# TOURVIEW: DETERMINING RATING OF TOURIST PLACES IN INDIA BASED ON PUBLIC OPINIONS

Md. Toukir Ahmed[1]*, Md. Niaz Imtiaz[1], Toyoba Islam Pobitra[1]
[1] Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh
*Correspondence: toukirahmedreal@gmail.com

**Abstract**— In modern world, people are becoming more sincere and work oriented to lead a successful life. But, maintaining daily routine repeteadly can be monotonous sometimes. Thus, people often tour to different places to remove their monotony and to have recess for few days. The practice of tourism is rapidly increasing nowadays. For visiting different places, tourists seek information about those particular places in advance for assurity. Thus, analyzing opinions and reviews of a particular tourist place helps tourists in decision making and better understanding about the place. In this work, a tourism information analysis system is proposed for the places of different states located in India analyzing the polarity of the positive or negative reviews. Based on polarity score on public opinion, the best and relatively low rated places are analyzed. Moreover, the accuracy of the data that is also predicted using some machine learning algorithms on tourists' online reviews acquired from TripAdvisor.

**Index Terms**— Machine Learning , Polarity Score, Sentiment Analysis, Tourism, Tour Review, Web Scraping.



## 1 INTRODUCTION

Tourism is important for those who are fond of travelling. tourists of different country visit tourist places every year and share their experience on various travel websites such as TripAdvisor [1, 2]. However, there are a big amount of opinions available on each particular place and which is very tough for a normal traveller to read these opinions and make decision on whether to visit a place or not. In Internet there are many information which contain tourist reviews and opinions about tourist destinations. The information of tourism is often presented as comments expressed in natural language that explain customer opinions about various tourist places. In this paper, we propose for the tourism information analysis system. First, we explain a basic idea for the data scraping process, data processing. Here, we talk about sentiment analysis. The sentiment analysis is one of the important topics in natural language processing. In this paper, we apply a combined machine learning approach [3]. Here we handle Vader sentiment analyzer to get polarity score of text. Then we calculate rating based on the scores so that it shows the place with top most review about a particular place [4]. We also established a model and apply machine learning algorithm to know the accuracy rate that perform on our real dataset.

## 2 Literature Review

### 2.1 Natural Language Processing

Natural Language Processing is a field of Artificial intelligence. It analysis of the written language of a user. If we have a lot of data written in plain text and for analyzing those text we need to use NLP [4]. The significance of sentiment in putting opinions of users expressed via social media is the most [5]. The emotions that are extremely useful, natural language processing focused on it for sentiment analysis. NLP is used in speech analyzing. Speech analysis is become easier by NLP. The application of NLP to indentify sentiment is speech analysis [6].

### 2.2 Machine Learning

Machine Learning is an area of Artificial Intelligence [4].Machine Learning is the field of study that gives computers capability to learn without being explicitly programmed let the machine adapt to the user using the data. It is not actually an Artificial Intelligence field in itself, but a way to

solve the real problems of Artificial Intelligence.Today Machine Learning is used in different field.for example that is NLP.In order to apply Machine Learning techniques to NLP problems, we need to convert the unlabled text into labled format.This approach, employes a machine learning technique to build a classifier that identify the text that expresses sentiment.

### 2.3 Deep Learning

Deep Learning is an extension of Neural Networks and sub sector of machine learning language.it is the algorithm inspired by structure and function of human behavior.it handle huge amount of training and testing data. Deep Learning is quite used for vision based classification [7]. Deep Learning is used for NLP tasks as well. The methods of deep learning fit on data learning representation which make it popular.

### 2.4 Sentiment Analysis

Now-a-days Sentiment Analysis and Opinion Mining is currently an active research area [3] .it is also widely known as opinion mining, is defined as the domain of research that evaluates public sentiments. Sentiment analysis detect polarity of a text whether it's a whole document, paragraph, sentence, or clause. it automatically analyze customer feedback, from social media conversation [4].



## 3 MODEL,DESIGN,ANALYSIS AND IMPLEMENTATION

### Step-1: Web Scraping

Web scraping is a method that is used to extract large amounts of data from websites [8]. The data on the websites are unlabled. Web scraping help to collect these unlabled data and store it in a labled form. There are different ways to scrape websites such as online Services, APIs. For our thesis work first we need a real dataset.for this purpose we have to scrape the reviews of the tourists given to TripAdvisor website. We scrape data over on India's Tourist place. We use visual studio code and written python code for scraping our dataset. We need scrapy package for this purpose.

**TABLE 1**

Dataset of Public Opinion about Places

| State | Place | Review | Reviewer |
|---|---|---|---|
| Assam | Agnigarh Hill | Park over the hill with go… | BabonBasu |
| Gujrat | Ahmedabad | a soulful sight | saurabhpraTimk |
| Himachal | Kangra Valley | Beautiful place | Goldygsp84 |
| Kashmir | Pahalgam | Can't enjoy full beauty during | RaamkumarL |
| Kerala | Kumarakom | No doubt about that the best | Saumitra |
| …… | ……. | ……. | ……. |

### Step-2: Pre-processing the Datasets

First, we are going to remove all the non character words, all the punctuation marks, all the ASCII symbols. Then, we convert the review to lower form. Then remove all tha single characters and extra spaces that are generated.

### Step-3: Calculating Mean Value and Finding Ratings

Statistics functions are part of the Python Standard Library in the statistics module. To access Python's statistics functions,we need to import the functions from the statistics module.Then we calculate the mean value for each place based on the compound score that we get before.After finding mean we need to find

the ratings for each place.We implement loop and conditions in python for finding the rating.

### Step-4: Creating The model

For creating BOW model we need to import CountVectorizer class. Transform our BOW model into Tf-idf model using the TfidfTransformer class.

### Step-5: Creating Training Dataset and Test Dataset

We have to split our dataset into training dataset and testing dataset.We have 2000 different reviews.Out of these 2000 documents we want to use some of these documents for training the whole model and the rest of the documents for testing whether it build a proper model or not.We use 1600 documents for training the model and 400 documents for testing the model.

### Step-6: Fitting the Model into Logistic Regression Algorithm

We are going to train our model using the text train, the sent train that we have created after splitting the whole dataset and we are going to use logistic regression to create the whole classifier.

First, we need to import logistic regression class from sklearn. Then we fit our text_train and sent_train data into a variable to get our classifier.

### Step-7: Fitting the Model into SVM Algorithm

We are going to train our model using the text train,the sent train that we have created after splitting the whole dataset and we are going to use svm to create the whole classifier.First we need to import svm class from sklearn.then we fit our text_train and sent_train data into a variable to get our classifier.

### Step-8: Fitting the Model into Random Forest Algorithm

We are going to train our model using the text train,the sent train that we have created after splitting the whole dataset and we are going to use random forest algorithm to create the whole classifier.First we need to import random forest class from sklearn.then we fit our text_train and sent_train data into a variable to get our classifier.

## 4 RESULT AND DISCUSSION

### Bar Chart of Tourist Spots

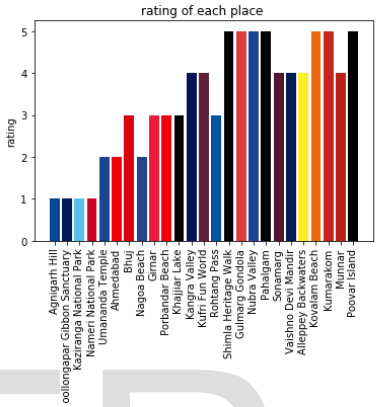


Fig. 1.Bar Chart of Tourist Spots

Here the tourist places that we analyzed is showing in a bar chart.where in the x-axis there is the name of the places and in y-axis there is ratings (starts from 0 to 5).the bar chart shows that the places named “Poovar Island”,”Kumarakom”,”Kovalam Beach”,”Nubra valley”,”Gulmarg Gondola” and “Shimla heritage” have the highest rating(rating with 5) among all the 25 tourist spot.and the places named “Nameri National Park”,”Kaziranga National park”,”Gibbon Santuary”,”Agrinath Hill” have the lowest rating(rating with 1).

### Bar Chart of Tourist States

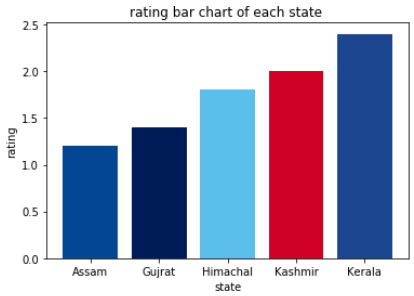


Fig. 2.Bar Chart of Tourist States

The above figure shows the bar chart for the states of India.here we saw that the state named “kerala” having the highest ratings among the all of five states.so we can say that kerala is most popular state for tourism.On the other side we see that the state named “Assam” having the lowest rating among all of them.so we say that this state is comparatively less popular state for tourism.

### Logistic Regression Algorithm

In previous chap we have used logistic regression class and we have created and object and trained our model.now we have our model so let’s find out how efficient or accurate our model performance is.

### TABLE 2

Confusion matrix of logistic regression algorithm

| | 0 | 1 |
|---|---|---|
| 0 | 143 | 14 |
| 1 | 37 | 206 |

The above table is the confusion matrix of logistic regression algorithm. Here for the columns we have the actual values and for the rows we he the predicted value

Here 143 predictions are really negative and our model also predict it negative and 37 predictions are really negative bt our model predict it as positive.so here this is the error and 143 is the accurate result.for positive results 14 of predictions are actually positive bt our model classifies it as negative.similarly 206 predictions are actually positive and our model also classify it as positive.Here we have 349 accurate prediction out of 400.and we got 87.25% accuracy.

### Precision-recall Curve of Logistic Regression Algorithm

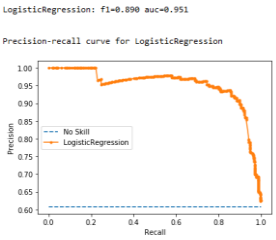


Fig.3. Precision-recall curve for logistic regression algorithm

A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds . The result is a value between 0.0 for no precision and recall and 1.0 for full or perfect precision and recall.

### ROC-curve of Logistic Regression Algorithm

The ROC curve shows the tradeoff between the true positive rate and the false positive rate of a classifier.

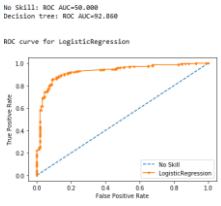


Fig.4. ROC- curve for logistic regression algorithm

The x-axis shows the false positive rate (FPR) from 0 to 1 and the y-axis shows the true positive rate (TPR) from 0 to 1. A perfect classifier would yield a true positive rate of 1

and a false positive rate of 0. In such an ideal case, the ROC curve would be a straight line from (0,0) to (0,1) and a horizontal line from (0,1) to (1,1).

### Support Vector Machine

In previous chap we have used SVM class and we have created and object and trained our model.now we have our model so let's find out how efficient or accurate our model performance is.

**TABLE 3**
Confusion matrix of svm algorithm

| | 0 | 1 |
|---|---|---|
| 0 | 148 | 9 |
| 1 | 35 | 208 |

The above figure is the confusion matrix of svm algorithm. Here for the columns we have the actual values and for the rows we have the predicted values.
Here 148 predictions are really negative and our model also predict it negative and 35 predictions are really negative bt our model predict it as positive.so here this is the error and 143 is the accurate result.for positive results 9 of predictions are actually positive bt our model classifies it as negative.similarly 208 predictions are actually positive and our model also classify it as positive.Here we have 356 accurate prediction out of 400.and we got 89.0% accuracy.

### Precision-recall Curve of Support Vector Machine

A system of high recall and low precision returns many results, but when compared to the training labels most of its predicted labels are incorrect. The system with high precision but low recall is just the opposite, but when compared to the training labels most of its predicted labels are correct. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

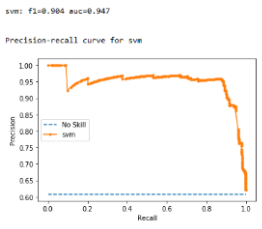


Fig.5. Precision-recall curve for SVM algorithm.

The precision may not decrease with recall.

### ROC-curve of Support Vector Machine

ROC curve features true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. ROC curves are typically used in binary classification.

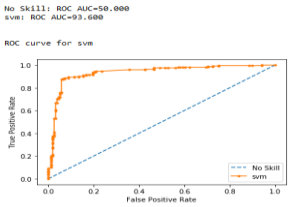


Fig.6. ROC- curve for logistic SVM algorithm

### Random Forest

In previous chap we have used Random Forest class and we have created and object and trained our model.now we have our model so let's find out how efficient or accurate our model performance is.

**TABLE 4**
Confusion matrix of Random Forest algorithm

| | 0 | 1 |
|---|---|---|
| 0 | 142 | 15 |
| 1 | 40 | 203 |

The above figure is the confusion matrix of random forest algorithm. Here for the columns we have the actual values and for the rows we have the predicted values.

Here 142 predictions are really negative and our model also predict it negative and 40 predictions are really negative bt our model predict it as positive.so here this is the error and 143 is the accurate result.for positive results 15 of predictions are actually positive bt our model classifies it as negative.similarly 203 predictions are actually positive and our model also classify it as positive.Here we have 345 accurate prediction out of 400.and we got 86.25% accuracy.

### Precision-recall curve of Random Forest

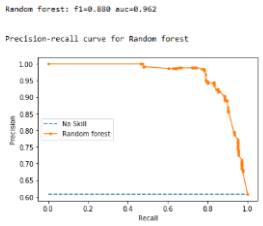


Fig.7. Precision-recall curve for Random Forest algorithm

### ROC-curve of Random Forest

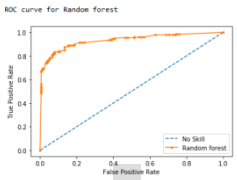


Fig.8. ROC- curve for logistic Random Forest algorithm.

The shape of the ROC curve value close to 90% shows that the performance of our building model is pretty good.

## 5 FUTURE WORK

We presented the analysis sentiment analysis for the textual reviews. In future work we should focus on the implementation of ensemble learning for gaining accurate result. On the short term we plan to strengthen our results by investigating the correlation between other user characteristics like sex and age interval.

## 6 CONCLUSIONS

The interest in sentiment analysis as a field of research that is growing rapidly. In this paper we investigated the application of opinion mining on real data extracted from a travel review site. We presented the results of a sentiment analysis algorithm for the textual reviews. It has been shown that conversion of the huge volume of textual data from the web into meaningful data which can be very useful. However, the task of accurate opinion extraction is very challenging. In our combined approach a lexicon score is being used as a feature in machine learning classification. This innovation allowed to improve the accuracy of predictive model.

## REFERENCES

[1] C. Vásquez, "Narrativity and involvement in online consumer reviews: The case of TripAdvisor.," *Narrative Inquiry,* vol. 22, no. 1, pp. 105-121, 2012.

[2] A. K. P. P. D. Sharma, Tourview: Sentiment Based Analysis on Tourist domain, vol. 6 (3), India: (IJCSIT) International Journal of Computer Science and Information Technologies, 2015, pp. 2318-2320.

[3] S. G. a. M. T. L. A. García, "A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain," *eRTR,* vol. 10, no. ENTER 2012 Idea Exchange , 2012.

[4] M.-F. M. Erik Boiy, Machine Learning Approach for Sentiment in multilingual web texts, Belgium: ResearchGate, october,2009.

[5] C. B. a. A. S. M. Colhon1, "Relating the Opinion Holder and the Review Accuracy," *Springer International Publishing Switzerland ,* no. 2014, p. 246–257.

[6] S. O. S. X. (. L. A. P. Kirilenko1), "Automated Sentiment Analysis in Tourism: Comparison of Approaches," 2017.

[7] S. B. a. B. S. A. R. Alaei, "Sentiment Analysis in Tourism:Capitalizing on Big Data," *SAGE,* pp. 1-17, 2017.

[8] T. T. P. S. C. T. Olga Kolchyna, "Methodology for Twitter Sentiment Analysis," vol. 2, p. 30, jul,2015.

[9] M. U. A. C. M. F. S. F. a. Y. Z. M. Afzaal, ""Fuzzy Aspect Based Opinion Classification System for mining tourist review," *hindawi,* vol. 2016, no. 25 July 2016, p. 14, 2016.

[10] B. Liu, "Sentiment Analysis and Opinion Mining," *ISBN,* 2012.

[11] F. K. a. F. B. A. Sadia, "An Overview of Lexicon-Based Approach For Sentiment Analysis," *IEEC ,* Feb, 2018.

[12] T. T. S. t. A. o. kolchayna, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," vol. 2, July 2015.

[13] M. Thelwall, "Sentiment Analysis for Tourism," *Springer,* p. 18, 2019.

[14] S. I. H. M. a. T. E. K. Shimada, "Analyzing Tourism Information on Twitter for a Local City," *IEEE,* 2011.

[15] A. K. A. D. INKPEN, "SENTIMENT CLASSIFICATION OF MOVIE REVIEWS USING CONTEXTUAL VALENCE SHIFTERS," vol. 22, 2006.

[16] P.-H. C. C.-J. Lin, "Tourism-Related Opinion Detection and Tourist-Attraction Target Identification," vol. 15, March 2010.

[17] J. D.-M. A. l. o. o. panelEdisonMarreseTaylo, "Identifying Customer. Preferences about Tourism Products Using an Aspectbased Opinion Mining Approach," vol. 22, 2013.

[18] p. a. mandira, "sentiment analysis to determie accomodation,shopping and culinary location on Foursquare," *ScienceDirect,* pp. 300-350, 2015.

[19] B. A. Andrea Ballatore, "Extracting Place Emotions from Travel Blogs," *AGILE,* p. 5, 2015.

[20] L. T. Holdings, "tripadvisor," 2000. [Online]. Available: https://www.tripadvisor.com/?fid=46788fc5-cad7-4a3f-b4db-d009c85fd47f.