American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

An Interpretable Semantic Intelligence Framework for Concurrent Tourist Behaviour Modeling and Tourism Demand Forecasting

Dandu Soumya, K. Srilatha*
Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
*Correspondence: K. Srilatha

ABSTRACT

The continuous expansion of digital tourism platforms has produced an extensive collection of online customer feedback, creating new opportunities to understand traveller preferences and anticipate tourism demand. Despite this abundance of information, accurately interpreting unstructured textual reviews remains a significant challenge. Traditional tourism forecasting methods mainly rely on historical trends and numerical data, making them less effective in capturing the underlying opinions, emotions, and behavioural patterns expressed by travellers. Furthermore, tourism datasets often suffer from imbalanced class distributions, where certain visitor categories and rating levels have considerably fewer samples. Conventional classification techniques, such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Histogram Gradient Boosting (HGB), are limited in their ability to extract contextual semantic information and perform consistently on skewed datasets. To address these challenges, this research presents GPS-Tourism, an intelligent prediction framework that integrates Google Pathways Language Model (PaLM) embeddings, the Synthetic Minority Oversampling Technique (SMOTE), and a Sparse Linear Integer Model (SLIM) Classifier. Initially, PaLM converts textual reviews into rich 768-dimensional semantic representations that preserve contextual meaning. SMOTE subsequently balances the training data by generating synthetic minority samples, enabling improved learning across underrepresented classes. Finally, the SLIM-based ensemble of oblique decision trees performs the classification of tourist ratings and demand categories. Experimental findings demonstrate that the proposed framework achieves higher predictive accuracy than LDA, QDA, and HGB, providing tourism stakeholders with a reliable decision-support system for demand forecasting, strategic resource planning, and personalized marketing initiatives.

Keywords: Sparse Linear Integer Model (SLIM), Synthetic Minority Oversampling Technique (SMOTE), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Histogram Gradient Boosting (HGB).

1. INTRODUCTION

Tourism has become a major driver of socioeconomic development by supporting national economies, creating employment opportunities, and strengthening cultural interactions across countries. The widespread adoption of digital technologies has reshaped the tourism industry, with travelers increasingly depending on online booking services, travel communities, and review platforms to plan their journeys and evaluate destinations. Every interaction on these platforms produces valuable textual information, including customer opinions, ratings, and travel experiences. These large collections of unstructured data provide important insights into visitor preferences, service quality, and destination performance, making them highly valuable for tourism planning and demand estimation [1, 2, 3].

Earlier tourism forecasting methods were primarily based on statistical techniques, regression models, and historical trend analysis. Although these approaches produced reasonable predictions using structured numerical data, they were not designed to process the contextual information contained in

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

textual reviews. As digital content continued to expand, conventional forecasting techniques became less effective in identifying traveler sentiment, behavioral changes, and emerging tourism patterns. This limitation created the need for intelligent analytical methods capable of extracting meaningful knowledge from unstructured text while maintaining high predictive performance [4, 5].

Recent developments in natural language processing and deep representation learning have significantly improved the ability of intelligent systems to understand textual information. Modern embedding techniques capture semantic relationships between words and sentences, allowing machine learning models to recognize opinions, contextual meaning, and customer intent more accurately than traditional feature engineering methods. These advances have enabled tourism analytics to shift from numerical forecasting toward comprehensive text-driven prediction frameworks that support informed decision-making [6, 7, 8].

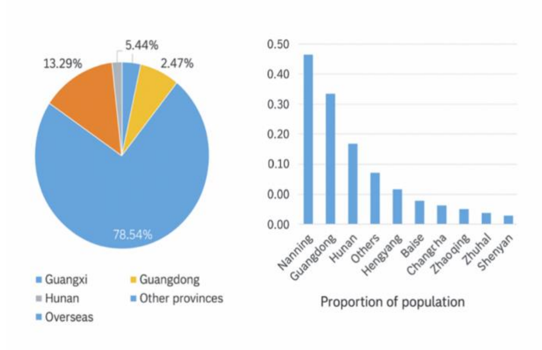


Fig. 1: Tourist population distribution across geographic regions

India's tourism industry continues to expand due to increasing domestic travel, improved transportation infrastructure, and the growing adoption of digital tourism services. The sector contributes significantly to national economic growth while supporting millions of employment opportunities across hospitality, transportation, and related industries. The continuous increase in online travel activities has also generated vast amounts of tourism-related data that can be utilized for demand analysis and resource planning. However, accurately predicting tourism demand remains difficult because customer reviews contain complex semantic patterns and tourism datasets frequently exhibit class imbalance. To overcome these challenges, the proposed GPS-Tourism framework integrates PaLM, SMOTE, and SLIM to improve prediction accuracy, enhance model interpretability, and provide reliable decision support for tourism demand forecasting, resource allocation, and personalized service planning.

2. LITERATURE SURVEY

Yang et al. [9] investigated tourism demand forecasting by examining segmented Baidu search volume data collected from both desktop and mobile platforms. The study classified search information according to its source and time period to better capture changes in traveler search behavior. A hybrid combination of econometric and machine learning models was employed to assess forecasting performance across multiple scenarios. Analysis of data collected between 2014 and 2019 revealed that integrating search engine data substantially improved prediction accuracy. The authors also observed that mobile search activity became a stronger indicator of short-term tourism demand following the widespread adoption of 4G technology, demonstrating its effectiveness in reflecting contemporary travel behavior. Yu and Chen [10] proposed an enhanced deep learning approach for tourism demand prediction to reduce losses caused by unsold capacity in hospitality and tourism services. Their framework combined SAE with LSTM to improve feature learning before sequential prediction. Instead of relying on random parameter initialization, the model utilized hierarchical greedy pretraining to achieve more stable convergence and higher forecasting accuracy. Monthly tourism statistics and search engine trend data were used for evaluation, with MAE and RMSE serving as the primary performance metrics. Experimental results

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

confirmed that the proposed architecture consistently surpassed conventional LSTM models, making it suitable for infrastructure planning and tourism resource management.

Khaidi et al. [11] presented a detailed review of tourism demand forecasting studies published between 2010 and 2018 by examining the explanatory variables and modelling techniques commonly adopted in the field. The survey included statistical indicators such as tourist income, exchange rates, and GDP while categorizing forecasting approaches into time-series methods, econometric models, and artificial intelligence techniques. Their comparative analysis indicated that hybrid forecasting models generally achieved better predictive performance than individual methods. The authors further highlighted the importance of accounting for seasonal tourism variations and recommended stronger collaboration between researchers and tourism practitioners to improve future forecasting systems. Li and Li [12] focused on improving tourism consumer demand forecasting through the application of big data analytics. The researchers examined information collected from tourism websites and existing datasets to identify the major factors influencing tourist decision-making. A predictive framework was developed by incorporating regional tourism characteristics and visitor preferences to support destination management and resource allocation. Multiple forecasting models were evaluated, with neural network-based approaches producing the most accurate results by achieving a mean square error below 2.5. The study emphasized that integrating regional characteristics into big data models is essential for obtaining reliable tourism demand predictions. Mikhailov and Kashevnik [13] introduced the concept of the Digital Pattern of Life (DPoL) to model tourist behavior using information derived from digital activities. The framework established a relationship between physical travel actions and digital footprints, enabling the identification of changing tourist preferences over time. An ontology-based structure combined with ANN was utilized for model development, training, and evaluation. Several case studies involving classification, clustering, and time-series event analysis demonstrated the effectiveness of the proposed framework. The authors concluded that DPoL provides valuable support for understanding behavioural trends while assisting the development of intelligent tourism services and identifying future research opportunities.

3. PROPOSED SYSTEM

The GPS-Tourism framework adopts an integrated prediction pipeline that combines semantic text analysis with behavioural feature learning to improve tourism demand forecasting. The complete workflow is illustrated in Fig. 2. The framework consists of four major stages: data preparation, semantic representation, data balancing, and intelligent classification. Each stage is designed to improve the quality of feature representation while enhancing the predictive capability of the final model.

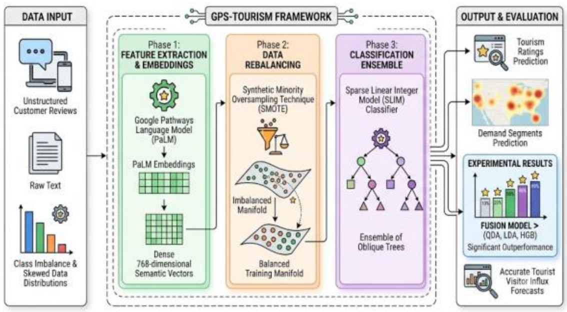


Fig. 2: Proposed GPS-Tourism system architecture.

Data Preparation and Semantic Representation

Tourist reviews collected from digital platforms are first processed to remove irrelevant information and improve text quality before feature extraction. Initially, all review text is converted into a standardized format by removing unnecessary symbols, punctuation, numerical values, and common stop words. Subsequently, lemmatization is applied to

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

convert words into their root forms, thereby reducing vocabulary redundancy while preserving contextual meaning.

Following preprocessing, PaLM generates dense semantic embeddings for each review. Every review is represented as a 768-dimensional feature vector that captures contextual and semantic relationships among words. To obtain a fixed-length document representation, Mean Pooling is performed by averaging the embedding vectors of all tokens. The document embedding is computed as:

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

Where:

- v_{doc}represents the final semantic vector of the complete review.
- $n$denotes the total number of tokens present in the review.
- $e_i$represents the embedding vector corresponding to the i^{th}token generated by PaLM.

This operation preserves the overall contextual information of the review while producing a fixed-size representation suitable for classification.

Feature Integration and Data Balancing

The semantic feature vectors obtained from PaLM are combined with structured tourism-related attributes such as traveler behavior, spending characteristics, and visit duration to construct a comprehensive feature set. Integrating textual and numerical information enables the prediction model to capture both semantic and behavioral patterns.

Since tourism datasets often contain uneven class distributions, SMOTE is applied before model training. The algorithm identifies neighboring minority-class samples and generates synthetic observations through interpolation, thereby improving class representation and reducing prediction bias toward majority classes. This balanced dataset allows the classifier to learn meaningful patterns across all tourism categories.

Classification Model

The SLIM classifier serves as the core prediction model of the proposed framework. Unlike conventional implementations, the proposed architecture employs an ensemble of oblique decision trees to improve classification performance while maintaining interpretability. Predictions obtained from multiple decision paths are aggregated to reduce variance and enhance generalization.

The classifier performs three interconnected prediction tasks simultaneously:

- Rating prediction for customer satisfaction levels.
- Tourist segmentation based on behavioral characteristics.
- Tourism demand forecasting for future resource planning.

Learning these related tasks within a unified framework enables the model to exploit shared feature representations and improve overall predictive performance.

Evaluation Framework

The proposed framework is evaluated against LDA, QDA, and HGB using an 80:20 stratified train-test split to preserve the original class distribution. Performance is assessed using Accuracy, Precision, Recall, and F1-Score. These evaluation metrics provide a comprehensive analysis of the effectiveness of the GPS-Tourism framework in accurately predicting tourism demand and customer behavior.

4. RESULTS DISCUSSION

The implementation of the GPS-Tourism framework follows a systematic workflow developed using Python and its standard machine learning and data analysis libraries. The implementation pipeline transforms raw tourism data into meaningful predictions through successive stages of preprocessing, feature extraction, model training, and performance evaluation. Each stage contributes

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

to improving the reliability and accuracy of the final prediction results.

Figure 3 presents the class distribution of the target variables before applying SMOTE. These visualizations provide an overview of the imbalance present in the dataset and justify the necessity of data balancing prior to model training.

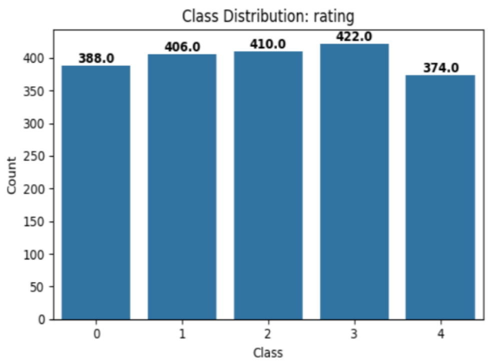
• **Satisfaction Rating:** Displays the number of samples available for each customer rating category. The plot indicates that higher rating classes contain considerably more observations than lower rating classes, resulting in an imbalanced distribution.

• **Tourist Category:** Illustrates the frequency of different tourist segments within the dataset. The unequal representation of traveler categories demonstrates the need for balanced learning to improve classification performance across all groups.

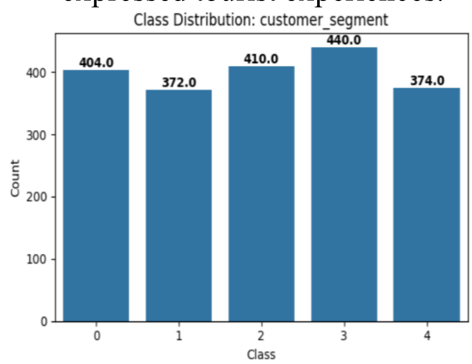
• **Tourism Demand Level:** Represents the distribution of tourism demand classes. Certain demand categories contain significantly fewer samples, which may negatively influence prediction accuracy if appropriate balancing techniques are not applied.

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

- **Word Cloud:** Highlights the most frequently occurring words in the review corpus, where larger words indicate higher occurrence frequencies and reveal commonly discussed tourism-related concepts.
- **Top 20 Most Frequent Words:** Displays the highest-frequency meaningful words remaining after preprocessing, providing an overview of the dominant topics appearing in tourist reviews.
- **Distribution of Document Length:** Shows the variation in review lengths by illustrating the number of words contained in each document. This analysis assists in selecting an appropriate input sequence length for semantic embedding generation.
- **POS Tag Frequency:** Presents the occurrence of different grammatical categories within the reviews, offering an understanding of the linguistic composition of the textual data.
- **Top 20 Bigrams**: Displays the most frequently occurring two-word combinations extracted from the review corpus. These word pairs provide richer contextual information than individual words and help identify commonly expressed tourist experiences.

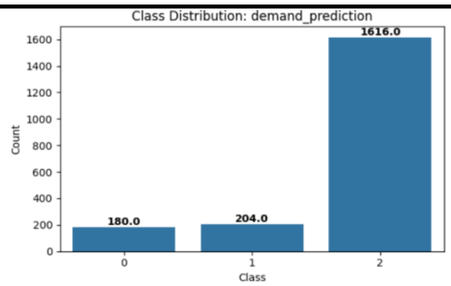


(a)



(b)

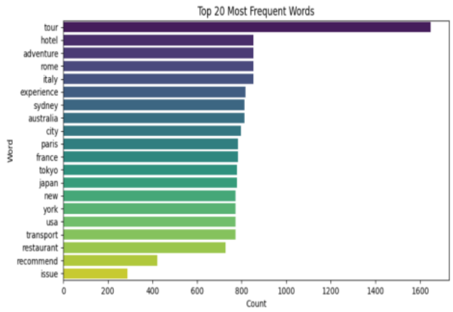
American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper



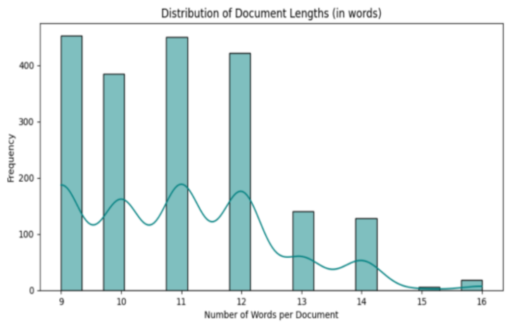
(c)

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

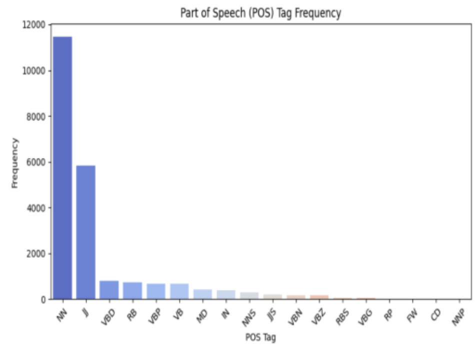
(a)



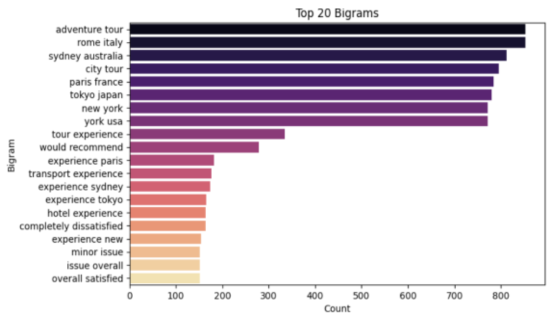
(b)



(c)



(d)



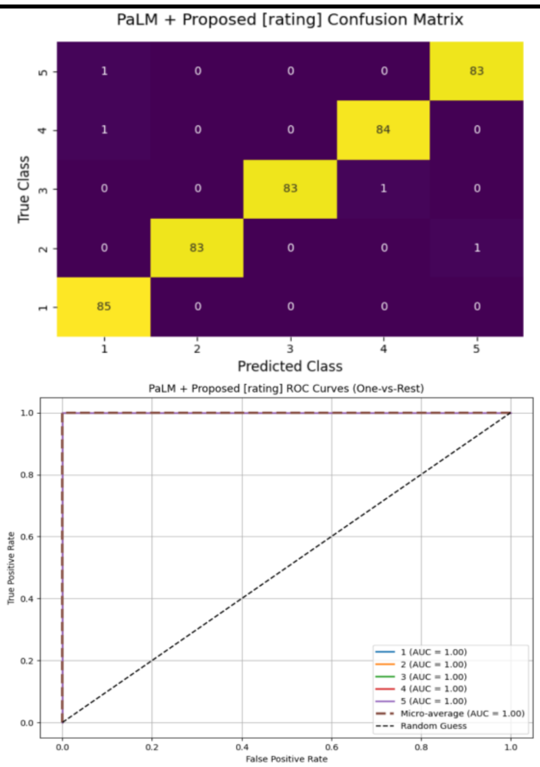
(e)

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.

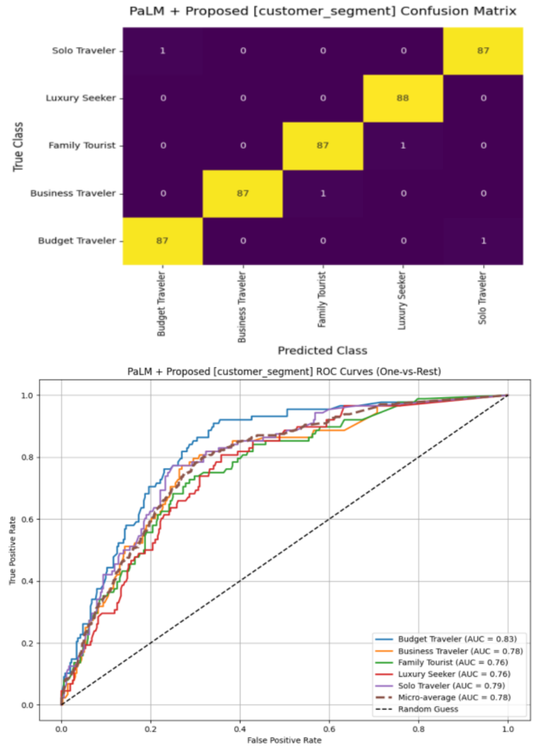
The proposed framework Fig. 5, utilizing the SLIM classifier, demonstrates near-ideal performance across all visual metrics.

- **Satisfaction Rating:** The confusion matrix indicates that most review ratings are correctly classified with only a limited number of misclassifications. The corresponding ROC curve demonstrates excellent discriminative capability across all rating categories.
- **Tourist Category:** The classification matrix shows highly accurate identification of different traveler segments, while the ROC curve confirms strong separability among the tourist categories.
- **Tourism Demand Level:** The confusion matrix illustrates precise classification of demand levels, and the ROC curves indicate consistently high prediction performance for all demand classes, demonstrating the robustness of the proposed framework.

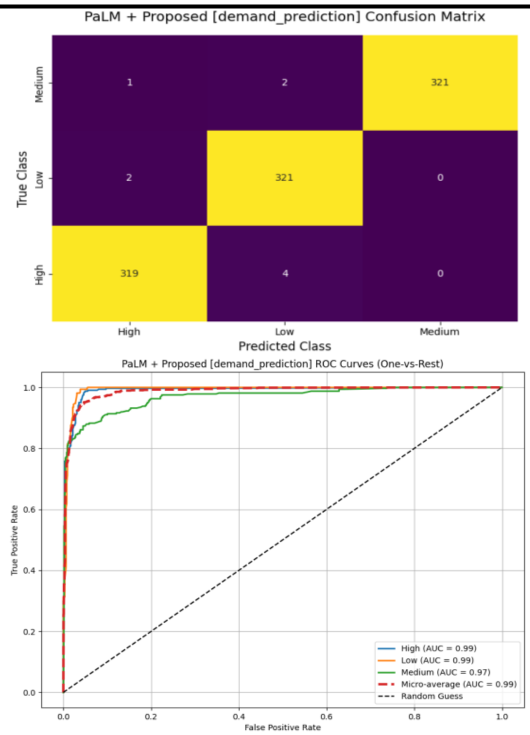
Zestera PUBLICATIONS
American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper



(a)



(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 assess the effectiveness of the proposed GPS-Tourism framework, its performance is compared with three baseline models: QDA, LDA, and HGB. Table 1 summarizes the Accuracy, Precision, Recall, and F1-Score obtained for the three prediction tasks, namely Tourist Rating, Customer Segment, and Demand Prediction.

The comparative analysis shows that the conventional models exhibit inconsistent performance across different prediction tasks. Although LDA achieves comparatively higher performance for tourist rating prediction, its effectiveness decreases considerably for customer segmentation. QDA produces moderate results but struggles to accurately classify complex behavioral patterns. Similarly, HGB provides limited performance, particularly for demand prediction, indicating its inability to fully exploit the semantic information extracted from tourism reviews.

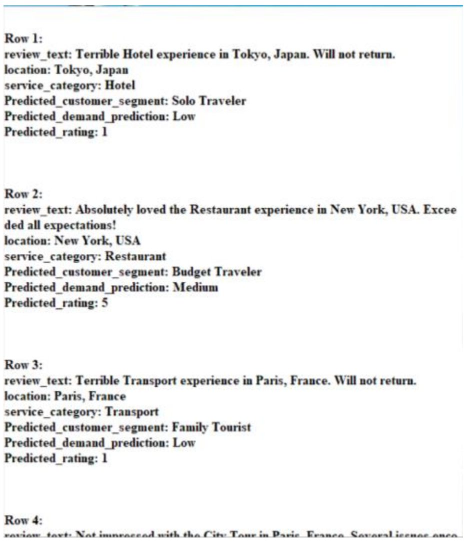
American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

In contrast, the proposed GPS-Tourism framework consistently achieves approximately 99% across all evaluation metrics and prediction tasks. These results demonstrate that the integration of PaLM, SMOTE, and SLIM provides superior feature representation, balanced learning, and highly accurate classification compared with the conventional baseline approaches.

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**

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.


Row 1:
review_text: Terrible Hotel experience in Tokyo, Japan. Will not return.
location: Tokyo, Japan
service_category: Hotel
Predicted_customer_segment: Solo Traveler
Predicted_demand_prediction: Low
Predicted_rating: 1

Row 2:
review_text: Absolutely loved the Restaurant experience in New York, USA. Exceeded all expectations!
location: New York, USA
service_category: Restaurant
Predicted_customer_segment: Budget Traveler
Predicted_demand_prediction: Medium
Predicted_rating: 5

Row 3:
review_text: Terrible Transport experience in Paris, France. Will not return.
location: Paris, France
service_category: Transport
Predicted_customer_segment: Family Tourist
Predicted_demand_prediction: Low
Predicted_rating: 1

Row 4:

Fig. 6: Sample predictions on test data.

5. CONCLUSION

The proposed GPS-Tourism framework offers an effective and intelligent approach for tourism demand forecasting by integrating semantic feature extraction, data balancing, and advanced classification into a unified prediction model. By utilizing PaLM, the framework successfully captures the contextual meaning of unstructured tourist reviews, enabling accurate analysis of customer opinions and travel behavior. The incorporation of SMOTE effectively addresses the problem of class imbalance by generating representative samples for underrepresented categories, thereby improving the learning capability of the prediction model. Furthermore, the SLIM classifier, implemented using an ensemble of oblique decision trees, efficiently models complex relationships within high-dimensional feature spaces while maintaining robust classification performance. Experimental results demonstrate that the proposed framework consistently outperforms QDA, LDA, and HGB across all prediction tasks, achieving accuracies of 99.05% for tourist rating prediction, 99.09% for customer segmentation, and 99.07% for tourism demand prediction, along with correspondingly high Precision, Recall, and F1-Score values. These findings confirm the effectiveness of combining semantic embeddings with balanced learning for accurate tourism analytics. Overall, the GPS-Tourism framework provides a reliable decision-support solution for tourism organizations by supporting demand forecasting, customer segmentation, resource planning, and personalized service delivery, making it well suited for modern data-driven tourism management applications.

REFERENCES

[1] World Tourism Organization (UNWTO), International Tourism Highlights, Madrid, Spain: UNWTO, 2023.

[2] U. Gretzel and D. R. Fesenmaier, "Persuasion in recommender systems," International Journal of Electronic Commerce, vol. 11, no. 2, pp. 81–100, 2006.

[3] X. Li, L. Wang, and Y. Zhang, "Online tourist reviews and tourism demand forecasting: A review," Tourism Management Perspectives, vol. 35, pp. 100710, 2020.

[4] H. Song and G. Li, "Tourism demand modelling and forecasting—A review of recent research," Tourism Management, vol. 29, no. 2, pp. 203–220, 2008.

[5] H. Song, G. Li, S. F. Witt, and B. Fei, "Tourism demand modelling and forecasting: How should demand be measured?" Tourism Economics, vol. 16, no. 1, pp. 63–81, 2010.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[8] Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[9] Yang, Y., Guo, J., and Sun, S., "Tourism demand forecasting and tourists' search behavior: Evidence from segmented Baidu search volume," Data Science and Management, vol. 4, pp. 1–12, 2021.

[10] Yu, Y., and Chen, X., "A tourism demand forecasting model based on Stacked Autoencoder and Long Short-Term Memory network," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 5, pp. 8991–9003, 2021.

[11] Khaidi, M., Mohamed, M. A. N., and Ahmad, N. A., "Tourism demand forecasting: A review on the variables and models," Journal of Physics:

Conference Series, vol. 1366, no. 1, Art. no. 012111, 2019.

[12] Li, Y., and Li, X., "Research on Tourism Consumer Demand Forecasting Based on Big Data Analysis," Journal of Physics: Conference Series, vol. 1744, Art. no. 032211, 2021.

[13] Mikhailov, S., and Kashevnik, A., "Tourist Behaviour Analysis Based on Digital Pattern of Life—An Approach and Case Study," Future Internet, vol. 12, no. 10, Art. no. 165, 2020.