HOTEL CLASSIFICATIONS IN TRAVEL REVIEW DATASET

Dr.Ayyappan.G, Dr.A.Kumaravel

Associate Professor, Machine Learning Group, Department of Information Technology, BIHER, Bharath Institute of Higher Education and Research, Chennai.

Professor, Machine Learning Group, Department of Information Technology, BIHER, Bharath Institute of Higher Education and Research, Chennai.

ayyappangmca@gmail.com

Abstract

Travel and Tourism sector has vast growth in a current situation. So Recommender models are embryonic as a necessary to travel and tourism sector. Due to usage of Social media in current situation, huge volumes of recorded fact have been developed. Traditional system may not be enough for managing these kinds of data. So new system need to be developed in this research work based of the hotel classification made in travel and review dataset. These researches focus on the various classifications and visualize threshold curve measurements and time taken to build the each model. More or less Meta and Rules classifier have been produced the highest accuracy in this research work.

Keywords: NaiveBayesMultinomialText, MultiScheme, ZeroR, SGDText, Travel Review

I INTRODUCTION

In this section presents introduction of this research work. In end user perspective, travel and tourism is mostly explorative in nature and repetitive travels to same locations are minimal. So, travelers have to take decisions regarding their destinations and associated facilities to be consumed without adequate prior or personal knowledge. The best option available is to leverage social media and internet, but the amount of time required to extract relevant information is too high. Tourism recommenders are the best solutions in this scenario. Recommender systems helps in terms of automated filtering, processing, personalization and contextualization of the huge volume of data that is available and growing on a daily basis on the internet and the social media.

In this paper presents section 2 of this paper explains the detail on the related works. In section 3 presents the materials and methods adopted and section 4 presents the details of the experiments and discussions. Finally section 5 concludes the paper by sharing our inferences and future plans.

II RELATED WORKS

In this section presents focuses the related works of this research work. A. Clustering in machine learning world is an unsupervised approach of grouping a set of entities together so that the entities in one group are more similar to each other than to the entities in another group. Unsupervised learning is applied while there is input data, but there is no corresponding output variables associated with it. Its goal is to understand and model the underlying distribution of data so as to learn more about it. Clustering has various applications like market segmentation for targeted advertisements and promotional offers, grouping of web contents in a search engines, text summarization, biological applications, astronomy, etc. Clustering reveals natural and meaningful groups among available data. Clustering algorithms aims to achieve highest intra-cluster similarity and least inter-cluster similarity. The concept of distance measure is used to calculate the similarity between objects. When the distance measure between two entities is very less, they are considered as similar. Based on the data under consideration appropriate distance measure can be chosen for clustering. A few of the most common distance measures include Euclidean, Manhattan, Cosine, Jaccard and Minkowski distances.

Clustering Algorithms can be generally categorized into three groups – partitioning [4], hierarchical and density based clustering. Partitioning clustering is used to categorize observations within a dataset based on their similarity. In this approach, the user has to identify the optimal count of clusters for the dataset in consideration and it need to be mentioned to the algorithm. The common partitioning clustering algorithms are k-means clustering [5][6], k- medoids clustering which is also known as Partitioning Around Medoids (PAM) [7][8], Clustering for Large Applications (CLARA) [8][9][10] .

III MATERIALS AND METHODS

In this section presents the materials and methods of this research work. Reviews on destinations in 10 categories mentioned across East Asia. and average rating is used. This data set is populated by capturing user ratings from Google reviews. In this research work has implemented in Weka3.8.3. version.

Dataset Description

Reviews on attractions from 24 categories across Europe are considered. Google user rating ranges from 1 to 5 and average user rating per category is calculated. Each traveler rating is mapped as Excellent(4), Very Good(3), Average(2), Poor(1), and Terrible(0)

Figure 1: Visualization of attributes

NaiveBayesMultinomialText: Multinomial naive bayes for text data.

MultiScheme: Class for selecting a classifier from among several using cross validation on the training data or the performance on the training data.

ZeroR: Class for building and using a 0-R classifier.

SGDText: Implements stochastic gradient descent for learning a linear binary class SVM or binary class logistic regression on text data.

IV RESULTS AND DISCUSSIONS

In this section focuses the results and discussions of this research work. The multi class has categorized in three namely food has 19%, health has 14% and entertainment has 67%.

Table 1: List of Classifiers

S.No	Category	Classifier	Accuracy	ROC	Time Taken to Build the model(In Seconds)
1	Bayes	NaiveBayesMultinomialText	87.39 %	0.64	0.07
2	Meta	MultiScheme	87.45%	0.50	0.06
3	Rules	ZeroR	87.45%	0.50	0.04
4	Function	SGDText	87.30%	0.50	8.92

This above table represents the various classifiers like NaiveBayesMultinomialText classifier belongs to Bayes, MultiScheme classifier belongs to Meta, ZeroR classifier from Rules and SGDText classifier from Function applied in this research work.

NaiveBayesMultinomialText classifier has 87.39% accuracy , MultiScheme classifer has 87.45% accuracy level, ZeroR classifier has 87.45% accuracy level, and SGDText classifier has 87.30% accuracy level.

Visualize Threshold Curve (ROC) values are NaiveBayesMultinomialText classifier has 0.64 accuracy , MultiScheme classifer has 0.50 accuracy, ZeroR classifier has 0.50 accuracy and SGDText classifier has 0.50 accuracy level.

These classifiers have taken time to build the models of NaiveBayesMultinomialText classifier has 0.07secondsMultischeme classifier has 0.06 seconds, ZeroR classifier has 0.04 seconds, SGDText has 8.92 seconds.

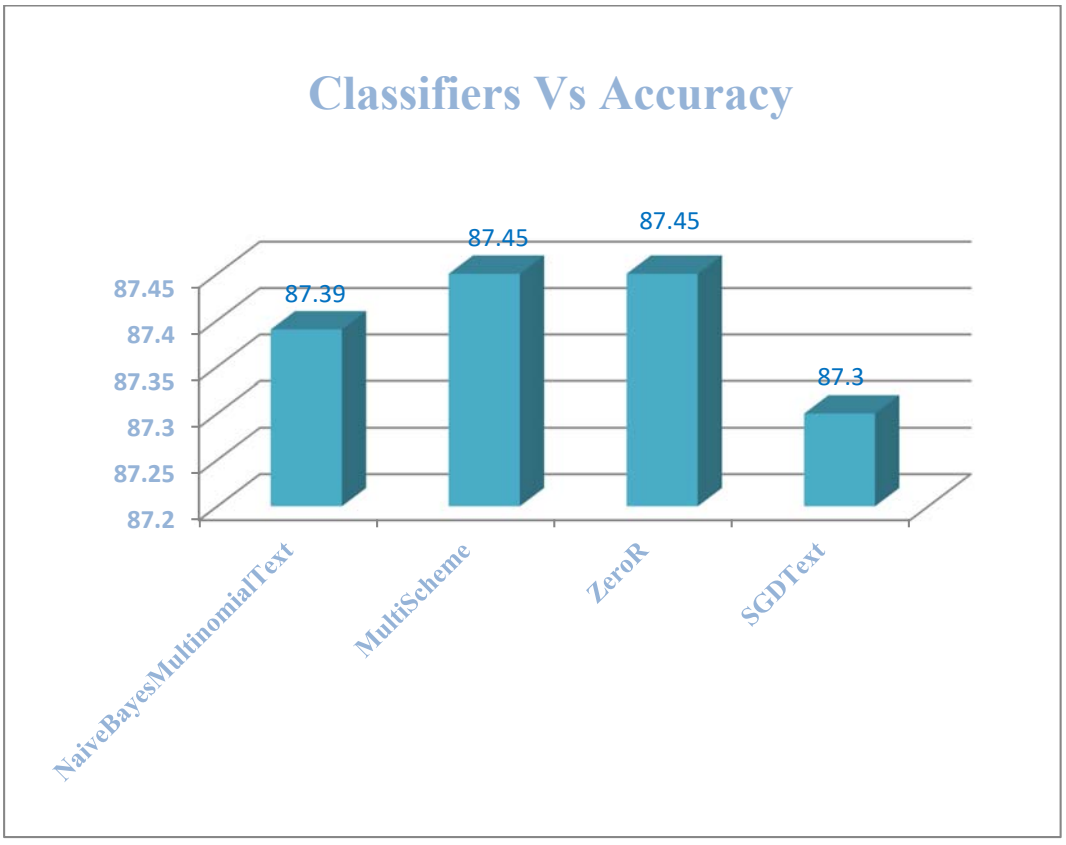

Figure 2: Graphical Representation of Various Classifiers and their accuracies

The above diagram clearly represents all the classifiers have above 87% accuracy. the NaiveBayesMultinomialText has 87.39% level of accuracy and SGDText has 87.3 accuracy level. The Multischeme and ZeroR have same accuracy which as 87.45% and highest accuracies.

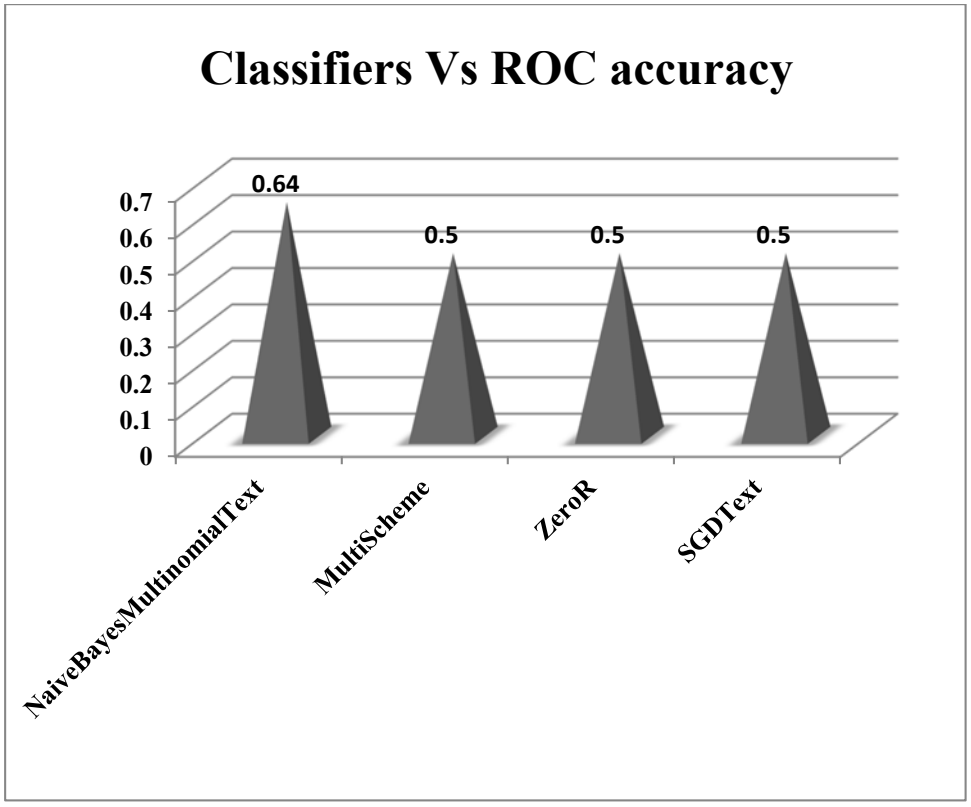

Figure 3: Graphical Representation of Various Classifiers and Their ROC Values

The above diagram represents the Various classifiers and Their ROC Values like Visualize threshold Curve values Multischeme ZeroR and SGDText classifiers have same ROC Value.i.e.0.50 accuracy level. The NaiveBayesMultinomialText has 0.64 accuracy. NaiveBayesMultinomialText has highest ROC values compare with other classifiers.

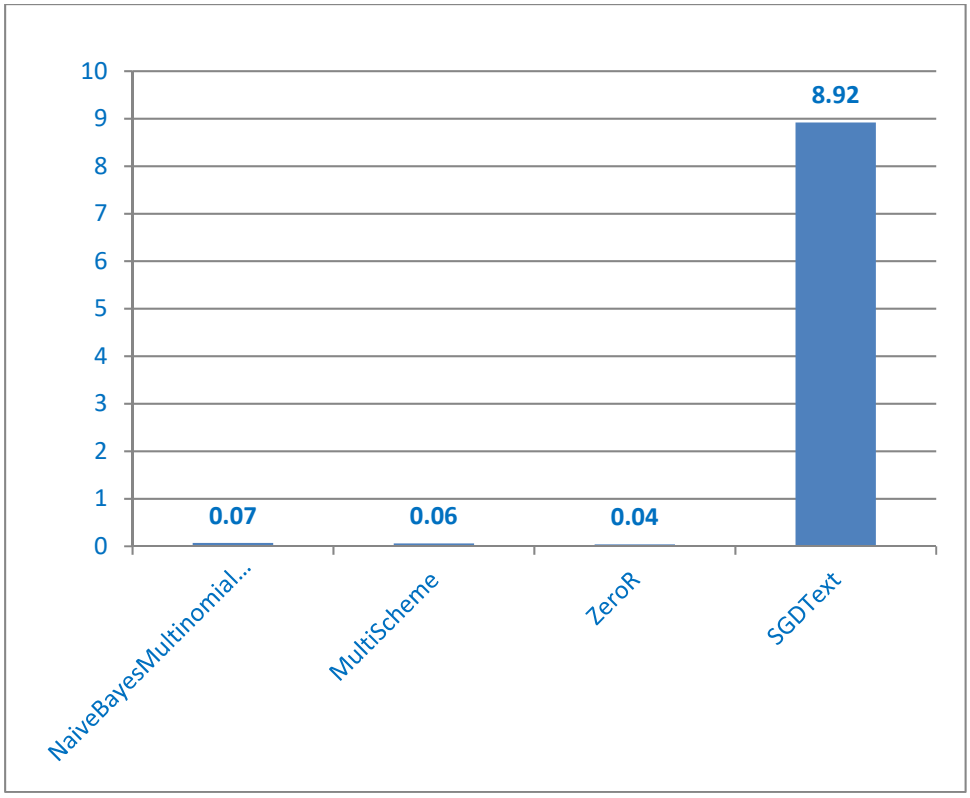

Figure 4 : Graphical Representation of Various Classifiers and Their ROC Values

The above diagram represents the time consumption of each model. NaiveBayesMultinomialText has taken to build the model has 0.07seconds and Multischeme has 0.06 seconds. ZeroR has the time to build the model 0.04 second. But SGDText has more time to taken build the model.It has taken the time 0.892 Seconds.

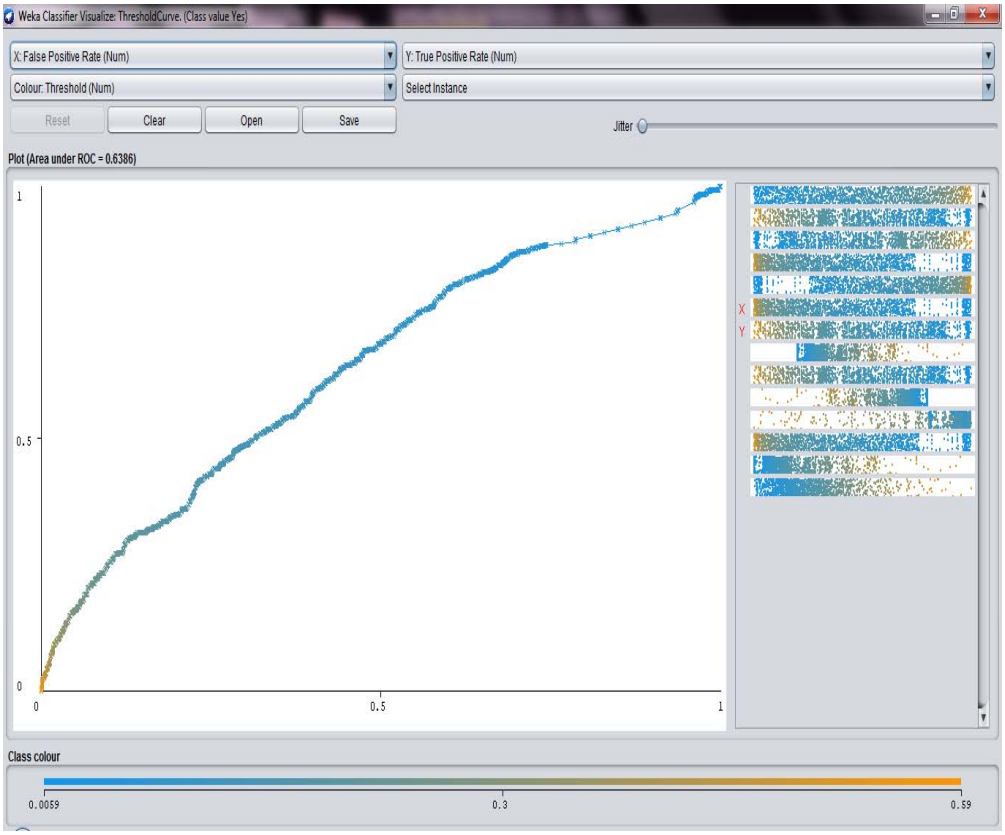

Figure 5: Representation of ROC for NaiveBayesMultinomialText Classifier

The above diagram represents ROC of NaiveBayes NaiveBayesMultinomialText Classifier .It has 0.6396 accuracy level.

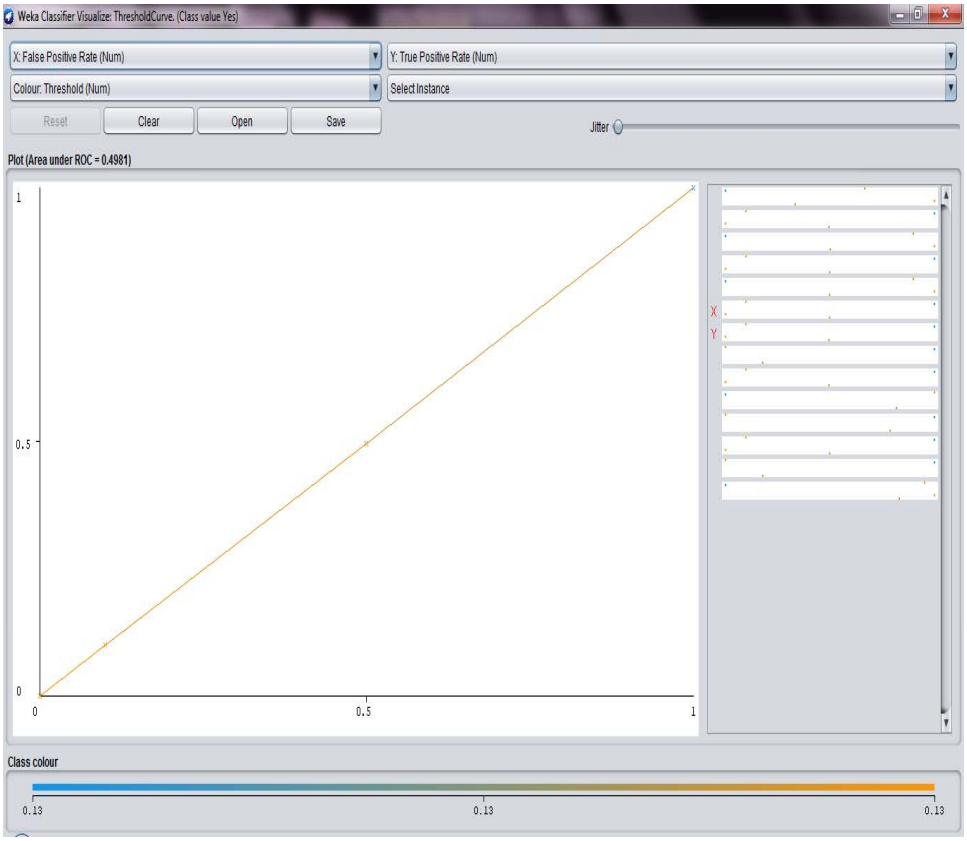

Figure 6: Representation of ROC for MultiScheme Classifier

The above diagram represents ROC of MultiScheme Classifier .It has 0.4981 accuracy level.

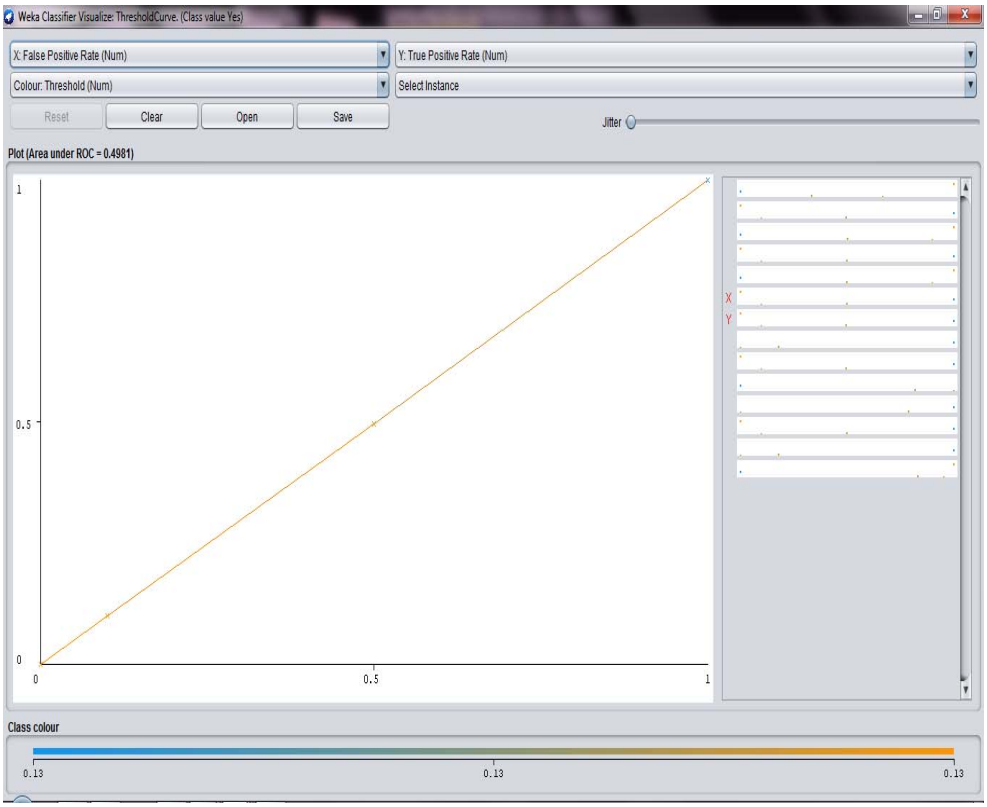

Figure 7: Representation of ROC for ZeroR Classifier

The above diagram represents ROC of ZeroR Classifier .It has 0.4981 accuracy level.

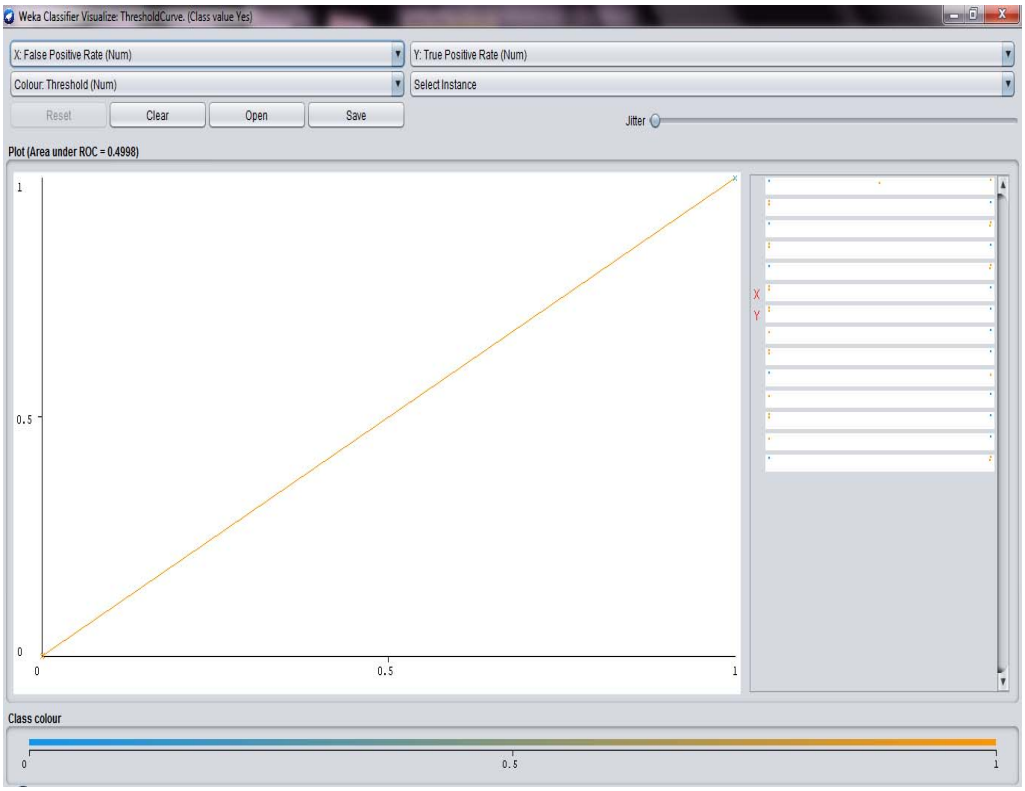

Figure 8: Representation of ROC for SGDText Classifier

The above diagram represents ROC of SGDText Classifier .It has 0.4998 accuracy level.

V CONCLUSION

Finally, this work concludes that based on the data volume and data distribution pattern in consideration, they can adopt appropriate clustering algorithms to segment their customer base so that targeted marketing strategy and/or travel solutions can be offered. These research work Multischeme and ZeroR belongs to Meta and Rules recommended for build the model based on the computation of various classifications and ROC results and time taken to build the model.

REFERENCES

[1] Renjith, Shini, A. Sreekumar, and M. Jathavedan. 2018. “Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domainâ€ . In 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 127” 31. IEEE.
[2] Renjith, Shini, and C. Anjali. "A personalized mobile travel recommender system using hybrid algorithm." In Computational Systems and Communications (ICCSC), 2014 First International Conference on, pp. 12-17. IEEE, 2014.
[3] Renjith, Shini, and C. Anjali. "A personalized travel recommender model based on content-based prediction and collaborative recommendation." International Journal of Computer Science and Mobile Computing, ICMIC13 (2013): 66-73.
[4] Estivill-Castro, Vladimir. "Why so many clustering algorithms: a position paper." ACM SIGKDD explorations newsletter 4, no. 1 (2002): 65-75.
[5] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297. 1967.
[6] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A kmeans clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, no. 1 (1979): 100-108.
[7] Kaufman, Leonard, and Peter Rousseeuw. Clustering by means of medoids. North-Holland, 1987.
[8] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.
[9] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert systems with applications 36, no. 2 (2009): 3336-3341.
[10] Wei, Chih-Ping, Yen-Hsien Lee, and Che-Ming Hsu. "Empirical comparison of fast clustering algorithms for large data sets." In System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, pp. 10-pp. IEEE, 2000.