Project Final Report

User Review Rating Prediction

TEAM MEMBERS:-

1) 140050044 - Naru Divakar Reddy
2) 140050060 - Gangam Rohith Reddy
3) 140050067 - Kilaru Manoj
4) 140050076 - Vakacharla Pramod

INTRODUCTION:-

To improve the quality of their products and services, many companies employ an user review system containing text reviews and overall ratings (in number of stars). However, the overall rating that usually accompanies online reviews cannot express the multiple or conflicting opinions that might be contained in the text. So, one solution to this is to convert the text reviews into numerical ratings that can be used to compare each service and product's quality. This is the main idea behind our project.

DESCRIPTION:-

Given a set of user review data each of which is associated with a user rating, the objective is to build a model that is able to predict numerical rating from textual data. We will try to do general review rating, but we mainly focus on restaurants as we got data on restaurants.

DATASETS:-

This data set has a file named yelp_academic_dataset_review.json which has a column called text which has user review and another column which has user rating we will use this file.

Each review contains a text sequence represents users review along with numerical user rating ranges from 1 to 5.

Link:-https://www.yelp.com/dataset_challenge.

Downloadable Link:-
https://www.dropbox.com/s/7iiasyripeciyxs/yelp_academic_dataset_review.json?dl=0

REFERENCES:-

Our main idea of doing this project is taken from the paper in the 1st link below. We are also using the ideas from the other links so as to have a more efficient idea which is more relevant to this Machine Learning course.

1. http://cs229.stanford.edu/proj2012/LeongBhatia-UserReviewRatingPrediction.pdf
2. http://arxiv.org/pdf/1511.05263.pdf
3. http://link.springer.com/article/10.1007%2Fs11257-015-9155-5#/page-1

IDEA:-

First we will divide the data from yelp into train data and test data. we will randomly select 80% of the data as train data. This will increase training accuracy and later will increase generalisation accuracy.

We intend to do feature selection using various methods like PCA or Forward Feature Selection etc.

We are planning to use learning algorithms like

1) Linear Regression
2) SVR
3) Logistic Regression
4) Naive Bayes ,on the raw data.

OUR PLAN TO SUBMIT FOR THE 21ST SUBMISSION:-

1. Details of the progress of our project till that day.
2. Pointers to literature that dealt with a similar problem of predicting the user review ratings.
3. Our next step in further progress towards the final submission from this stage.

OUR PLAN TO SUBMIT FOR THE FINAL SUBMISSION:-

The final report will describe the following broadly:

1. What is the problem that we worked on?
2. Why is the problem of user review rating interesting?
3. How is it related to the material covered in class.
4. Ideas we used in our project from the class.
5. What did we learn from the project?
6. Final results of the project.
7. How can we improve upon our project and future plans.

HOW OUR PROJECT IS RELEVANT TO CLASS:-

The problem addressed in our project has the data of the users' reviews. The output we desire is converting the text reviews into numerical ratings. We are planning to do this using the ideas of regression and the various methods of feature selection for the problem. The types of regression we are considering for our problem are Support Vector Regression (SVM), Ridge regression and others. We are also considering to do clustering to improve our methods, if necessary. As we learned these ideas from the class, we can say this project is completely relevant to machine learning taught in class.

TALKS DURING THE MEETS:-

We were just thinking of scraping the data set and just use linear regression on the data to establish a relationship between ratings and reviews.

Instead, they have advised us to use feature selection, clustering or other regression techniques on the data sets and choose the best technique among them that fits the test data. This enables us to have the best technique that suits the problem.