Statistical Methods in Artificial Intelligence

Project REPORT

UTILITY SCORING OF PRODUCT REVIEWS

TEAM-1 MARS

Professor: Vineeth Gandhi **Teaching Assistant**:Aryaman Gupta

Introduction

We identify a new task in the ongoing research in text sentiment analysis: predicting utility of product reviews, which is orthogonal to polarity classification and opinion extraction. We build regression models and classification models by incorporating a diverse set of features, and achieve highly competitive performance for utility scoring and review classification on real-world data sets.

Datasets:

We downloaded reviews and related data in 22 different domains like Electronics, Clothing, Sports,Video games etc from amazon product data(http://jmcauley.ucsd.edu/data/amazon/).

We considered only reviews with denominator scores greater than 10. There are various features in the dataset namely:

1. Reviewer ID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
2. asin - ID of the product, e.g. 0000013714
3. Reviewer Name - name of the reviewer
4. helpful - helpfulness rating of the review, e.g. ⅔
   a. Numerator: Number of persons upvoted the review
   b. Denominator:Number of persons the review
5. Review Text - text of the review

6. overall - rating of the product
7. summary - summary of the review
8. Unix Review Time - time of the review (unix time)
9. Review Time - time of the review (raw)

Problem Definition:

Utility (or, reliability, usefulness, informativeness) is not the same as indifference. Totally indifferent or neutral reviews are useless; well-grounded subjective opinions can be convincing and illuminating. In other words, the utility of a product review is a property orthogonal to its polarity or embedded opinions. Our goal in this research is to build a computational model to predict the utility of reviews. We view the problem as one of regression. Formally, given a product review T, a number of features f1(T),...,fj(T),...,fPT) can be computed. Our task is to approximate a function.
The output $u \in [0,1]$ should reflect the real utility of T as accurately as possible.

But the other problem is that we do not know the category of the review then calculating its utility becomes difficult. So this task is divided into two parts:

1. **Utility score prediction of a review belonging to particular category**
2. **Utility score prediction of a review of unknown category**
   a. **Classification:**
      Given an unknown review, predict its category.
   b. **Utility Score prediction:**
      Train a model to predict the utility score of a review.

Approaches

Features:

1. **Lexical Similarity Features:**
   Clearly an informative customer review should not be a literal copy of another review because it should reflect the customer's own experience with the product, not the manufacturer's or other users description or expectation.
   For this we use **GMM clustering** and cluster the reviews based on the Proper Nouns

and their positions in text , number of other words between them and use distance of a review from the nearest cluster as feature.

2. **Shallow Syntactic Features:** We compute counts of words with the following part-of-speech tags in T , in order to characterize the subjectivity-objectivity mixture of the text at a shallow syntactic level:
    a. Proper nouns
    b. Numbers
    c. Modal Verbs
    d. Interjections
    e. Comparative and Superlative adjectives
    f. Comparative and superlative adverbs
    g. Wh Words

3. **Shallow Syntactic Features:** To capture the subjectivity-objectivity-mixture at lexical semantic level we calculate the counts of words in the following clue lists respectively
    a. Dynamic adjectives (e.g., “careful”, “serious”).
    b. Polarity Plus and Minus adjectives (e.g., “amusing”, “awful”)
    c. Gradability Plus and Minus adjectives (e.g., “appropriate”)
    d. Weak and strong Subjective Nouns
    e. Weak and strong Subjective

Classification of an Unknown review:

We have a set of 22 categories of product reviews. Now the task is to predict a category of review , so for that we used three different methods:

1. **Neural Networks:**
   We use one-hot vector representation of the review as an input to neural network with embedding layer. We used the following architecture of the neural network implemented in keras:
   Architecture:
    a. Embedding layer with the size of vocabulary.
    b. 1-D convolutional layer with 32 filters of size 3 with relu activation.
    c. Max pooling 1-D layer with pool size 2
    d. Fully connected layer with relu activation function.
    e. Fully connected layer with 22 nodes and softmax classifier.

2. **Adjectives:**
   Certain adjectives are used to describe specific categories. Generated adjectives for each category using a seed set and recursively adding all its synonyms. Using these adjective list we predict the category of a review using the count of the adjectives or its synonyms as the only feature. The category with maximum count is the class of the given review.
3. **Voted Classifier:**
   We classify a review based on the maximum vote among 5 classifiers used:
   a. Decision Tree
   b. Naive Bayes
   c. SVM
   d. Logistic Regression
   e. Random forests

   The ensemble model of all these is used to predict the model.

Now as we have predicted the model it becomes just utility score prediction of a given review with known category.

Learning Algorithms:

We experiment with two types of regression algorithms:

- **Support Vector Regression (SVR):**
  SVR is attractive due to the power of different kernel functions. In this study, we use the radial basis kernel function (RBF), which handles the potentially nonlinear relationship between the target value and features.
- **Simple linear regression (SLR):** SLR is a classical regression tool and has been widely used in numerous applications. It serves as a reasonable baseline.

Evaluating Measures

we can use the following metrics to evaluate its quality, both of which are standard in regression analysis:

- **Squared correlation coefficient:**
  Squared correlation coefficient is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data

points. An R^2= 1indicates that the regression line perfectly fits the data.

$$r^2 = \frac{\left(\sum_{i=1}^{n}(u_i - \overline{u_i})(\hat{u}_i - \overline{\hat{u}_i})\right)^2}{\sum_{i=1}^{n}(u_i - \overline{u_i})^2 \sum_{i=1}^{n}(\hat{u}_i - \overline{\hat{u}_i})^2}$$

u_i - real utility score
$\overline{u_i}$ - mean of real utility score
$\hat{u}_i$ - predicted utility score
$\overline{\hat{u}_i}$ - mean of predicted utility score

- **Mean Square Error:**
  In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations that is, the difference between the estimator and what is estimated.

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}(u_i - \hat{u}_i)^2$$

u_i - real utility score
$\hat{u}_i$ - predicted utility score

Experiments

- **Utility Score Prediction with known category**
  - Trained the models to predict utility scores using three types of features separately:
    - Lexical similarity
    - Shallow syntactic
    - lexical subjective
  - Combined Shallow syntactic and lexical subjective features to predict the score of a review.

- Combined all the above features to predict the score of a review with known category.

- **Prediction of category of an unknown review**
    - Train a neural network to predict the category of a review with unknown category.
    - Using specific adjectives and their count as features to predict category.
    - Voted Classifier of models to predict category:
        - Decision Tree
        - Naive Bayes
        - SVM
        - Logistic Regression
- Once The category is predicted, it is considered as a known category and the previous experiments i.e Utility score prediction with known category are repeated.
- **Prediction of category of an unknown review:**
  Instead of predicting the category of review , we trained a regression model using reviews from all the categories.

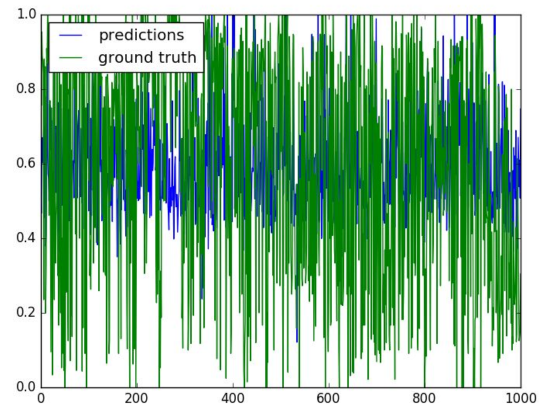
RESULTS

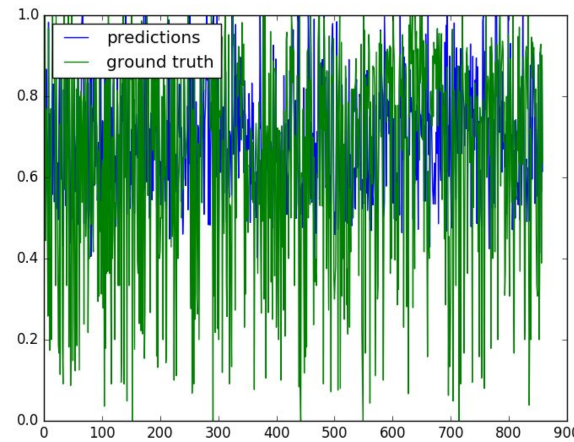
Results for the experiment 1 i.e **Utility Score Prediction with known category**

Type	SLR(SCC)	SLR(MSE)	SVR(SCC)	SVR(MSE)
Apps for android	0.0588	0.0604	4.94e-05	0.0668
Automotives	0.0135	0.0518	0.0081	0.0458
Baby	0.0559	0.0376	1.01e-05	0.0384
Beauty	0.0308	0.0386	0.0056	0.0403
CDs and Vinyl	0.2353	0.0896	0.0159	0.1155
Cell Phones and accessories	0.0433	0.0419	0.0058	0.0436
Clothing Shoes	0.0032	0.0156	0.0005	0.019

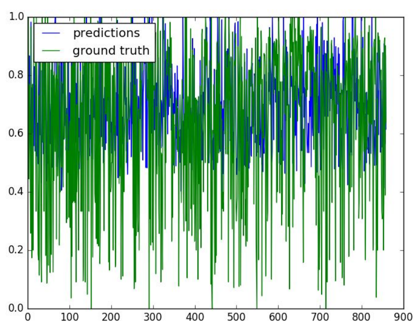
and Jewellery				
Digital Music	0.1753	0.0613	0.0076	0.07121
Electronics	0.0654	0.0494	0.0018	0.0537
Grocery_Gourment_Food	0.0143	0.0488	5.07e-05	0.0504
Health and Personal Care	0.04575	0.04625	0.004	0.0498
Home and Kitchen	0.049	0.0414	0.0009	0.0044
Kindle Store	0.0345	0.0475	0.00855	0.05095
Musical Instruments	0.022	0.049	0.223	0.0516
Office Products	0.027	0.013	0.0006	0.02
Patio_Lawn and garden	0.0737	0.0155	0.0075	0.0160
Pet Supplies	0.025	0.027	0.0001	0.02976
Sports and Outdoors	0.0650	0.0463	1.7e-06	0.051
Tools and Home improvement	0.056	0.0440	0.0016	0.04535
Toys and games	0.045	0.060	0.0004	0.055
Video Games	0.1698	0.0787	0.0072	0.0924

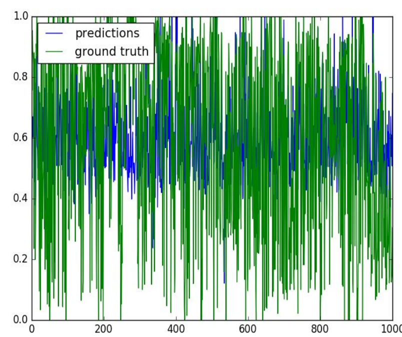
Some of the images are shown here, **rest are attached in the folder**

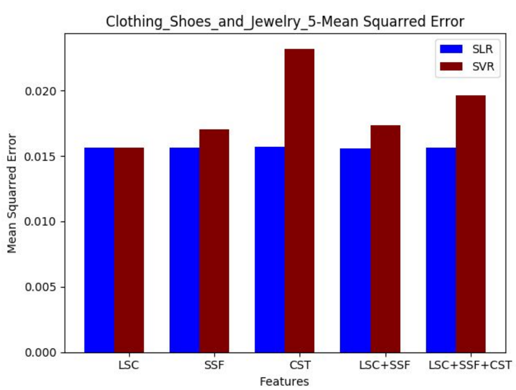


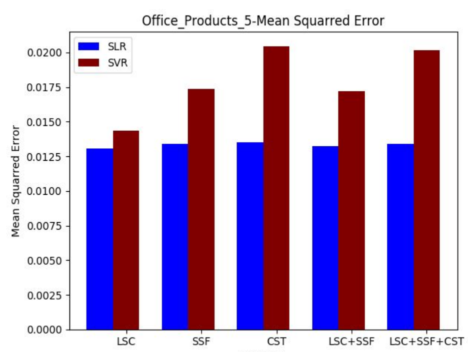


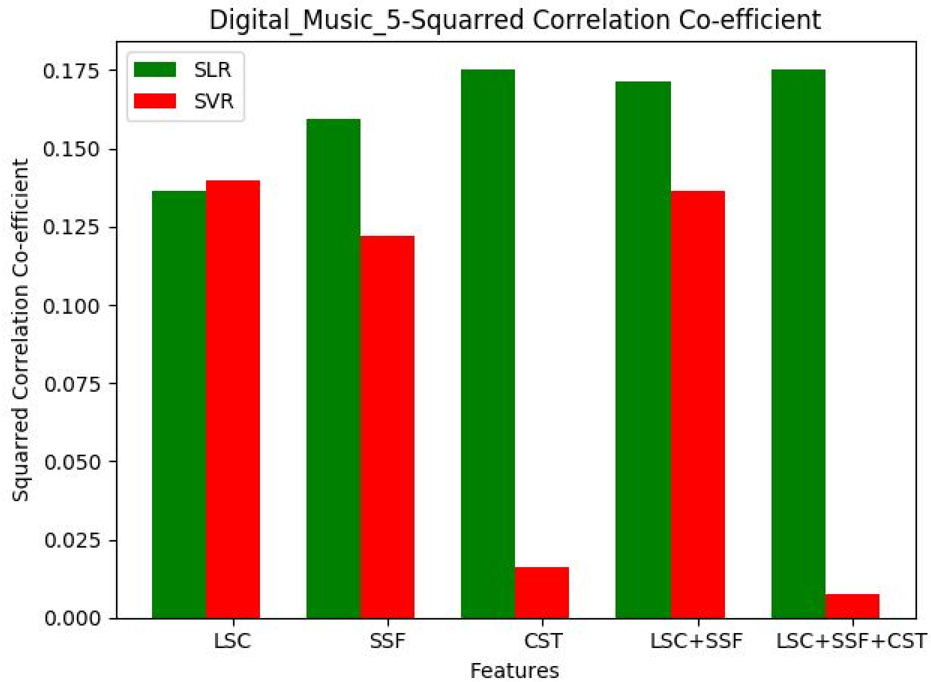
Digital Music (SLR)   Movies and TV(SLR)

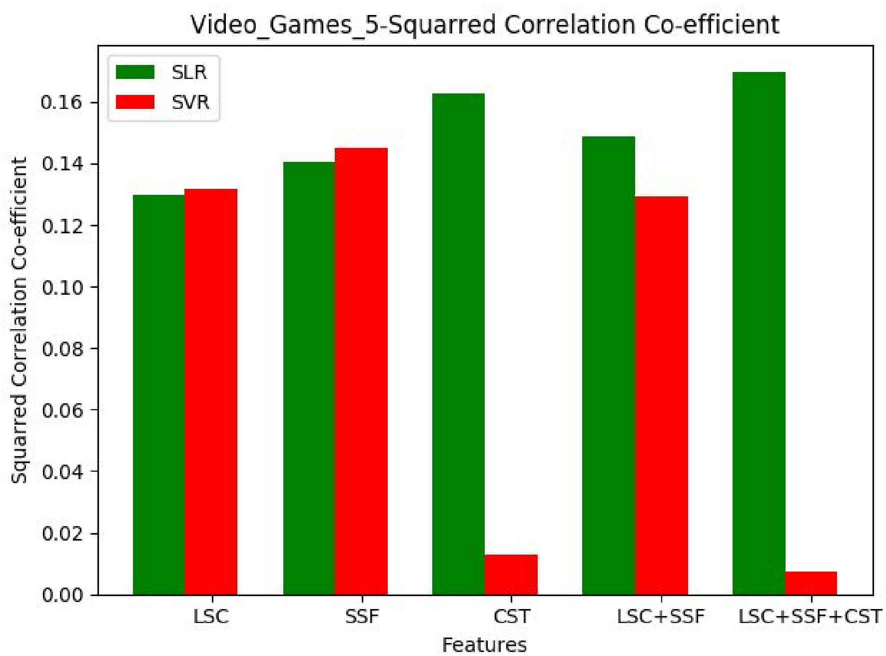






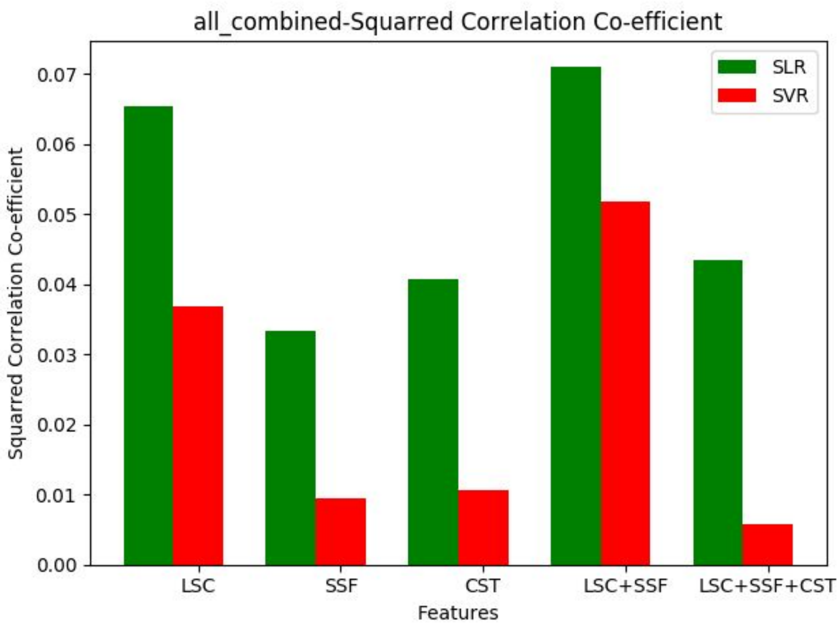


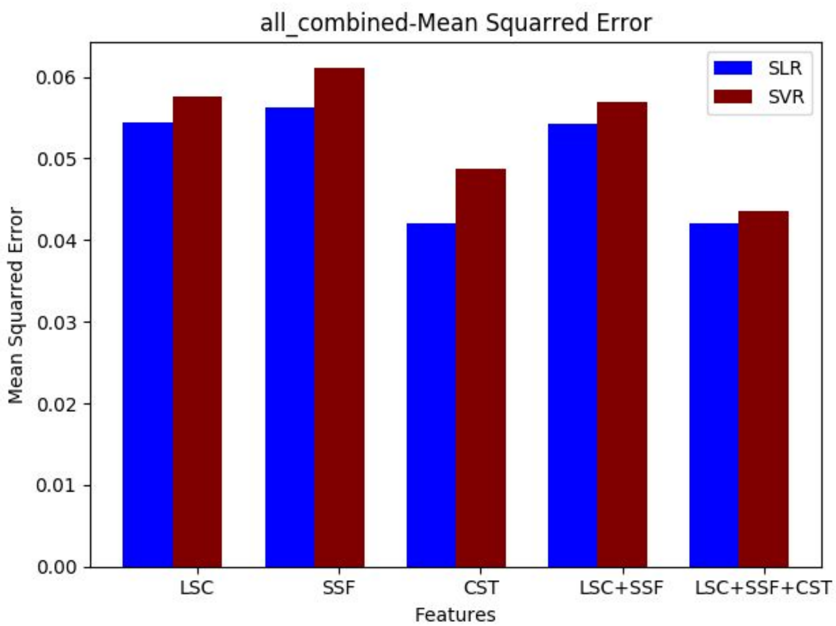




Results for the experiment 2 i.e Results of regression model trained using reviews from all the categories, instead of predicting the category of review ,

Features	SLR(SCC)	SLR(MSE)	SVR(SCC)	SVR(MSE)
Shallow syntactic Features	0.065393	0.0544	0.0368	0.575
Lexical Subjective Clues	0.0332	0.0562	0.0095	0.0611
Clustering	0.040	0.042	0.0106	0.0486
Shallow Synatactic + Lexical Subjective	0.0710	0.0541	0.0518	0.05690
Shallow Syntactic +Lexical Subjective + Clustering	0.0433	0.04199	0.058	0.0436





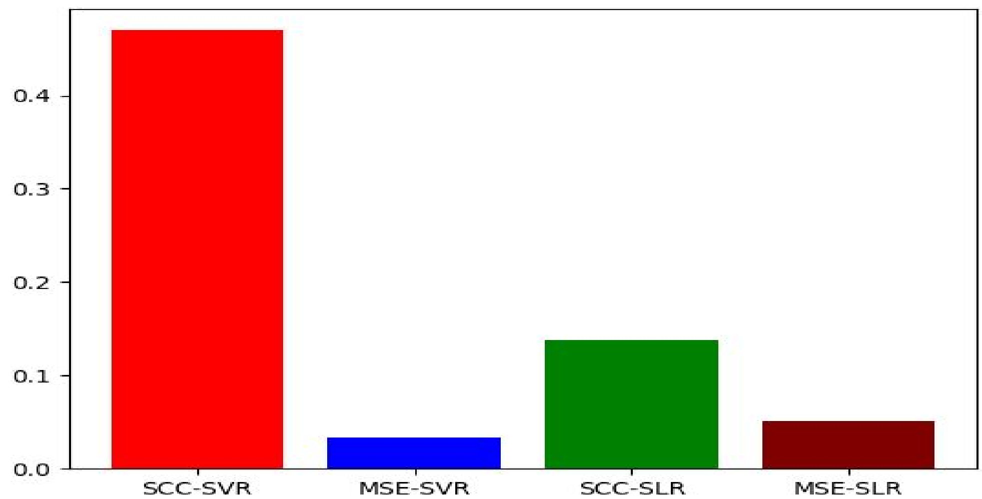
Prediction of category of an unknown review

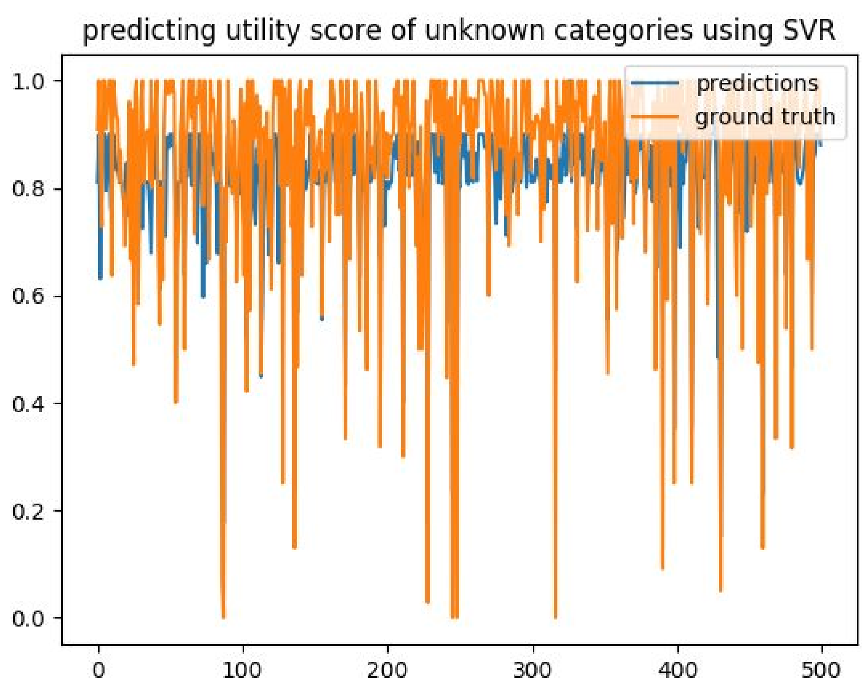
1. **Accuracies of various models used for predicting the category:**
   a. Neural Networks: **70.036**
   b. Using Adjectives: **45.36**
   c. Voted Classifier: **82.32**

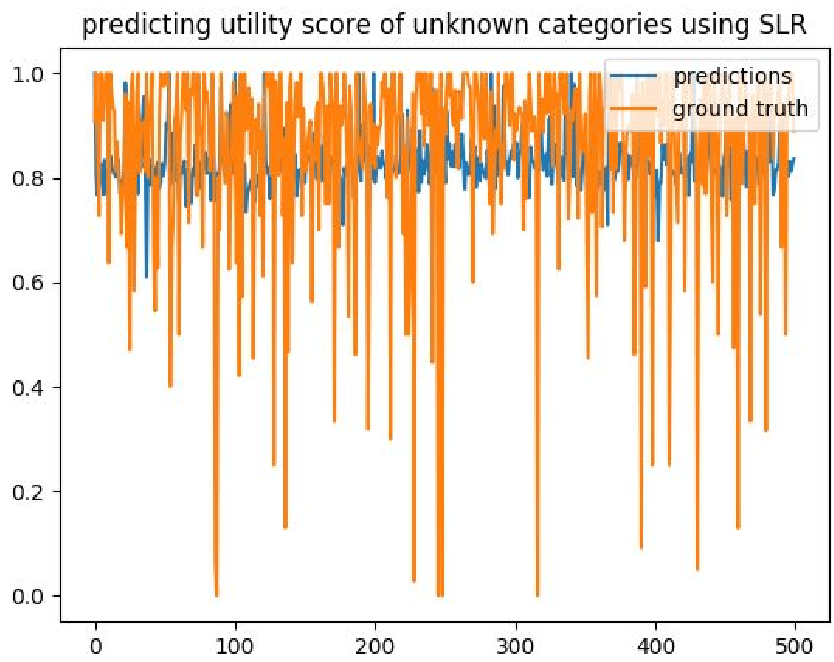
2. **Results of utility score predictions:**

SLR(SCC)	SLR(MSE)	SVR(SCC)	SVR(MSE)
0.4698	0.0330	0.1380	0.0515

MSE and SCC after classifying reviews
and predicting utility scores using corresponding models







Analysis and Conclusions

- **Classification**
  - Accuracy in classification is highest in the case of Voted Classifier.
  - Voted classifier has the advantage of testing a review with 5 different classifiers.
  - Neural networks also works good but less accurate than voted classifier.
  - The naive method of counting number of matches in adjectives does not work well as there is no learning based in this method and also it is ambiguous.
- **Score Predictions**
  - Each of the 5 models used for predicting the scores individually gives satisfactory results on particular type of data i.e they are data dependent.
  - Since the description and grammar used by users vary, the accuracies of classifier will depend highly on the data given. Thus, we can see that the predictor with combination of all features gives better accuracy than single feature predictors.

- Adding lexical similarity features was effective.

REFERENCES:

1. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6726818&tag=1
2. http://scikit-learn.org/stable/documentation.html
3. https://www.sciencedirect.com/science/article/pii/S1567422311000639