International Journal of Computational Science and Engineering.
ISSN 2249-4251 Volume 10, Number 1 (2020), pp. 25-34

http://www.ripublication.com

Opinion Mining on Restaurant Rating Based on Aspects

Mr. KAVIN PRAKASH.M
(UG Student)
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.

Mr. ARAVINTH.S
(UG Student)
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.

Ms. HARI NESHA. D
(UG Student)
Vellalar Engineering College
Erode, Tamil Nadu, India.

Ms .MONICA.M
(PG Student)
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.

ABSTRACT

All-most the users rating about a restaurant is shared on the internet about food, service, environment and so on, to show their humanity. Nowadays opinions are expressed through the rating are increased day-by-day on the web. There is a large number of reviews and ratings are available on different aspects, to analyze and extract these ratings and reviews manually is difficult. To solve this problem some technique is needed. Opinion mining or sentiment analysis is such a technique. Opinion mining can extract the polarity of positive, negative. By examining these reviews and rating the positive, negative polarity can be found. In our work, we have developed an overall process of restaurant rating based opinion mining using the Decision Tree Algorithm. To improve the accuracy and finding positive and negative ratings about a restaurant.

Keywords: component; opinion mining, sentimental analysis, Decision Tree Algorithm

I. INTRODUCTION

People bring their attention to know the opinion of the customer about a restaurant which they have visited. Customer review helps to know the information about a restaurant in crystal clear. Now a day's social media has become one of the sharing resources to express everything in a smart way, people use the internet to access and update every happy and sad moment to express their opinion as either positive or negative review about something. These opinions are valuable to make a Decision–making the process, Social media has become a repository which stores all kinds of reviews about a product or service with a different opinion. According to restaurant domain, we focus on the entities of "food", "service", and "environment" and analyzed those opinions which are given by customer about a restaurant. Hence people cannot read billions of opinions about a restaurant manually and it is difficult to extract the important ideas from them. Data Mining [1] techniques provide promising solutions to resolve the aforementioned issues.

Sentimental analysis is also known as Opinion Mining, where opinion mining is one type of web repository for mining the opinion of the people. Web repository (opinion mining) aims at extracting useful information from online reviews by the customer. Opinion mining is defined as the application of data mining techniques to discover patterns from the Web and apply

Natural Language Processing (NLP) to track the opinion of the public about a particular product or a service.

Customer rating about a restaurant plays an important role in the process of decision making. When the customer decides a restaurant, the most important aspect that they consider is "Type of food the restaurant serves", " the quality of food". The online rating about a restaurant is done on the food items, service of the barer, events, cost of the food and environment of the restaurant and so on. The Aspect-level sentimental analysis performed for food items depends on the ability to identify the food names appearing in the review and rating of those food items.

These customer reviews will help us to analyze and can give a polarity value to find out the best restaurants. A most important feature of sentimental analysis where the user's interests can be extracted. It determines the polarity of positive and negative or the emotional attitude ratings towards any interaction. A happy customer thought can be judged by a sentimental analysis and also the angry customer thought. A sentiment analysis system can help you immediately to identify these kinds of situations and take action.

II. RELATED WORK

Opinion mining can be performed in the following levels [2]: "Document-level", "Sentence level" and "Aspect level". Document-level opinion mining categorizes the overall opinion polarity of the document as positive or negative [3]. The document level approach helps the users in decision making by providing a summary of the total number of positive and negative documents.

Sentence level opinion mining analyses every sentence of the document and categorizes the sentence in to positive or negative. It improves the fine-grained of extraction by extracting comments from sentence-level and makes feature selection based on word frequency statistics. Sentence level opinion mining allows the learner to maintain word frequency statistics when mining the opinion. As a result, it becomes easier to do feature construction and modeling, in contrast, to document level opinion mining. [4].

Aspect level opinion mining classifies customer reviews based on important features. In the aspect level of opinion mining, there are two parts. The first part is the opinion of word identification. The second part is orientation detection. For example, a restaurant review can be "Environment is bad, but foods are delicious". Here the review comments are on two aspects. One aspect is the "environment" and the other is "food". There are two opinion words in this example. They are "bad" and "delicious". Orientation detection is finding whether the opinion word is positive or negative to the aspects. In the above example, the environment has a negative opinion and the food has a positive opinion. SentiWordNet [5] is a lexical dictionary which helps to find the positive and negative polarity value of each opinion word.

An early model, incremental induction decision tree [6], reconstructs a decision tree by determining a feasible split after each incoming data instance arrives. The downside of this approach is that it is possible to produce an unstable tree in some rare cases when the splitting feature may be shuffled repeatedly as a result of incoming data. Furthermore, a single decision tree has been known to be outperformed by a forest of decision trees (an ensemble model) that uses consensus opinion.

QiweiGan and Yang Yu [7] have presented three factors through which a restaurant is rated. Those three factors are food quality, ambiance, and service of the restaurant. They have also considered two other factors like "cost" and "special context" in it as it also plays an equally important role. They have collected 268,442 feedbacks of 7,508 restaurants from Yelp.com, which a digitized site for a word of mouth.

S. Prakash et al.[8] have suggested that in spite of so many factors are there for rating a restaurant, food type and cost of the food is more important to rate a restaurant. According to this paper, a customer will not look for the ranking of a particular food item but for the classification of the particular food item category.

III. PROPOSED WORK

A. Data Collections

The proposed system is tested with the data's which are collected from the kaggle datasets. The kaggle datasets contain plenty of data and information about the various different applications, wherein these kaggle datasets we focus on a dataset based on restaurant rating. Where these records are collected and made use of it.

	A	B	C	D	E	F	G	H	I	J
1	userID	restaurantID	hours	restaurant_name	address	country	rating	food	service	Class
2	U1077	135085	00:00-23:	Kiku Cuernavaca	Revolucion	Mexico	2	2	2	1
3	U1077	135038	00:00-23:	puesto de tacos	esquina santos degollado y leon g	mexico	2	2	1	0
4	U1077	132825	00:00-23:	El Rincïn de San F	Universidad 169	Mexico	2	2	2	1
5	U1068	135104	00:00-00:	carnitas_mata	lic. Emilio portes gil	Mexico	1	1	2	0
6	U1068	132740	00:00-00:	Restaurant los Cor	Camino a Simon Diaz 155 Centrc	Mexico	0	0	0	0
7	U1068	132663	08:00-21:	Taqueria EL amigc	Calle Mezquite Fracc Framboyan	Mexico	1	1	1	0
8	U1068	132630	08:00-21:	Pollo_Frito_Buen	tampico	Mexico	1	1	1	0
9	U1067	132584	00:00-23:	la Estrella de Dim	Villa de Pozos 192 Villa de Pozo	Mexico	2	2	2	1
10	U1067	132733	00:00-23:	Restaurante 75	Villa de Pozos 4497 Villa de Poz	Mexico	1	1	1	0
11	U1067	132732	00:00-23:	Abondance Restau	Industrias 908 Valle Dorado	Mexico	1	2	2	0
12	U1067	132630	07:00-23:	El angel Restauran	Venustiano Carranza 1625 Jardin	Mexico	1	0	1	0
13	U1067	135104	07:00-23:	Restaurante Puebl	Mexico 2015 Providencia	Mexico	0	0	0	0
14	U1067	132560	07:00-23:	Mcdonalds Parque	Lateral Salvador Nava Martinez 3	Mexico	1	0	0	0
15	U1103	132584	18:00-23:	Tortas y hamburgu	Ricardo B. Anaya	Mexico	1	2	1	0
16	U1103	132732	18:00-23:	Sirlone	carr. mexico	Mexico	0	0	2	0
17	U1103	132630	18:00-21:	rockabilly	agustin de iturbide	mexico	1	2	0	0
18	U1103	132613	00:00-23:	Unicols Pizza	Plaza del Carmen	Mexico	2	2	2	1
19	U1103	135104	00:00-23:	Restaurant El Mul	De Guadalupe 460 San Miguelito	Mexico	1	2	0	0
20	U1103	132663	00:00-23:	La Posada del Virr	Av. V. Carranza	Mexico	1	0	2	0
21	U1103	132733	00:00-23:	Restaurant and Bar	Domingo 10 711 El Empleado	Mexico	2	2	2	1

Figure 1. Restaurant dataset

To trust the information given by the customer about a restaurant, the data consideration is around 1,162 data, from this collection 70% is for training and 30% is for testing the data is done. The restaurant in several categories like Café, fast-food, restaurant, barbeque, star-hotels, lodge, and etc… The rating about a restaurant is getting extracted from the dataset and the testing of the proposed system is performed in it.

B. Data Preprocessing

As the dataset from the kaggle.com, where the data is in the form of numeric values where there are no empty values or NA is there. So the data which is collected for our work is cleaned and a relevant ambiance is provided, from the trained data we can make a decision-making process.

C. Algorithm For Decision Tree

- Create the root node for a tree
- If all examples are positive, return leaf node ‘positive’
- Else if all examples are negative, return leaf node ‘negative’

- Calculate the entropy of current state H(S)
- For each attribute, calculate the entropy with respect to the attribute 'X' denoted by H(S, X)
- Select the attribute which has a maximum value of IG(S, X)
- Remove the attribute that offers highest IG from the set of attributes
- Repeat until we run out of all attributes, or the decision tree has all leaf nodes

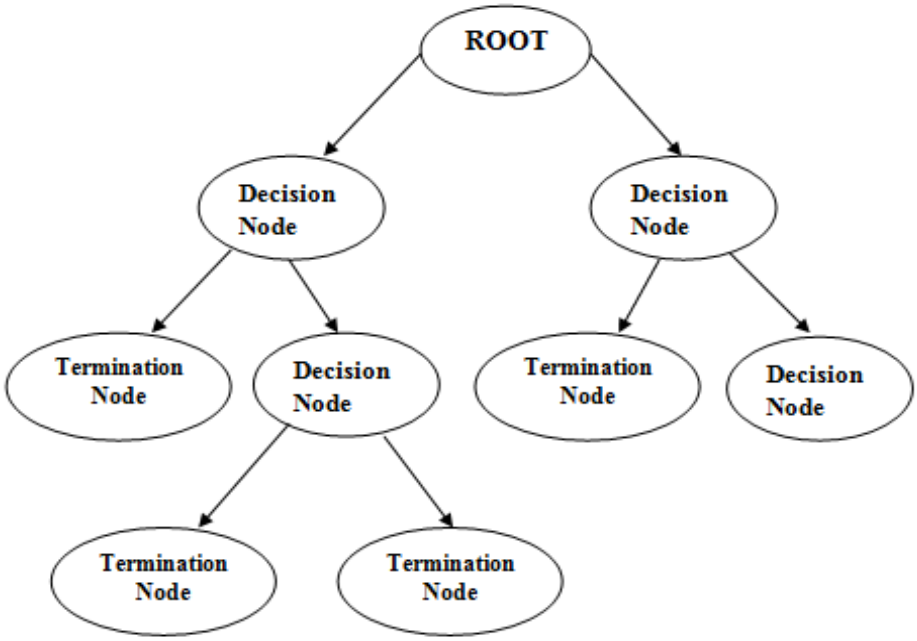

Figure 2. General Decision Tree

Common terms used in the decision tree:

1. ROOT NODE: It represents the entire population or sample and their further gets divided into two or more homogeneous sets.
2. SPLITTING: It is a process of dividing a node into two or more sub-nodes.
3. DECISION NODE: When a sub-node splits into further sub-nodes, then it is called a decision tree.
4. LEAF/TERMINAL NODE: Nodes do not split is called leaf/terminal node
5. PRUNING: When we remove a sub-nodes of a decision node, this process is called pruning. (this can say the opposite process of splitting)
6. BRANCH/SUB-TREE: A subsection of the entire tree is called branch or sub-tree
7. PARENT AND CHILD NODE: A node that is divided into sub-nodes is called the parent node of sub-nodes whereas sub-nodes are the child of a parent node.

D. Attribute Selection Measures

In Decision Tree, the major challenge is to the identification of the attribute for the root node at each level. This process is known as attribute selection. We have two popular attribute selection measures such as Entropy, Information Gain and Gini Index**.**

Entropy is a measure of randomness in the information being processed, in another term it can be mentioned as a disorder. Entropy is measured between 0 and 1 (depend on the number of classes in the dataset) Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information

ENTROPY = - {Summation of (fraction of each class. Log base2 of that fraction)}

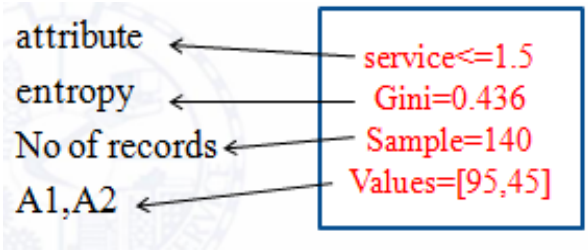

Figure 3. Entropy Calculation

Information Gain is when we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.

INFORMATION GAIN = entropy (parent) – [weighted average] * entropy (child)

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with a lower Gini index should be preferred. Sk-learn supports the “Gini” criteria for Gini Index and by default, it takes the “Gini” value.

GINI INDEX = Parent entropy- child entropy with a weighted average

Child entropy with weighted Average= [no of example in left child node) /total no of example in Parent node)*(entropy of left node)] + [no of example in right child node) /total no of example in Parent node)*(entropy of right node)]

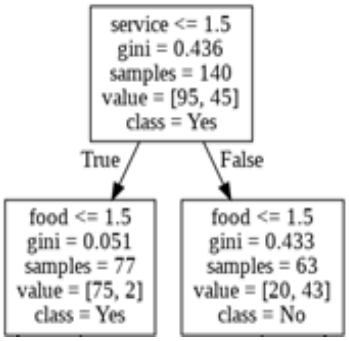

Figure 4. Gini calculation

E. Decision Tree For Restaurant Rating

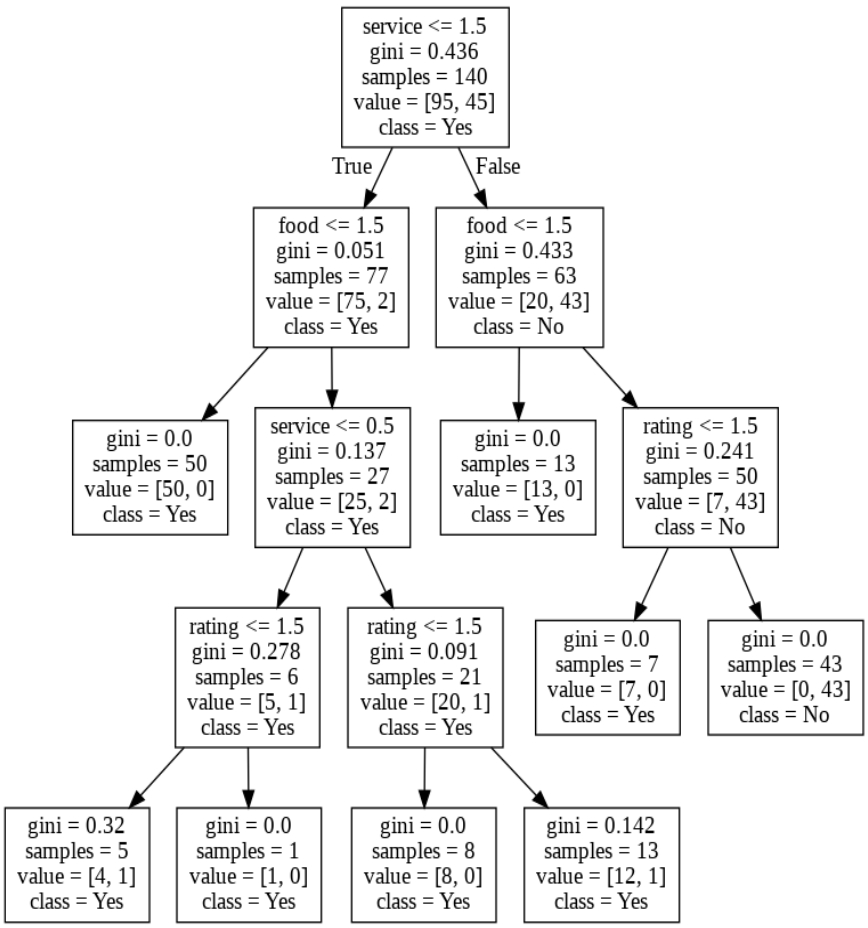

Figure 5. Decision Tree for Restaurant Rating

F. Measurement Factors

The metrics that are chosen to evaluate your machine learning model are very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. some of the measurement factors are Accuracy, Precision, Recall.

Accuracy which brings up to the closeness of estimated value with standard or known value. Confusion matrix accuracy is calculated and the result is analyzed. Accuracy in classification problems is the number of correct predictions made by the model over all kinds of predictions made.

ACCURACY = TP + TN / TP + TN + FP + FN

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative

PRECISION = TP/TP+FP

Recall helps when the cost of false negative is high. The recall is defined as the number of true positives divided by the number of true positives plus the number of a false negatives.

RECALL = TP/TP+FN

```
[ ]  from sklearn.metrics import confusion_matrix
     from sklearn.metrics import classification_report
     confusion_matrix(y_test, y_pred)
     print("Report : ",classification_report(y_test, y_pred))
```

```
Report :               precision    recall  f1-score   support

         0.0       1.00      0.98      0.99        49
         1.0       0.92      1.00      0.96        11

    accuracy                           0.98        60
   macro avg       0.96      0.99      0.97        60
weighted avg       0.98      0.98      0.98        60
```

Figure 6. Overall Structure

IV. RESULT & DISCUSSION

A model is trained to determine the opinion about a restaurant with relatively high accuracy using the kaggle restaurant rating dataset. The system can handle nearly thousands of users' ratings and review posted online. The result from this system shows the relationship between the customer and the three aspects such as (food_rating, service_rating, and environment_rating).

The outcome of the system is the rating of the restaurant based on the opinion of the customer. Where the positive and negative polarity scores are found and collected. Where the sum of these scores provides 98% accuracy for the given aspects.

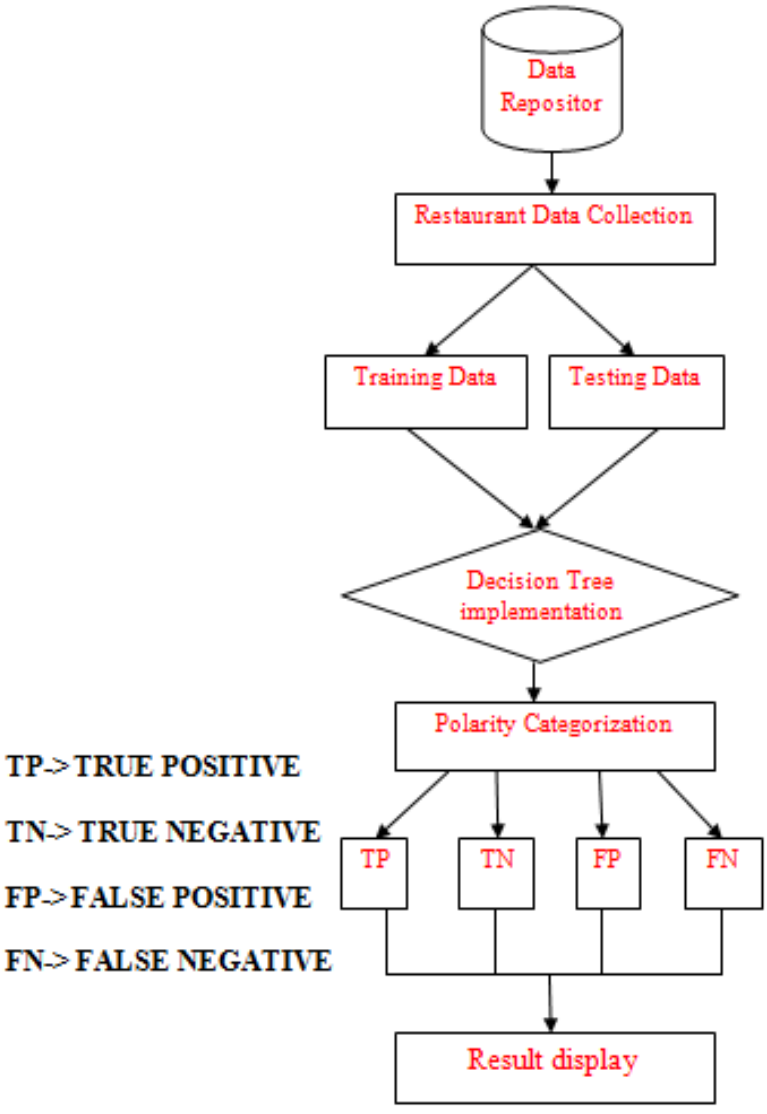

Figure 7. Result for Restaurant Rating

V. CONCLUSION AND FUTURE WORK

Opinion mining or Sentimental Analysis is a technique that determines whether the opinion of a customer is positive or negative based on the writing. The system can withstand huge volumes of customer ratings about a restaurant and can provide better accuracy. This study and work are based on the data extracted from the kaggle restaurant rating dataset.

This paper could be further studied for improvements. where the decision tree algorithm over feting occurs so testing with some other algorithm such as Random Forest can provide better accuracy. By considering some other aspects like restaurant locality, season, price, etc.. The proposed system can also be analyzed with classifiers for better results.

REFERENCES

[1] Gary M. Weiss, Brian D. Davison, Data Mining : Handbook of Technology Management, 2010.

[2] B. Liu, "Sentiment Analysis and Opinion Mining," in Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[3] Richa Sharma,Shweta Nigam, Rekha Jain, "OPINION MINING OF MOVIE REVIEWS AT DOCUMENT LEVEL," in International Journal on Information Theory (IJIT), Vol.3, No.3, India, 2014

[4] Hongting Li, Qinke Peng, Xinyu Guan, "Sentence Level Opinion Mining of Hotel Comm," in IEEE International Conference on Information and Automation, Ningbo, China, 2016.

[5] "SentiWordNet," SentiWordNet, 2010. [Online]. Available: http://sentiwordnet.isti.cnr.it/. [Accessed 09 March 2017].

[6] P. E. Utgoff, "Incremental induction of decision trees," Machine Learning, vol. 4, no. 2, pp. 161–186, 1989.

[7] QiweiGan and Yang Yu "Restaurant Rating:Industrial Standard and Word-of-Mouth A Text Mining and Multi-dimensional Sentimental Analysis" 2015 IEEE 48th Hawaii International Conference on System Science

[8] S. Prakash, A. Nazick, R. Panchendrarajan, A. Pemasiri , M. Brunthavan, and S. Ranathunga " Categorizing Food Names in Restaurant Reviews" 2016 IEEE, pp:1-5