Hybrid Machine Learning model for A Next-Generation Ecofriendly Travelling and Guides to Reduce Carbon Emissions

Preprint · December 2022

DOI: 10.13140/RG.2.2.10030.18247

10 authors, including:

Elias Hossain
Mississippi State University
29 PUBLICATIONS **413** CITATIONS
SEE PROFILE

Wahidur Rahman
Uttara University
52 PUBLICATIONS **558** CITATIONS
SEE PROFILE

Redwan Abedin
Time research & innovation
6 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Nasim Ahmed Roni
Daffodil International University
9 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Hybrid Machine Learning model for A Next-Generation Ecofriendly Travelling and Guides to Reduce Carbon Emissions

Elias Hossain
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
0000-0002-9364-7916

Wahidur Rahman
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
0000-0001-6115-2364

Alif B Ekram
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
alifbekram@gmail.com

Redwan Abedin
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
redwanabedin@gmail.com

Nasim Ahmed Roni
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
nasimahmedroni@gmail.com

Erfanul Haque
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
ehaquedipto@gmail.com

S M Asaduzzaman
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
0000-0001-7058-2606

Md. Shazzad Hossain
Research and Development
Time Research and Innovation
Dhaka, Bangladesh
0000-0001-9507-5395

Dr Shah Siddiqui
Faculty of Technology
University of portsmouth
Portsmouth, United Kingdom
0000-0003-3958-9572

Dr Shamsul Masum
Energy and Electronic Engineering
University of Portsmouth
Portsmouth (UK), PO1 3DJ
0000-0001-8489-9356

***Abstract*—In this modern era, travelling has become an inevitable activity for medium-high income people and students to energise their mental strength. As a result, they tried to have a budget tour, such as finding a cheap place to stay or purchasing a cheap ticket. Although, numerous systems have been developed to target travel, yet, innumerable gaps exist between them. For instance, reducing carbon emissions and identifying human needs automatically have not been applied to date. Hence, it is crucial to comprehend the needs of individuals through opinion mining and to design the system accordingly. In this study, we analysed human sentiment through Natural Language Processing (NLP) and Machine Learning (ML) to propose a hybrid software framework (HSF) that will automatically fulfil people's needs and reduce carbon emissions. Latent Dirichlet Allocation (LDA) techniques are used to analyse the sentiment regarding tourism, including locations, costs, hotels, the nearest food, etc. Positive and negative opinions were classified using the Naive Bayes (NB), Logistic Regression (LR), and Gradient Boosting (GB) models. In contrast, the HSF describes a hybrid system designed as the next generation of travelling software. The accuracy of LR and GB models achieved 95.98% and 95.88%, respectively. These models are further evaluated with the performance of the Receiver Operating Characteristic Curve (ROC) and Average Precision (AP). They enumerated the value of 0.91 and 0.40 in Area-Under the Curve-Precision Recall (AUC-PR). This study even utilised the most dominant words in the sentiment analysis from the hotel review. Therefore, the proposed model can be one of the possible solutions to reduce carbon emissions, minimise plastic use, and be a game-changer for green-ecofriendly living.**

***Index Terms*—Gradient Boosting, Human sentiment, Latent Dirichlet Allocation (LDA); Logistic Regression and Sentiment Analysis Approach**

I. Introduction

The tourism industry has widened its paradigms, and the global Gross domestic product (GDP) has increased dramatically over the last few decades [1]. More specifically, the tourist sector has become essential for long-term economic growth in most countries. Therefore, it contributes 11% to the current world GDP, and an average of one billion people travel internationally annually [2]. It, additionally, contributes significantly to creating jobs, alleviating poverty, boosting income distribution, raising demand for goods and services, increasing tax revenue, and providing foreign exchange reserves [2]. Hence, it is significant for any country's long-term economic success.

Besides all of destructive impacts, tourism has more significant successes in shaping community and culture [3]. Cultural exchange programmes, educational developments, knowledge exchange programmes and many more milestones are achieved through community cohesion [4]. Additionally, when a traveller thinks of travelling somewhere, the first thought that comes to mind is accommodation, safety, cultural differences, language and bus/train /flight ticket booking [5]. It is noteworthy that the coronavirus pandemic has become a matter of concern worldwide. This caused a lot of worries to travellers as finding a safe place seemed a challenging issue, as well as entry clearance to specific destinations [6].

Therefore, we have come up with a solution that can

979-8-3503-4602-2/22/$31.00 ©2022 IEEE

initially answer some questions. Our tourism model can bring an outstanding revolution to reduce carbon emissions and gain travellers confidence by enhancing awareness of modern technology. Furthermore, it will help them solve and overcome travel-related complications while planning a tour and saving the environment proportionally. Through sentiment analysis, with a specific dataset and ML/AI algorithms, the green travelling scheme can help tourists find suitable locations, routes, accommodation, food, and other related information to enjoy their holiday.

Simultaneously, they will be offered to choose and donate to tree plantations by calculating their carbon footprint. It also lets them know the uses of plastic and utilise available resources to save the environment. In the twenty-first century, when the globe is experiencing a tremendous pollution imbalance, creating a green ecosystem that reduces CO2 emissions, and minimises the use of plastic and plants trees is essential. Therefore, this research aims to develop an intelligent application based on the sentiment analysis of hotel reviews to integrate the artificially green eco-friendly recommendation system for determining the customer's needs.

There are eight interconnected sections in this article. Following the literature review in section two, sections three, four, and five discuss research methodology and systems analysis. The results and discussion are summarized in Sections 6 and 7. We have described the study's conclusion and future work in section eight.

II. Literature Review

This research has observed significant contributions in implementing ML and Artificial Intelligence (AI) in travel and tourism systems worldwide. Many great contributors have made to eco-friendly travelling and tourism. Shafqat et al. [7], integrated place recommendations, food quality, clean environment, opening and closing hours, and suggestions on the under-emphasis places and their previous histories. Ghani et al. [8], proposed a smartphone-based ML system with a Dijkstra algorithm, Google mapping API, and a real-time database to suggest the most exciting proximity location, including all routes. Although they have a user-friendly cost-effective theoretical proposed system, the practical implementation is not validated with a real user perspective.

Bodhankar et al. [9], provided a descriptive study based on the challenge of recommendation systems in social networks. Their theoretical concept is clear, but maintaining a relationship between these two systems is not well defined. Kyaw et al. [10], have estimated the travel speed of public transportation in road networks using GPS data and analyzed the trajectory data by applying ML techniques. The data of GPS is based on a single bus line, indicating that if the user needs to know other courses' situations, it will not be accessible. Furthermore, this research is based only on Yangon Route's map, which means it cannot be used for other areas.

Ganga et al. [11], developed a system where necessary information can be observed about a tourist spot without having an app through AI components. The system is based on GPS and geodata handling, so a data connection is required indicating if the internet connection falls or disrupts, it can not be possible to find the route.

Parvez et al. [12], explained that the application of ML in tourism makes travel more accessible, more comfortable, and refreshing, which helps service recipients and service providers. The main limitation is that this study lacks qualitative or quantitative analysis.

Clarizia et al. [13], designed a chatbot for tourist destinations to increase processing efficiency and serve users related to contextual aspects, such as position and time. However, new heterogeneous data sources and services can be used for interaction augmentation of the system. Li et al. [14], described a task-oriented chatbot system that provides hotel deals and recommendations through third-party messaging platforms such as Facebook Messenger to book hotels through text messaging. Deep language models are memory-intensive, and memory share across different models is significant.

III. Proposed Methodology

The proposed research methodology has been divided into two phases: Opinion Mining (OM) and Hybrid Software Framework (HSF). The HSF consists of three steps, the first of which includes data integration and data cleansing. In the second stage, the carbon footprint is calculated using machine learning, and in the last stage, opinion mining is used to make decisions.

IV. Opinion Mining

The following section is classified into five subsections. They are Data integration, Pre-processing, Feature Extraction, Model Selection, and Topic Modeling Intuition.

A. Data Integration

We have collected this dataset from booking.com [15], which contains customers (N=515K) and luxury hotel reviews (N=1493) in Europe [16]. It has 17 fields like hotel address, positive or negative reviews, etc. We classified the reviews as positive and negative, where the positive review is marked as 0, and the negative review is marked as 1. Table 1 shows the insights of the dataset.

B. Text Pre-Processing

To eliminate hotel reviews' unstructured existence, we have used preprocessing to establish the right sentiment for efficient decision-making, as shown in Fig. 2. To achieve correct results, we have included converting HTML entities, removing "@user" from all the tweets, changing all the tweets into lowercase, apostrophe lookup, short word lookup, emoticon Lookup, replacing special characters with space, replacing numbers (integers) with space, removing words whom length is 1, tokenization, remove stop words, lemmatization/stemming, removal of URLS, removal of HTML tags [17].

979-8-3503-4602-2/22/$31.00 ©2022 IEEE

TABLE I: INSIGHT INTO THE RESEARCH DATASET

Insight	Value	Label
A total positive review (length)	412601	-
A total negative review (length)	330011	-
Review total negative word count (length)	402	-
Review total positive word count (length)	365	-
Total number of the review that the reviewer has given	3695779	-
Positive review	The rooms are nice but for the elderly a bit difficult	0
Negative review	My room was dirty, and I was afraid to walk	1

C. Feature Extraction

We have used the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert the text data to a numeric value [18]. TF helps identify the similarity between documents, and the exact length vector represents each document containing the words counted. Then, they are normalized and added to the sum of their components. Here groups of words represent each record. Therefore, if a word is in a document, it will be described as one; if it is not, it will be set to zero. The TF-IDF is a weighting parameter often used to process natural language and information retrieval. A statistical metric is used in a dataset to calculate a word's importance to a text. The value of a phrase increases with the number of times a word appears in the text, but the word's frequency in the corpus counteracts this. One of the IDF's key features is that the term frequency is weighted down while the uncommon ones are scaled up. For instance, words such as "the" and "then" frequently appear in the text, and terms like these would dominate the frequency count if we only use TF [19]. Nevertheless, using IDF scales down the effect of these words.

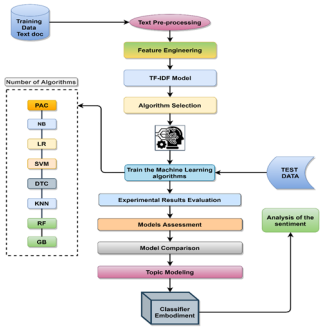

Fig. 1: Overall Steps of Opinion Mining

$$\log \log \frac{P(x)}{1-P(x)} = \sum_{j=0}^{K} b_j x_j$$

$$\frac{P(x)}{1-P(x)} = \exp\left(\sum_{j=0}^{K} b_j x_j\right) = \prod_{j=0}^{K} \exp\left(b_j x_j\right)$$

Algorithm 1. An algorithm for Gradient Boosting(GB) 1 Input: Import the word reference of performance words and the file of the contact numbers. 2 The model initializes with a constant value, as follows:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$$

3 For $m = 1$ to M : a) $\quad r_{1m} = -\left[\frac{\partial L(y_i F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots, nCo$ impute the following expression which is known as so-called pseudo-residuals: b) Fit, $h_m(x)$ presents a base learner for pseudo residuals with the training set. c) Here, γ_m is presented as a multiplier and computed through the following equation:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma h_m\left(x_i\right)\right)$$

d) Updating with the model with the following equation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

4 Result: Finally, we got the output as $F_M(x)$.

D. Prediction Strategy

We have recognized that the Logistic Regression (LR) and Gradient Boosting (GB) are suitable; their sensitivity and specificity were promising. GB is a popular conventional machine learning model for classification and regression problems [20]. It more significantly analyses a prediction strategy in an ensemble prediction model architecture by generating decision trees shown in Algorithm 1.

LR is one of the standard algorithms for fitting a model, especially binary responsive data. The similarities of LR are diverse from the conventional Linear Regression algorithm. It can predict the corresponding probabilities where the values lie between 0 and 1 [21]. Again, it has the capabilities of the training data of a model to conserve the marginal probabilities. In our discussion, we assume the binary responsive data where 0 represents false and 1 for true. It can be used in the following Eq. 1, where the observations of y can be expressed as the K number of the variables of x. The equation is called the log of P. Here, we can find b0 if the values of X0 = 1. If we present the equation to both side's exponent, the equation will be converted into Eq. 2.

979-8-3503-4602-2/22/$31.00 ©2022 IEEE

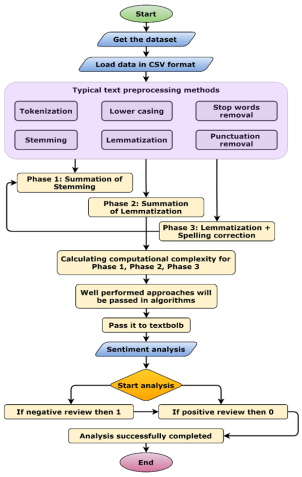

Fig. 2: Illustration of the Text Pre-processing

E. Topic Modeling Intuition (TMI)

We have utilised Latent Dirichlet Allocation (LDA) as TMA [22]. LDA is a very interpretable and widespread architecture for analysing topics in the Text data set. In LDA, a layer of complexity is added and assumed K, which presents a list of topics. In a particular document, m offers the probability distribution over k topics. Again, the probability distribution of each specific topics is called vocabulary V. The formula required for LDA is segmented into five interconnected modules shown in Fig. 3.

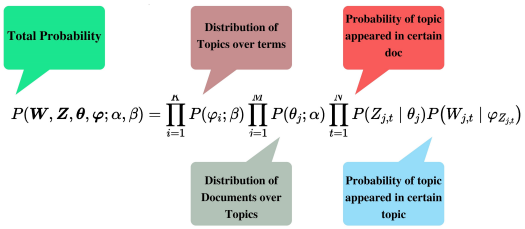

Fig. 3: The Details of Latent Dirichlet Allocation (LDA)

V. Hybrid Software Framework (HSF)

The Hybrid Software Framework (HSF) explains the overall proposed system. The HSF is connected to four significant components: The user Input Layer (UIL), Input Communication Layer (ICL), Recommendation Layer (RL), and Payment Gateway Module (PGM). The UIL module is intended to maintain engagement between the end user and the scheme, as well as propose news and articles regarding carbon emissions to raise the end user's awareness of their carbon footprint. This layer will provide input for a destination, and system users must select a landing location on the software system. After that, the UIL input will be connected to the communication layer (ICL) to start the next operation. The RL module will start its activities when receiving input requests from system users. The RL component is designed for recommendations where a user will be suggested the nearest hotels and local tourist places and help to decide by comparing the price.

In addition, the RL component will assist users in calculating the carbon footprint based on traveling vehicles for a specific location by using their vehicle's information. It will also suggest the use of the shortest route to make the least use of the fuel vehicle, thereby achieving our goal of creating an eco-friendly system that protects the environment. Since a certain quantity of plastic and water is required to go to a specific location, our system will advise users of the quantity of plastic and water necessary. In addition, numerous tree-planting groups will be integrated into our system, and users will be able to give to promote environmental safety. We have implemented a payment gateway mechanism via which donations can be made in order to accept contributions. Fig. 4 depicts the appropriate graphic of the suggested model in its entirety.

The RL component is designed through the Collaborative Filtering Method (CFM). This filtering approach generally collects and analyses user experience information, behaviors, or interests and predicts what they would like based on similarities with other users. The collaborative filtering approach's advantage includes that it does not rely on machine-analyzable content and can correctly recommend complex items without requiring an "understanding" of the the item itself [23]. A typical recommendation engine processes data over the following four steps: selection, storage, analysis, and filtering [24]. We have applied the K-Nearest Algorithm (KNN), the Jaccard's coefficient, the Dijkstra algorithm, and the cosine similarity to forecast the shortest path from the user's current location to the user's desired destination and as well as to suggest places to the users based on the rating.

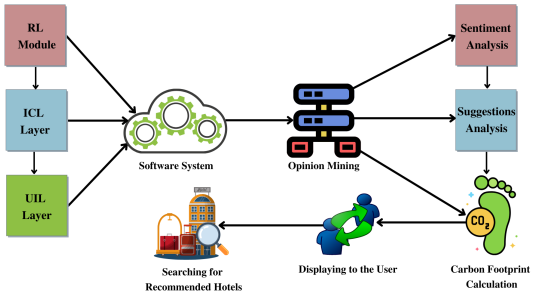

Fig. 4: Architecture Diagram of the Proposed Software System

VI. Result and Discussion

The Results Analysis has been classified into five phases: Experimental Results, Model Assessment, Comparative Analysis, and Sentiment Analysis & Topic Modelling.

A. Experimental Results

The Experimental Results section narrated the practical consequence found after executing several models. In this the study, various ML algorithms are utilized, namely, Passive

Aggressive Classifier (PAC), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree Classifier (DTC), Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN) algorithm. Consequently, the LR and GB algorithm was satisfactory to classify negative and positive opinions, with 95.98% and 95.88% accuracy. This section displays the accuracy, precision, recall, and f1 scores of the leading ML models on a per-class basis. The matrix is calculated using true and false positives and true and false negatives. We have followed relevant equations to determine the accuracy, recall, and f1 ranking. Fig.5, and 6 illustrate the precision and accuracy of visual representation.

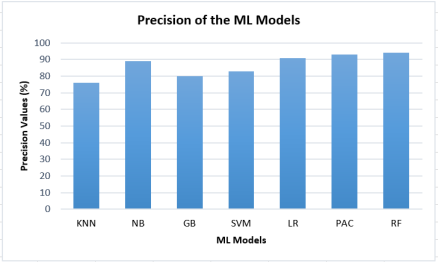

Fig. 5: Visual Representation of the Precision Report Obtained by Applying Various ML Models

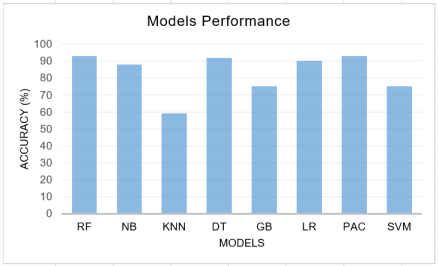

Fig. 6: Accuracy Interpretation of the Accepted ML Models

B. Model Assessment

Two methods of evaluation are described in the model assessment section, namely, Receiver Operating Characteristics (ROC) curve and K-folds Cross-Validation (KCV). The ROC is a graphical approach to evaluating a classifier's output. A classifier's efficiency uses a pair of statistics - true-positive and false-positive rates. The Fig. 7 are plotted on a two-dimensional graph, with a false positive rate on the X-axis and a true positive rate on the Y-axis. The resulting plot can evaluate the relative output of various classifiers and determine if a classifier performs better than random guessing. KCV is a method by which machine learning models are created by making multiple subsets of the same dataset, resulting in different model prediction accuracy for different subsets. Fig. 7 exhibited the ROC curve with KCV for the logistic regression classifier.

Precision-Recall is a helpful prediction indicator of performance when the classes are imbalanced. Accuracy is a measure of outcome validity in data retrieval, while recall measures how many genuinely valid results are retrieved. The precision-recall curve indicates the tradeoff between precision and recall for various thresholds. A high area represents high recall and high accuracy under the curve. High accuracy refers to a low false-positive rate, and high recall refers to a low false-negative rate. By observing Fig. 8, it can be observed that the value of AP was found to be satisfactory, which indicates that the model has a strong capacity for prediction.

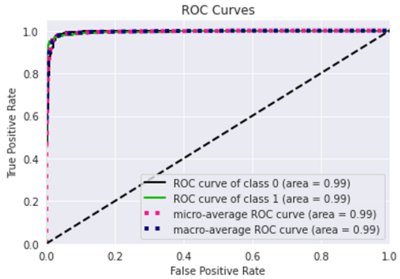

Fig. 7: Receiver Operating Characteristics (ROC) Curve with Cross-Validation for the Top ML model

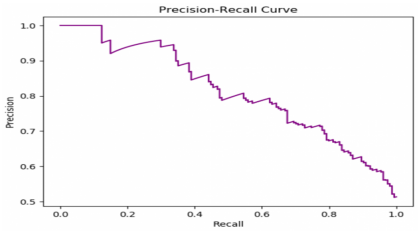

Fig. 8: Visualization of the Precision Recall Curve on the Accepted Model

C. Sentiment Analysis

In this study, we can consider that there are many unsatisfactory opinions besides good reviews. This indicates that the number of people who have traveled and received accommodation assistance in recent years is not happy. It also picks up on the most harmful words, including public opinion on important issues like location, hotel, bathroom, reception, and services. Since people have these problems, we have taken them into our observation. However, it was essential to investigate these issues before creating an effective system because it is necessary to find out the users of the proposed approach, identify their problems and fully meet their needs.

D. Comperative Analysis

After carefully reviewing the literature, it has been discovered that no adequate ML implementation is carried out in traveling systems that can ensure multiple functionalities. Therefore, Ghani et al. [8], and Coelho et al. [25] have proposed a traveling system that has compatible with our system. However, these papers deal with only four categories-historical buildings, museums, parks, and restaurants. On the contrary, our green eco-friendly traveling system will suggest

979-8-3503-4602-2/22/$31.00 ©2022 IEEE

hotels, food, cost, nearest location (such as historical places, museums, parks, restaurants, mountains, sea beaches), ticket booking, and tourist guides. The authors Clarizia et al. [13], Shafqat et al. [7], Gomathi et al. [26], used a smaller dataset for their systems. In contrast, we used a dataset from booking.com [15], which contains 515K customers and 1493 luxury hotel reviews in Europe [16]. The proposed system by Ganga et al. [11], need a data connection for GPS tracking and geodata handling. Furthermore, Thura Kyaw et al. [10], collected data on GPS based on a single bus line route. In contrast, our system can be used effectively and will show the shortest routes for any destination by using the shortest path findings algorithm. Tasfiqul Ghani et al. [8], and Thennakoon et al. [27], developed a system that did not use real-time users; also Bai Li et al. [14], proposed a system that could not handle new types of queries, whereas our system could solve queries of real-time travelers.

E. Discussion

Our proposed green eco-tourism model will suggest how to reduce carbon emissions from traveling. Furthermore, our system will suggest a cost-effective hotel with affordable prices and solve all travelers' queries. In addition to these, sentiment analysis from the hotel review also utilized the most dominant words in our study to understand any traveler's sentiments better. Our research aims to build a green eco-system-based application that eases the inconvenience of traveling worldwide with the help of hotel booking, cost, nearest location, meals, ticket booking, and tourist guide.

VII. Conclusion With Future Work

Travel is one of the most successful ways of personal development. It individually encourages us to do things that are different from everyday activities. Many environmental interactions can be found with traveling because enormous ecological pollution is caused by traveling. In this research, a next-generation web application is proposed based on Machine Learning, which is straightforward to recommend several things for making a green ecosystem. Our proposed system will play an essential role in creating a green ecosystem.

This research conducted a sentiment analysis on top of the hotel review and extracted the most dominant keywords from the dataset. With Logistic Regression (LR) and Gradient Boosting (GB) algorithms, we have analyzed sentiment with 95.98% and 95.88% accuracy. Architecture diagram of the proposed software system Fig. 4. A Positive review was found by applying lemmatization proposed hybrid software system, the carbon footprint can be calculated, which is an optimal contribution to creating a green environment. In the future, we will create a complete software system integrated with machine learning and solve research gaps. Users will find our system in their daily lifestyle towards traveling abroad.

References

[1] "Travel & tourism economic impact: World travel & tourism council (wttc)."

[2] "Travel amp; tourism economic impact: World travel amp; tourism council (wttc)."

[3] S. D. Kahramonovna, "Event tourism is a significant part of cultural tourism," *Central Asian journal of innovations on tourism management and finance*, vol. 2, no. 6, pp. 45–53, 2021.

[4] M. Gupta and A. Hasnain, "Sustainable tourism: Elevating collaboration between the hospitality industry and local citizens," *ECS Transactions*, vol. 107, no. 1, p. 18611, 2022.

[5] J. Duan, C. Xie, and A. M. Morrison, "Tourism crises and impacts on destinations: A systematic review of the tourism and hospitality literature," *Journal of Hospitality & Tourism Research*, vol. 46, no. 4, pp. 667–695, 2022.

[6] J. Ap, "Residents' perceptions on tourism impacts," *Annals of tourism Research*, vol. 19, no. 4, pp. 665–690, 1992.

[7] W. Shafqat and Y.-C. Byun, "A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis," *Sustainability*, vol. 12, no. 1, p. 320, 2019.

[8] T. Ghani, N. Jahan, S. H. Ridoy, A. T. Khan, S. Khan, and M. M. Khan, "Amar bangladesh-a machine learning based smart tourist guidance system," in *2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp. 1–5, IEEE, 2018.

[9] P. A. Bodhankar, R. K. Nasare, and G. Yenurkar, "Designing a sales prediction model in tourism industry and hotel recommendation based on hybrid recommendation," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1224–1228, IEEE, 2019.

[10] T. Kyaw, N. N. Oo, and W. Zaw, "Estimating travel speed of yangon road network using gps data and machine learning techniques," in *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 102–105, IEEE, 2018.

[11] R. S. Ganga, P. C. P. Reddy, and B. C. Mohan, "System for intelligent tourist information using machine learning techniques," *International Journal of Applied Engineering Research*, vol. 13, no. 7, pp. 5321–5327, 2018.

[12] M. O. Parvez, "Use of machine learning technology for tourist and organizational services: high-tech innovation in the hospitality industry," *Journal of Tourism Futures*, 2020.

[13] F. Clarizia, F. Colace, M. De Santo, M. Lombardi, F. Pascale, and D. Santaniello, "A context-aware chatbot for tourist destinations," in *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 348–354, IEEE, 2019.

[14] B. Li, N. Jiang, J. Sham, H. Shi, and H. Fazal, "Real-world conversational ai for hotel bookings," in *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 58–62, IEEE, 2019.

[15] "The largest selection of hotels, homes, and holiday rentals."

[16] J. Liu, "515k hotel reviews data in europe," Aug 2017.

[17] V. A. Kozhevnikov and E. S. Pankratova, "Research of text preprocessing methods for preparing data in russian for machine learning.," *Theoretical & Applied Science*, no. 4, pp. 313–320, 2020.

[18] W. Scott, "Tf-idf from scratch in python on real world dataset," *Towards Data Science*, vol. 15, 2019.

[19] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pp. 127–138, Springer, 2017.

[20] S. Ray, "Commonly used machine learning algorithms: Data science," Nov 2022.

[21] A. T. Arnholt, "Predictive model building - github pages."

[22] S. Li, "Topic modeling and latent dirichlet allocation (lda) in python," Jun 2018.

[23] B. Rocca, "Introduction to recommender systems," Jun 2019.

[24] "Retail documentation nbsp;—nbsp; google cloud."

[25] J. Coelho, P. Nitu, and P. Madiraju, "A personalized travel recommendation system using social media analysis," in *2018 IEEE International Congress on Big Data (BigData Congress)*, pp. 260–263, IEEE, 2018.

[26] R. Gomathi, P. Ajitha, G. H. S. Krishna, and I. H. Pranay, "Restaurant recommendation system for user preference and services based on rating and amenities," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. 1–6, IEEE, 2019.

[27] M. Thennakoon, R. Rajarathna, S. Jayawickrama, M. Kumara, A. Imbulpitiya, and N. Kodagoda, "Tourguru: Tour guide mobile application for tourists," in *2019 International Conference on Advancements in Computing (ICAC)*, pp. 133–138, IEEE, 2019.

979-8-3503-4602-2/22/$31.00 ©2022 IEEE