**Book Reviews Ratings Prediction**

**1 Project goal**

The problem of predicting a user’s rating for a product, given the user’s review for that product, is called Review Rating Prediction and has lately become a popular problem in machine learning and NLP.

Considering the above problem, I plan to analyze Amazon book reviews using NLP techniques to help predict book ratings. Each book review in the data set has a rating but there are many such reviews in Internet that have no ratings. This project aims to quantify such reviews in term of ratings and help determine the quality of items. I plan to use language models initially to get a baseline and improve upon it, and later enhance the project using semantic/sentiment analysis methods.

**2 Programming Languages**

I plan to use python as a major programming language for this project.

**3 Natural languages**

The book reviews are in English so it’s the natural language that I will be processing and analyzing.

**4 Data Set**

I will be using book reviews data set from Amazon that is in json format. This dataset contains book reviews and metadata from Amazon, including 12.4 millions reviews spanning May 1996 to July 2014. The data set contains the reviewer id, amazon standard identification number, helpfulness rating, review text, overall rating by the reviewer, summary of the review, and review time.

**5 Machine Learning Techniques and libraries**

I will be using a subset of machine learning techniques such as Language Models, Max Entropy Classifiers, Naïve Bayes Classifier, etc.

**6 Evaluation**

I will evaluate the results by comparing predicted ratings and actual ratings. I will be splitting the data set into train-test sets and train the model/classifier with reviews and corresponding labeled ratings. I will use the test data set to test the accuracy of my model. The goal will be to beat the baseline based on initial features.

**7 Project Timeline**

|  |  |
| --- | --- |
| First week: | Process the data and learn the techniques to create the models |
| Second week: | Create a baseline and start improving the baseline with additional features |
| Third week: | Conduct test and experiments |
| Fourth week: | Analyze the results and prepare the report |