IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 23, Issue 6. Ser. VIII (June 2021), PP 59-66
www.iosrjournals.org

Effective Marketing Strategies for the Tourism Industry using Sentiment Analysis

Aman Jangid[1], Ayushi Mishra[2], Ashutosh Sharma[1]

[1]*Aman Jangid & Ashutosh Sharma are undergraduate students with a Minor in Industrial Management and Engineering (IME) at the Indian Institute of Technology Kanpur*

[2]*Ayushi Mishra is Doctor of Philosophy (PhD) student in Industrial and Management Engineering(IME) at the Indian Institute of Technology Kanpur*

Abstract*—Websites like Tripadvisor, where people share their hospitality and tourism opinions generally have more than 2 million reviews and keep getting updated within minutes. They have the potential to create a great impact on new customers' sentiments. Therefore, we have developed a model for the tourism industry registered on social network platform Tripadvisor and analyzed the reviews of customers through sentiment analysis using Natural Language Processing and Machine Learning methods. We have also suggested marketing strategies to improve and strengthen the relationship of tourism companies with customers. We have extracted thousands of customer reviews about the industry on Tripadvisor using web scraping software Octoparse.*

Index Terms*—Sentiment Analysis, Social Network, User Behaviour, Marketing Strategy, NLP, ML, User Problems*

Date of Submission: 16-06-2021 Date of Acceptance: 01-07-2021

I. Introduction

Rapid development in information technology and com-munication has opened up new opportunities for marketing tourism and hospitality products, leading to significant com- petition amongst industries. Tourism industries are now more concerned about their services, products and thinking about the possibilities to enhance productivity, efficiency, and increasing competitiveness. Hence, it is imperative to take crucial steps in increasing customer satisfaction, building up market share and the industry's reputation. Some of the necessary actions taken by the sector are analyzing the behavior of users, resolving their critical problems with the industry, and developing the marketing strategies/campaigns to improve and strengthen their relationship with customers and thus reach their primary aim.

In the tourism industry, websites like Trip-advisor and social media provide an abundance of information regarding experiences and reviews about the destination, services, restaurant, and property. When tourists want to choose a comfortable hotel for their trip, they will look for reviews from other travelers [1]. Almost 93% of hotel management states that online tourist reviews are critical to the future of their property [2]. Thus customer reviews on Trip-advisor will be a potential source of information to analyze their sentiments. Traditionally, customer feedback is analyzed manually, and it is challenging to read individual comments and respond based on their needs since it requires human labor and is time-consuming.

With the advancements in technology, we are using Natu- ral Language Processing and Machine Learning techniques to get insights from thousands of reviews by customers about the industry. We are majorly focusing on three marketing research problems/opportunities in the tourism industry:

- Identifying the sentiments and behavior of the users about the industry.
- Developing different marketing strategies/campaigns based on the research.
- Identifying key problems users are facing with the tourism industry.

Although there has been significant research in sentiment analysis, most of the study is based on the polarity scores, i.e., providing an overall rating to the industry that did not capture the user's behavior on a minute scale. Our model focuses more on segregating different users based on their sentiment, i.e., loyal fans, unhappy user, quiet followers, cheerleaders [3] etc., and then target brand building will be done by industry in the following way:

- Loyal Fans - Useful to spread positive word of mouth (+WOM), so the industry appreciates them, should reward their behavior.
- Unhappy Users - can cause harm to the tourism industry reputation (-WOM), so the industry immediately solves their problems.
- Quiet Followers - Neutral, are there because of their friends, so the company puts efforts in engaging

them and making them active users.

- Cheerleaders - Top class fans of the brand and likes everything a company does, hence the tourism industry should keep them updated and inspired.

We are providing marketing and campaign strategies to the industry using the data visualization tool "Word Cloud" that extracts the most frequent words used by users in reviews. Finally, discover key users' problems with the industry through feature extraction with Word2Vec and TF- IDF (Term Frequency-Inverse Document Frequency) numer- ical statistics and obtained crucial features responsible for user problems.

II. Literature Review

There has been significant research on the role of social media platforms in tourism marketing and sentiment analysis of customer reviews. We have read some papers aligning with our interests.

Role of Social Media in Tourism Marketing [4] This paper furnishes a thorough understanding of the various theoretical models of social media, how marketers can create and implement marketing campaigns using social media, in what manner social media is used in decision making, analysis of different kind of social media users and their behavior, how the industry should target different users based on their requirements, what are the social media marketing strategies for the business, and by what means industry measure social network platform performance and effectiveness.

In [5] covers crucial ways to enhance competitiveness, efficiency, and productivity in tourism services: collection, analysis, and application of data of the targeted groups. The researcher collects data from various internet and web-based services and solutions, including Google Trends, Google search engine, Google Analytics, Flickr, social networks. Comparative analyses will then determine which source of information provides the most exciting data to be mined in the area of the tourist industry by analyzing the searcha- bility of the tourist destination through a set of predefined keywords. Their study shows that the most significant impact the searchability of a tourist destination has on the tourist location is virtual communities and reviews found on the internet.

In [6], analyzing hotel reviews written by tourists on social network platforms can significantly improve and evaluate the hotel; therefore, they have conducted sentiment anal- ysis to detect whether reviews have positive or negative opinions. Researchers provide a solution by classifying positive opinion reviews and negative opinions using the Multinomial Naïve Bayes Classifier and compare models us- ing pre-processing, feature extraction, and feature selection. Lemmatization, Tokenization, and Stopwords removal were done for pre-processing. In contrast, a used bag of words for feature extraction, two feature selections to be used to compare model performance. The first feature is frequency- based or deletes features with the lowest word occurrence frequency. The second feature is to remove features that have a minimum difference of positive and negative probability values.

The factor analysis study of social media in the tourism industry [7] identifies factors that determine the purpose of social media in Tourism Marketing, i.e., sharing the infor- mation, gathering customer feedback, building a network, brand endorsement, or increasing customer engagement. The results mainly focus on the information about the type of operation tourism enterprises concentrate. It may be accommodation, a tourist attraction, a restaurant, a spa, a tour company, the number of employees a tourism company has, and how long they have been in operation. Should they only cater to locals, or do they even cater to a foreign client. The descriptive statistic also provides the answer to the tourism business company on which type of social media (Facebook, Youtube, Line, Twitter, TripAdvisor) is mainly being used. The study's primary goal is to determine factors that significantly impact Social Media in Tourism Businesses and how they compete with one another.

In [8] Social Media can be helpful at various stages of planning, implementing, and tracking the impact of tourism growth on a local level. The planning process of tourism involves different steps like making a strategic goal, collecting primary data, reviewing secondary data to determine the critical components of the destination, finally arriving at the recommendation stage by discussing among authorities in the field of tourism. Text mining tools are used to evaluate reviews published on social media. It aims to extract keywords from unstructured data such as user comments and determine if there's a connection between them. It reveals specific visitor behavior patterns as well as their interpersonal interactions. Social Media is an important channel to communicate with the consumers for improving the product offers. Interactive dialogues can help in the process of designing new solutions. To reach the needs of consumers, Social Media makes necessary aspects in developing product innovation.

III. Study Data

There are tens of thousands of customer reviews on Tripadvisor about the tourism industry [9]. It is imperative to extract a good amount of reviews for sentiment analysis to get better results, so we have extracted

1000 reviews randomly using web scraping software Octoparse that allows quick retrieval of data in .xls file from any website without having to code. There is a loop on the "Next" button of the page and the "Read more" button of review in the Octoparse algorithm chart.

TABLE I
CUSTOMER REVIEWS

	REVIEWS
1	I have now waited a year to get back my deposit which was meant to have been a flexi rate. Each email they send me says they will return my deposit, so far nothing. This is the only hotel that has done this to me. If this happened in Australia they would be in a lot of trouble. I find this illegal that they can just steal my money.
2	Had a great stay. Loved the room and design. Hotel staff was very friendly and made me feel safe during COVID. Overall, I would stay at this hotel again. I enjoyed my NYC visit as a result of my stay at this hotel.

IV. RESEARCH METHODS

A. Proposed Method

During the study, the type of research conducted is Exploratory Research. Pre-processing, Feature Extraction, and Feature Selection executed using the NLP library, i.e.,

NLTK. The ML model proposed in this research is Random Forest and expected to provide the best accuracy. Fig 1 shows the description of the stages of the proposed method. Evaluated the model to identify the performance of the model. All models were trained and tested on an open-source GPU resource provided by Google Colaboratory [10].

B. Pre-processing Data

Data pre-processing is done to remove unwanted and noisy data in the dataset to get the best results. During pre- processing and cleaning of text, the steps taken are case nor- malization, Tokenization, stopwords removal, POS tagging, and Lemmatization.The stages involved in the processing of NLP text are described in Table 1.1. The procedure of turning all of the letters to lowercase is known as case normalisation.Tokenization is the method of splitting the input strings into tokens based on the compiler words. Its principle works by separating each word in a text. The method of eliminating less relevant word that often appears in the document is known as 'stopword removal'. Stop words like 'which', 'the', 'and can be removed, speed up the classification problem. POS (part of speech) tagging is the process of marking up a word in a text as corresponding to a specific part of speech based on both its meaning and context. For each tokenized word, lemmatization is the process of converting the word to a root word. Each affixed word will be extracted and converted to a simple word during the lemmatization process.

C. Feature Engineering

Feature engineering is the process of using domain knowl- edge to extract features from raw data. We first start by including sentiment analysis features because we believe that customer reviews are strongly related to how they feel about their hotel stay. Vader(part of the NLTK module) uses four values for each review, a positivity score, a negativity score, a neutrality score, an aggregate score that summarizes the previous score. The Gensim module uses the contexts in which words appear to generate a numerical vector representation of each word in the corpus (Word2Vec). The word vectors can also be used to convert any text into numerical vectors (Doc2Vec). Finally, we add the TF-IDF (Term frequency-inverse Document Frequency) values for each word and review. The number of times a word appears in the review is calculated by TF, whereas IDF calculates the relative importance of this word based on how many reviews it appears in.[11]

D. Exploratory Data Analysis

We applied some data visualization methods like Word Cloud and Seaborn library based on matplotlib to provide attractive and informative statistical graphs that can finally help spot specific patterns or trends in customer reviews and identify anomalies in the data.

E. Machine Learning Algorithm

Obtained features in feature engineering are then applied to a supervised machine learning classifier Random Forest. It is an ensemble learning method that creates a set of decision trees from a randomly selected subset of training sets [12]. The random forest can produce great results most of the time. It is the most flexible classification algorithm that focuses on determining feature importance.

V. Results

A. User Behaviour & Sentiments

1) Most Positive Reviews: **Keywords** - Perfect location of the hotel (Hotel Giraffe) in midtown in New York, reasonably sized rooms & storage space, staff (waitress, housekeeping, bartender) is friendly warm, helpful and cheerful, cleanliness, healthy & safe food (best breakfast, evening wine, and cheese), delivery services neat & on time, bed & pillow very comfortable, best customer service compared to the range of other hotels in the locality, easy access to subway stations.

Analysis - These are the loyal or cheerleader users for the tourism industry, and hotels should appreciate them and reward their behavior. Their reviews will play a significant role in deciding further customers' conversion, thus increasing their customer base.

2) Most Negative Reviews: **Keywords** - Small rooms, small bathroom lacking a stall shower, the window has no sound protection & hotel faces busy Park Avenue South, loud truck noises at 5.30 am, lobby/cafeteria dull and cold, no coffee machine at the room, bathroom sink was not draining correctly, the noise of tap and bathroom from an adjacent room, No of Deposit Money for a year, red wine horrible, bugs in bed.

Analysis - Industry-first fix the problems arising out of these unhappy users (i.e., soundproof window, early deposit, resolving bugs in beds, bathroom taps, the shower of some rooms), as these reviews can cause harm to the tourism industry reputation (-WOM), so should immediately solve the problems and reply to their reviews after they get resolved.

3) Most Neutral Reviews: **Keywords** - The experience was the average, good and friendly staff, no soundproofing, traffic noise; breakfast is basic, closes quite early, good places near for morning walk, TV is not connected correctly, not a pot of coffee, lack of an onsite gym, 4 to 6 subway near, close to Madison Square Park area.

Analysis - These neutral users may be mood changing, the industry should engage them, motivate them and resolve their problems so that they are connected with the industry, conversion of these doubtful customers into permanent will be the main aim, and replying to their reviews.

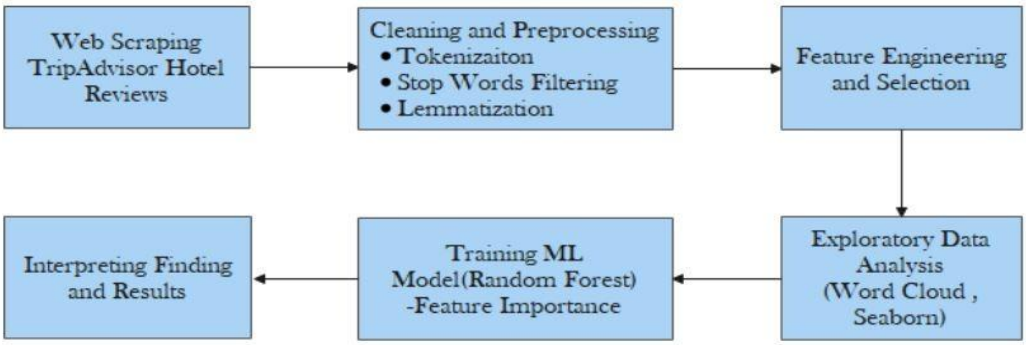


Fig 1. Proposed Method

Review	Had a great stay. Loved the room and design. Hotel staff was very friendly and made me feel safe during COVID. Overall, I would stay at this hotel again. I enjoyed my NYC visit as a result of my stay at this hotel
Case Normalization	had a great stay. loved the room and design. hotel staff was very friendly and made me feel safe during covid. overall, i would stay at this hotel again. i enjoyed my nyc visit as a result of my stay at this hotel
Tokenization	'had', 'a', 'great', 'stay', '.', 'loved', 'the', 'room', 'and', 'design', '.', 'hotel', 'staff', 'was', 'very', 'friendly', 'and', 'made', 'me', 'feel', 'safe', 'during', 'covid', '.', 'overall', ',', 'i', 'would', 'stay', 'at', 'this', 'hotel', 'again', '.', 'i', 'enjoyed', 'my', 'nyc', 'visit', 'as', 'a', 'result', 'of', 'my', 'stay', 'at', 'this', 'hotel'
Stopword Removal	'a', 'the', 'and', 'was', 'very', 'and', 'me', 'during', 'at', 'this', 'again', 'my', 'as', 'a', 'of', 'my', 'at', 'this'
POS Tagging	('great', 'JJ'), ('stay', 'NN'), ('.', '.'), ('loved', 'JJ'), ('room', 'NN'), ('design', 'NN'), ('.', '.'), ('hotel', 'NN'), ('staff', 'NN'), ('friendly', 'RB'), ('made', 'VBD'), ('feel', 'NN'), ('safe', 'JJ'), ('stay', 'VB'), ('hotel', 'NN')
Lemmatization	'great', 'stay', '.', 'loved', 'room', 'design', '.', 'hotel', 'staff', 'friendly', 'made', 'feel', 'safe', 'covid', '.', 'overall', ',', 'would', 'stay', 'hotel', '.', 'enjoyed', 'nyc', 'visit', 'result', 'stay', 'hotel'
review_clean	great stay love room design hotel staff friendly make feel safe covid overall would stay hotel enjoy nyc visit result stay hotel

Table 1.1 Data Preprocessing

B. Marketing Strategy

To develop different marketing strategies, we build a word cloud of top review scorers in various categories to see the most commonly listed words, understand better customer opinion, and other vital features such as mood and behavior. Words like friendly, wonderful, enjoy, treat, best, pleas- antly, and fantastic appear in the positive word cloud (Fig 2). These terms describe customer feelings about the hotel: they found the staff welcoming, their vacation in that hotel was great, they enjoyed their stay, staying in that hotel was a treat for them, it was their best option, and the hotel's atmosphere was good. Positive words are assisting the tourism industry in developing innovative marketing techniques and creat- ing new profit-generating concepts. These optimistic word clouds are used in marketing campaigns to draw users and improve search engine rankings. (For example, want to stay in a fantastic location in New York? Then stay at the Hotel Giraffe for a fun and enjoyable experience.)

Words like luxurious, service, management, room, dirty appear in the negative word cloud (Fig 3). They portrayed customers' negative opinions, such as the hotel being un- affordable, lousy management and customer service, and small and untidy hotel rooms. These negative words will help the tourism industry win in marketing by reacting favorably to negative feedback, increasing their confidence. The tourism industry should solve the user's issue and establish courtship and friendship with them, responding to customers individually to prove that they care about their opinions. There are some positive words in the negative word cloud because the data-set is highly imbalanced; only 2% of data is bad reviews.

C. Key Users Problems

When customers engage with a tourism company by leaving their reviews, whether it is a complaint or positive feedback, there is always an underlying emotion. We can capture this information to understand our customers better. For analyzing the feelings (attitude, opinions, emotions)

positive-wordcloud

Fig 2

negative-wordcloud

Fig 3

of customers, understanding the importance of features is a necessary task. It helps the industry learn what makes customers happy and what frustrates them.
Feature selection helps identify some critical problems faced by the customers. Words like money, small room, bathroom & shower, trouble, bad given high importance shown in the bar chart below and the industry should consider the worry of these unhappy customers.

D. Model Evaluation

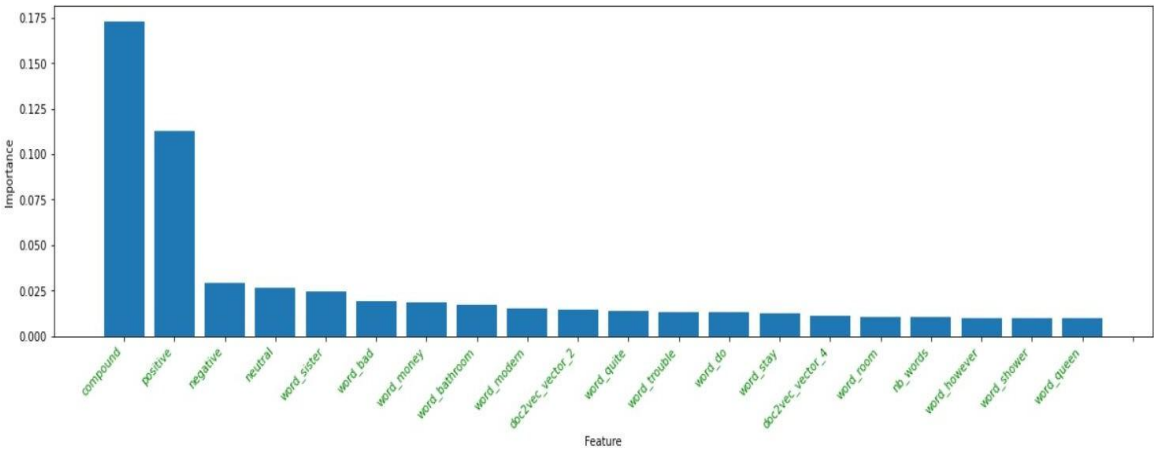
We have used the Random Forest Classifier for our prediction. Data is split into an 80/20 ratio for training and testing the model. Independent variables are the features created using WORD2VEC and TF-IDF (Term Frequency- Inverse Document Frequency) numerical statistics. In contrast, the dependent variable is a column created by assigning binary numbers to each review, i.e., 1 for bad review and 0 for good review using a compound score. Through the sentiment distribution graph (fig 4), we can see that good reviews for most of them are considered very positive. On the contrary, bad reviews tend to have lower compound sentiment scores. The dataset is highly skewed because just about 2 percent of our reviews are considered flawed.

A decent graph to summarise the consistency of our classifier is the ROC (Receiver Operating Characteristic) curve. The better the predictions, the higher the curve is above the diagonal baseline.
As the density graph indicates that almost 98% of the reviews about the industry are positive, only a few reviews are wrong, which we analyzed in critical user problems using essential features. Checking the performance metric on this imbalanced dataset may give biased results. We have obtained the confusion matrix and ROC curve to check the consistency of our classifier. As shown in the table, out of 142 datapoints in y test 141 are good reviews, the rest one bad review, and the model provides an accuracy of 99.3%. As the dataset is highly imbalanced, we cannot rely on the accuracy, but our main aim to find user problems is fulfilled using feature importance. For the ROC-AUC curve, the better the predictions, the higher the curve is above the diagonal baseline. Since our dataset is unbalanced, the number of Negatives(TN + FP) correlates to our positive feedback, which is quite large. As a result, our FPR (False Positive Rate) will remain low even though any False Positives occur.

Performance Measure		
Case	Actual/Reality	Test-Outcome /Predicted
True Positive 0 True Negative 141 False Positive 1 False Negative 0	Bad Review Good Review Bad Review Good Review	Bad Review Good Review Good Review Bad Review

$$FPR = FP / TN + FP$$

In our case, the ROC-AUC curve for a Random Forest classifier is about 1, which is quite large, and therefore ROC-AUC curve can be misleading for the imbalanced classification of the few examples of the bad review class because a small number of correct or incorrect predictions can significantly change the ROC Curve or ROC-AUC score. ROC graphs have one flaw: they are unreliable in the presence of class rarity or class imbalance accompanied by a limited sample size of minority occurrences.



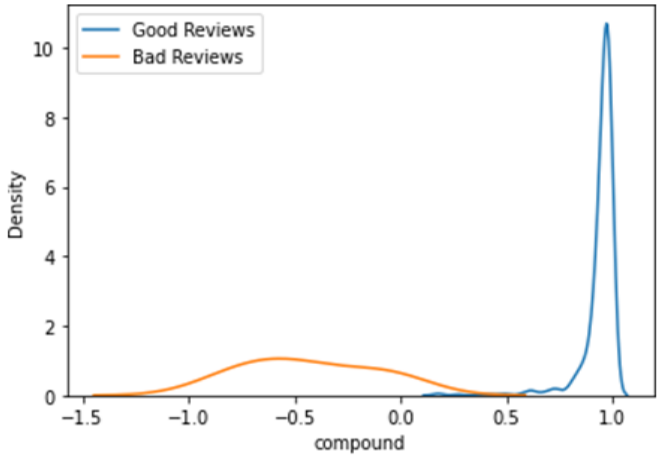
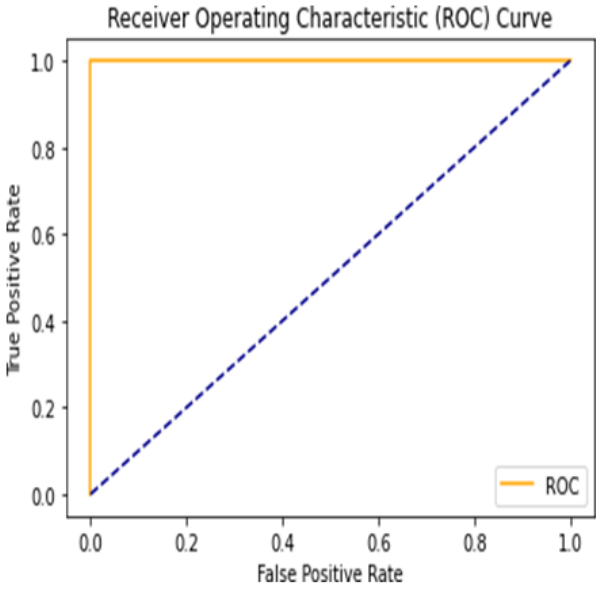


Fig 4



VI. Conclusion

Analyzing user behavior and resolving fundamental user problems using text analysis will significantly impact new customers' sentiments and strengthen the relationship be- tween the tourism industry and customers. The hospitality industry should take the same direction, which will help them learn about a hotel's image, consider its shortcomings, analyze rivals, and review all data in one go, saving time and effort. Tourism Business can turn to its advantage using Sentiment Analysis:

- **Revising marketing strategy** - addressing consumers' complaints and attracting them to the goods and ser- vices.
- **Increase product quality** - Customers also address and highlight key points that can be changed, resulting in the resolution of customer problems and the improve- ment of product quality.
- **Lead generation** - Information that we gain from performing accurate sentiment analysis or by analyzing user behavior and key user problems helps the industry come up with the best ways to reach out and attract new customers their brand.
- **Crisis management** - Observing what people are talk- ing about in the industry will assist in identifying possible crises and addressing them before they become a major issue.
- **Creating effective marketing campaigns** - Creating marketing campaigns using a word cloud to enhance search engine rankings.
- **Boost Sales revenue** - The advantages of sentiment analysis lead to a rise in sales revenue.

FUTURE SCOPE

As in our research, the tourism industry we selected has a very high positive response from users, i.e., 98% are good reviews. In this case, in the future, we may do the following task:-

- Extracting more user review data from Tripadvisor website to compensate class imbalance.
- Deal with high class imbalance dataset by Random Under-Sampling, Near-Miss, and SMOTE techniques.

References

[1]. "Research paper," https://ieeexplore.ieee.org/document/6016866.
[2]. "major trends," https://www.tripadvisor.co.id/TripAdvisorInsights/w665.
[3]. "Research paper," https://www.researchgate.net/publication/297767250.
[4]. "Literature review," https://www.researchgate.net/publication/297767250.
[5]. "Literature review," https://www.longdom.org/open-access/the-use-of-social-media-and-internet-datamining-for-the-tourist-industry-2167- 0269-1000197.pdf.
[6]. "Literature review," https://www.researchgate.net/publication/333168276.
[7]. "Literature review," http://www.ijimt.org/vol11/871-NT028.pdf.
[8]. "Literature review," https://www.researchgate.net/publication/345066142.
[9]. "Tripadvisor reviews," https://www.tripadvisor.in/Hotels.
[10]. "Google colaboraory," http://colab.research.google.com/.
[11]. "Tf-idf," https://scikit-learn.org/stable/modules/generated/sklearn. feature extraction.text.TfidfVectorizer.html.
[12]. "Random forest," https://scikit-learn.org/stable/modules/generated/sklearn .ensemble.RandomForestClassifier.html.

Aman Jangid, et. al. "Effective Marketing Strategies for the Tourism Industry using Sentiment Analysis." *IOSR Journal of Business and Management (IOSR-JBM)*, 23(06), 2021, pp. 59-66.