Comprehensive Restaurant Rating System

Abhishek Rath
Department of Computer Engineering
San Jose State University
San Jose, CA - 95192
abhishek.rath@sjsu.edu
(013718107)

Chaitanya Valluri
Department of Computer Engineering
San Jose State University
San Jose, CA - 95192
chaitanya.valluri@sjsu.edu
(08507291)

Smitha Eshwarahalli
Department of Computer Engineering
San Jose State University
San Jose, CA – 95192
smitha.eshwarahallli@sjsu.edu
(013714545)

Sudarshan Aithal
Department of Computer Engineering
San Jose State University
San Jose, CA - 95192
sudarshan.aithal@sjsu.edu
(013638703)

***Abstract* - This paper focuses on the project performed under Machine Learning course. The project involves Comprehensive Restaurant Rating System. Businesses around the world need how to improve themselves and ratings help them do that. Here, we have made a comprehensive ratings systems for restaurants based on Yelp dataset which has numerous review and various other parameters and using Support Vector Machines to give a comprehensive rating. This project focuses on giving an overview of the ratings that the businesses and even customers can use to improve their experience.**

Keywords - Machine Learning, Rating System, Support Vector Machine

I. INTRODUCTION

In today's world where analytics play an important role to help businesses succeed and not only that but help the customers or users to improve their experience, Machine learning helps to improve this process. As businesses focus on higher ratings they can target focused groups and get to know the fields they have to work on to improve. This also decides their offerings. This does not only apply to restaurants as seen in this case of project but to every business existing out there. The amenities provided by businesses can vary from Wi-Fi, parking lot to hygiene and immediate response time. Every businesses has different parameters. Machine learning can help businesses to notify them on what parameters to focus and which parameters are to be ignored based on analysis of reviews. We are focusing on getting a comprehensive rating of every restaurant available on Yelp which can give the customers a uniform and unbiased review. The project can also help the people related to restaurant business to realize and focus on parameters that need improvement or can be worked on and what parameters to not focus on which can improve their performance and reduce workload.

II. BACKGROUND WORK

Many journals and research papers were referred so that we can make this model and also know about the work that has been till now. Some of the papers guided us in making the rating system and some papers were useful in deciding the model which we should use for our comprehensive recommendation system. Some were helpful in deciding the process that we might have to go through to get an efficient model.

a) The paper that was most useful was by Wael Farhan[1]. It was the paper that made us go for this project. The author focuses on developing a linear regression predictor which will help restaurant owners to get insight on their performance and improve customer experience and satisfaction based on various attributes of restaurant. He has done an analysis of Yelp dataset consisting of restaurant reviews and suggested various methods to predict ratings of the restaurant. He has gone through following models in his paper:

i) Neural Networks
ii) Naïve Bayes
iii) Decision Trees (Random Forest)
iv) Linear Regression

b) Next paper which was helpful was by Peter Mark[2] from University of New Hampshire where he predicted Food estabilishments ratings only based on business attributes. He has proposed a linear regression model based on various business amenities and other attributes. He was able to predict the behaviour of consumers. He especially focussed on the attributes that had significant importance based on the correlation testing performed by him. He found a state based consumer behaviour of people going to restaurants which will be very helpful to businesses.

c) The paper by Aansi Kothari and Warish Patel[3] was responsible for encouraging us to go for Support vector mcahines for Recommendation or Review system. He has described various examples where SVM has helped in implementation. He has implemented his proposed SVM based model on Trip Advisor review and rating system. He also used collabarative filtering to get better results. He has mentioned that using Support vector Machine model helps in reducing the chances of misclassification. He has used mean Reciprocal Rank and for und hit ratio for top recommendation for the dataset he found. They also realized that this increases the speed of training and testing the dataset also making the whole system more efficient. He compared the performances with MI, CHI and IG models to prove the point. The reviews were used to detect aspect-level context-dependent preferences. The model combines above parameters with independent preferences to get review or recommendation. It was concluded that customers opinions on certain attributes are meaningful to correlate with the other factors. He realized that this correlation holds importance as it aids in classifying users preferences under different contexts. Results showed that this method significantly perfomed better than the related context based recommendation method.

d) Paper by Yichuan Tang[4] focuses on Support Vector Machine. The focus is on exploring various aspects of Support Vector Machine. The aspects that were discussed by Yichuan were Softmax and Multiclass Support Vector Machine. Data mining was used in developing potential model. The model was successfully developed to create a image recognition system using Support Vector machine. Various iterations were done using numerous methods to get better results.

e) Paper by Ngo Ye[5], Sinha and Sen focused on using script analysis to predict helpfulness rating of online reviews. People were asked to highlight parts of reviews that they consisdered as important and then the phrases were added to a text dataset. Regression method was applied on text to predict review helpfulness based on words in the dataset. It was compared against models such as Baseline model and bag-of-words model. It was found that the scripts enriched model was found to be most accuaret among all the models.

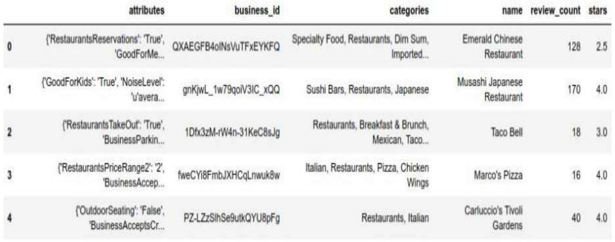
f) Analysis[6] was performed on the Yelp dataset to identify a correlation between rating and a customer's location.This was done to find any relation that might be existing between rating and the distance of a customer from their house. They also found out that there is a relation between the ratings of items that are bought by cutomers are common with the customers' friends. They also observed that customers rate products higher those are available farther from home. Similarity was found between the customers and their friends' reviews even in this aspect.

III. DATA PROCESSING

The dataset which is used in this project is provided by Yelp. The business review dataset which was to be used for analysis originally contained 192609 records with 14 variables. The users review dataset originally contained approximately 5,200,000 user reviews. To reduce the computation time, TF/IDF is used to vectorize the text data.

A. Feature Extraction

We realized that the data had many variables that were not required for this project. We decided to clean the data so that the sample can be drawn for analysis. First of all, the variables that were found to be not useful or unrelated were dropped using Correlation Matrix. The sample was narrowed down to 163773 records with 6 variables after dropping records with any of the 6 variables missing. As the goal of our project is to get ratings for restaurants, dataset was filtered to only include records belonging to "Restaurants" categories. This narrowed the dataset considerably, bringing it down from the initial 163773 records to 57176. The user review dataset was processed to only have restaurant reviews which narrowed down the dataset to 164045 reviews with 5 variables of interest, i.e. the star ratings (1 to 5). This reduced the dataset so that it reduces unnecessary computing.

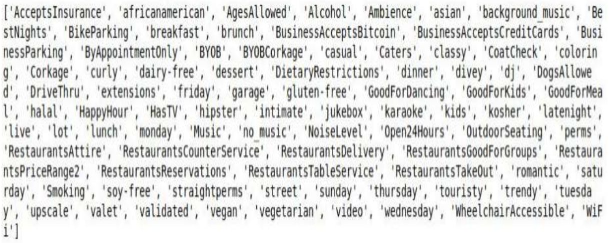


	attributes	business_id	categories	name	review_count	stars
0	{'RestaurantsReservations': 'True', 'GoodForMe...	QXAEGFB4oINsVuTFxEYKFQ	Specialty Food, Restaurants, Dim Sum, Imported...	Emerald Chinese Restaurant	128	2.5
1	{'GoodForKids': 'True', 'NoiseLevel': 'u'avera...	gnKjwL_1w79qoiV3IC_xQQ	Sushi Bars, Restaurants, Japanese	Musashi Japanese Restaurant	170	4.0
2	{'RestaurantsTakeOut': 'True', 'BusinessParkin...	1Dfx3zM-rW4n-31KeC8sJg	Restaurants, Breakfast & Brunch, Mexican, Taco...	Taco Bell	18	3.0
3	{'RestaurantsPriceRange2': '2', 'BusinessAccep...	fweCYi8FmbJXHCqLnwuk8w	Italian, Restaurants, Pizza, Chicken Wings	Marco's Pizza	16	4.0
4	{'OutdoorSeating': 'False', 'BusinessAcceptsCr...	PZ-LZzSlhSe9utkQYU8pFg	Restaurants, Italian	Carluccio's Tivoli Gardens	40	4.0

Fig. 1: Dataset Sample

B. Data Analysis

After getting all the data related to restaurant reviews which are based on both business and customer reviews as we need the analysis of both the sides of business. All the attributes in the dataset were extracted. Certain attributes such as alcohol, were split up into types. For instance, the variable, "Alcohol", would have the type, "none", "beer and wine", or "full bar". Some variables, such as "Ambience", had subtypes, like "hipster", "divvy", and "trendy". Subtypes listed within these variables had binary options, true or false. An individual column for each variable was created. Each column would portray the variables as either having or lacking the certain attribute in a binary format (0 or 1). Finally, this resulted in 87 attributes



```
['AcceptsInsurance', 'africanamerican', 'AgesAllowed', 'Alcohol', 'Ambience', 'asian', 'background_music', 'Be
stNights', 'BikeParking', 'breakfast', 'brunch', 'BusinessAcceptsBitcoin', 'BusinessAcceptsCreditCards', 'Busi
nessParking', 'ByAppointmentOnly', 'BYOB', 'BYOBCorkage', 'casual', 'Caters', 'classy', 'CoatCheck', 'colorin
g', 'Corkage', 'curly', 'dairy-free', 'dessert', 'DietaryRestrictions', 'dinner', 'divey', 'dj', 'DogsAllowe
d', 'DriveThru', 'extensions', 'friday', 'garage', 'gluten-free', 'GoodForDancing', 'GoodForKids', 'GoodForMea
l', 'halal', 'HappyHour', 'HasTV', 'hipster', 'intimate', 'jukebox', 'karaoke', 'kids', 'kosher', 'latenight',
'live', 'lot', 'lunch', 'monday', 'Music', 'no_music', 'NoiseLevel', 'Open24Hours', 'OutdoorSeating', 'perms',
'RestaurantsAttire', 'RestaurantsCounterService', 'RestaurantsDelivery', 'RestaurantsGoodForGroups', 'Restaura
ntsPriceRange2', 'RestaurantsReservations', 'RestaurantsTableService', 'RestaurantsTakeOut', 'romantic', 'satu
rday', 'Smoking', 'soy-free', 'straightperms', 'street', 'sunday', 'thursday', 'touristy', 'trendy', 'tuesda
y', 'upscale', 'valet', 'validated', 'vegan', 'vegetarian', 'video', 'wednesday', 'WheelchairAccessible', 'WiF
i']
```

Fig. 1: List of attributes as per dataset

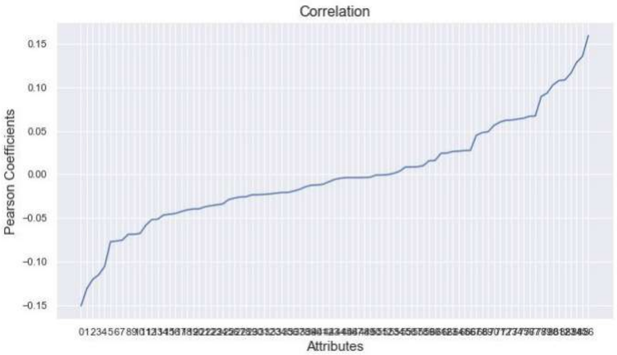


Fig. 2: Correlation graph of all the variables before feature Data Analysis

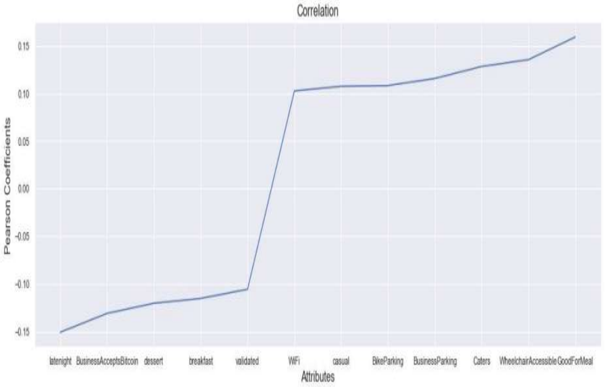


Fig. 2: Correlation graph of all the variables after Data Analysis

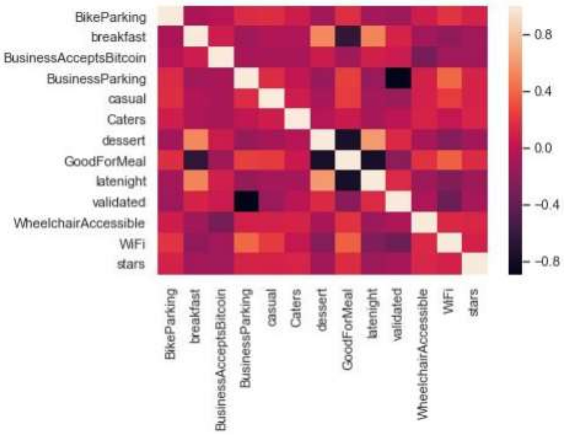


Fig. 3: Correlation Matrix of the variables after Data Analysis

C. Balancing the Reviews

The dataset do not contain uniform number of reviews across every star ratings. So we are balancing the dataset to make sure no biased selection occurs during testing and training. The dataset is scaled to 15267 i.e. total number of 2 star ratings which is the least.

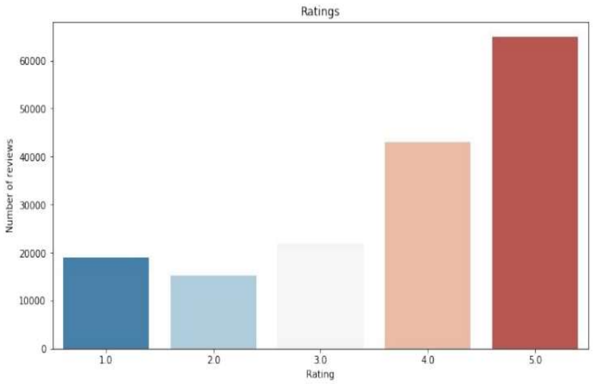


Fig. 4: Ratings vs Number of Reviews Graph before balancing

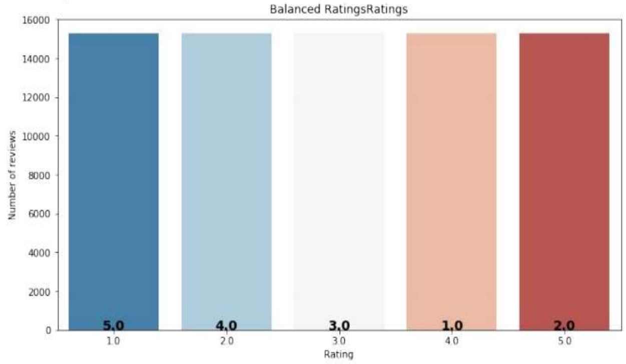


Fig. 5: Ratings vs Number of Reviews Graph after balancing

IV. THEOROTICAL APPROACH

We went through various models to solve this problem. But after performing various trial-and-error we decided to go for Linear Support Vector Machine (LSVM). Decision Tree was also used in this model. It is used in this project to classify and predict the outcomes.

A. Linear Support Vector Machine

- Goal of any support vector machine (SVM) algorithm is to find a hyperplane in a N-dimensional space, where N is the number of parameters or features, that can distinctly classify the given data points.
- There are various models to classify data points but we need to find a hyperplane that has maximum margin so as to provide reinforcement to the data points that might be added in future.
- Support vectors play a crucial role in SVM. They are the data points which are closest to the hyperplane. They can influence the position and orientation of the hyperplane. We can use these support vectors to manipulate and maximize the margin of the classifier.
- Linear support vector machine is mainly used for classification purpose.

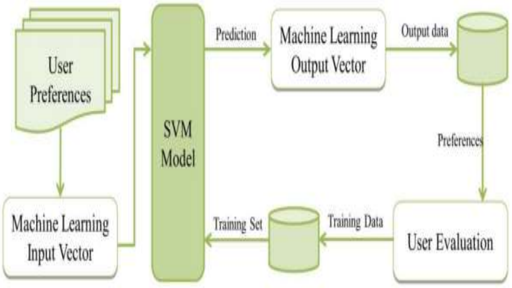


Fig. 6: Basis Working of SVM Model

B. Decision Tree

Decision tree is used as a support tool for the decisions taken. It is a tree-like model that represents all the decisions that were taken and possible consequences that may result in. It includes chance event outcomes, resource costs, and utility. It is a flowchart-like structure wherein each internal node represents a parameter of the model. Each branch is the result of the tests performed and each leaf node consist of computed decision of attributes.

Here, the expected values of alternatives that are competing are calculated.

It consists of three types of nodes:

- Decision nodes
- Chance nodes
- End nodes

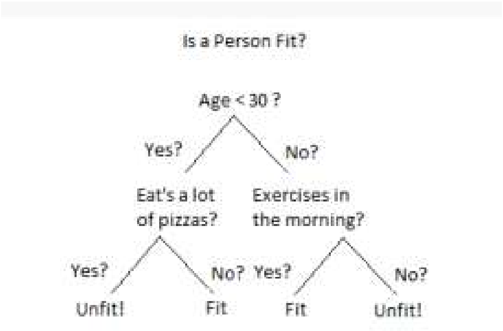


Fig. 7: Example of Decision Tree

C. Equations

Various factors combine together to help us in building a SVM like Large Margin Intuition, Cost Function and Gradient Updates.

a) Large Margin Intuition

In SVM, linear function's output is taken. Now if the output is greater than 1, it is considered to be in a particular class and in case output is -1, it can be considered to be in another class. Threshold values changes in between 1 and -1. This can be used to get reinforcement range of values [-1,1] that can act as margin.

b) Cost Function and Gradient Updates

Cost or Loss function that helps us in maximizing the hyperplane is known as hinge loss. The hinge loss is represented as:

$$\min_w \lambda \| w \|^2 + (1 - \sum_{i=1}^{n} y_i < x_i, w >) \quad □□□$$

Regularization is an essential addition in order to balance margin maximization. Adding regularization gives us below equation

$$\delta\lambda \| w \|^2 / \delta w_k = 2 \lambda w_k \quad (2)$$

Taking partial derivatives of above equation with respect to the weights so as to find gradient. We can use the gradients to update weights. In case of misclassification, our updated weight will be:

$$w = w + \alpha . (y_i . x_i - 2 \lambda w) \quad (3)$$

Otherwise, in case of no misclassification, the updated weight will be:

$$w = w + \alpha . (2 \lambda w) \quad (4)$$

V. PRACTICAL APPROACH

Linear regression and Decision tree regression models were built. The variables used in modeling were limited to those identified with a statistically significant correlation to business rating after correlation analysis. The figure given below is the representation of the decision tree built for the project:

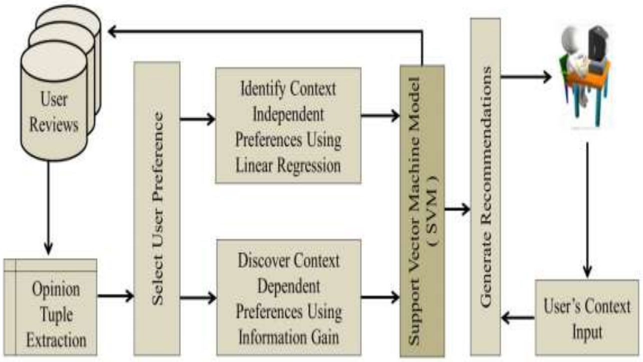

User Reviews
Select User Preference
Identify Context Independent Preferences Using Linear Regression
Discover Context Dependent Preferences Using Information Gain
Opinion Tuple Extraction
Support Vector Machine Model (SVM)
Generate Recommendations
User's Context Input

Fig. 8: Flow of Model

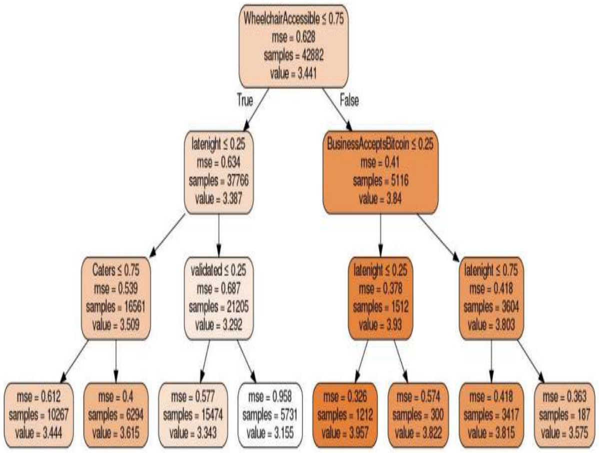
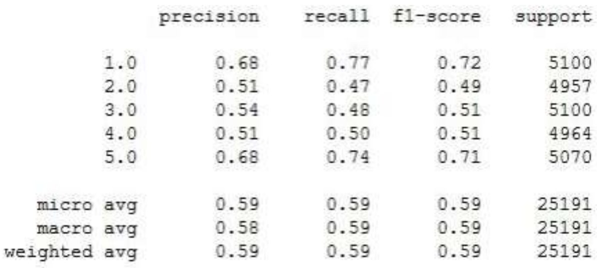


Fig. 9: Decision Tree representing every feature

VI. RESULTS

A model that could give a fair comprehensive rating and fairly predict business rating for restaurants based on business attribute was developed. It is also able to help the businesses in finding attributes that have a significant correlation to business rating. We were able to get Root Mean Square Error (RMSE) of about 0.74 that is 74%. The model is able to predict restaurant ratings using Support Vector Machine (SVM) classifier with an accuracy of 59.21%.



```
              precision    recall  f1-score   support

        1.0       0.68      0.77      0.72      5100
        2.0       0.51      0.47      0.49      4957
        3.0       0.54      0.48      0.51      5100
        4.0       0.51      0.50      0.51      4964
        5.0       0.68      0.74      0.71      5070

  micro avg       0.59      0.59      0.59     25191
  macro avg       0.58      0.59      0.59     25191
weighted avg       0.59      0.59      0.59     25191
```

Fig. 8: Classification Report

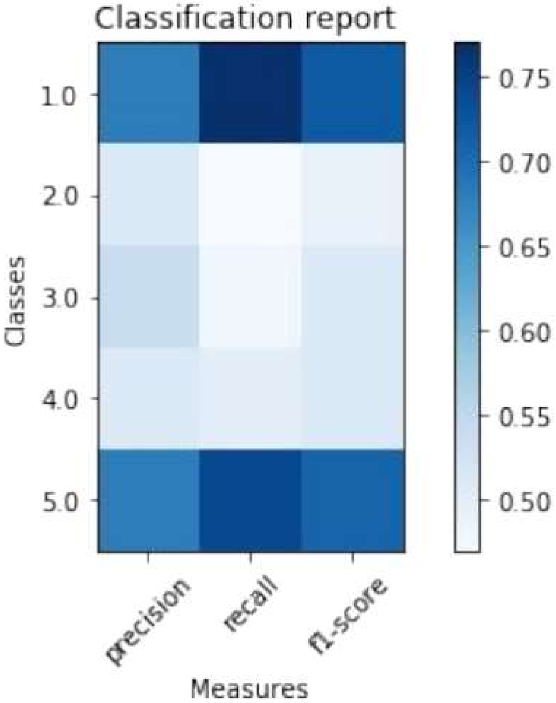


Fig. 9: Classification Report Matrix

VII. CONCLUSION

Extracting restaurant ratings from was difficult and time consuming. Also, determining the influencing factors for the model was tricky. But after overcoming these difficulties we were able to develop a comprehensive ratings system. Analysis of a pooled set of restaurant records, reviews and their attributes produces important insights about the general rating behavior of consumer. This information is valuable to businesses, as they may be able to identify the most impactful features on rating, effectively implement or remove them, and potentially raise future ratings. This will also be helpful to customers looking for a good experience. As it provides an unbiased rating it is more efficient than the ratings that are based on just average ratings of reviews. This model can be used to improve the performance of not only restaurants but also various areas of businesses. This system can be used in any situation that requires one to give an unbiased review.

ACKNOWLEDGMENT

Deep gratitude to Professor Mahima Agumbe Suresh for guiding and providing very of the basic information to Machine Learning Concepts and their implementation. Her classes on topics such as Support Vector Machines, PCA and various other topics were very helpful and resourceful.

REFERENCES

[1] Wael Farhan, Predicting Yelp Restaurant Reviews.

[2] Peter Mark Shellenberger Jr., Predicting Yelp Food Establishment Ratings Based on Business Attributes.

[3] Aansi A. Kothari, Warish D. Patel, A Novel Approach Towards Context Based Recommendations using Support Vector Machine Methodology.

[4] Yichuan Tang, Deep Learning using Linear Support Vector Machines.

[5] Ngo-Ye, Sinha, Sen, Predicting the Helpfulness of Online Reviews using a Scripts - Enriched Text Regression Model

[6] Sunil, M, Bari, Shetty ,Prediction of Rating by Using Users' Geographical Social Factors