize texts. We can compare between models by the accuracy. We got a clear idea of model accuracy by plotting a horizontal bar graph.