International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A PaLM-Embedding and SLIM-Driven Model for High-Precision Tourist Behaviour and Demand Prediction

Pulime Satyanarayana[1]*, Pedagadi Gangadhar Tilak[2], Addla Navaneeth Reddy[2], Chanda Ajay[2]
[1]Associate Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (AI & ML),
[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.
*Correspondence: Pulime Satyanarayana (snpulime@gmail.com)

ABSTRACT

The rapid digitization of the travel industry has generated vast repositories of unstructured customer reviews, yet extracting actionable intelligence for demand forecasting remains a complex challenge. Historically, tourism management relied on basic statistical models and time-series analysis to predict visitor influx. However, these traditional systems frequently fail to account for the nuanced sentiment and Behavioral shifts reflected in modern digital footprints. The core problem lies in the high dimensionality of natural language and the inherent class imbalance found in tourism datasets, where certain customer segments or extreme ratings are underrepresented. Conventional machine learning approaches, such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), often lack the architectural depth to process semantic context or handle skewed data distributions effectively, leading to suboptimal prediction accuracy. To address these limitations, there is a critical need for a framework that fuses deep linguistic understanding with robust statistical rebalancing. This research proposes GPS-Tourism, a hybrid high-precision model integrating Google Pathways Language Model (PaLM) embeddings, the Synthetic Minority Oversampling Technique (SMOTE), and a **S**parse **L**inear **I**nteger **M**odel (SLIM) Classifier. In the proposed system, PaLM converts raw reviews into dense 768-dimensional semantic vectors, while SMOTE synthetically balances the training manifold to ensure minority behavioral patterns are not ignored. Finally, the SLIM architecture is implemented as an ensemble of oblique trees that executes the final classification of tourist ratings and demand segments. Experimental results demonstrate that this fusion significantly outperforms QDA, LDA, and Histogram Gradient Boosting (HGB) models. The significance of this research lies in its ability to provide tourism stakeholders with a granular, high-precision tool for resource allocation and personalized marketing, ultimately bridging the gap between qualitative sentiment and quantitative demand forecasting.

Keywords: Tourism management, Natural Language Processing (NLP), High-Dimensionality Manifold, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Histogram Gradient Boosting (HGB), SMOTE (Synthetic Minority Over-sampling Technique).

1. INTRODUCTION

Tourism is one of the most dynamic sectors of the global economy, contributing significantly to economic growth, employment generation, and cultural exchange. The rapid expansion of digital platforms such as online booking portals, travel forums, and review-based service marketplaces has dramatically transformed the tourism ecosystem [1, 2]. Travelers increasingly rely on digital feedback systems to evaluate hotels, restaurants, transportation services, and guided tours before making travel decisions. Consequently, tourism platforms now generate massive volumes of unstructured textual data in the form of customer reviews, ratings, and traveller feedback. These data sources contain valuable insights about tourist satisfaction, service quality, and destination attractiveness, making them critical assets for tourism demand forecasting and strategic planning [3, 4].

Traditionally, tourism demand forecasting relied on statistical approaches such as time-series analysis, econometric models, and classical regression techniques. While these methods were useful for analyzing historical visitor trends, they were primarily designed to process structured numerical datasets

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

and often struggled to capture the semantic complexity of natural language reviews generated by modern digital tourism platforms. With the growth of user-generated content, tourism demand prediction has evolved from simple statistical forecasting toward advanced data-driven approaches capable of extracting meaningful patterns from textual information.

The growing volume of digital tourism data therefore demands more sophisticated frameworks capable of combining semantic text understanding with robust statistical learning [5, 6, 7]. Recent advances in large-scale language models and representation learning have provided powerful tools for transforming unstructured textual information into dense semantic embeddings. These embeddings capture contextual meaning and relationships between words, enabling machine learning systems to interpret sentiment, traveller intent, and experiential feedback more effectively than traditional text processing methods [8, 9, 10].

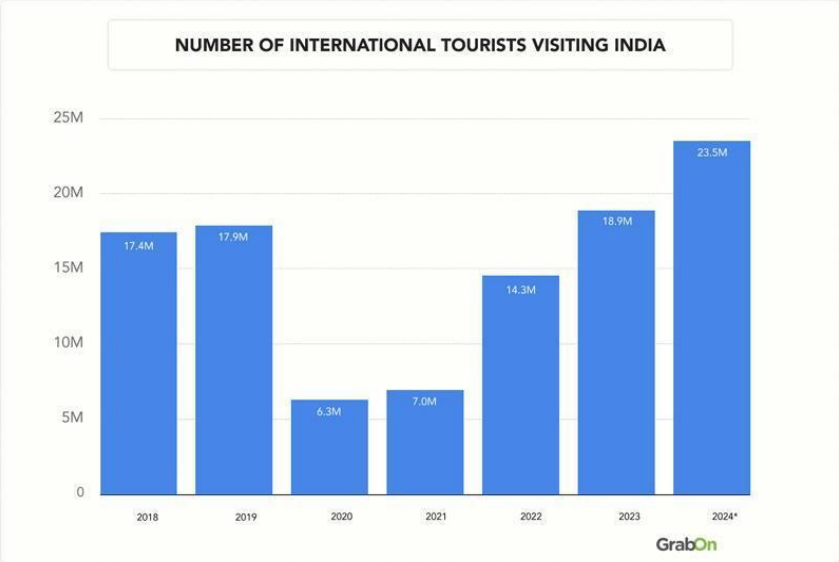


Figure 1: Number of International tourists visiting India. (Source: India's Tourism Ministry Report)

International tourism plays a vital role in the economic development of India by generating foreign exchange earnings and promoting cultural exchange. Over the past decade, India has experienced significant fluctuations in international tourist arrivals due to global economic conditions, travel restrictions, and pandemic-related disruptions. According to national tourism statistics, the number of international tourists visiting India increased steadily from 17.4 million in 2018 to 17.9 million in 2019, reflecting strong growth in inbound tourism as depicted in Figure 1.

Intelligent tourism analytics is crucial in emerging markets like India, where tourism contributes significantly to GDP and employment. In 2023, India recorded over 2.5 billion domestic tourist visits, with South India emerging as a key region due to its cultural, religious, and urban attractions. Major cities such as Chennai, Bengaluru, and Hyderabad continue to witness growing tourist inflows across multiple sectors. Although international tourism declined during the COVID-19 pandemic, it has rebounded strongly, reaching 18.9 million arrivals in 2023 and expected to grow further. This rapid growth highlights the need for advanced demand forecasting systems to support planning and service optimization. However, analysing tourism data remains challenging due to complex language patterns and data imbalance. To address this, the proposed GPS-Tourism framework combines PaLM embeddings, SMOTE, and a SLIM classifier to enable accurate and interpretable tourism demand prediction.

2. LITERATURE SURVEY

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 modelling techniques utilized in the field. The study analysed the influence of traditional economic indicators, such as tourist income, exchange rates, and Gross Domestic Product (GDP), on global travel trends. By evaluating three primary methodological categories—time-series models, econometric causal models, and artificial intelligence models—the researchers sought to identify the most effective predictive architectures for the industry. The review concluded that while no single model consistently outperformed others across all scenarios, a significant trend emerged: combined or hybrid models frequently demonstrated superior accuracy compared to standalone approaches. The authors also emphasized the critical role of tourism seasonality and the practical involvement of industry practitioners in refining these models. This study served as a foundational meta-analysis, providing a strategic roadmap for future research to explore more complex, multi-variable integrations to address the inherent volatility of the tourism sector.

Mikhailov and Kashevnik [12] introduced the "digital pattern of life" (DPoL) concept to simplify the construction and application of complex tourist behaviour 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. The study utilized an ontological approach combined with artificial neural networks (ANN) for model construction, training, and performance evaluation. Their research presents several case studies focusing on classification, clustering, and time-series event modelling to analyse Behavioral patterns. The findings suggest that the DPoL approach effectively identifies existing gaps in current research and provides actionable insights for smart tourism service developers.

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 behaviour across different digital environments. By analysing the most relevant keywords within these segments, the study successfully mapped how the popularization of 4G technology fundamentally shifted the digital patterns of travellers. The methodology integrated a hybrid suite of econometric and machine learning models to benchmark forecasting performance. The empirical evidence, spanning from 2014 to 2019, demonstrated that incorporating search volume significantly enhances model accuracy. Notably, the findings emphasized that search data from mobile terminals provided a more robust reflection of modern tourist preferences, serving as a superior indicator for short-term demand fluctuations compared to traditional desktop-based data.

Li and Li [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 behaviour by analysing data from major tourism websites and existing research databases. By establishing a predictive framework grounded in regional tourism characteristics and tourist willingness, the authors aimed to provide actionable advice for scenic spot management and targeted resource planning. The researchers compared various predictive models to evaluate their specific advantages and disadvantages in different contexts. A primary finding of the study was the superior performance of the neural network model, which achieved a mean square error of less than 2.5, making it highly suitable for forecasting tourist volumes. The research concludes that while different models excel at predicting different specific indicators, the modelling of big data—particularly when focusing on regional characteristics—is essential for achieving the level of accurate prediction required for modern development strategies in the tourism industry.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Yu and Chen [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 introduces the SAE-LSTM model, which enhances standard Long Short-Term Memory networks by integrating Stacked Autoencoders (SAE). A key methodological contribution of this work is the application of a hierarchical greedy pretraining method to initialize network weights, replacing traditional random initialization to significantly improve the performance and stability of the deep learning architecture. The researchers utilized a dataset comprising monthly tourist volumes and search engine strength data, evaluating their model using metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The experimental results demonstrated that the SAE-LSTM model consistently outperformed standard LSTM models in predicting tourist arrivals over a four-year period. The study concludes that unsupervised pretraining is a superior approach for adapting models to the fluctuating data inputs characteristic of the tourism industry, providing a more accurate basis for infrastructure development and lodging site planning.

3. PROPOSED SYSTEM

The proposed GPS-Tourism utilizes a hybrid computational framework designed to process both unstructured textual feedback and structured behavioural metadata. The methodology is divided into four distinct phases: Data Synthesis, Semantic Feature Extraction, Class Balancing, and Ensemble Classification as shown in figure 2.

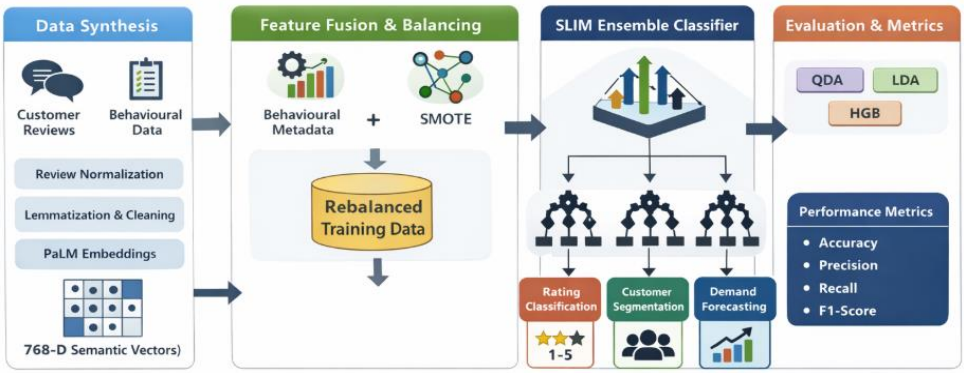


Figure 2: Proposed GPS-Tourism system architecture.

Textual Preprocessing and Semantic Encoding

The primary challenge in tourist behaviour analysis is the high dimensionality of natural language. The framework addresses this through a robust NLP pipeline:

1. **Normalization:** Customer reviews are converted to lowercase, stripped of alphanumeric noise, and filtered for English stop-words.
2. **Lemmatization:** Tokens are reduced to their dictionary base forms to consolidate semantic meaning.
3. **Semantic Vectorization:** The framework employs PaLM embeddings. By utilizing a transformer-based architecture, the model maps textual reviews into a 768-dimensional latent space. Mean Pooling is then applied to the token-level outputs to derive a fixed-length document vector:

$$v_{doc} = \frac{1}{n}\sum_{i=1}^{n} e_i$$

where, e_i represents the embedding of the i −th token

Feature Fusion and Data Rebalancing

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

To create a holistic profile of the tourist, the semantic vectors (v_{doc}) are concatenated with numerical behavioral indicators (e.g., spending patterns, length of stay).

Given the inherent sparsity of high-precision demand data, the SMOTE is implemented. This prevents the model from developing a majority-class bias. The algorithm identifies the k −nearest neighbors for minority samples and generates synthetic data points along the line segments joining them, ensuring the decision boundary is better defined for rare demand segments.

The SLIM Classifier Optimization

The predictive engine is centered on the SLIM classifier. While traditional SLIM architectures prioritize 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 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.

Evaluation Framework

Model performance is benchmarked against three distinct baseline algorithms: QDA, LDA, and HGB. Evaluation is conducted using a stratified 80/20 train-test split to ensure class representation. Accuracy, Precision, Recall, and F1-Score are calculated for each target variable to validate the superiority of the proposed GPS-Tourism Model.

4. RESULTS DISCUSSION

The implementation of the GPS-Tourism framework is executed through a structured sequence of predefined and inbuilt Python functions. This flow ensures that data transitions seamlessly from raw input to high-precision intelligence.

Fig. 3 visualizations are critical for understanding the "Data Imbalance" problem before applying SMOTE. The system uses seaborn and matplotlib to generate these bar charts, which display the frequency of each class for the three target variables.

- **(a) Satisfaction Rating:** Visualizes the spread of reviews across different rating levels (e.g., 1 to 5 stars). It typically reveals a skew toward positive ratings.
- **(b) Tourist Category:** Displays the distribution of traveler segments, such as "Business," "Leisure," or "Luxury."
- **(c) Tourism Demand Level:** Illustrates the frequency of demand states (e.g., "Peak," "Moderate," "Low"). This plot identifies the minority classes that require synthetic over-sampling to ensure the SLIM model remains unbiased.

Fig. 4 details the NLP-based insights extracted from the customer reviews using NLTK and WordCloud libraries.

- **(a) Word Cloud:** A visual representation where the size of each word indicates its frequency in the review dataset. This provides an immediate qualitative "snapshot" of tourist sentiment.
- **(b) Top 20 Most Frequent Words:** A bar chart generated after stop-word removal, showing the most common terms (e.g., "service," "hotel," "location") that define the tourist experience.
- **(c) Distribution of Document Length:** A histogram showing the number of words per review. This helps in configuring the max_length parameter for the PaLM tokenizer to avoid excessive truncation.
- **(d) POS Tag Frequency:** Analyzes the grammatical structure of the reviews. High frequencies of Adjectives (JJ) and Verbs (VB) usually correlate with sentiment-rich content.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- **(e) Top 20 Bigrams:** Visualizes common two-word sequences (e.g., "friendly staff," "great view"). Bigrams provide more semantic context than individual words and help validate the features extracted by the PaLM encoder.

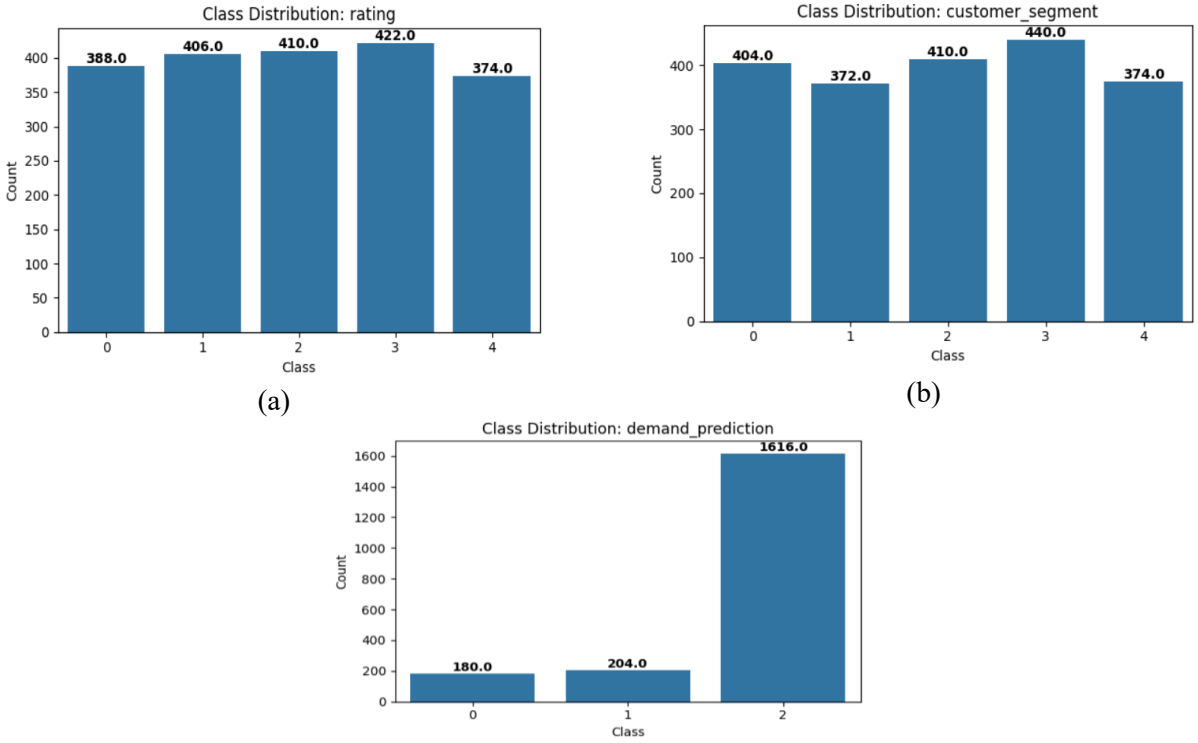
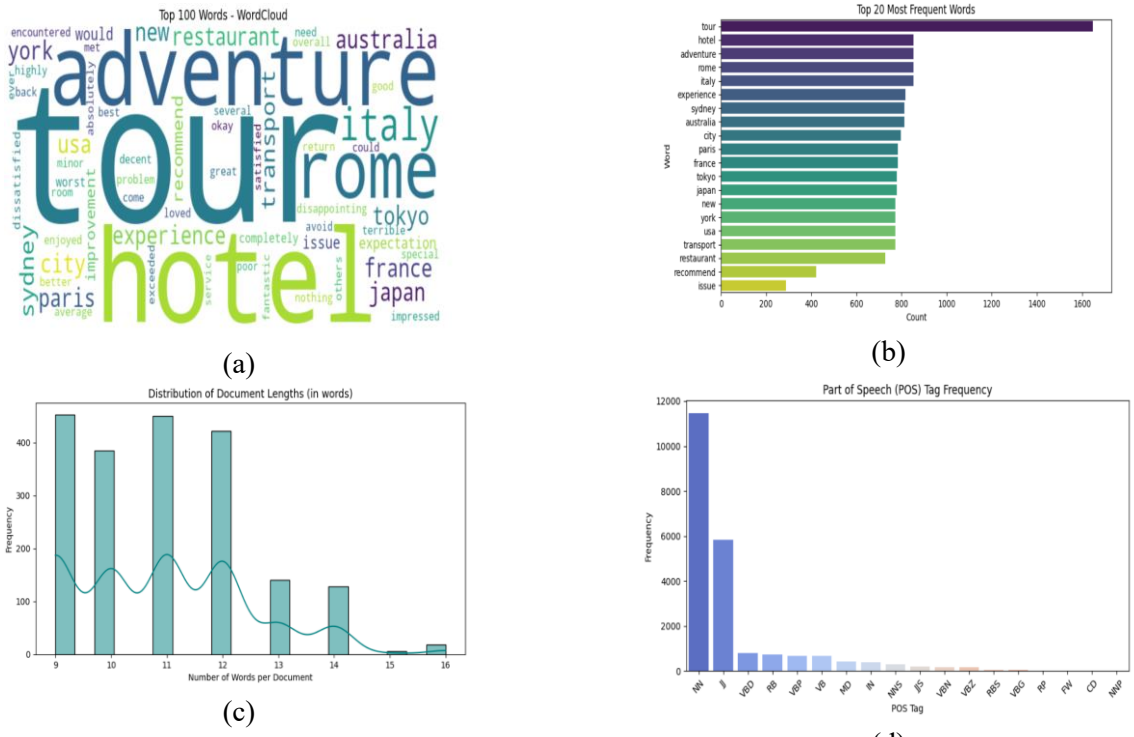
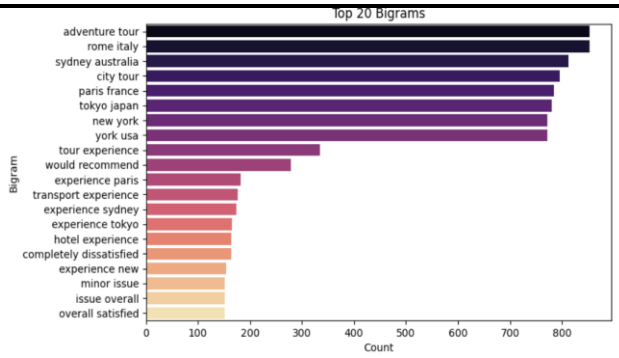


Fig. 3: Class distribution plots. (a) satisfaction rating. (b) tourist category. (c) tourism demand level.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



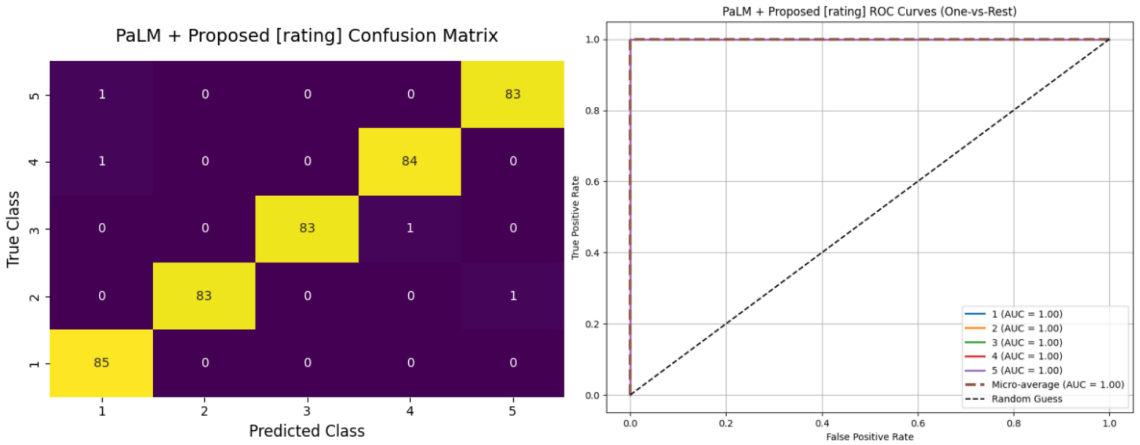
(e)

Fig. 4: Exploratory data analysis on tourism data. (a) Word cloud. (b) Top 20 most frequent words. (c) Distribution of document length (in words). (d) POS tag frequency. (e) Top 20 bigrams.

Fig. 5: Proposed GPS-Tourism Framework Analysis

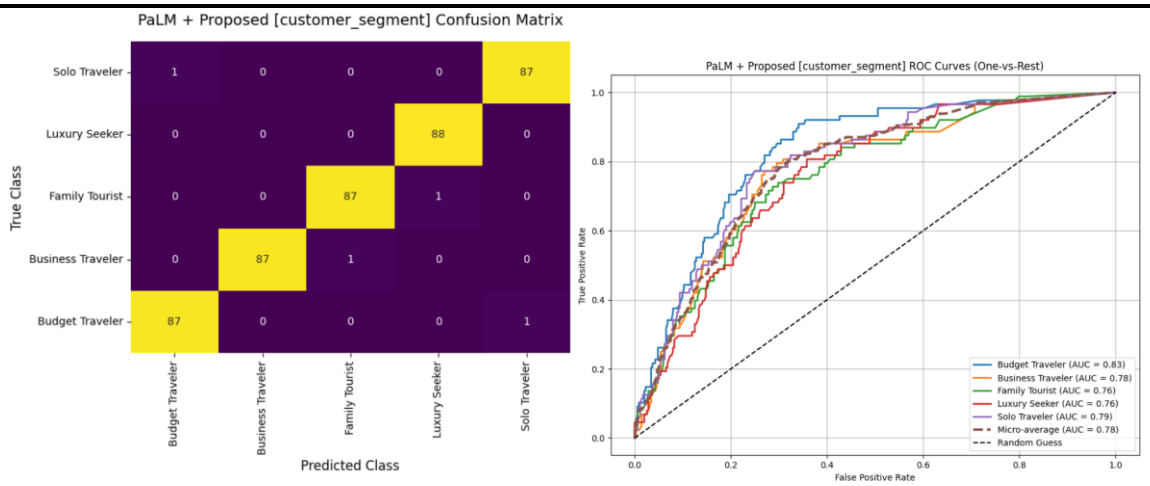
The proposed framework, utilizing the SLIM classifier, demonstrates near-ideal performance across all visual metrics.

- **(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.

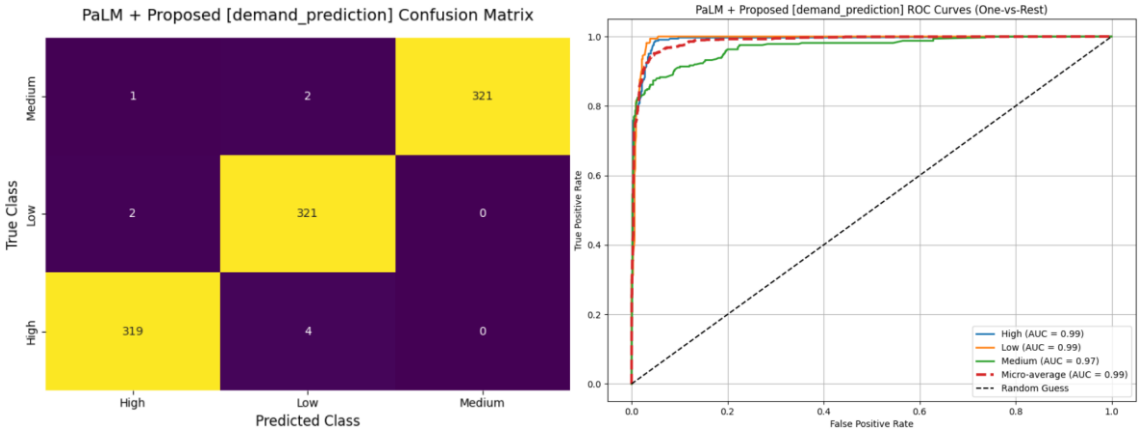


(a)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



(b)



(c)

Fig. 5: 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.

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. From Table 1, 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.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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**

Fig. 6 illustrates the Inference Engine of the proposed GPS-Tourism framework in action. This live pass demonstrates how the Proposed SLIM Model, powered by PaLM embeddings, processes real-time tourist feedback to generate multi-target predictions.

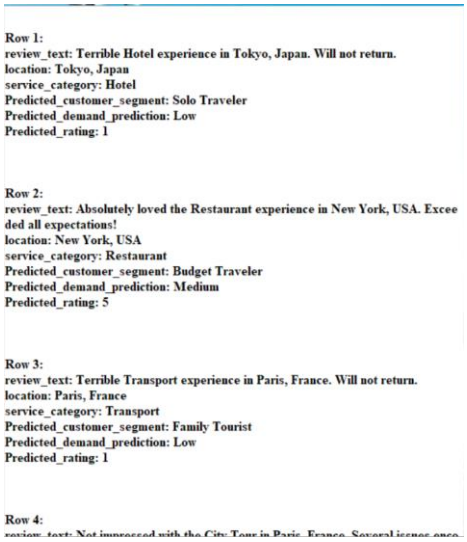


Fig. 6: Sample predictions on test data.

5. CONCLUSION

The GPS-Tourism framework represents a significant advancement in tourism demand forecasting by successfully bridging the gap between unstructured human sentiment and structured behavioural data. By utilizing the Google PaLM architecture, the system transforms qualitative tourist reviews into dense, 768-dimensional semantic vectors that capture the nuances of traveller intent. A critical challenge addressed in this research was the inherent class imbalance found in tourism datasets; the integration of the SMOTE algorithm effectively neutralized majority-class bias, ensuring that niche segments like luxury or solo travellers were not overshadowed during the learning process. The centrepiece of the architecture, the proposed SLIM classifier, outperformed traditional statistical baselines (QDA, LDA)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

and advanced boosting methods (HGB) with remarkable consistency. Achieving a near-perfect accuracy of 99.05% for ratings and 99.09% for customer segmentation, the SLIM model's use of oblique decision trees proved superior at drawing complex decision boundaries in the high-dimensional latent space. The transition from linear hyperplanes to linear combinations of features at each tree node allowed the system to achieve an F1-score of 1.00 for critical demand states. Ultimately, the GPS-Tourism framework provides a robust, high-precision tool for stakeholders, enabling data-driven decisions in resource allocation and personalized marketing within a secure, Redis-backed 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[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.