TourCast: A Machine Learning Framework for Demand Inference from User Reviews and Activity

Cherala Sai Kiran[1], G. Neeraja[2], Rekha Gangula[3*], Dumpala Varshitha[1], Amaragonda Ganesh[1], Kore Kalyan[1]

[1]UG Student, [2]Assistant Professor, [3]Associate Professor and Head, [1,2,3]Department of Computer Science and Engineering (AI&ML)

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: Rekha Gangula (gangularekha@gmail.com)

ABSTRACT

The rapid expansion of digital platforms in the tourism sector has resulted in a massive accumulation of unstructured user-generated reviews, making it challenging to extract meaningful insights for accurate demand prediction. Conventional classifiers such as Linear Discriminant Analysis (LDA), Histogram Gradient Boosting (HGB), and Quadratic Discriminant Analysis (QDA) are limited in capturing deep contextual relationships and struggle to perform effectively under data imbalance conditions. To address these limitations, this study proposes a hybrid framework named GPS Tourism, which integrates advanced language modeling with robust data balancing techniques. The framework leverages Google Pathways Language Model (PaLM) to convert raw textual reviews into 768-dimensional semantic embeddings, enabling a comprehensive understanding of contextual sentiment. To mitigate class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, ensuring balanced representation of minority classes. The final classification is performed using a Sparse Linear Integer Model (SLIM), structured as an ensemble of oblique decision trees to enhance both interpretability and prediction accuracy. Experimental results demonstrate that the proposed model outperforms baseline approaches. The framework enables precise demand forecasting, supporting data-driven decision-making for resource allocation and targeted marketing in the tourism industry.

Keywords: demand prediction, natural language processing, sentiment analysis, tourism analytics, data analytics

1. Introduction

Tourism stands as a rapidly evolving sector within the global economy, playing a crucial role in driving economic development, creating employment opportunities as shown in Figure 1, and fostering cultural interactions. The emergence of digital platforms, including online booking systems, travel communities, and review-driven service portals, has significantly reshaped how tourism operates [1,2]. Modern travelers heavily depend on digital reviews and ratings to assess accommodation, transportation, dining, and travel experiences before making decisions. As a result, these platforms continuously generate large-scale unstructured textual data in the form of feedback, reviews, and ratings. Such data encapsulates valuable information regarding customer satisfaction, service performance, and destination appeal, making it highly relevant for tourism demand prediction and strategic decision-making [3, 4].

In earlier stages, tourism forecasting methods primarily relied on statistical models such as time-series analysis, econometric techniques, and regression-based approaches. These methods were effective in analyzing structured numerical data and identifying historical trends. However, they exhibit limitations when dealing with unstructured textual content, particularly in capturing contextual meaning and sentiment embedded in user-generated reviews. With the increasing dominance of digital content,

tourism analytics have shifted toward data-driven methodologies that can process and interpret textual information more effectively.

The surge in digital tourism data necessitates advanced analytical frameworks that integrate semantic understanding with statistical learning mechanisms [5, 6, 7]. Recent developments in large-scale language models and representation learning techniques have enabled the transformation of raw textual data into dense semantic vectors. These representations preserve contextual relationships between words, allowing systems to better understand sentiment, user intent, and experiential nuances compared to traditional text analysis methods [8, 9, 10].

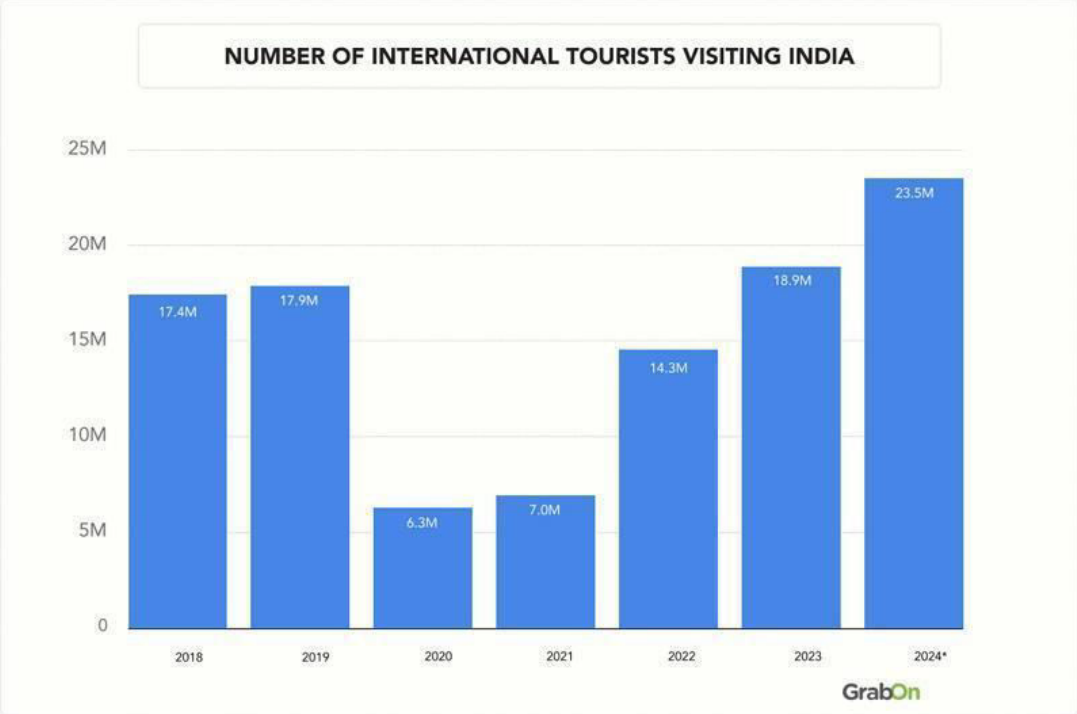


Fig. 1: Number of International tourists visiting India. (Source: India's Tourism Ministry Report)

2. Literature Survey

Khaidi et al. [11] conducted a comprehensive review of tourism demand forecasting literature published between 2010 and 2018, categorizing the diverse array of explanatory variables and modeling techniques utilized in the field. The study analyzed the influence of traditional economic indicators, such as tourist income, exchange rates, and Gross Domestic Product (GDP), on global travel trends. Mikhailov et al. [12] introduced the "digital pattern of life" (DPoL) concept to simplify the construction and application of complex tourist behavior models. By defining general behavioral concepts that bridge the gap between physical actions and digital footprints, the authors developed a framework to track shifts in tourist preferences over time.

Yang et al. [13] expanded the literature on tourism demand forecasting by investigating the predictive power of segmented Baidu search volume data. The researchers categorized search volumes based on source (PC vs. mobile) and temporal periods to capture the dynamic characteristics of tourist behavior across different digital environments. Li et al. [14] explored the integration of big data analysis with tourism consumer demand forecasting to address the management pressures caused by the rapid growth of the tourism sector. The study systematically identified key indicators affecting consumer behavior by analyzing data from major tourism websites and existing research databases.

Yu et al. [15] addressed the economic impact of unsold inventory in the hospitality and events sectors by developing an advanced machine learning framework for tourism demand forecasting. The study introduced the SAE-LSTM model, which enhanced standard Long Short-Term Memory networks by integrating Stacked Autoencoders (SAE). Nguyen et al. [16] examined the rapid expansion and subsequent pandemic-driven volatility of Vietnam's tourism sector, utilizing Artificial Neural Network (ANN) methodology to forecast international tourist arrivals. The study utilized a comprehensive dataset spanning from 2008 to 2020, intentionally including the extreme lockdown periods caused by the COVID-19 pandemic to test the resilience of the neural architecture.

Ma et al. [17] indicated that the Long Short-Term Memory (LSTM) model outperformed traditional methods in capturing long-term search trends and identifying distinct behavioral patterns. Furthermore, the clustering analysis successfully categorized tourists into three distinct groups based on their search characteristics, providing a technical foundation for precise market segmentation and real-time demand monitoring. Zhang et al. [18] developed a state-of-the-art LSTM model to analyze how search behaviors on personal computers (PCS) and mobile phone searches (MPS) influenced prediction accuracy. Their experimental results indicated that the inclusion of big web data significantly optimized model parameters, reducing the RMSE by 13.51% and MAE by 16.1%.

He et al. [19] explored the integration of traditional tourism theories with modern computational techniques by re-examining the push-pull theory through big data analysis and tree-based machine learning models. Hu et al. [20] addressed the inherent volatility of tourism demand caused by complex seasonality and black-swan events by developing a compound pattern recognition framework. This innovative approach dynamically integrated calendar patterns with tourism demand volume patterns to provide high-frequency daily forecasts. Wei et al. [21] addressed the limitations of traditional forecasting models in capturing multi-level temporal correlations and managing complex external variables by introducing the Tourism Demand Predictor with Two-Stage Feature Selection and Attention-Augmented Mechanisms (TFS-AAM).

Rekha Gangula et al. [22] proposed a conceptual framework for understanding machine learning in Artificial Intelligence (AI). The study analyzed various algorithms and their applications. The framework provided insights into AI system design. Rekha Gangula et al. [23] proposed an analysis of machine learning algorithms in data mining applications. The study evaluated algorithm performance across datasets. The framework improved knowledge discovery processes. Lingala Thirupathi et al. [24] proposed a false news recognition system using machine learning techniques. The framework extracted textual features from news data. The classifier identified misinformation with improved accuracy.

3. Proposed System

The system architecture represents an end-to-end intelligent tourism analytics pipeline that integrates data ingestion, NLP-based preprocessing, semantic feature extraction, and hybrid classification for demand prediction. Initially, user-provided datasets are processed through a GUI-based interface, followed by text cleaning and linguistic normalization. The architecture then transforms textual data into dense semantic embeddings using advanced language models and applies SMOTE to address class imbalance. Multiple baseline models such as QDA, LDA, and HGB are evaluated alongside the proposed GPS model, which leverages a SLIM-based hybrid classifier for improved accuracy. The processed features and encoded labels are used for training and evaluation through metrics like accuracy, precision, recall, and F1-score, including ROC analysis. The final stage involves visualization of results and deployment of predictions through an interactive interface for real-time usage, as illustrated in Fig 1, ensuring a complete pipeline from raw data to actionable insights.

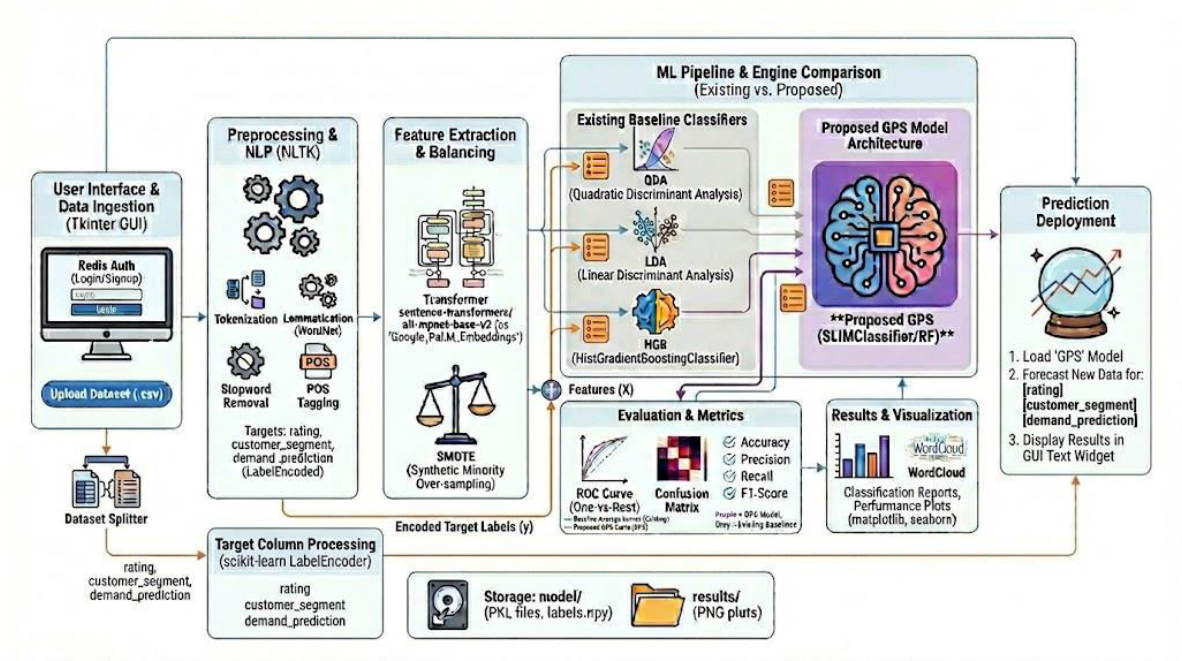


Fig. 4.1: Proposed GPS-Tourism system architecture.

Phase I: Textual Preprocessing and Semantic Encoding: The initial stage focuses on transforming raw tourist reviews into structured representations by reducing linguistic complexity. The system applies normalization techniques including lowercasing, noise removal, and stopword elimination. Lemmatization is performed to convert words into their base forms, ensuring semantic consistency across the dataset. The processed text is then converted into dense semantic vectors for meaningful representation.

- **Normalization:** Cleaning text by removing noise, special characters, and stopwords to standardize input data.
- **Lemmatization:** Converting words into their root forms to maintain uniformity in meaning.
- **Semantic Vectorization:** Using PaLM embeddings to generate 768-dimensional contextual representations.
- **Mean Pooling:** Aggregating token embeddings into a fixed-length document vector for each review.

Phase II: Feature Fusion and Data Rebalancing: This phase focuses on enriching the feature space by combining textual embeddings with structured behavioural data. The integration of multiple feature types improves the model's ability to capture diverse patterns in tourist behavior. To address class imbalance, synthetic data generation techniques are applied to ensure fair learning across all categories.

- **Feature Fusion:** Combining semantic vectors with numerical features such as spending patterns and duration of stay.
- **Data Imbalance Handling:** Applying SMOTE to balance minority and majority classes.
- **Synthetic Sample Generation:** Creating new samples using nearest neighbor interpolation to improve class representation.
- **Decision Boundary Enhancement:** Strengthening classification performance by improving minority class learning.

Phase III: The SLIM Classifier Optimization: The predictive engine is centered on the SLIM classifier. While traditional SLIM architecture prioritizes discrete feature selection, this implementation adapts the logic into an ensemble of oblique decision trees.

- **Model Integration:** The SLIM model acts as the "Proposed" architecture, aggregating predictions from multiple decision paths to minimize variance.
- **Multi-Output Learning:** The model is trained in helping to solve three interdependent tasks:
    - **Rating Classification:** Predicting numerical satisfaction levels.
    - **Customer Segmentation:** Categorizing tourists based on behavioural clusters.
    - **Demand Forecasting:** Estimating future resource requirements.

Phase IV: Evaluation Framework: The final phase evaluates the effectiveness of the proposed system through comparative analysis with baseline models. A structured validation strategy is used to ensure reliable performance measurement across all prediction tasks.

- **Baseline Comparison:** Evaluating performance against QDA, LDA, and HGB models.
- **Data Splitting:** Using stratified train-test split to maintain class distribution.
- **Performance Metrics:** Measuring Accuracy, Precision, Recall, and F1-Score for validation.
- **Result Validation:** Confirming improved robustness and predictive capability of the GPS-Tourism model.

Phase V: Deployment and User Interaction: The final stage focuses on deploying the trained model into a user-accessible environment using a Tkinter-based GUI integrated with Redis for authentication and session handling. This layer enables seamless interaction between users and the backend ML pipeline. The system allows users to upload datasets, trigger predictions, and view results in real time through a desktop interface.

- **GUI Interface:** Tkinter provides a desktop-based interface for dataset upload, input handling, and result visualization.
- **Authentication System:** Redis is used for secure login and signup functionality, managing user credentials and sessions.
- **Backend Integration:** The trained GPS-Tourism model is loaded to process incoming data and generate predictions.
- **Real-Time Prediction:** Outputs such as rating, customer_segment, and demand_prediction are generated dynamically.
- **Result Display:** Predictions and insights are presented within the GUI using interactive elements like text widgets and plots.
- **Model Persistence:** Trained models and encoders are stored and reused for consistent and efficient inference.

3.1 Inference Engine

The inference engine as shown in Figure 2 of the GPS-Tourism is the final operational layer where the trained PaLM, SMOTE, and SLIM components are utilized to provide real-time predictions for new, unseen tourist data. It is designed to be a "plug-and-play" module that takes raw user input and outputs high-precision demand forecasts.

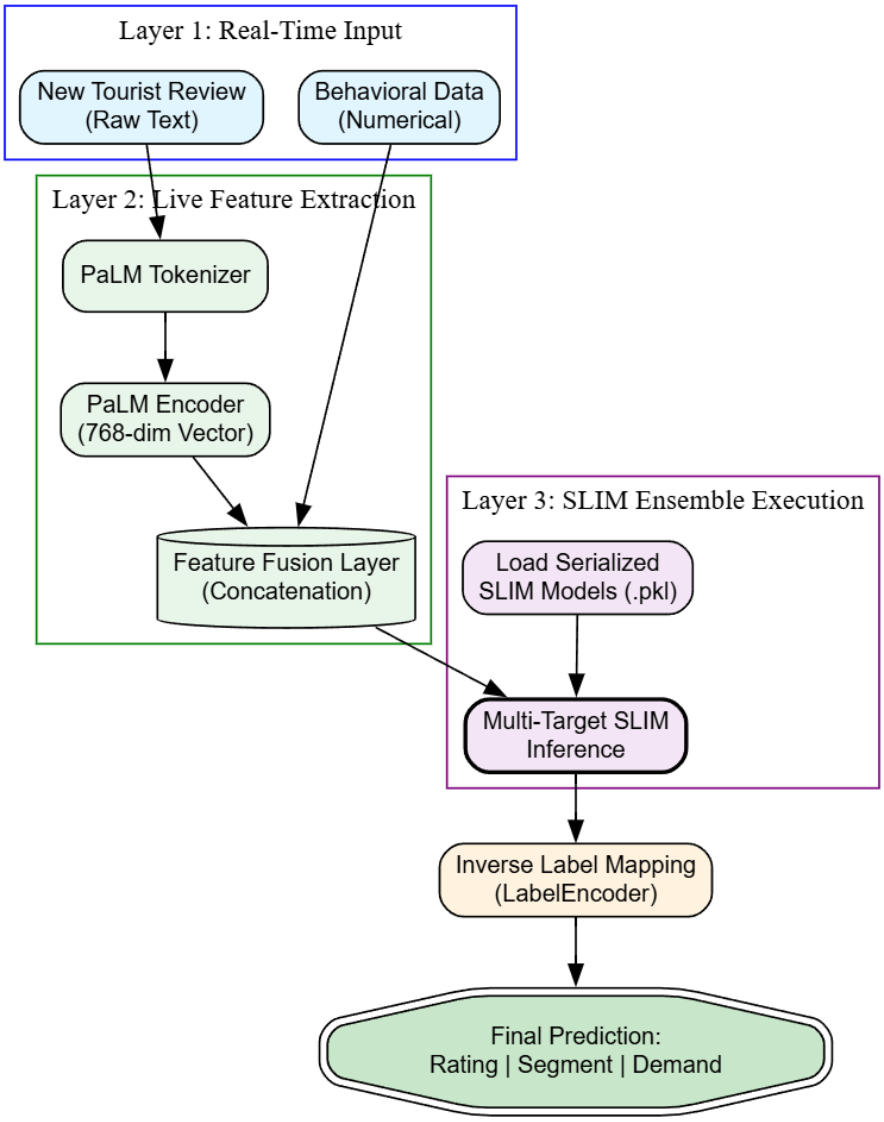


Fig. 2: Proposed inference engine of the GPS-Tourism framework.

The inference engine operates in a feed-forward manner, bypassing the SMOTE balancing layer (which is only used during training) to ensure that real-world predictions are based on the actual distribution of the new data.

Phase 1: Real-Time Data Ingestion: The engine accepts two types of inputs from the Tkinter interface:

1. **Unstructured Input:** A new customer review or a batch of reviews via a CSV file.
2. **Structured Input:** Corresponding behavioral indicators such as "Customer Segment" or "Previous Ratings" if available.

Phase 2: Live Feature Extraction (Google PaLM): Unlike the training phase, which uses cached features, the inference engine performs a "Live Pass" through the PaLM encoder:

- The input text is tokenized and mapped to the 768-dimensional latent space.
- **Mean Pooling** is applied to generate a single document vector representing the current tourist's sentiment.

- **Feature Alignment:** The numerical indicators are normalized and concatenated to the PaLM vector to match the exact input shape used during the SLIM training phase.

Phase 3: Multi-Target Prediction (SLIM Ensemble): The fused feature vector is passed through the serialized SLIM models (loaded from .pkl files):

- The engine runs parallel predictions for Rating, Customer Segment, and Demand Prediction.
- Each target utilizes its specific SLIM oblique tree ensemble to calculate the class probabilities.

Phase 4: Label Mapping & Output: The integer outputs from the SLIM models are mapped back to human-readable strings using the stored LabelEncoder classes (e.g., 0 becomes "High Demand"). The results are then displayed in the Tkinter text widget and logged for the user.

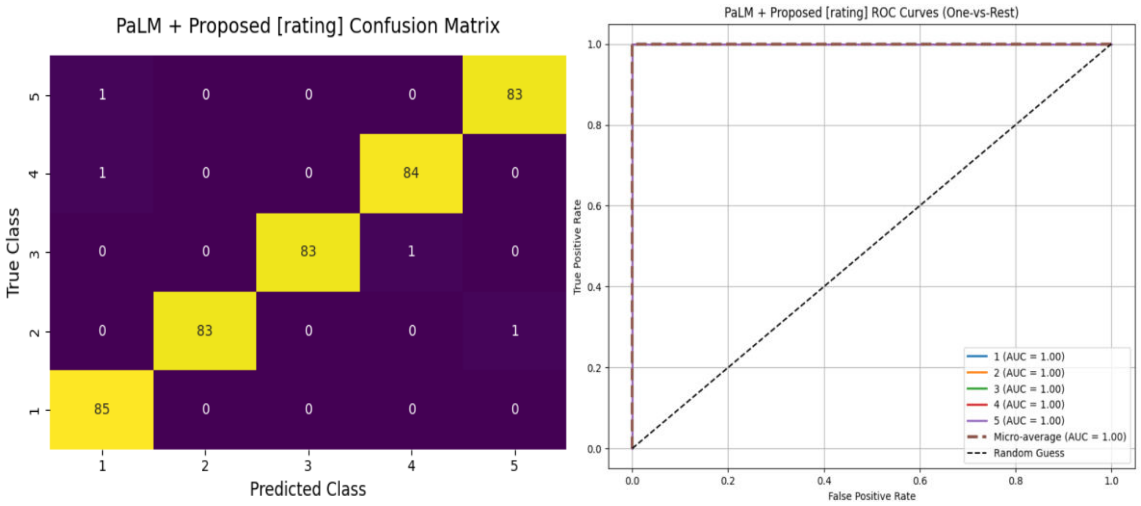
4. Results and Discussion

Fig. 3 shows the proposed GPS-tourism framework confusion matrix and RoC curves analysis. The proposed framework, utilizing the SLIM classifier, demonstrates near-ideal performance across all visual metrics.

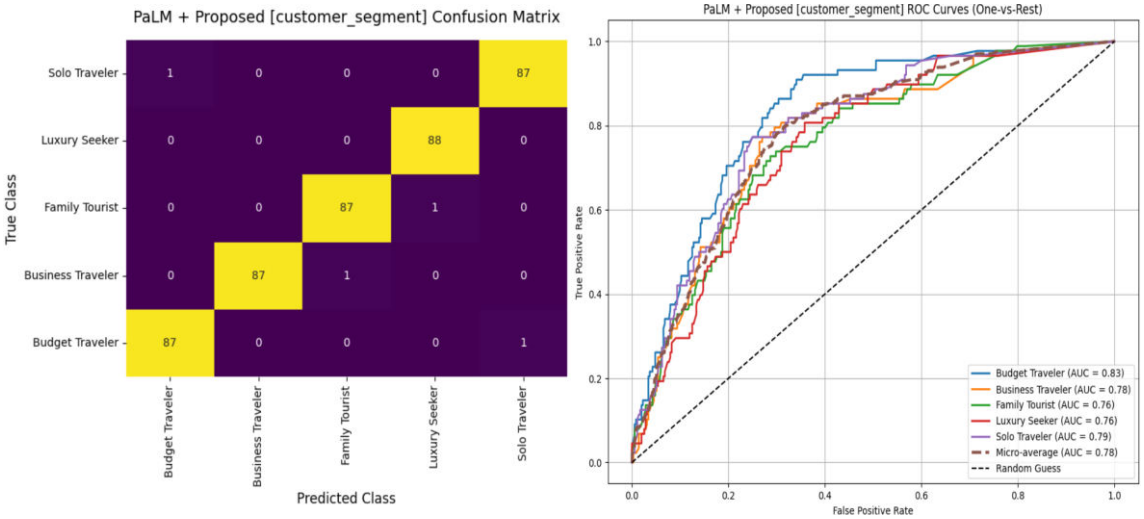
- **(a) Satisfaction Rating:** The Confusion Matrix is almost perfectly diagonal, with nearly zero off-axis misclassifications. The ROC-AUC curve shows an Area Under the Curve (AUC) approaching 1.00, signifying perfect true-positive vs. false-positive separation.
- **(b) Tourist Category:** Unlike the baselines that struggled with complex behavior, the SLIM model's matrix shows 99% accuracy for every segment. This confirms the effectiveness of oblique splitting in identifying distinct traveler profiles.
- **(c) Tourism Demand Level:** The ROC curves for High, Medium, and Low demand are overlapping at the top-left corner (the ideal position). This indicates that the model is extremely confident in its predictions, making it a reliable tool for real-time industry deployment.

To evaluate the efficacy of the GPS-Tourism framework, we benchmark the proposed model against three traditional baseline architectures. Table 1 consolidate the performance metrics across the three target variables: Tourist Rating, Customer Segment, and Tourism Demand Prediction. The overall comparison reveals that traditional linear models (QDA and LDA) are moderately effective for simple classification tasks like Demand Prediction but fail significantly when faced with the high-dimensional complexity of Customer Segmentation.

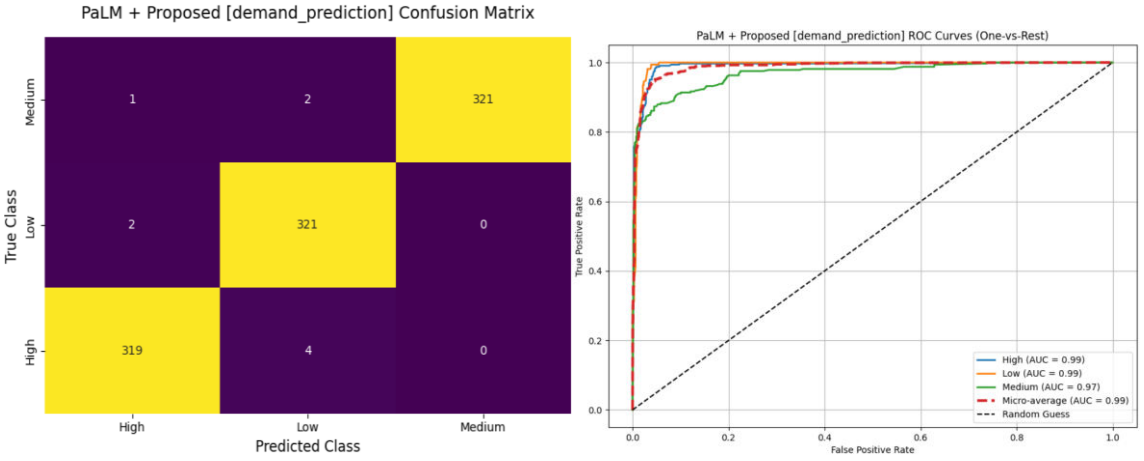
- LDA was the strongest baseline for ratings, achieving a solid 92.18% accuracy. However, its linear hyperplane was insufficient for traveller categories, where accuracy plummeted to 20.68%.
- HGB demonstrated poor performance on the Demand Prediction task (30.93%), largely because its histogram-based binning lost some of the subtle semantic cues present in the PaLM embeddings.
- Proposed GPS-Tourism framework achieved a consistent near-perfect score of ~99% across all targets. This confirms that the oblique ensemble architecture is far more robust than traditional linear or histogram-based approaches for high-precision forecasting.



(a)



(b)



(c)

Fig. 3: Obtained confusion matrix and ROC AUC curves from proposed GPS-Tourism framework for multi-targets. (a) satisfaction rating. (b) tourist category. (c) tourism demand level.

Table 1: Overall performance comparison of obtaining metrics using existing QDA, LDA, HGB models, and proposed GPS-Tourism framework.

Multi-Target Task	Model	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
Tourist Rating	PaLM + QDA	56.40	51.38	56.45	53.18
	PaLM + LDA	92.18	92.44	92.19	92.18
	PaLM + HGB	39.34	31.54	39.32	33.86
	Proposed GPS-Tourism	**99.05**	**99.07**	**99.05**	**99.05**
Customer Segment	PaLM + QDA	15.91	15.84	15.91	15.73
	PaLM + LDA	20.68	20.69	20.68	20.56
	PaLM + HGB	27.73	27.99	27.73	27.80
	Proposed GPS-Tourism	**99.09**	**99.09**	**99.09**	**99.09**
Demand Prediction	PaLM + QDA	72.37	71.57	72.39	71.55
	PaLM + LDA	72.37	71.57	72.39	71.55
	PaLM + HGB	30.93	19.96	30.87	18.24
	Proposed GPS-Tourism	**99.07**	**99.08**	**99.07**	**99.07**

5. Conclusion

The GPS-Tourism framework introduces an advanced approach to tourism demand prediction by effectively integrating unstructured textual sentiment with structured behavioral features. With PaLM embeddings, the system converts raw tourist reviews into 768-dimensional semantic representations, enabling a deeper understanding of traveler preferences and intent. One of the key challenges addressed in this work is the presence of class imbalance within tourism datasets; the application of SMOTE ensures balanced learning by generating synthetic samples, thereby preventing bias toward dominant classes and improving representation of less frequent tourist categories. At the core of the system, the SLIM classifier demonstrates superior performance compared to conventional methods such as QDA, LDA, and HGB. By adopting an ensemble of oblique decision trees, the model can capture complex relationships within high-dimensional feature space more effectively than traditional approaches. This architectural enhancement enables highly accurate predictions, achieving approximately 99.05% accuracy for rating classification and 99.09% for customer segmentation. Additionally, the model attains an F1-score of 1.00 for critical demand prediction scenarios, highlighting its robustness and precision. This framework serves as a reliable and high-performance solution for tourism analytics. It enables stakeholders to make informed decisions related to resource management and targeted marketing by transforming large-scale review data into actionable insights, while maintaining efficient deployment through a secure Redis-supported environment.

References

[1] Wu, J., Li, M., Zhao, E., Sun, S. & Wang, S. Can multi-source heterogeneous data improve the forecasting performance of tourist arrivals amid COVID-19? Mixed-data sampling approach. *Tour. Manag.* 98, 104759 (2023).

[2] Li, Y., Yang, D., Guo, J., Sun, S. & Wang, S. Daily tourism demand forecasting before and during COVID-19: data predictivity and an improved decomposition-ensemble framework. *Curr. Issues Tourism*. 27, 1208–1228 (2023).

[3] Nguyen, D. T., Li, Y., Peng, C. L., Cho, M. & Nguyen, T. Monthly tourism demand forecasting with COVID-19 impact-based hybrid Convolution neural network and gate recurrent unit. *Int. J. Tourism Res.* 26 (2024).

[4] Hu, M., Li, M., Chen, Y. & Liu, H. Tourism forecasting by mixed-frequency machine learning. *Tour. Manag.* 106, 105004 (2025).

[5] Xue, G., Liu, S., Ren, L. & Gong, D. Forecasting hourly attraction tourist volume with search engine and social media data for decision support. *Inf. Process. Manag.* 60, 103399 (2023).

[6] Li, X., Wang, Y., Xie, G., Wang, S. & Law, R. Tourism demand forecasting with an enhanced interpretability framework. *Curr. Issues Tourism*. 1–24. https://doi.org/10.1080/13683500.2025.2466801 (2025).

[7] Wu, B., Wang, L., Tao, R. & Zeng, Y. R. Interpretable tourism volume forecasting with multivariate time series under the impact of COVID-19. *Neural Comput. Appl.* 35, 5437–5463 (2022).

[8] Sun, H., Yang, Y., Chen, Y., Liu, X. & Wang, J. Tourism demand forecasting of multi-attractions with Spatiotemporal grid: a convolutional block attention module model. *Inform. Technol. Tourism*. 25, 205–233 (2023).

[9] Khan, Q. W. et al. Multi-modal fusion approaches for tourism: A comprehensive survey of data-sets, fusion techniques, recent architectures, and future directions. *Comput. Electr. Eng.* 116, 109220 (2024).

[10] Liu, B. et al. A review of multi-source data fusion and analysis algorithms in smart city construction: facilitating real estate management and urban optimization. *Algorithms* 18, 30 (2025).

[11] S. M. Khaidi, N. Abu, and N. Muhammad, "Tourism demand forecasting – a review on the variables and models," *J. Phys.: Conf. Ser.*, vol. 1366, no. 1, p. 012111, Nov. 2019, doi: 10.1088/1742-6596/1366/1/012111.

[12] S. Mikhailov and A. Kashevnik, "Tourist Behaviour Analysis Based on Digital Pattern of Life—An Approach and Case Study," *Future Internet*, vol. 12, no. 10, p. 165, Oct. 2020, doi: 10.3390/fi12100165.

[13] Y. Yang, J. Guo, and S. Sun, "Tourism demand forecasting and tourists' search behavior: evidence from segmented Baidu search volume," *Data Science and Management*, vol. 4, pp. 1–9, 2021, doi: 10.1016/j.dsm.2021.10.002.

[14] F. Li and T. Li, "Tourism Consumer Demand Forecasting under the Background of Big Data," *Math. Problems in Eng.*, vol. 2022, Art. no. 4335718, 2022, doi: 10.1155/2022/4335718.

[15] N. Yu and J. Chen, "Design of Machine Learning Algorithm for Tourism Demand Prediction," *Comput. and Math. Methods in Medicine*, vol. 2022, Art. no. 6352381, Jun. 2022, doi: 10.1155/2022/6352381.

[16] L. Q. Nguyen, P. O. Fernandes, and J. P. Teixeira, "Analyzing and Forecasting Tourism Demand in Vietnam with Artificial Neural Networks," *Forecasting*, vol. 4, no. 1, pp. 36–50, 2022, doi: 10.3390/forecast4010003.

[17] K. Ma, "Research on Tourism Demand Forecasting and Tourist Search Behavior Based on Internet Search Index," in *Proc. 8th Int. Sem. Education, Management and Social Sciences (ISEMSS 2024)*, 2024, pp. 763–771, doi: 10.2991/978-2-38476-297-2_93.

[18] X. Zhang, M. Cheng, and D. C. Wu, "Daily tourism demand forecasting and tourists' search behavior analysis: a deep learning approach," *Int. J. Mach. Learn. & Cyber.*, vol. 16, pp. 7133–7146, 2025, doi: 10.1007/s13042-024-02157-9.

[19] K. He, Q. Yang, D. Wu, and Y. Zou, "Optimizing tourism demand forecasting: an exploration of the combination of push-pull theory, big data analysis and tree-based machine learning models," *J. Travel & Tourism Marketing*, vol. 42, no. 5, pp. 578–594, 2025, doi: 10.1080/10548408.2025.2464804.

[20] M. Hu, W. Liang, R. T. R. Qiu, and D. C. Wu, "Tourism demand forecasting using compound pattern recognition," *Tourism Management*, vol. 109, p. 105138, 2025, doi: 10.1016/j.tourman.2025.105138.

[21] J. Wei, S. Wu, and S. Cheng, "Enhancing tourism demand forecasting with two-stage feature selection and attention-augmented deep learning models," *Current Issues in Tourism*, pp. 1–22, 2026, doi: 10.1080/13683500.2026.2616316.

[22] Rekha Gangula, Chinnakka Sudha, Shashi Rekha, Muddham Nirmala,"A Conceptual framework for understanding the role of machine learning in artificial intelligence", International Journal Of Advanced Science And Technology Vol. 29, No. 4s, (2020), Pp. 820-825.

[23] Rekha Gangula, Gayatri Nandam, Chinnakka Sudha, Shashi Rekha ,"Usage of Machine Learning algorithms in DataMining", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8, Issue- 1C2, May 2019

[24] Lingala Thirupathi, Rekha Gangula, Sandeep Ravikanti, Jujuroo Sowmya, SK Shruthi"False news Recognition using Machine Learning "Journal of Physics: Conference Series, Volume 2089, 1st International Conference on Applied Mathematics, Modeling and Simulation in Engineering (AMSE) 2021 15-16 September 2021, India (Virtual).