Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 717- 722

Research Article

Opinion Mining On Rural Tourism In India- Qualitative Perspective

[*1]Garima Verma, [2]Hemraj Verma

[*1]School of Computing, DIT University, Dehradun, INDIA

Mobile: 9410148069, *Email:garimaverma.research@gmail.com

[2]Faculty of Management Studies, DIT University, Dehradun, INDIA

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Tourism sector is one of very important sector in the economy of any country as it provides them a chance to earn foreign exchange at a low social cost. In India, tourism contributed around 9.2% to its GDP in 2018. There is always an effort to increase tourist inflow in a country as there are several related businesses that flourish because of inflow of tourists. Therefore, it becomes critical that tourist expectations are met to their satisfaction and they plan a revisit and/or recommend the same to their friends and family members thus bring in more revenue. To know customer feedbacks, a formal feedback is/can be taken. However, tourists do give write-ups and reviews about a place, hotel, and restaurants that they have visited in a country on travel related websites such as trip advisor, blogs on incredible India etc.. Since these reviews and write ups are generally assumed to be written by a neutral travellers and provided first-hand account of their experiences, therefore, tourists all over the world take these reviews seriously to plan their future travel, stay and decide other preferences. Hotel authorities and other travel related businesses can use these reviews to improvise their services to meet customer expectation better. If they could analyse all the reviews given at different places about their services, then they can possibly provide best services leading to better ratings for their businesses. In this paper the effort has been made to analyse the reviews given by international tourists at various hotels and other travel related businesses about the destinations of rural India. The data about online reviews has been scrapped from various travel related websites and blogs. Further, a supervised machine learning technique has been used to classify the sentiments of the and hence suggest the areas of improvement for future business.

Keywords: Travel blogs, reviews, scrapping, sentiments.

1. Introduction

Tourism is one of the sectors that generally contributes heavily in the growth of economy of a country. It is a low investment, labour demanding industry with economic multiplier and always provides an opportunity to earn foreign exchange at low social cost [1]. The world travel and tourism council (WTTC) evaluated that in 2018, tourism in India generated approximately 9.2% of GDP[1]. It has also provided 43 million jobs approximately [2], which is 8.1% of overall employment. The predictions done by WTTC says that the tourism sector in India will grow at 6.9% by 2028, which will be 9.9% of GDP. According to the Indian Medical tourism sector (IMTS) , there were more than 1.5 lakhs patients from foreign countries who travelled to India for seeking natural health treatments in 2014 [3], [4]. With the advent of new technologies supported by web 2.0, millions of people share their thoughts and memories about different places they travel and explore. These thoughts are present in the form of reviews, blogs or comments on different travel websites. The analysis of these text reviews can provide insights for improvement in the services of hotels and travel agencies, which provides guides, taxies, travel vouchers, etc [3]. The owners of the services can take positive and negative comments from these blogs and reviews and analyse their services and space of improvements. This study aims to explore the tourism reviews of India and perform the text analytics on scrapped data for better understanding of the facts about the places, services, hotels, restaurants, etc [4]. The main websites considered for the study are tripadvisor.com, blog incredible India, incredible rural India, etc. The study will generate some new facts and insights through keywords in collected data [5], [6]. The special challenges that are associated with study are how to find the semantic meaning of a particular keyword i.e. whether it is said in a positive or negative sense. For example if there is a word “unpredictable” in a review and is used with nature (unpredictable nature), then it can be used for negative sense. However, if somewhere it is used with experience (unpredictable experience) then it has to be used in positive sense [5], [7], [8].

2. Literature Review

Sentimental analysis is used as a prominent research tool which comes under natural language processing (NLP). The opinion mining model for classification of hotel reviews presented by [9]. In this study, reviews were taken from tourism websites for hotels in Rome[9]. In another study, sentimental analysis was done by [10] in which

trip advisor was used as a data source. The analysis model has been developed for finding the similarity between sentiments given by users and automatic sentiment detection algorithms such as – sentiStrength, Bing, Syuzhet, etc.. In a study by [11], common modern framework for sentimental analysis has been discussed. The entire process of analysis contained three steps- processing of data, classification and validation of data. Various methods for processing were used such as- tokenization, filtering of stop words and stemming. Similarly, most popular classification algorithm used were support vector machine (SVM), naïve bayes, k-NN, etc. [11]. Tyagi et al. [12], conducted study on sentimental analysis for microblogging site twitter. N-gram technique has been used for feature extraction and the k-nearest neighbour (k-NN) algorithm has been used for classification of tweets. Lansono et al. [13] performed sentimental analysis of customer reviews of restaurants based on trip advisor. The naïve bayes and textblog algorithm has been used for classification. The naïve bayes gave better performance in comparison to the textblog algorithm. A survey on sentimental analysis based on twitter data is done by [14]. The techniques of extracting tweets has been explored and comparison performed using different sentimental analysis techniques. The Arabic sentiment analysis is done by [15]. Various limitations have been highlighted in the steps of pre-processing, feature engineering in the existing literature. In yet another study, various types of machine learning techniques have been explored by [16] for performing sentimental analysis. In this, approximately 20 research works have been evaluated to check the type of performance measure used by various researchers. Encouraged by the stated studies, a sentimental model has been proposed to classify the sentiments of the tourist who have visited India to explore various places.

2.1 Our Contribution

The detailed contribution of the work is as follows:

- The work uses a sentimental analysis for generating text classification by using reviews of the tourists who posted on webpages of tourism blogs and sites of tourism in Indian.
- The negative, positive, neutral and overall score is calculated by analysing the context of a sentence.
- Term Frequency - Inverse Document Frequency (TF-IDF) [17] metrics is calculated to find the relevance of the word.
- For proving the effectiveness of the work, random forest model has been implemented and quantitative measure have been calculated in terms of AUC-ROC curve and average precision (AP) metrics.

3. Methodology

The current study uses sentimental analysis technique to analyse reviews and other write ups written by tourists who have visited India in the past. Sentimental analysis is a technique, which comes under the models of Natural Language Processing (NLP). It is used to find the sentiments from the raw text data. This raw text data can be extracted from the social media sites such as- twitter, facebook, youtube, etc. This type of analysis is very helpful for a business organization to analyse what type of reviews and opinions are given by the customers (negative or positive) and where they need to improvise their services. The proposed model analyses the reviews given by foreign tourists about the tourism agencies and hotels in India. The Fig-1 shown the block diagram of the proposed model. The whole study was conducted in four phases. First phase was collection of data, 2nd phase cleaning of data or preparation of data. In the second phase, apart from normal cleaning, cleaning of stop words and extra words such as “no negative”, “no positive”, etc was performed. Three new features have been created in the dataset. These features are polarity score, numeric vector representation of each review and Term Frequency-Inverse Document Frequency (TF-IDF) for word. TF-IDF is a numerical statistic that is projected to imitate the importance of a word to a document in a pool or corpus [17]. The value of TF-IDF is proportional to the number of times word appears in a document. In the 3rd phase, a sentimental analysis with a rule-based model has been performed. After this, the implementation of random forest classifier [18] and visualization of results has been done in the last phase.

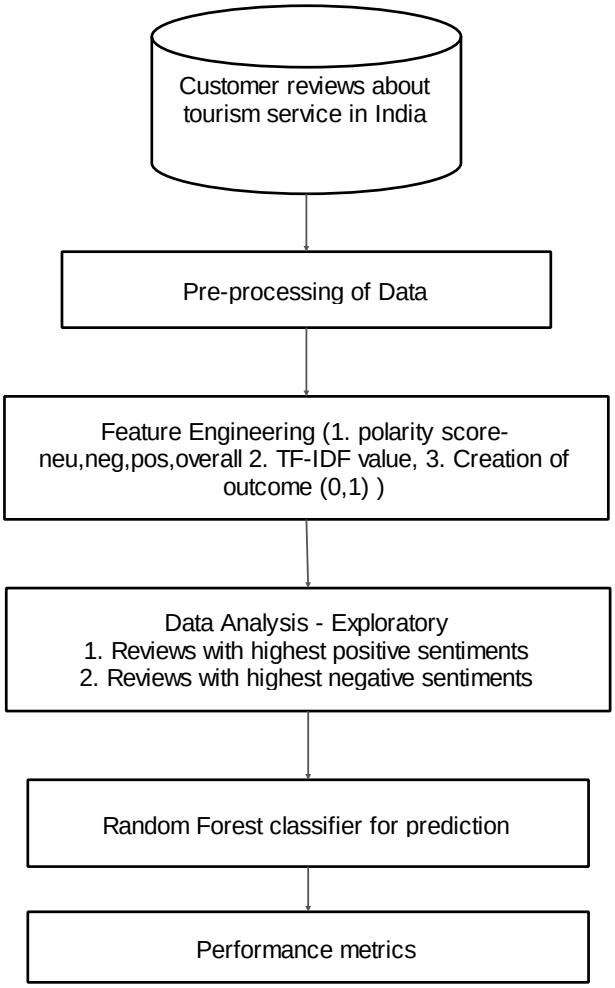

Fig-1 Block Diagram of Proposed system

Data collection and cleaning

For collecting the data of text reviews, popular travel sites such as tripadvisor.com, incredible India, and incredible rural India blogs have been used. The scrapper tool has been used to scrap all the reviews from the webpages and same were collected in the form of a excel sheet. The main travel locations of India covered in the reviews are – Agra, Mathura, Banaras, Uttarakhand, Jaipur, Udaipur, Gwalior, etc. Approximately 833 reviews are collected for the study. For cleaning of data, some words are removed from the collected data such as- no negative or no positive. These are the reviews given by some tourists, who do not have neither any negative feedback nor any positive feedback. After this some more operations have been performed for cleaning such as converting of all text in lowercase, removal of all punctuations, removal of useless stop words, and lemmatization of text.

Feature Engineering

For doing robust analysis, three features have been created and added to the dataset. First feature added is 'polarity score' of reviews. This score categorized all reviews in four category-neutral, negative, positive and overall. Second feature added is a vector representation of each review in numerical form [19]. The main benefit of adding this feature is that the same text values have same numerical representation, which can be used as a training feature. The 3rd feature added in dataset is Term frequency-Inverse Document Frequency (TF-IDF) for word [17]. The TF-IDF value for every word appeared in minimum 10 texts.

Data analysis

After creation of features, an exploratory data analysis has been performed in the form of a distribution graph and word cloud. The distribution graph shows plot between positive and negative reviews, as given in Fig-2. The positive reviews are approximately 89% in the dataset. This can be seen clearly from distribution graph as positive reviews tend to have higher compound score of sentiments. The word cloud has been generated of approximately

200 words from the reviews to check what kind of positive and negative words have appeared in the review. From the word cloud shown in Fig-3, it can be found that, most of the words are either related to the facilities such as – hotel, breakfast, room, etc. or are related to experience such as- amazing, great, excellent, etc.

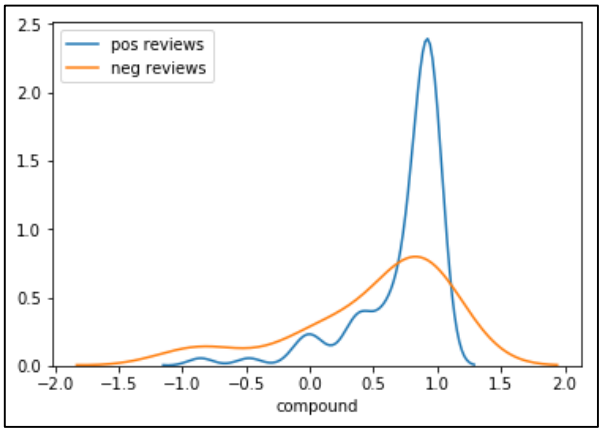

Fig-2 Distribution positive reviews vs. negative reviews

Fig-3 Word cloud of the most appearing words in the review

4. Results and Discussions

Experimental Setup

Simulation of the proposed sentimental analysis work has been done using PyCharm 2019.1.2 IDE using Python 3.7 in a workstation with Intel coreTM i7 2.21 GHz processor. The proposed model has been applied for the dataset of 833 tourist agencies in India. The main purpose of developing this model is to help various travel organizations in analysing reviews and opinions given by the customers (negative or positive) and further suggest where they need to improvise their services so as to meet tourist expectations better.

Performance Measures and Evaluation

The reviewer score in the dataset given by the tourist has been used to check the performance. For modelling of the reviewer score, random forest classifier has been implemented [18]. The whole dataset in the experiment has been divided into two parts - training and testing. 80% of data has been taken for the training of the model and the remaining 20% used for testing. The importance of features has been calculated with top five features shown in Table-1.

Table -1 Top five features with their importance

Feature	Importance
doc2vec_vector_2	0.080702
neu	0.052885
nb_chars	0.052752
doc2vec_vector_0	0.049660
pos	0.049596

Area under the Curve- Receiver Operating Characteristic (AUC-ROC Curve)

The ROC curve is generally a good metric in the form of a graph to digest the quality of the classifier [20]. It shows better predictions as the line goes above the diagonal base line. The Fig-4 shows the ROC curve of the proposed model. It shows the 82% of the area is covered.

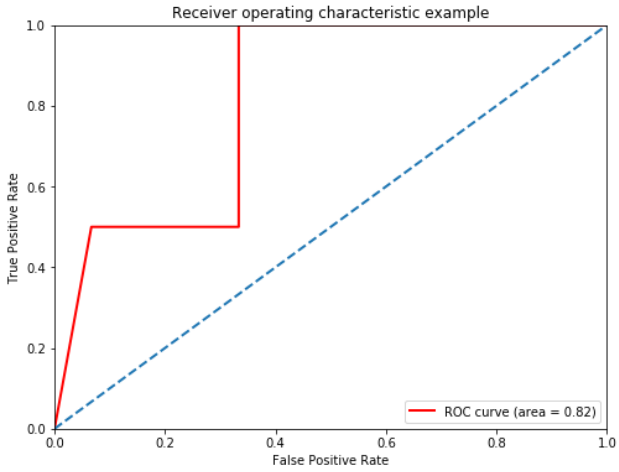

Fig-4 ROC Curve

Precision Recall Curve (PRC)

There are two reasons of considering this metrics while checking the performance of the model. First, in the PRC curve, the true negative values are not used. As our dataset in quite imbalanced with 89% reviews being positive and rest as negative, so there are chances of wrong prediction of positive review. Also, there are some words which can be predicted as negative while used in terms of positive and vice versa. Second, it is also used to cross verify the performance of the model. The PRC curve of the model is shown in Fig-5. The value of average precision (AP) is basically a value called as area under the precision recall curve [21]. AP of the model is approximately 0.39, which is considered as eight times better than random method.

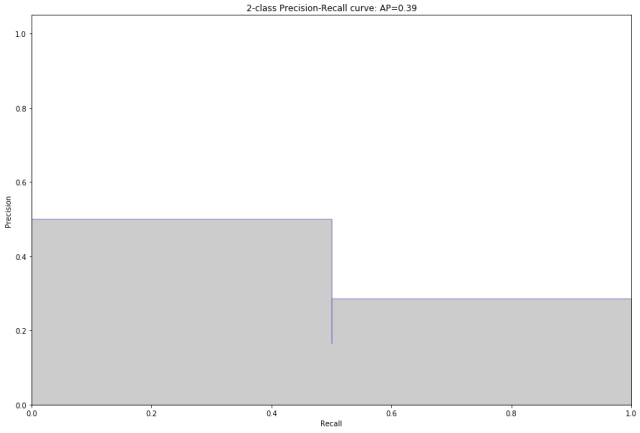

Fig-5 PRC Curve

5. Conclusion

The model for opinion mining using sentimental analysis is proposed in this paper. The analysis done by the model is based on the reviews of tourist given by them at various travel web sites and blogs. In this model, feature engineering has been performed to create three new variables for the purpose of providing a robust model. The reviews score, ROC curve and Precision recall curve are calculated to provide insight about the reviews. With this model, hotels authorities or travel agencies can analyse their respective tourist reviews and improvise their services as per the customer expectations. The main limitation of the model is the size of the dataset and only seven tourist places in India are considered. In future the large dataset can be taken with more tourist places.

References

1. Guliani LK, editor. Corporate social responsibility in the hospitality and tourism industry. IGI Global; 2016 Mar 4
2. W. T. and T. Council, "WTTC_India2019," p. 2019. Accessed on 14th July 2020
3. "Performance of Tourism Sector during December, 2016". Ministry of Tourism. Retrieved 28 February 2017.
4. Ohlan R. The relationship between tourism, financial development and economic growth in India. Future Business Journal. 2017 Jun 1;3(1):9-22.
5. N. H. Frijda, The Emotions, Cambridge University Press, 1986.
6. C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eight International AAAI conference on weblogs and social media, 2014.
7. Salvador Anton Clave´ Planeta, Lessons on tourism: the challenge of reinvesting destinations, Annals of Tourism Research, vol. 41, pp.249- 250, 2013.
8. Mostafa MM. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications. 2013 Aug 1, 40(10):4241-51.
9. Bucur C. Using opinion mining techniques in tourism. Procedia Economics and Finance. 2015 Jan 1;23:1666-73.
10. Valdivia A, Luzón MV, Herrera F. Sentiment analysis in tripadvisor. IEEE Intelligent Systems. 2017 Aug 17;32(4):72-7.
11. Karmaniolos S, Skinner G. A Literature Review on Sentiment Analysis and its Foundational Technologies. In2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) 2019 Feb 23 (pp. 91-95). IEEE.
12. Tyagi P, Chakraborty S, Tripathi RC, Choudhury T. Literature Review of Sentiment Analysis Techniques for Microblogging Site. InInternational Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India 2019 Mar 15.
13. Laksono RA, Sungkono KR, Sarno R, Wahyuni CS. Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. In2019 12th International Conference on Information & Communication Technology and System (ICTS) 2019 Jul 18 (pp. 49-54). IEEE.
14. Mittal A, Patidar S. Sentiment Analysis on Twitter Data: A Survey. InProceedings of the 2019 7th International Conference on Computer and Communications Management 2019 Jul 27 (pp. 91-95).
15. Ghallab A, Mohsen A, Ali Y. Arabic Sentiment Analysis: A Systematic Literature Review. Applied Computational Intelligence and Soft Computing. 2020 Jan 29;2020.
16. Devi GD, Kamalakkannan S. Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications, International journal of Advanced Science and Technology, 2020, 29(7), pp. 1462-1471.
17. Christian H, Agus MP, Suhartono D. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications. 2016 Dec 31;7(4):285-94.
18. Hegde Y, Padma SK. Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. In2017 IEEE 7th International Advance Computing Conference (IACC) 2017 Jan 5 (pp. 777-782). IEEE.
19. Giatsoglou M, Vozalis MG, Diamantaras K, Vakali A, Sarigiannidis G, Chatzisavvas KC. Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications. 2017 Mar 1;69:214-24.
20. Bhutani B, Rastogi N, Sehgal P, Purwar A. Fake news detection using sentiment analysis. In2019 Twelfth International Conference on Contemporary Computing (IC3) 2019 Aug 8 (pp. 1-5). IEEE.
21. Tartir S, Abdul-Nabi I. Semantic sentiment analysis in Arabic social media. Journal of King Saud University-Computer and Information Sciences. 2017 Apr 1;29(2):229-33.