Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Development of a Sentiment Analysis-Based Proposal Scheme for Tourist Destinations Using a Novel Random Forest and Support Vector Regression (RF-SVR)

Aseem Purohit[1], Jaspreet Sidhu[2], Bhawana Saraswat[3], Dr. Batani Raghavendra Rao[4], Yuvraj Parmar[5], Hansika Disawala[6]

[1]Professor of Practice, Department of Management Studies, Vivekananda Global University, Jaipur, India, Email Id- aseem.purohit@vgu.ac.in
[2]Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India, Email Id- jaspreet.sidhu.orp@chitkara.edu.in
[3]Scholar, Department of Computer Science & Engineering, Sanskriti University, Mathura, Uttar Pradesh, India, Email Id- bhawnasaraswatphd@sanskriti.edu.in
[4]Professor, Department of Finance, JAIN (Deemed-to-be Univesity), Bangalore, Karnataka, India, Email Id- brr@cms.ac.in
[5]Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh-174103 India, Email Id- yuvraj.parmar.orp@chitkara.edu.in
[6]Faculty, Department of ISME, ATLAS SkillTech University, Mumbai, Maharashta, India, Email Id- hansika.disawala@atlasuniversity.edu.in

Introduction: Tourism is a vital sector that relies on customer satisfaction and feedback; therefore examining tourists' opinions of various areas may benefit local governments and businesses. Decision-making systems in numerous sectors need sentiment analysis approaches.
Methods: This research describes a novel technique for sentiment analysis for tourist destinations that combines two Machine Learning (ML) approaches: Stochastic Random Forest-Dynamic Support Vector Regression (SRF-DSVR). This scheme attempts to deliver personalized location suggestions by analyzing traveler evaluations, collecting sightseeing reviews, ratings, and weather data for varied classifications, and improving sentiment analysis through data collection and augmentation. We use tokenization and stop word removal to clean and prepare the text data. The Term Frequency-Inverse Document Frequency (TF-IDF) approach converts text input into numerical vectors, allowing useful characteristics for ML algorithms to be extracted. The SRF-DSVR system successfully determines tourist preference uncertainty and variability by offering robust sentiment analysis and reacting to changing sentiment patterns, allowing the system to deliver current suggestions that correspond with current tourist preferences.
Results: The findings reveal that the SRF-DSVR combination is superior to standard sentiment analysis algorithms in f1-score of 94,2 %, accuracy of 95, 8 %, precision of 94, 9 %, and recall of

95, 3 %. We also performed a detailed assessment utilizing confusion matrices, which revealed the model's efficiency in sentiment classification tasks.
Conclusion: The research demonstrates the potential of the SRF-DSVR technique for sentiment analysis and suggestion schemes in tourist sites, therefore improving visitor experiences and allowing local governments and companies to make data-driven decisions.

Keywords: Tourist Destinations, Sentiment Analysis, Machine Learning (ML), Stochastic Random Forest-Dynamic Support Vector Regression (SRF-DSVR).

1. Introduction

Travel and tourism academics have always been interested in the effects and changes that occur in tourist destinations as it is well acknowledged that tourism is a dynamic phenomenon. Due to the tourist industry's economic significance and quick expansion, new destinations, attractions, and facilities are constantly being introduced as modern societies and global markets increasingly depend on them. [(1)] Plans and renovations are being made to certain locations to attract more travelers and foreign investment in the travel industry. [(2)] The needs of modern tourists and the growth of the travel industry might have a big influence on the economies of whole nations, making tourism and its development highly political and social undertakings. [(3)] The physical settings that we and "others" occupy and experience, and our views of other people and cultures are all impacted by growing tourism. [(4)] Since the beginning of modern tourism, exchanges of people, money, commodities, ideas, and values have made it possible for remote areas to work together to build their tourist industries while also increasing their interdependence. [(5)] Since the tourism industry accounts for 9.8 % of the global GDP and contributed $7.61 trillion to the world economy in 2016, it is an excellent instrument for promoting economic growth in nations that are industrialized. [(6, 7)] Technological developments significantly alter the tourist business by enabling players in the sector to create new markets, management approaches, and competitive tactics. [(8)] Sustainable development and information technology are two elements in smart tourist destinations. For tourists, consolidated and functionally integrated data indicates smart tourism places.[(9)] When it comes to trying to make their destinations more competitive in light of the trend toward smarter travel destinations, a deeper comprehension of the components of smart tourist destinations would be beneficial for leaders and organizers of travel.[(10)] Achieve sustainable tourism;in a similar vein, determining communities' long-term interests requires the participation of a wide variety of stakeholders in planning and decision-making. The study [(11)] proposed globally, and governments now consider the tourism sector among the most significant economic areas. Because it gives governments and other relevant businesses vital information and allows stakeholders to make necessary plans and policy adjustments, accurately anticipating tourist demand is essential. The article [(12)] described one important tactic for fostering rural rejuvenation in China that has long been acknowledged: rural tourism. The research [(13)] addressed the important touristic settlements have a dense spatial distribution with noticeable regional variations. The study [(14)] addressed the identities of tourist locations, and the changes they undergo have long piqued the attention of tourism scholars. This essay aims to advance the knowledge of research on

tourist destinations and their transformations in light of the "specialization of social theory" and previous conversations on locality studies in human geography.

The article [15] evaluated travel destinations as dynamic, historical entities with distinct identities defined by hegemonic and other discourses, all of which contribute to the idea of what the destination is and symbolizes at the moment. The study [16] analyzed that even though tourism is crucial for both driving regional development and serving as a competitive weapon, some occurrences happen outside of the control of tourist locations. The article [17] described the primary contribution of this research as the strategies and operational guidelines for managing tourist destinations after natural disasters. The research [18] addressed other things; the data let us determine that several separate natural catastrophes occur in nations throughout the world's continents, which has seriously detrimental effects on any destination's appeal to tourists. The study [19] addressed that the environment is changing more quickly in tourist areas. Nevertheless, current scientific knowledge is too stagnant, conceptual, or aggregate to be very helpful in real-world situations. The article [20] proposed that an essential step towards going beyond aggregate and static evaluations is the methodological tools offered by the methodology, which facilitate system integration and provide researchers and stakeholders with the chance to experience and experiment with dynamic vulnerabilities.

The goal of this study is to create a novel sentiment analysis tool, SRF-DSVR, for tourist sites that uses machine learning methods to deliver personalized location recommendations on the basis of traveler evaluations, ratings and reviews, and meteorological data.

2. Methodology

The Stochastic Random Forest- Dynamic Support Vector Regression (SRF-DSVR) method is used to determine the tourism destination. After gathering the dataset for the tourist destination, the research employed two pre-processing steps: tokenization and stop word removal. For feature extraction, Term Frequency Inverse Document Frequency (TF-IDF) is used. The figure 1 shows the system architecture of SRF-DSVR, which was used to collect the data about the tourist destinations.

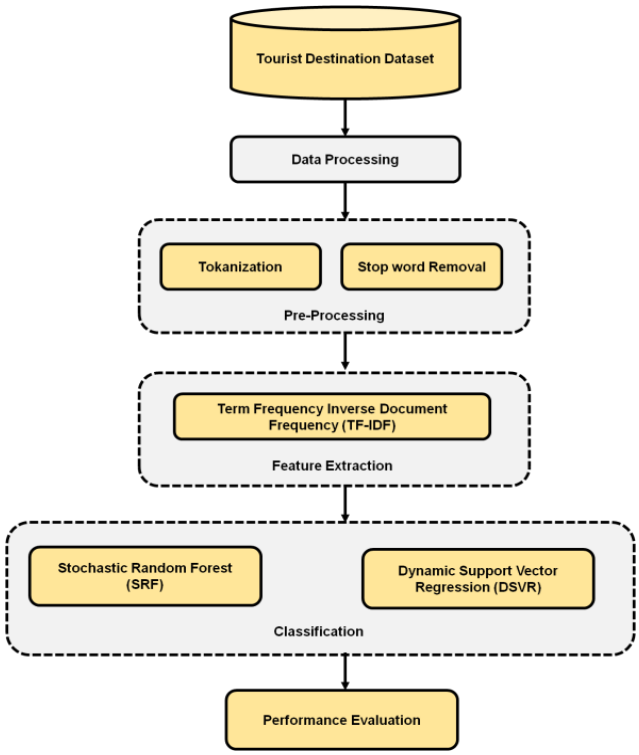

Figure 1.System Architecture [Source: Author]

Data Collection

The dataset was collected in India from a variety of sources. India Tourism Statistics (ITS), World Development Indicators (WDI), Statistical Year Book (SYB), Ministry of Civil Aviation in India. [22] Information gathered on the number of visitors visiting India for tourism. The Indian government's Ministry of Tourism publishes ITS every year, which documents and provides information on a variety of tourist-related concerns.

Preprocessing

An important step in the data mining technique is preprocessing the data. It explains the procedures for integrating, transforming, and cleaning data in order to get it ready for analysis. Enhancing the data's quality and adapting it to the particular data mining job is the aim of data preparation.

Tokenization

Tokenization separates words, phrases, representations, etc. Tokenizing a proposition evaluates its terms. Text parsing and mining use tokens. Tokenization helps language and CS. Original text is blocks. Data collection is needed for any information retrieval. Text tokenization for parsing. A machine-readable reader finds this easy. Adjust punctuation. Deal with brackets/hyphens. Tokenization includes document consistency. Find meaningful phrases are tokenization's key goal. Time and number formats vary. Standards

for acronyms and abbreviations are another matter.

Stop Word Removal

Stop words split natural language. Stop words make text heavier and less important to analysts, so remove them. Eliminating stop words reduces space's dimension. The most common words in text documents that don't communicate the meaning include articles, prepositions, pronouns, etc. These phrases are illegal. Stop words are not keywords, thus text mining systems ignore them.

Features Extraction Using TF-IDF

The AI extracts features. Word uses natural language processing. Name also means token. Vector files for machine-learning classifiers. Characterize tokens. Different methods can extract textual features. Vectorization typically employs TF-IDF/Count. Paper and corpus word frequency are used in TF-IDF vectorization to weight texts. Document word matrix as record. The equations (1-3) compute TF-IDF:

$$\mathrm{TF} = \frac{\mathrm{Noodtimesparticularword\,,u'occursinadocument}}{\mathrm{Totalnoofwordsinadocument}} \quad (1)$$

$$\mathrm{IDF} = \log\left(\frac{\mathrm{Totalnoofdocument}}{\mathrm{ofdocumentscontainsparticularWord}}\right) \quad (2)$$

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} \quad (3)$$

Equation 1 calculates word frequency. Equation 2 calculates Word inverse document frequency. Equation 3 calculates a word's phrase frequency or inverse document frequency score.

3. Classification Using Stochastic Random Forest- Dynamic Support Vector Regression (SRF-DSVR)

A machine learning model that combines the advantages of both stochasticity for enhanced variety and dynamic adaptation to shifting patterns or trends of one was to combine stochastic random forest with dynamic support vector regression.

Stochastic Random Forest (SRF)

A prominent machine-learning method Breiman's binary decision tree ensemble is used. Bagging developed an SRF algorithm training decision tree. Unsupervised learning, regression, and classification employ SRF. SRF minimizes generalization error more than other machine learning approaches. Since statistical approaches fail for huge data sets, we used the SRF model's Mean Decrease Accuracy (MDA) index to estimate variable relevance. MDA index and equation 4 determined variable importance.

$$\mathrm{VI_i} = \frac{1}{\mathrm{ntree}} \sum_{\mathrm{d=1}}^{\mathrm{ntree}} \mathrm{AB_{di}} - \mathrm{A_{di}} \quad (4)$$

Where $\mathrm{A_{di}}$shows the OOB (out-of-bag) problem on node t before rearranging the numbers of$\mathrm{Y_i}$, or $\mathrm{AB_{di}}$indicates the OOB error on tree t after permuting the values of$\mathrm{Y_i}$. VI denotes

the relative significance of variables.

Dynamic Support Vector Regression (DSVR)

A supervised machine learning algorithm created DSVR. The complex control function and system architecture were built utilizing this technology. DSVR optimizes nominal margin using regression. Complex training datasets are frequent for DSVR models. It curves many edges. DSVR shows input-output correlations using the structural risk minimization (SRM) norm. Equations (5) and (6) compute DSRM, an important DSVR model phase.

$$x = r(h) = c\emptyset(h) + v \tag{5}$$

Where $x_p \in K^1$displays the resulting value represents the input data. Furthermore, in each model, V denotes the data size, $v \in K^1$, Shows the continuous number for the based on numbers function. The algebraic function's constant number and $c \in K^1$indicate the weightage factor. The irregular process used to map the input dataset is represented by ∅ (h). The following equation, constructed using SRM may declare v and c:

$$\text{Minimize: } \left[\frac{1}{2}\|c\|^2 + b\sum_{p=1}^{1}(\zeta_p + \zeta_p^*)\right]$$

$$\text{Subjectto: } \begin{cases} x_p - (c\phi(h_p) + v_p) \le \varepsilon + \zeta_p \\ (c\phi(h_p) + v_p) - x_p \le \varepsilon + \zeta_p^* \\ \zeta_p, \zeta_p^* \ge 0 \end{cases} \tag{6}$$

Where, ζ_p, ζ_p^*denotedynamically variable, ε signifies the model's optimal execution, and their penalty factor p, strikes a compromise between the model's flatness and risk. The Lagrangian function was used to solve the optimization issue using the following equation (7-8):

$$,\zeta_p, \zeta_p^*, \beta_p, \beta_p^*, \delta_p, \delta_p^*)$$

$$= \frac{1}{2}\left|c|^2 + B\sum_{p=1}^{1}(\zeta_p + \zeta_p^* - \sum_{p=1}^{1}\beta_p(\zeta_p + \varepsilon - x_p + c\emptyset(h_p h) + v\right)$$

$$-\sum_{p=1}^{1}\beta_p(\zeta_p^* + \varepsilon - x_p + c\emptyset(h_p h) + v) - \sum_{p=1}^{1}(\delta_p\zeta_p + \zeta_p^*\delta_p^*) \tag{7}$$

Where, δp, δ∗p, βp, and β∗p are the Lagrangian multipliers. Thus, SVR may be computed using:

$$R(h) = \sum_{p=1}^{1}(\beta_p - \beta_p^*)n(h, x_p) + v \tag{8}$$

Where, the expression of the kernel function is $n(h, x_p) = \langle\emptyset(h), \emptyset(h_p)\rangle$.

4. Result and Discussion

In this study we utilized Python 3.11 for our investigation. Core i7 laptops with 32GB

SSDs run Windows 10. Including industry-wide decision-making systems, many applications need sentiment analysis. Tourism sentiment analysis employing two ML algorithms is innovative with Stochastic Random Forest-Dynamic Support Vector Regression (SRF-DSVR). SRF-DSVR outperforms generic sentiment analysis in accuracy, precision, recall, and F1 score. Tourist destinations may employ SRF-DSVR for sentiment analysis and recommendation systems to improve visitor experiences and assist local governments and businesses make data-driven choices.

Accuracy

Performance measures should focus on accuracy, which is the ratio of correctly predicted observations to all seen. See Table 1 and Figure 2 for accuracy comparisons. Multiple classification algorithms were tested, including CV+NB [(21)] at 82,71 %, TFIDF+SVM[(21)] at 80,21 %, and CV+RF[(21)] at 85,51 %. A unique approach dubbed SRF-DSVR worked well with 95,8 % accuracy. The above data demonstrate the advantages of the proposed SRF-DSVR strategy over standard approaches and its potential for improved performance within the study aim.

Table 1. Numerical Outcomes of Accuracy

Methods	Accuracy (%)
CV+NB	82,71
TFIDF+SVM	80,21
CV+RF	85,51
SRF-DSVR (Proposed)	95,8

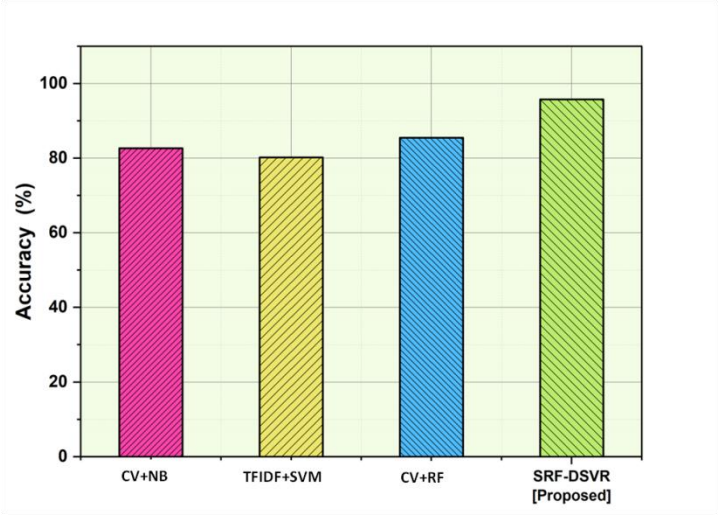

Figure 2.Comparison of Accuracy (Source: Author)

Precision

It is a positive predictive value. Table 2 and Figure 3 show the precision algorithm comparison. CV+NB had 82 % accuracy, TFIDF+SVM 79 %, and CV+RF 85 % in classification precision study. The anticipated SRF-DSVR was 94,9 % more accurate.

These findings suggest SRF-DSVR can accurately anticipate and categorize data, decreasing false positives.

Table 2. Numerical Outcomes of Precision (source: author)

Methods	Precision (%)
CV+NB	82
TFIDF+SVM	79
CV+RF	85
SRF-DSVR [Proposed]	94,9

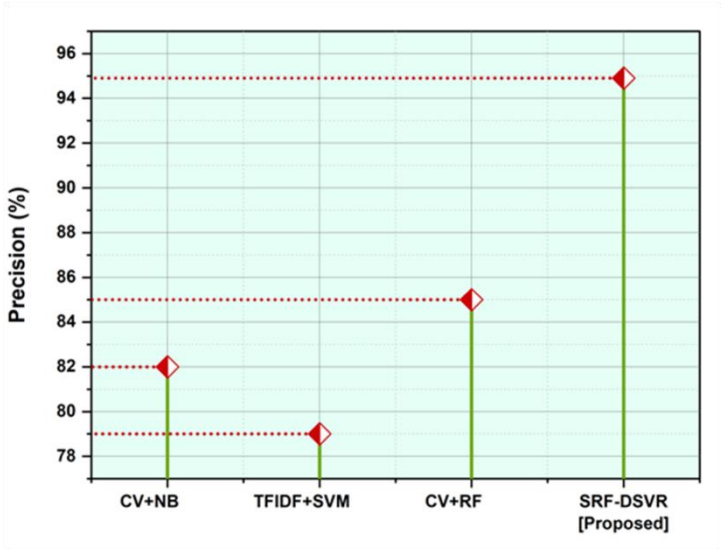

Figure 3.Comparison of Precision (Source: Author)

Recall

Table 3 and Figure 4 compare techniques. TFIDF+SVM 80 %, CV+NB 83 %, CV+RF 86 %. SRF-DSVR recall estimated 95,3 %. The SRF-DSVR discovers suitable dataset instances. Due to its high recall rate, the model may benefit from extensive coverage and minimal false negatives. It also shows how well the model handles multiple positives.

Table 3.Numerical Outcomes of Recall (Source: Author)

Methods	Recall (%)
CV+NB	83
TFIDF+SVM	80
CV+RF	86
SRF-DSVR [Proposed]	95,3

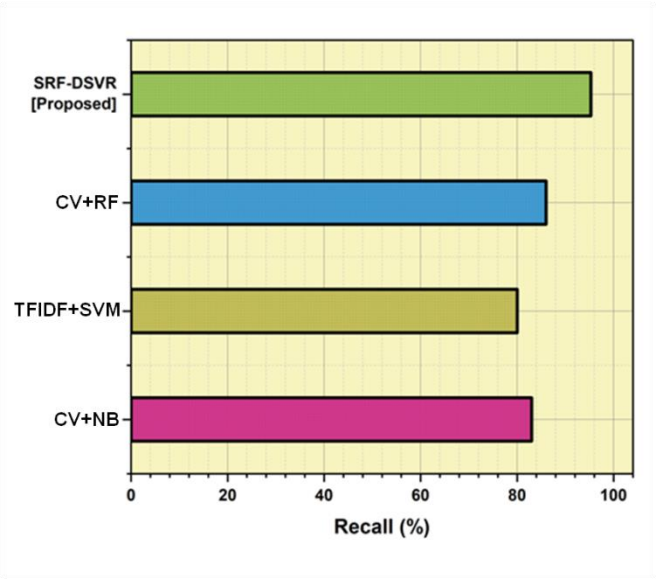

Figure 4. Comparison of Recall (Source: Author)

F1-Score

F1-Score balances accuracy and memory. Compare method F1-Scores in Figure 5 and Table 4. TFIDF+SVM, CV+NB, and CV+RF had 82 %, 84 %, and 84 % F1-Scores. SRF-DSVR F1-Score may raise 94,2 %. Model recall and accuracy are good. High F1-Score makes SRF-DSVR a good classification model for broad coverage and accurate predictions.

Table 4. Numerical Outcomes of F1-Score [Source: Author]

Methods	F1-Score (%)
CV+NB	82
TFIDF+SVM	84
CV+RF	84
SRF-DSVR [Proposed]	94,2

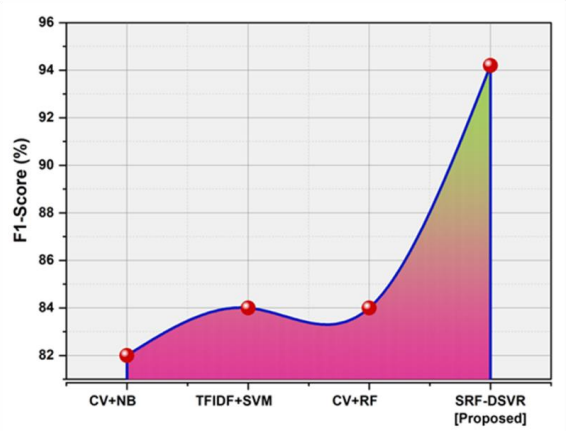

Figure 5.Comparison of F1-Score (Source: Author)

5. Conclusion

The Stochastic Random Forest- Dynamic Support Vector Regression (SRF-DSVR) beats TFIDF Vectorization in feature extraction, according to the research. SRF-DSVR's better classification accuracy makes TFIDF worth the extra time, according to studies. Significantly, SRF-DSVR outperformed rival classification algorithms in many assessment parameters. SRF-DSVR's f1-score of 94,2 %, accuracy of 95,8 %, precision of 94,9 %, and recall of 95,3 % demonstrated its robustness and dependability. Even though it requires computational trade-offs, complex classification algorithms like SRF-DSVR improve results. This emphasizes the importance of accuracy and efficiency while picking a research algorithm. The Sentiment Analysis-Based Proposal Scheme for Tourist Destinations may benefit from improving the Random Forest and Support Vector Regression (RF-SVR) model. Ensemble approaches and deep learning architectures may improve prediction accuracy. Additional data sources like social media or real-time reviews may strengthen the model.

References

1. Njoya ET, Seetaram N. Tourism contribution to poverty alleviation in Kenya: A dynamic computable general equilibrium analysis. Journal of travel research. 2018 Apr;57(4):513-24. DOI: https://doi.org/10.1177/0047287517700317.
2. Sotiriadis M. Tourism destination marketing: Academic knowledge. Encyclopedia. 2020 Dec 23;1(1):42-56. DOI: https://doi.org/ 10.3390/encyclopedia1010007.
3. Wei C, Dai S, Xu H, Wang H. Cultural worldview and cultural experience in natural tourism sites. Journal of Hospitality and Tourism Management. 2020 Jun 1;43:241-9. DOI: https://doi.org/10.1016/j.jhtm.2020.04.011.
4. Theocharous AL, Zopiatis A, Lambertides N, Savva CS, Mansfeld Y. Tourism, instability and regional interdependency: Evidence from the Eastern-Mediterranean. Defence and Peace Economics. 2020 Apr 2;31(3):245-68. DOI: https://doi.org/10.1080/10242694.2018.1501531.
5. Shafiee S, Ghatari AR, Hasanzadeh A, Jahanyan S. Developing a model for sustainable smart tourism destinations: A systematic review. Tourism Management Perspectives. 2019 Jul 1;31:287-300. DOI: https://doi.org/10.1016/j.tmp.2019.06.002.
6. Appiah, M., Gyamfi, B.A., Adebayo, T.S. and Bekun, F.V., Do financial development, foreign direct investment, and economic growth enhance industrial development? Fresh evidence from Sub-Sahara African countries. Portuguese Economic Journal, 2023. 22(2), pp.203-227. DOI: https://doi.org/10.1007/s10258-022-00207-0.
7. Xu, J., She, S., Gao, P. and Sun, Y.Role of green finance in resource efficiency and green economic growth. Resources Policy, 2023. 81, p.103349. DOI: https://doi.org/10.1016/j.resourpol.2023.103349.
8. Lee P, Hunter WC, Chung N. Smart tourism city: Developments and transformations. Sustainability. 2020 May 12;12(10):3958. DOI: https://doi.org/10.3390/su12103958.
9. Bhuiyan KH, Jahan I, Zayed NM, Islam KM, Suyaiya S, Tkachenko O, et al. Smart Tourism Ecosystem: A New Dimension toward Sustainable Value Co-Creation. Sustainability. 2022 Nov 14;14(22):15043. DOI: https://doi.org/10.3390/su142215043.
10. Roxas FM, Rivera JP, Gutierrez EL. Mapping stakeholders' roles in governing sustainable tourism destinations. Journal of Hospitality and Tourism Management. 2020 Dec 1;45:387-98. DOI: https://doi.org/10.1016/j.jhtm.2020.09.005.
11. Liu HH, Chang LC, Li CW, Yang CH. Particle swarm optimization-based support vector

regression for tourist arrivals forecasting. Computational Intelligence and Neuroscience. 2018 Sep 19;2018. DOI: https://doi.org/10.1155/2018/6076475.

12. Du X, Wang Z, Wang Y. The spatial mechanism and predication of rural tourism development in China: a random forest regression analysis. ISPRS International Journal of Geo-Information. 2023 Aug 2;12(8):321. DOI: https://doi.org/10.3390/ijgi12080321.
13. Femenia-Serra F, Gretzel U. Influencer marketing for tourism destinations: Lessons from a mature destination. InInformation and Communication Technologies in Tourism 2020: Proceedings of the International Conference in Surrey, United Kingdom, January 08–10, 2020 2020 (pp. 65-78). Springer International Publishing. DOI: https://doi.org/10.4398/encyclopedia1010087.
14. Croes R, Ridderstaat J, Bąk M, Zientara P. Tourism specialization, economic growth, human development and transition economies: The case of Poland. Tourism Management. 2021 Feb 1;82:104181. DOI: https://doi.org/10.1016/j.tourman.2020.104181.
15. Duan Z, Zhang K, Chen Z, Liu Z, Tang L, Yang Y, et al. Prediction of city-scale dynamic taxi origin-destination flows using a hybrid deep neural network combined with travel time. IEEE Access. 2019 Sep 6;7:127816-32. DOI: https://doi.org/10.1109/ACCESS.2019.2939902.
16. Estevão C, Costa C. Natural disaster management in tourist destinations: a systematic literature review. European Journal of Tourism Research. 2020 May 1;25:2502-. DOI: https://doi.org/10.54055/ejtr.v25i.417.
17. Naeem N, Rana IA. Tourism and disasters: a systematic review from 2010–2019. Journal of Extreme Events. 2020 Mar 27;7(01n02):2030001. DOI: https://doi.org/10.1142/S234573762030001X.
18. Rosselló J, Becken S, Santana-Gallego M. The effects of natural disasters on international tourism: A global analysis. Tourism management. 2020 Aug 1;79:104080. DOI: https://doi.org/10.1016/j.tourman.2020.104080.
19. Student J, Lamers M, Amelung B. A dynamic vulnerability approach for tourism destinations. Journal of Sustainable Tourism. 2020 Mar 3;28(3):475-96. DOI: https://doi.org/10.1080/09669582.2019.1682593.
20. Spencer N, Strobl E, Campbell A. Sea level rise under climate change: Implications for beach tourism in the Caribbean. Ocean & Coastal Management. 2022 Jun 15;225:106207. DOI: https://doi.org/10.1016/j.ocecoaman.2022.106207.
21. Wadhe AA, Suratkar SS. Tourist place reviews sentiment classification using machine learning techniques. In2020 international conference on Industry 4.0 Technology (I4Tech) 2020 Feb 13 (pp. 1-6). IEEE. DOI:- https://doi.org/10.1109/I4Tech48345.2020.9102673.
22. Barman H, Nath HK. What determines international tourist arrivals in India?. Asia Pacific Journal of Tourism Research. 2019 Feb 1;24(2):180-90. DOI: https://doi.org/10.1080/10941665.2018.1556712.