IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Tour Package Recommendations through Integrated Machine Learning and Text Vectorization: A Personalized Approach

[1]Vikas Chaurasiya, [2]Avinash Chaurasia, [3]Prathamesh Amrutkar, [4]Omkar Tawade, [5]Sonali Dhamele
[1,2,3,4]Student, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India
[5]Professor, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India

Abstract: This paper presents an innovative approach to enhancing tour package recommendations through the integration of various Machine Learning algorithms and vectorization methods for text processing. The proposed approach is designed to offer personalized tour package suggestions tailored to user specified destinations and preferred travel styles. Leveraging the Machine Learning algorithms alongside different vectorization enables the system to analyze textual data efficiently and extract meaningful features for recommendation generation. Experimental evaluations conducted on real-world datasets demonstrate the efficiency of the proposed method in delivering pertinent and diverse tour recommendations. The results underscore the capability of our approach to enhance the user experience by providing personalized travel suggestions that align closely with individual preferences and interests. This research contributes to advancing the field of tour recommendation systems by offering a robust and effective framework for generating tailored tour packages.

Index Terms **- Random Forest, Count Vectorization, Tour Packages, Content-based Recommendation.**

I. INTRODUCTION

In an age where travel is not just a journey but an experience, the quest for personalized recommendations has become paramount. Amidst a sea of options, travelers seek guidance that resonates with their unique interests and desires. Traditional recommendation systems, although effective to an extent, often falter in capturing the intricacies of individual preferences, resorting to generic approaches that overlook the subtleties that define a truly tailored travel experience. Our study addresses this challenge head-on, presenting a groundbreaking methodology that marries the robustness of the Machine learning algorithm with the semantic richness of TF-IDF (Term Frequency-Inverse Document Frequency) and Count vectorization. By fusing these advanced techniques, we aim to revolutionize tour recommendations, offering users a bespoke selection of tour packages that transcend the conventional boundaries of relevance and diversity. At the heart of our approach lies the recognition that travel preferences are multifaceted and deeply personal, shaped by a myriad of factors ranging from destination aspirations to preferred travel styles and budget considerations. By harnessing the power of the Machine learning algorithms, we endeavour to decode the complexities of individual preferences, identifying similar packages based on their unique travel introduction and recommending packages that align seamlessly with their aspirations.

Furthermore, the integration of TF-IDF and Count vectorization adds a layer of semantic understanding to our recommendation engine, enabling it to discern the nuanced nuances embedded within travel descriptions. By transforming textual data into numerical representations, TF-IDF prioritizes terms that are both distinctive and indicative of a particular travel experience, ensuring that recommendations are not only relevant but also rich in context and meaning. Count vectorization stands out as a pivotal technique employed to transform textual data into numerical vectors, leveraging word frequencies as its foundation.

Through this synergistic fusion of algorithmic prowess and semantic sophistication, our methodology promises to redefine the landscape of tour recommendations. By transcending the limitations of traditional approaches, we aspire to empower travellers with recommendations that not only meet their expectations but exceed them, guiding them towards experiences that are as unique and unforgettable as they are. In the ensuing sections, we delve into the intricacies of our methodology, elucidating the implementation details and presenting empirical evidence of its efficacy in generating personalized tour recommendations. By bridging the gap between technology and human-centricity, our study heralds a new era in tour recommendation systems, where every journey is not just recommended but crafted with care and precision to reflect the individuality of each traveler.

II. RELATED WORK

In the study [1] Abbasi and Alesheikh’s pioneering work introduces a groundbreaking approach that utilizes word embeddings to revolutionize place recommendations. By harnessing the semantic relationships embedded in textual data, their method provides more nuanced and contextually relevant suggestions to users, considering not only geographical proximity but also semantic similarities between locations. This innovative fusion of geographical and semantic data enriches the accuracy and relevance of place recommendations, offering users a more tailored and personalized travel experience.

[2] Anand et al.’s research focuses on leveraging chatbot technology to enhance tourism services in Indian cities. Their study underscores the transformative potential of AI driven chatbots in providing personalized recommendations and real-time assistance to tourists. By analyzing user queries and historical data, chatbots offer tailored suggestions for attractions, accommodations, dining options, and activities, thereby facilitating seamless trip planning and exploration. Through interactive conversations, chatbots not only improve the user experience but also empower tourists with personalized guidance throughout their travel journey.

[3] Stefanovicˇ and Ramanauskaite propose an innovative model for travel˙ direction recommendation that capitalizes on analyzing user generated photos. By extracting visual cues and preferences from images shared on social network profiles, their method aims to deliver highly personalized tour recommendations tailored to individual interests. By deciphering the content of photos, including landmarks, scenery, and activities, the model infers users’ travel preferences and aspirations, enriching the recommendation process with a more immersive and engaging user experience. This integration of multimedia data into recommendation systems marks a significant advancement in trip planning and exploration, offering users a deeper level of personalization.

[4] Amara and Subramanian’s research delves into the synergy between personalized and content-based recommender systems in the travel domain. Their study proposes a collaborative filtering approach that combines user preferences with content-based analysis to generate contextually relevant recommendations. By leveraging textual data from travel descriptions, reviews, and user profiles, their model crafts personalized recommendations that closely align with users’ preferences and interests. This hybrid approach enhances recommendation accuracy and user satisfaction, offering travelers tailored suggestions that cater to their individualized needs and preferences.

[5] Gharibi et al. present a cutting-edge content-based model for tag recommendation in software information sites, with implications for enhancing content discoverability and user engagement in the travel domain. Their study employs advanced recommendation techniques to analyze textual descriptions and user interactions, generating relevant tags that improve content organization and searchability. By leveraging content-based recommendation methodologies, the model enriches metadata and enhances the discoverability of tour packages based on descriptive content and user interactions, ultimately enhancing the overall user experience.

[6] Chen, Jang, and Lee’s research introduces a sophisticated kernel framework for content-based recommendation systems, drawing parallels from the music domain to enrich tour Package recommendations. Their method emphasizes feature representation and similarity computation to capture complex relationships between content features, enhancing recommendation accuracy and diversity. By learning feature representations from textual descriptions, audio features, and user interactions, their model generates personalized recommendations that cater to individual preferences and interests, elevating the relevance and quality of tour recommendations.

[7] Firdaus et al.'s comprehensive analysis of existing movie recommendation systems offers valuable insights into the challenges and opportunities in recommendation system research, with direct implications for the travel domain. Their study sheds light on key challenges such as data sparsity, cold start problems, and user diversity, providing a roadmap for addressing these obstacles in the development of robust tour recommendation systems. By identifying emerging trends and future directions in recommendation system research, their work offers invaluable guidance for advancing the field of tour recommendation, paving the way for more personalized and tailored travel experiences.

[9] Le and Pishva(2016) Developed a system that personalizes recommendations based on user preferences, interests, and past travel behavior. Padia et al. Focused on sentiment-aware recommendations, incorporating user opinions and emotions expressed in online reviews.

[10] U.M and Y.C (2021) Employed a hybrid model combining collaborative filtering and content-based filtering for place recommendations. Kulkarni et al. Utilized sentiment analysis to understand user opinions and tailor recommendations accordingly. Cumlievski et al. Employed Machine Learningˇ models to classify accommodation types based on online guest reviews.

[11] Priya and Rao (2023) Conducted an aspect-based sentiment analysis of online reviews to understand user experiences and preferences. Alharbi et al. Classified tourist review sentiment using deep learning techniques, potentially providing insights into user satisfaction.

III. PROPOSED METHODOLOGY

The objective of the proposed methodology is to develop a recommendation model tailored to user preferences. Our approach involves the construction of a recommendation model derived from the vectorized representation of a corpus. Figure 1 provides an overview of the workflow employed in our methodology. The subsequent subsections delineate the distinct components comprising the workflow.

A. Dataset

1) Data Collection: From the perspective of tourists, tours serve a dual purpose, extending beyond mere visits to places. They are intricately intertwined with specialized activities and features such as cultural immersion, exploration, and wildlife encounters in the search for package-based tours, considerations go beyond just duration and price; the unique characteristics of destinations, including cultural significance, opportunities for active exploration, and wildlife richness, are also taken into account. This research methodology is centered on processing the distinctive features of destinations within tour schedules. The initial step involves gathering textual documents related to tourism packages that encompass these attributes. After extensive browsing of various websites to collect data, a final selection of the data source was made, and web scraping techniques were employed to compile a dataset. It is imperative for this dataset to encapsulate the special characteristics of the destinations featured in the overall tour packages to facilitate the evaluation of the study.

Table 1. Dataset Attributes

Variable Name	Description	Data Type
Package Number	A unique identifier for each tour package	Integer
Package Name	The name/title of the tour package	String
Travel Style	The style or theme of the tour package	String
Operator	The company or entity organizing the tour package	String
Operated In	Languages in which the tour is conducted or supported	String
Price	The cost of the tour package	Integer
Days	Description of activities for each day of the tour	String
Introduction	A brief overview of the tour package	String
Reviews	Number of reviews per Package	Integer
Rating	The overall rating of the tour package based on reviews	Integer

2) *Data Cleaning:* In the context of data-driven studies, validating the dataset is imperative to ensure its reliability and suitability for analysis. This validation process involves scrutinizing the extracted textual information to ascertain whether they align with the research objectives. Specifically, packages lacking descriptions should be excluded from the dataset, as they may not provide sufficient information for meaningful analysis. Additionally, even if descriptions are present but exceedingly brief, they may not contribute significantly to understanding the package and can be omitted. Furthermore, it is essential to consider the possibility of incorrect package information, necessitating correction of structural errors within the dataset. Addressing missing data using various imputation methods is also crucial in ensuring dataset completeness. This validation step holds significant importance as it directly influences the effectiveness of the recommendation process.

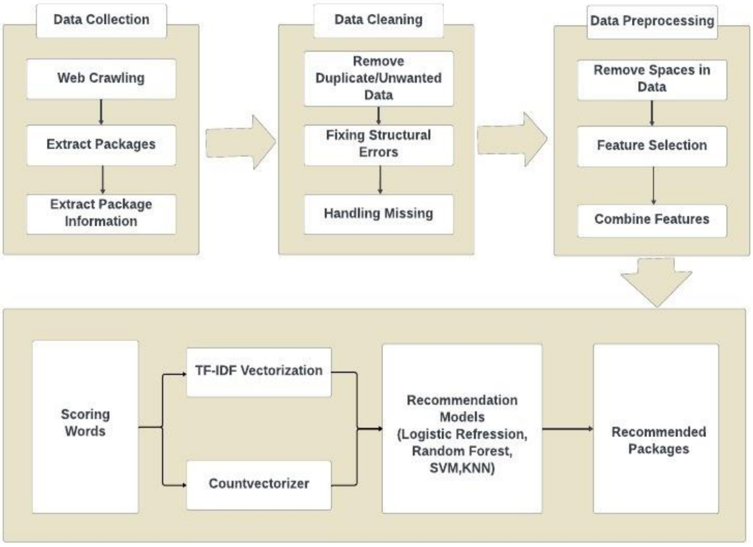


Fig. 1: Work flow of Proposed approach towards recommending Tour Packages

3) *Data Pre-processing:* During the pre-processing stage, documents undergo refinement to eliminate unwanted content about package information. This typically involves the removal of punctuation, alphanumeric characters, and stop words, as they do not contribute meaningfully to the analysis. Additionally,

superfluous spaces are eliminated to mitigate potential discrepancies in data interpretation. Feature selection is another pivotal aspect of this phase, wherein variables with a significant potential to impact the research outcome are identified and preserved. Integration of features may also be warranted to enhance the dataset's richness and facilitate the development of a more realistic recommendation model.

B. Scoring of words

The process of converting textual data into numerical format is crucial for utilizing them in machine learning algorithms. To achieve this, techniques such as Count-Vectorizer and TF-IDF vectorizer are employed. These techniques transform textual data into vectorized representations, enabling the combination of features and facilitating model building within machine learning algorithms. Both Count-Vectorizer and TF-IDF vectorizer play a vital role in quantifying documents, thereby enabling the integration of textual data into the analytical framework.

1) **Count-Vectorizer:** It tokenizes the input text corpus into individual words, phrases, or characters, and subsequently constructing a matrix where each row corresponds to a document and each column represents a unique token present in the corpus. The entries in this matrix denote the frequency of occurrence of each token within the respective document, effectively encoding the document's textual information into a structured numerical format.

2) **TF-IDF vectorization**: Given a set D of documents, TF-IDF scores the word w in the document d as Eq (1)

$$\text{TF-IDF}(w,d,D) = \text{TF}(w,d) \times \text{IDF}(w,D) \qquad (1)$$

where,

$$TF\,(w,d,D) = \frac{c\,(w,d)}{|d|} \qquad (2)$$

$$IDF\,(w,D) = \log\left(\frac{|D|}{|\{d \in D | w \in d\}|}\right) \qquad (3)$$

here c(w, d) is the number of times term w occurs in document d. Also, |*d*| and |*D*| denote the number of words in document d and the number of documents in corpus D, respectively.

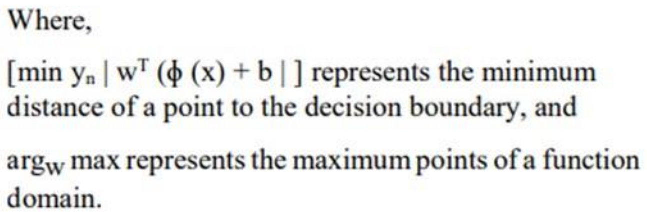
C. Model Development

1) ***Logistic Regression Classifier:*** The Logistic Regression (LR) algorithm serves as a valuable tool in binary classification tasks as well as in multi classification tasks, where the objective is to predict outcomes falling into one of two categories, typically denoted as '0' or '1'. Its mathematical formula leverages the logistic function to create a model that estimates the probability of an instance being associated with a specific class. The logistic function is given by:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x)}}$$

2) ***Support vector machine(SVM) Classifier:*** The formula of the Support Vector Machine (SVM) is designed to determine the best possible hyperplane for binary classification tasks. SVM strives to maximize the margin between data points belonging to distinct classes, thus ensuring a robust separation. This algorithm is renowned for its versatility and potency in classification, capable of effectively dealing with non-linear data patterns through the utilization of kernel functions. Consequently, SVM finds widespread utility in diverse applications, including image recognition and text classification.

$$w *= argmax[\min \mathrm{yn} \mid \mathrm{wT}\ (\phi\ (\mathrm{x})\ +\ \mathrm{b}\ |]$$



Where,

[min y_n | w^T (ϕ (x) + b |] represents the minimum distance of a point to the decision boundary, and

$\arg_W$ max represents the maximum points of a function domain.

3) ***Random Forest Classifier:*** The Random Forest algorithm is an ensemble technique employed in both classification and regression tasks. Its operation involves the creation of numerous decision trees during the training phase, and subsequently, the amalgamation of their predictions to yield a more resilient and precise outcome. Although there isn't a singular formula, the fundamental idea of the algorithm revolves around the amalgamation of multiple decision trees. Random Forest effectively mitigates over-fitting concerns by introducing an element of randomness into the process of constructing these trees, making it exceptionally adept at handling intricate, high-dimensional datasets.

4) ***K-Nearest Neighbor (KNN) Classifier:*** The k-Nearest Neighbors (KNN) algorithm categorizes data points by scrutinizing the 'k' closest neighbors within the training dataset. Its mathematical expression entails the computation of a data point's class by identifying the prevailing class among its 'k' nearest neighbors, typically determined using the Euclidean distance metric: Class=mode(closest neighbors' classes). KNN represents a straightforward yet efficient algorithm that finds application in both classification and regression tasks. It offers particular value when dealing with datasets exhibiting distinct clusters and is commonly harnessed in recommendation systems and image classification due to its user-friendly nature and adaptability.

$$\text{dist}((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2}$$

D. Performance Measure

Performance Measure such as Accuracy and Cross-Validation Score (CV Score) are commonly utilized metrics to evaluate the effectiveness of machine learning models.

- Accuracy: Accuracy measures the proportion of correctly predicted instances out of the total instances in the dataset. It is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \times 100$$

Accuracy provides a straightforward assessment of the model's overall correctness but may not always be the most reliable metric, especially in scenarios with imbalanced classes.

- Cross-Validation Score (CV Score): Cross-validation is a technique used to assess the performance of a model by splitting the dataset into multiple subsets, training the model on a subset, and testing it on the remaining subsets. CV Score typically refers to the average accuracy obtained across all folds in a cross-validation process. Cross-validation helps to mitigate issues such as over-fitting and provides a more robust estimate of the model's performance on unseen data. Common cross-validation techniques include k-fold cross-validation and stratified k-fold cross-validation. Both Accuracy and CV Score are essential for evaluating the performance of machine learning models, with Accuracy offering a direct measure of correctness and CV Score providing insights into the model's generalization capability.

E. Result

Table 2. Model Accuracy

Model Name	**Accuracy**	**CV Score**
KNN Classifier	68.7	69.2
SVM	84.8	85.1
Logistic Regression	86.2	77.9
Random Forest Classifier	89.4	89.3

In the evaluation of tour recommendation models, spanning accuracy and cross-validation score (CV Score), distinct patterns emerged among various algorithms. The Random Forest Classifier showcased the highest accuracy, boasting an impressive 89.4 Accuracy and a CV Score of 89.3. This exemplary performance can be attributed to the Random Forest's adeptness in managing high-dimensional data and intricate decision boundaries. By amalgamating predictions from multiple decision trees, the Random Forest model effectively combats over-fitting, rendering it proficient in generating robust recommendations for tour packages. Similarly, the Support Vector Machine (SVM) also exhibited commendable accuracy, registering an accuracy of 84.8 alongside a CV Score of 85.1. SVM's robustness lies in its capability to delineate complex decision boundaries by maximizing the margin between different classes in the feature space. This inherent ability enables SVM to excel in scenarios where data exhibits non-linearity and high dimensionality, contributing to its reliable performance in tour package recommendation.

Conversely, the Logistic Regression model displayed an accuracy of 86.2 with a CV Score of 77.9. While Logistic Regression is a versatile and interpretable algorithm, its performance may be hindered by the assumption of linear relationships between features and the target variable. In scenarios where data is highly nonlinear or contains complex interactions, Logistic Regression may struggle to capture nuanced patterns, potentially leading to slightly diminished performance compared to more flexible algorithms like Random Forest and SVM. KNN Classifier demonstrated relatively lower accuracy, recording an accuracy of 68.7 and a CV Score of 69.2. This diminished performance may stem from the model's simplistic approach, which relies on the similarity of instances to make predictions. Unlike the Random Forest Classifier, the KNN algorithm may struggle with high dimensional data and noisy features, leading to suboptimal performance in tour recommendation tasks where nuanced patterns are prevalent.

Additionally, the KNN Classifier's reliance on a single hyperparameter (k) for determining nearest neighbors could limit its flexibility in capturing complex relationships within the dataset. In summary, the Random Forest Classifier and SVM demonstrated robust performance in recommending tour packages, owing to their ability to handle complex data structures and nonlinear relationships effectively. Conversely, while Logistic Regression remains a valuable tool, its performance may be comparatively lower in datasets characterized by intricate patterns and high dimensionality.

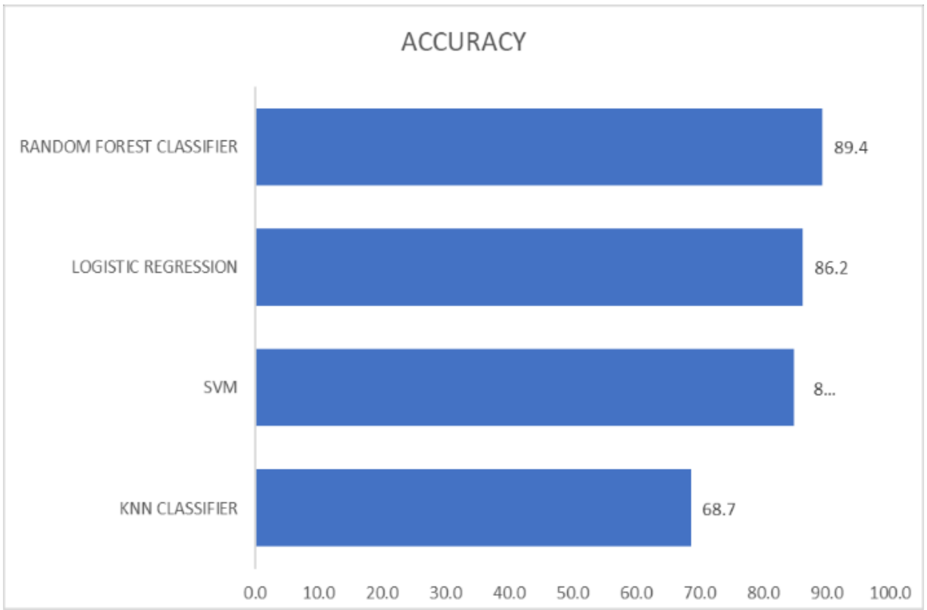


Fig. 2: Accuracy Comparison of Different models

IV. CONCLUSION AND FUTURE SCOPE

The implementation of a tour recommendation system utilizing machine learning techniques marks a significant advancement in the tourism industry. By leveraging a combination of different machine learning algorithms like K Nearest Neighbors (KNN), Logistic Regression, SVM, Random Forest and different vectorization methods like TF-IDF, count vectorizer, the proposed system effectively analyzes user preferences and textual data from tour packages. This allows for the generation of personalized suggestions tailored to user specified destinations and desired travel styles. Overall, the tour recommendation system represents a significant step forward in optimizing the tourism experience, offering travelers a seamless and personalized journey from discovery to booking.

As technology continues to evolve, such innovative solutions have the potential to reshape the landscape of travel, empowering individuals to embark on adventures perfectly tailored to their preferences and aspirations. Another potential direction for future work is Multi-modal Recommendations. Expanding the recommendation system to encompass various modes of transportation, accommodations, and activities allows for comprehensive trip planning. By considering factors like travel time, budget constraints, and traveler preferences across different facets of the journey, the system can offer holistic and seamless travel solutions. Scaling the recommendation system to cover a broader range of destinations worldwide enables travelers to explore diverse cultures and landscapes. Partnering with local tour operators and aggregating data from global sources can facilitate the inclusion of lesser-known destinations and niche travel experiences, catering to a wider audience of travelers.

REFERENCES

[1] O. R. Abbasi and A. A. Alesheikh, ”A Place Recommendation Approach Using Word Embeddings in Conceptual Spaces,” in IEEE Access, vol. 11, pp. 11871-11879, 2023, doi: 10.1109/ACCESS.2023.3241806.

[2] S. Anand, A. M. Abhishek Sai and M. Karthikeya, ”Chatbot Enabled Smart Tourism Service for Indian Cities: An AI Approach,” 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), Jaipur, India, 2023, pp. 1-7, doi: 10.1109/IEMECON56962.2023.10092286

[3] P. Stefanovic and S. Ramanauskaitˇ e, ”Travel Direction Recommen-˙ dation Model Based on Photos of User Social Network Profile,” in IEEE Access, vol. 11, pp. 28252-28262, 2023, doi: 10.1109/ACCESS.2023.3260103.

[4] S. Amara and R. R. Subramanian, ”Collaborating personalized recommender system and content-based recommender system using TextCorpus,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 105109, doi: 10.1109/ICACCS48705.2020.9074360.

[5] R. Gharibi, A. Safdel, S. M. Fakhrahmad and M. H. Sadreddini, ”A Content-Based Model for Tag Recommendation in Software Information Sites,” in The Computer Journal, vol. 64, no. 11, pp. 1680-1691, Nov. 2019, doi: 10.1093/comjnl/bxz144.

[6] Z. -S. Chen, J. -S. R. Jang and C. -H. Lee, ”A Kernel Framework for Content-Based Artist Recommendation System in Music,” in IEEE Transactions on Multimedia, vol. 13, no. 6, pp. 1371-1380, Dec. 2011, doi: 10.1109/TMM.2011.2166380.

[7] JOUR,Firdaus, Muhammad,Latt, Cho Aguilar, Mariz,Rhee, Kyung Hyune,2023/01/01,25,40,’A Prospective Extension Through an Analysis of the Existing Movie Recommendation Systems and Their Challenges’,12,DOI-10.3745/KTCCS.2023.12.1.25

[8] G. Linden, B. Smith and J. York, ”Amazon.com recommendations: itemto-item collaborative filtering,” in IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan.-Feb. 2003, doi: 10.1109/MIC.2003.1167344.

[9] Q. T. Le and D. Pishva, ”An innovative tour recommendation system for tourists in Japan,” 2016 18th International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea (South), 2016, pp. 717-729, doi: 10.1109/ICACT.2016.7423607.

[10] U. M and Y. C, ”COLPOUSIT: A Hybrid Model for Tourist Place Recommendation based on Machine Learning Algorithms,” 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1743-1750, doi: 10.1109/ICOEI51242.2021.9452746.

[11] Priya, P. S. and Malleswara Rao, N. N. . (2023). An Aspect based Sentiment Analysis of Tour and tour recommendation Approach using Machine Learning. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 754–762.

[12] Kulkarni, A.Barve, P. Phade, A. (2019). A machine learning approach to building a tourism recommendation system using sentiment analysis.
International Journal of Computer Applications, 178(19), 48-51.

[13] P. Padia, K. H. Lim, J. Cha and A. Harwood, ”Sentiment-Aware and Personalized Tour Recommendation,” 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 900-909, doi: 10.1109/BigData47090.2019.9006442.

[14] Cumlievski, N.; Brkiˇ c Bakari´ c, M.; Mateti´ c, M. A Smart Tourism´ Case Study: Classification of Accommodation Using Machine Learning Models Based on Accommodation Characteristics and Online Guest Reviews. Electronics 2022, 11, 913.
https://doi.org/10.3390/electronics11060913

[15] Alharbi, Banan A., Mohammad A. Mezher, and Abdullah M. Barakeh. ”Tourist reviews sentiment classification using deep learning techniques: A case study in saudi arabia.” International Journal of Advanced Computer Science and Applications 13.6 (2022).