ISSN No:-2456-2165

Review Rating Prediction Using Yelp Dataset

[1]Abhishek K B, [2]Divyashree M S, [3]Keerthana C Shekar, [4]Meghana M Koti, [5]Karthik V
Assistant Professor
Vidyavardhaka College of Engineering, Mysuru, India

Abstract:- The customer review is important to improve services for company. Websites allow users to write opinion about various businesses in the form of textual review and star rating in online platform such as yelp and these reviews affect the customer's shopping behaviour. Semantic analysis can be used for predicting the star rating which extract important information from textual review and convert this information into star rating by using machine learning algorithms. The yelp dataset is divided into training and test data for modelling.

Keywords:- *Feature Extraction, Semantic Analysis, Supervised Learning Algorithms.*

I. INTRODUCTION

With the emergence of internet era, millions of user reviews networking through social media platforms has become more significant. Yelp is a review website that allows users to rate various business categories.

The number of customer reviews is increasing or huge from websites, blogs, forums and social media, which the services or product is interesting. Therefore to prevent the customers from reading review randomly and making a bias decision on services or products, star rating is predicted for the same by using models and machine learning algorithms. Semantic analyses conducted on user reviews are the main approach for predicting class label which is star rating.

II. LITERATURE SURVEY

Yelp dataset challenge for review rating prediction is to build a good predictor that efficiently selects vital features of the product from reviews to assess the importance with respect to rating quantitatively. N-grams model are used to collect useful n-grams from text review which includes unigrams (one word), bigrams (two words) and trigrams (three words). To train our prediction model, linear classification algorithms such as logistic regression, support vector classification and naïve Bayes multiclass classification algorithm are used. The RMSE and Accuracy graphs are used to measure the performance of each of the above algorithms. Among which the Logistic Regression performed the best with the highest RMSE and accuracy score of 64% [1]

Opinion mining for analysis and prediction of rating is performed for 400 opinion texts. Frequent words are selected as feature vectors into classifier models. The classifier models classified texts as class labels: positive or negative. The models comprise of naïve Bayes and decision tree. Naive Bayes is a classification technique that creates a probabilistic model. Posterior probability is computed for each class label and class with the maximum probability is the class label. Cross validation techniques such as k fold is used for evaluating accuracy. Naïve Bayes performed better with an accuracy (93.61) compared to decision tree (92.84).their approach to general star rating given predicted class target was inconsistent with the opinion.[2]

Yelp dataset challenge to predict a business star from customer reviews can be implemented using feature generation method. The dataset contains 3500 text reviews parts of speech analysis is used to construct feature vector. Different feature, vectors from high frequency words and top frequent adjectives are selected and applied on linear regression and decision tree regression. Linear regression performed the best among different feature vectors with average RMSE of 0.6.Decision tree regression is robust with respect to number of features. [3]

Another approach to semantic analysis for yelp review star rating was introduced. Positive and negative word frequencies are selected as attributes. The learning method to train and test data are random forests, decision tree and multinomial naïve Bayes. The winners are decision tree and random forest with up to 51% accuracy. [4]

Belief propagation method is developed to predict the rating of such users.in belief propagation method, a bipartite graph is generated where user nodes are connected with reviews and stars as messages. Marginal distribution in the graph is calculated for every user to provide fair evaluation. This method shows around 10% improvement on prediction accuracy. Limitation of Belief propagation is that implementation of bipartite graph for all users and restaurants is not possible for limited resources.[4]

Some of the user reviews is contrast to the star rating given to any business platform. To overcome this vulnerability a sentiment analysis with different approach was made which categorizes the user reviews as positive and negative. The data set obtained from the yelp data set is used to classify the user reviews. The average rating was about 3.6, the ratings below this value is considered to be the negative responses set and the rating above average belongs to positive response set. First, the entire pool of document was divided into the corresponding ranking where the

occurrence of certain words for the star rating will be dumped to the same bin. Next a few comparison models were applied. First, they applied Naive Bayes classifier where this model calculates the probability of each word and the accuracy is about 78%.Support Vector Machine was implemented later which involves textual analysis and the accuracy obtained is of 73.50%. [5]

Recommended system is now a key tool for providing personalized recommendations on items such as music, books, news etc. The three different models or approaches are made to predict the future star ratings for the users. Content-Based filtering, Collaborative-Filtering and Hybrid approach are used where the methods are based on user's interest with description, analyzing the information of user's and a combination of both respectively. This approach was applied for few data sets like 20 reviews which got less error than other approaches. The best algorithm which got less error was Binary Decisions tree regression of about 0.78MAE (Mean Absolute Error).Since this method used just 20 reviews, the future work is to use more data sets and to obtain best accuracy. [6]

III. COMPARITIVE ANALYSIS

The results of different classifiers with the feature extraction techniques are analysed. N-gram model and bag of words model can be used as feature extraction method to select feature attributes. In an average logistic regression and linear support vector classifier give high accuracy scores of 64%(logistic regression) and 63%(support vector classification) for large number of features. Naïve Bayes and decision tree algorithms are effective in classification of text review as positive and negative based on probabilistic model when there are less than hundreds of features. Naive Bayes classifier provides better accuracy of 94% compared to decision tree with 92% accuracy score in predicting the star rating. For hundreds of features linear regression performs the best with least RMSE 0.6 among other regression model

IV. CONCLUSION

Review rating prediction is effective in business to effectively understand the opinion of customers and improve the services being provided. A classification model for predicting the star rating of a given review has been proposed. The necessary features are extracted using n-gram model or bag of words model to build feature vector. The feature vectors are combined with supervised learning algorithms for constructing model. For the classification model logistic regression, naïve bayes and support vector algorithms perform with better accuracy. k fold cross validation can be used as evaluation model. Therefore from the study the former approach is preferred.

REFERENCES

[1]. Mingming Fan and Maryam Khademi. Predicting a business star in yelp from its reviews text alone. arXiv preprint arXiv:1401.0864, 2014.
[2]. Wararat Songpan, The Analysis and Prediction of Customer Review Rating Using Opinion Mining, 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA) 2017.
[3]. Nabiha Asghar, Yelp Dataset Challenge: Review Rating Prediction, Published 2016 in ArXiv.
[4]. "Yelp Dataset Challenge." [Online]. Available:http://www.yelp.com/dataset_challenge/.
[5]. Jason Jong, Predicting Rating with Sentiment Analysis, December 16, 2011.
[6]. Shuang Wu, Xiaodong Wang, Bozhao Qi, A New Semantic Approach on Yelp Review-star Rating Classification.