STAP
Smart Technologies Academic Press

STAP Journal of Security Risk Management

https://www.jsrm.thestap.com/

A Smart Dashboard Framework for Urban Tourism Risk Analysis Using Deep Learning and Machine Learning

Adona Kulathinal Josephi[1], Mahmud Maqsood[2]

[1] *School of Computing, Ulster University, Belfast, Northern Ireland, United Kingdom*
[2] *School of Computing, Ulster University, Belfast, Northern Ireland, United Kingdom*

ARTICLE INFO

Article History

Received: 01-09-2025
Revised: 01-01-2026
Accepted: 12-01-2026
Published: 14-01-2026

Vol.2026, No.1

DOI:

***Corresponding author.**
Email: *Kulathinal_Joseph-A@ulster.ac.uk*

Orcid:

https://orcid.org/0000-0002-8478-6930

ABSTRACT

The study proposes an intelligent framework to evaluate tourism safety in Indian cities by integrating diverse, real-world data source, including crime statistics, hotel ratings, and user reviews. The methodology employs advanced artificial intelligence techniques, notably a fine-tuned BERT model for classifying user reviews into safety-related sentiment categories and XGBoost for predicting crime pattern prediction and city-level safety score computation. The analytical pipeline includes comprehensive data preprocessing, sentiment classification, predictive modeling, and cluster analysis to uncover patterns and associations. Cities are segmented into distinct risk categories based on crime density and public sentiment, enabling nuanced safety profiling. A noteworthy finding is the strong inverse relationship between tourist satisfaction and crime rates, underscoring the significant influence of safety perceptions on a destination's attractiveness. The final output is an interactive Power BI dashboard that supports real-time filtering, geospatial analysis, sentiment mapping, and predictive insights. This decision-support system enables travelers to make informed choices, assists policymakers in identifying high-risk areas, and assists urban planners in designing targeted safety interventions. Overall, the research addresses a critical gap in tourism safety information and demonstrates the potential of AI in developing data-driven, transparent, and responsive tools for smart tourism management.

Keywords: Crime analytics; tourism safety; sentiment analysis; BERT; XGBoost; predictive modelling; clustering; urban intelligence; Power BI dashboard; smart tourism.

How to cite the article

1. Introduction

The hospitality and tourism sector in India is increasingly being affected by rising crime rates, which pose significant risks to both tourist and local communities. while platforms such as TripAdvisor and Airbnb offer user reviews, they typically lack integration with comprehensive, real time crime data. As a result, tourists often make travel decisions without clear understanding of local safety conditions. There is an evident connection between the tourism and victimization of tourists. This is due to significant factors that influence the tourist choice [1].

This research addresses the critical gap in travel safety information by proposing a data-driven framework that combines tourism and accommodation review data with official crime reports, and temporal-spatial trends to analyze correlation between crime patterns and tourism activity. Using the merged dataset of tourism, accommodation and crime, the study applies data analysis, machine learning and deep learning techniques to uncover key insights and patterns and presents the results through an interactive dashboard.

Several issues emerged because of tourism industries explosive growth. Overpopulation at tourist destinations affects the environment and leads to safety concerns [2]. Unlike conventional travel advisory platforms, this system allows users to filter data by crime type, city and time, thereby offering a more nuanced and actionable view of destination safety. The objective is to empower travellers to make informed decision based on authentic, data-backed insights, while also offering a tool that can support tourism boards, local authorities, and smart city initiatives in enhancing public safety. This project not only contributes to safer and more informed tourism but also represents a novel step in integrating heterogeneous datasets for public benefits through intuitive visual analytics.

2. Literature Review

The hospitality and tourism sector holds a vital position within the economy. Tourism serves as a driving force for economic growth, a stabilizing factor, and a unifying element in society by establishing direct and indirect links with various industries such as agriculture, manufacturing, transportation, trade, and others [3]. Tourism is a widespread and significant phenomenon, particularly for developing regions, where it is often seen as beneficial. However, recent studies highlight a different perspective. The temporary influx of large numbers of visitors into a concentrated area introduces outsiders to local communities, which can sometimes lead to challenges, including an increase in crime [4]. Most of the research bulk in tourism and crime is targeted at incidents of crimes by locals against tourists, which referred to as tourist-oriented crime, and has been well charted in quantitative and qualitative research. In contrast, tourist crimes against locals and those offenses with fellow tourists, are considered largely uncharted areas of concern within the academic literature. It also includes, within the broader context of the study of crime and tourism, crimes among locals that are related to tourism but do not directly involve visitors [5].

The connection between tourism, safety, and security has been examined by both tourism researchers and criminologists. Crime rates in tourist resorts tend to be elevated, particularly during peak travel seasons [6]. In 2016, India's tourism sector contributed approximately 208.9 billion U.S. dollars to the country's GDP, making it the second-largest contributor to GDP from tourism in the Asia-Pacific region, following China. Given its significant economic impact, tourism plays a crucial role in economic growth. Studies highlight that various criminal activities and terrorism influence tourists' decisions when choosing a destination. Violent crimes, such as murder and theft, have been identified as key factors affecting tourist preferences [7]. Academic studies have definitively established the link between tourism and crime. Since high criminal levels lead fewer people and lower hotel occupancy, travel destinations and lodging businesses are negatively affected by crime [8]. Early data shows that during the most travelled seasons, property crime rates in Miami surged. Research often points to tourist destinations as hotspots for criminal activities; for example, studies show how perceived and actual safety risks affect consumers' choices. Similar patterns were observed in Mexico, there was a correlation of tourist and crime. Research carried out in locations such Hawaii and Tonga has also shown that unexpected consequence of tourism growth is a rise in crime rates [9]. Furthermore, present tourism sites including TripAdvisor and Airbnb offer useful reviews but do not include statistics on crime rates, therefore restricting the extent of information open to users. Different studies have investigated visualizing and analyzing crime patterns. For example, research investigates how visitors affect crime rates,

thereby exploring the link between travel and crime. The author explored using empirical techniques how many different crimes across 103 Italian regions in 2005 were impacted by tourist numbers given sociodemographic and economic variables. it accounts for crime spillover effects, and it uses two geographic models: the spatial error and the spatial lag model. This study also investigates the effect of multiple tourist destinations urban, mountain and coastal areas on crime rates. An evaluation is done of the social costs of crime associated with the rise in tourism [10].

A case study of St. Lucia identified that among the property-related crimes against tourists, theft and burglary were relatively common, especially in those areas with significant tourism development. It reiterated that even though the citizens are more often victims of crime, incidents in tourist areas may have consequences for the tourism industry [11]. Another research investigates how the demand for tourists in Indian states is affected by crime and terrorist activity. and the results validate that violent and non-violent crimes have differing effects on India's tourism demand. Tourist arrivals decrease by 5.2% and 1.6% for every 10% rise in violent and non-violent crimes, respectively [12]. This project attempts to examine the effects of crime at different spatial range. To achieve this purpose, to analyze many data sources, such as travel evaluations, housing ratings, and local crime statistics in India. The findings of this study help us to understand the effects of crime and tourism industry and provide insightful suggestions for effective management of tourism destinations to ensure the safety of both traveller's and hosting communities.

3. Research Methodology

The methodical procedures used to examine the effect of crime on Indian tourism are described in this section. The composition of methodology means data collection, preprocessing of the data analysis development of dashboard, and implementation of respective features to get comprehensive insights regarding the effect of crime on tourism in India.

3.1 Data Acquisition and Preprocessing

This study collects both structured and unstructured data from various sources, including The India Crime Dataset, OYO Hotel Reviews and Indian Place to Visit Review dataset available on Kaggle dataset repository. The crime dataset gives comprehensive insight into criminal activities in several Indian cities from 2020 to 2024, with a breakdown by type of crime, date and time of occurrence of crime, victim description such as age and gender of the victim. The OYO Hotel review dataset contain ratings and review for different hotel, beneficial in assessing how the user sentiments affect visitors view on safety [13], thus helping the business in enhancing the services and improving customer satisfaction. The tourism dataset encompasses city-based information, reviews and ratings left by visitors at various locations. Data preprocessing was performed to provide data quality, homogeneity and preparation for subsequent analysis [14], employing R with relevant libraries such as dplyr, data.table and tidytext. It comprised several steps such as cleansing, feature engineering, and normalization of three fundamental datasets. Irrelevant columns were identified and removed from each dataset to improve model performance. Important preprocessing task was conducting sentiment analysis with the use of Bidirectional Encoder Representations from Transformers (BERT), employs a transformer model to completely grasp the context of a word in a sentence in both directions, thus being more successful in classification tasks compared to conventional Natural Language Processing (NLP) models [15]. Using meticulously fine-tuned BERT models, text reviews collected from tourism as well as hotel datasets were labelled into four sentiments. To ensure consistency during merging, city names were standardized using string normalization techniques. A left join was applied based on city names as the key to combine the cleaned Crime, Tourism, and Hotel datasets. The approach enabled the retention of all the cities in the crime dataset.

3.2 Descriptive Statistics and Exploratory Data Analysis

Descriptive and Exploratory Data Analysis (EDA) were performed to check preliminary analysis on data to identify patterns, identify anomalies and to check assumptions using summary statistics and visualization [16]. Descriptive statistics were obtained using the summary () function to have an initial view of numerical variables. Visualization including histogram, boxplots, bar charts were used for understanding data distribution. Use of boxplot is essential to identify any outliers of the primary numeric fields. Heatmap of correlation by using the ggcorrplot package used to visualize an image

of pairwise correlation strength and direction between variables. These tests permitted simple comprehension of the data, shape, and for variable interactions.

3.3 Normality Test

To verify the dataset were normally distributed, a complete normality analysis was conducted. The skewness and kurtosis descriptive statistics were calculated for all the numeric variables to compare the symmetry and peakiness of the distribution of data respectively. The Anderson–Darling test was selected as a more suitable alternative to the Shapiro–Wilk test for determining the normality of continuous variables. The Anderson–Darling test is more sensitive towards the tails of the distribution [17], best suited for large data sets where small non-normalities are common but not consequential. The test was applied to all numeric fields of the dataset. The findings were that all variables possessed p-values < 0.05, leading to the null hypothesis of normality being rejected. Visual inspection was also conducted with quantile-quantile (Q-Q) plots, which enabled comparison between the sample distribution and an imaginary normal distribution. This is a vital step because most statistical methods, such as linear regression assume normality of the data. Therefore, understanding the distribution of variables directs the selection of appropriate statistical models or transformation techniques in the analysis.

3.4 Imputation Techniques for Missing Values

To ensure dataset completeness and to prevent data loss [18], a targeted imputation strategy was adopted according to variable type. For numeric missing values were imputed using column wise median imputation as it is robust to skewness and better preserve the central tendency in non-normal data. This method will maintain the overall statistical distribution of data. For categorical missing values were tagged as unknown. This strategy prevents the removal of potentially useful records. Additionally, it also enables clear accounting for data gaps downstream models and visualisation. This dual strategy guaranteed that the dataset integrity was maintained across all merged sources and no rows were removed unnecessarily, supporting reliable downstream analysis and modelling.

3.5 Feature Engineering

A comprehensive feature engineering framework was utilized to transform raw datasets into structured and meaningful features suitable for predictive modelling and analysis. Aggregated city-based features were first extracted, including the number of reported crimes by city, average tourism rating derived from reviews of individual tourist destinations, and average hotel rating derived from reviews of individual OYO hotels. Moreover, sentiment analysis of user created textual reviews was classified into four thematic categories of sentiment: bad review, good review, street scam and fraud, and drug safety and violent crimes. The number of reviews within each sentiment category was computed separately for both tourism and hotel datasets for each city, thereby enabling fine-grained analysis of public opinion and user experience across different domains. Additionally, offences were categorized into three risk levels high, medium, and low based on the seriousness and nature of the crimes. High-risk crimes included such categories as homicide, sexual assault, kidnapping, and gun crimes; medium risk crimes were such as assault, burglary, fraud, and drug crimes, while low-risk crimes were primarily traffic cases. This risk-based classification was used to categorize cities according to their dominant crime profile, which assigned each city a categorical risk rating: high, medium, or low. These constructed attributes in combination with each other present a good approximation of each city's safety, reputation, and quality of service, forming an important foundation for further analysis and modelling.

Table 1. Merged Dataset before Feature Engineering

City	VictimGender	Victim_Age	Crime_Description	Review_Text	Review_Rating
Delhi		25	Assault	Felt unsafe in the lobby	3
Agra	Male	30	Scam	Great room but scam outside	4.2
Chennai	F	46	Drug Use	Drugs seen near hotel	2
Mumbai	-	40	Pickpocketing	Pickpocketed at market	1.5

Table 2. Merged Dataset after Feature Engineering

City	VictimGender	VictimAge	CrimeDescription	Review_Text	Sentiment_Label	Avg_Rating	Risk level
Delhi	Unknown	25	Unknown	Felt unsafe in the lobby	Bad Review	3.5	Low
Agra	Male	30	Scam	Great room but scam outside	Street Scam & Fraud	4.2	Medium
Chennai	Female	46	Drug Use	Drugs seen near hotel	Drug Safety & Violent Crime	4	High
Mumbai	Unknown	40	Pickpocketing	Pickpocketed at market	Street Scam & Fraud	3.8	Medium

3.6 Cluster Analysis

Cluster analysis was implemented to identify natural patterns among Indian cities according to crime records, tourism satisfaction and accommodation quality. Due to the skewed distribution of several ratings and review variables, log transformation was applied to reduce skewness. z-score normalization was applied after transformation to guarantee that each numeric variable made an equal contribution to the clustering process. K-mean clustering is applied to the normalized dataset. To balance within cluster variation and interpretability, the ideal number of clusters is determined through Elbow method. K-Mean clustering was then conducted with three centres, and results were visualized using Principal Component Analysis (PCA) to capture the majority of features variance. The resulting clusters revealed distinct patterns among cities with differing crime profiles and safety ratings, enabling meaningful classification for subsequent analysis.

3.7 Machine Learning and Deep Learning Techniques Used

This research applies various machine learning techniques to classify user satisfaction and crime related information from user reviews through sentiment analysis, crime trend forecasts for the upcoming years, and safety score prediction of cities. Each task uses different data preprocessing, model training, and evaluation strategies, demonstrating the flexibility of machine learning to deal with actual problems.

3.7.1 BERT-Based Sentiment Analysis and Dataset creation

A key aspect of this project was the labelling of tourist and accommodation reviews into sentiment categories that are quite domain-specific in relation to tourism safety. Because readily available datasets often only include very broad sentiment labels such as "positive," "negative," or "neutral," they do not capture domain-specific sentiment such as concerns about fraud, crime, or unhappiness related to safety in the tourism sector. To address this issue, a custom dataset was constructed using an LLM which was aimed at reflecting the type of sentiment that would be expressed by tourists about safety in Indian cities.

The information was artificially created with the deepseek-r1:1.5b language model through the Ollama API. For each category, prompts were carefully crafted to instruct the model to generate one-line reviews in plain English, without the use of emojis or subjective exaggeration. The model output was cleaned using regular expressions to remove extraneous tags such as <think>, ensuring that only relevant, natural-language content was retained. To generate synthetic one-line reviews corresponding to various real-world tourism scenarios.

- Good Review
- Bad Review
- Street Scams & Fraud
- Drug Safety & Violent Crimes

Every review was labeled systematically according to the provided input prompt, and the whole collection of generated examples was pooled and incorporatedVS. In a well-organized unified dataset. The dataset was kept in CSV format with two kinds of fields: review text and sentiment classification. In this way, supervised control, class balance, and accurate semantic correlation with the safety-topical themes of the undertaking were provided.

4. Dataset evaluation

To ensure the quality and robustness of the dataset produced by LLM before fine-tuning the BERT model, a comprehensive assessment was conducted on the most critical aspects of label consistency, linguistic quality, semantic alignment, and class distribution. Among the primary quality metrics was model agreement confidence, which assesses classification consistency according to multiple inferences. The dataset achieved an agreement of 76.01%, which indicates internal label coherence and semantic clarity.The data maintained balanced class distribution, with each of the four classes accounting for approximately 25% of the whole, ensuring fairness during training.

Text quality tests included:

• A Flesch Reading Ease score of 60.26, reaffirming accessibility and natural language.

• GPT-2 perplexity testing, with most reviews scoring below 100, reaffirming human-like fluency.

• TextBlob sentiment analysis, which reaffirmed high correlation between review sentiment and class labels given.

For structural and lexical analysis, the dataset had an average lexical diversity of 0.025, and an average review length of 38.5 words, both of which are within standard online review norms. Redundancy checks via string matching and cosine similarity confirmed the absence of duplicate content.

Finally, a grammar quality test with language_tool_python revealed just slight errors, with no critical problems impacting readability or usefulness of the data.These assessments validate that the dataset is semantically correct, structurally healthy, and suitable for supervised learning tasks in sentiment and risk classification.

4.1 model training

To classify user generated text reviews into significant categories, the BertForSequenceClassification model from Hugging Faces Transformers library was fine tuned. This model extends the pertained bert-base-uncased architecture by adding a classification layer to output predictions for four sentiment classes: Street Scams& Fraud, Drug Safety and Violent Crimes, Good Reviews and Bad Reviews. A maximum sequence length of 128 tokens were used to encode the reviews after tokenised using the BERT tokenizer. The model was fine-tuned from the BERT based model Bert-base-uncased by adding a classification layer to output four sentiment classes. “BERT consists of Transformer encoder, which receives the input

from embedding layer and flows through the multiple self-attention layer to identify the appropriate relations between words in a text" [19]. The CrossEntropyLoss function and AdamW optimiser were used to train the model across five epochs with a batch size of 32. A training-validation ratio of 80-20 is used to the dataset. To guarantee the correct convergence, accuracy and loss measures were monitored throughout the training. The fine-tuned model is then used to classify reviews from accommodation and tourism datasets. To create a structured sentiment distribution that was incorporated into the master dataset, categorised sentiments were then combined at the city level. This contributed to the evaluation of each city's risk profiles, user satisfaction and safety.

(A) Crime Trend Forecasting

To model the temporal fluctuation of crimes within Indian cities, adopted a machine learning approach centred on Extreme Gradient Boosting (XGBoost). Crime forecasting has become widely recognised in the past few years as it helps authorities to address crimes analytically [20]. This approach is chosen because it is reliable when dealing with non-linear connections and can use to incorporate historical trends through engineered features. City and risk level specific annual crime figures were combined. The key features used for modelling included the prior year crime count, risk category, year and city. For smooth integration into XGBoost framework, Categorical variables are one-hot encoded. Standard metrics, such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) is used to assess the model after the model was trained using squared error loss function. Future crime trend for 2025 and 2026 were forecast recursively with the recent data available. Additionally, linear regression with lag and temporal factors and Exponential Smoothing (ETS) used to test XGBoost performance comparison. The forecasted data is then implemented in dashboard for future crime trend for all cities as inputs to downstream risk mapping

(B) Safety Score Predictions

To construct a quantitative city level safety score, a composite overall safety score is developed by using heterogeneous data source encompassing crime statistics, tourism experience and accommodation reviews. A structured methodology was employed to preprocess, normalize and aggregate these indicates into unified safety score. All Significant variables were normalized using min-max scaling to ensure compatibility across variables. This approach constrained input values to the [0, 1] range, reducing scale-induced bias in subsequent computations. Domain knowledge weights were then assigned to all variables to represent their relative importance in perceived safety. These three scores were then linearly combined with the following formula. Each city was assigned three domain-specific sub-scores Crime Score, Tourism Score, and Accommodation Score based on normalized features and expert-defined weights reflecting their influence on safety perception.

- Crime Score (Cs) was computed as:

$$Cs = 3 * TC + 2 * DVCT + 1.5 * SSFT + 2 * DVCH \quad (1)$$

where TC denotes normalized total crime counts, $DVCT$ is normalized drug/violent crimes in tourism reviews, $SSFT$ represents normalized tourism-related scam and fraud incidents, and $DVCH$ refers to normalized hotel-related drug/violent crime occurrences.

- Tourism Score (Ts) rewarded positive tourist sentiment while penalizing safety-related concerns:

$$Ts = 2 * ART - 1 * BRT - 1.5 * SSFT \ (2)$$

Where ART is the average normalized tourist rating, BRT is the count of negative tourism reviews.

- Accommodation Score (As) incorporated user satisfaction and safety indicators in hotel reviews:

$$As = 2 * ARH - 1 * BRH - 1.5 * SSFH \ (3)$$

Where ARH denotes average normalized hotel rating, BRH is the count of bad hotel reviews, and $SSFH$ captures normalized fraud/scam reports in hotel data.

- Overall Safety Score Aggregation

The three component scores were linearly combined to generate a raw safety score (Rs), which integrates crime, tourism, and accommodation dimensions using:

$$Rs = 2 * Ts + 2 * As - 3 * Cs \quad (4)$$

The final Overall Safety Score (S) was then rescaled into a bounded interpretability range [1.1, 9.9] using:

$$S = 1.1 + (Rs - \frac{\min(Rs)}{\max(Rs)} - \min(Rs)) \text{ *(9.9-1.1)} \quad (5)$$

This normalization enables safety scores to be visually comparable across different urban contexts.

While this formula was rational weighting of relevant characteristics based on domain experience, it was still heuristic. To validate and potentially refine this technique, regression models of XGBoost and a deep learning neural network were created to learn the same Overall Safety Score on the same characteristics. These models identified whether data-driven approaches could discover higher-order, nonlinear dependencies outside the ability of the static formula to detect. By comparing prediction performance through RMSE and R^2, this two-tiered strategy not only established the validity of the weighted score but also provided a predictive model template to be employed in the future against unseen cities or novel datasets.

In this estimation module of safety score, two advanced regression modelsXGBoost Regressor and a Deep Learning Neural Network (DNN) were used. The dataset was divided into input variables (excluding city name and target column) and the target variable (Overall_Safety_Score). 80-20 split was performed to achieve training and test datasets, deserving robust model evaluation.

The first model, XGBoost, is a gradient-boosted decision tree model with good speed and prediction performance, especially in structured/tabular data. It was trained on 500 estimators, a learning rate of 0.05, and a depth of 6 for maximum to achieve a balance between complexity and generalization. After training, the model was evaluated using the test set based on Mean Squared Error (MSE) and R-squared (R^2) as the key performance indicators. These values demonstrated how well the estimated safety scores conformed to actual values, where lower MSE and greater R^2 reflected higher performance.

To compare this, a Deep Learning Neural Network was constructed using TensorFlow/Keras to see whether deeper nonlinear patterns could be learned using the same features. The three hidden layers of the neural network consisted of two hidden layers of 64 units and one hidden layer of 32 units using ReLU activation, and one output layer for regression. Before input data was fed into the network, it was normalized using StandardScaler to ensure all features had an equal contribution. The model was then compiled using the Adam optimizer and trained for 30 epochs while observing training and validation loss. Evaluation was again on MSE and R^2 with addition of a training curve plot to verify model convergence. Both models possess varying strengths XGBoost possesses great interpretability and ranking of feature importance, while the neural network utilizes depth to learn fine patterns. Having both models allows comparative analysis and gives robustness to the predicted safety scores. The methodology ensures that the strength of both tree-based and neural architecture is utilized in safety assessment.

5. Result and Discussion

5.1 Exploratory Data Analysis

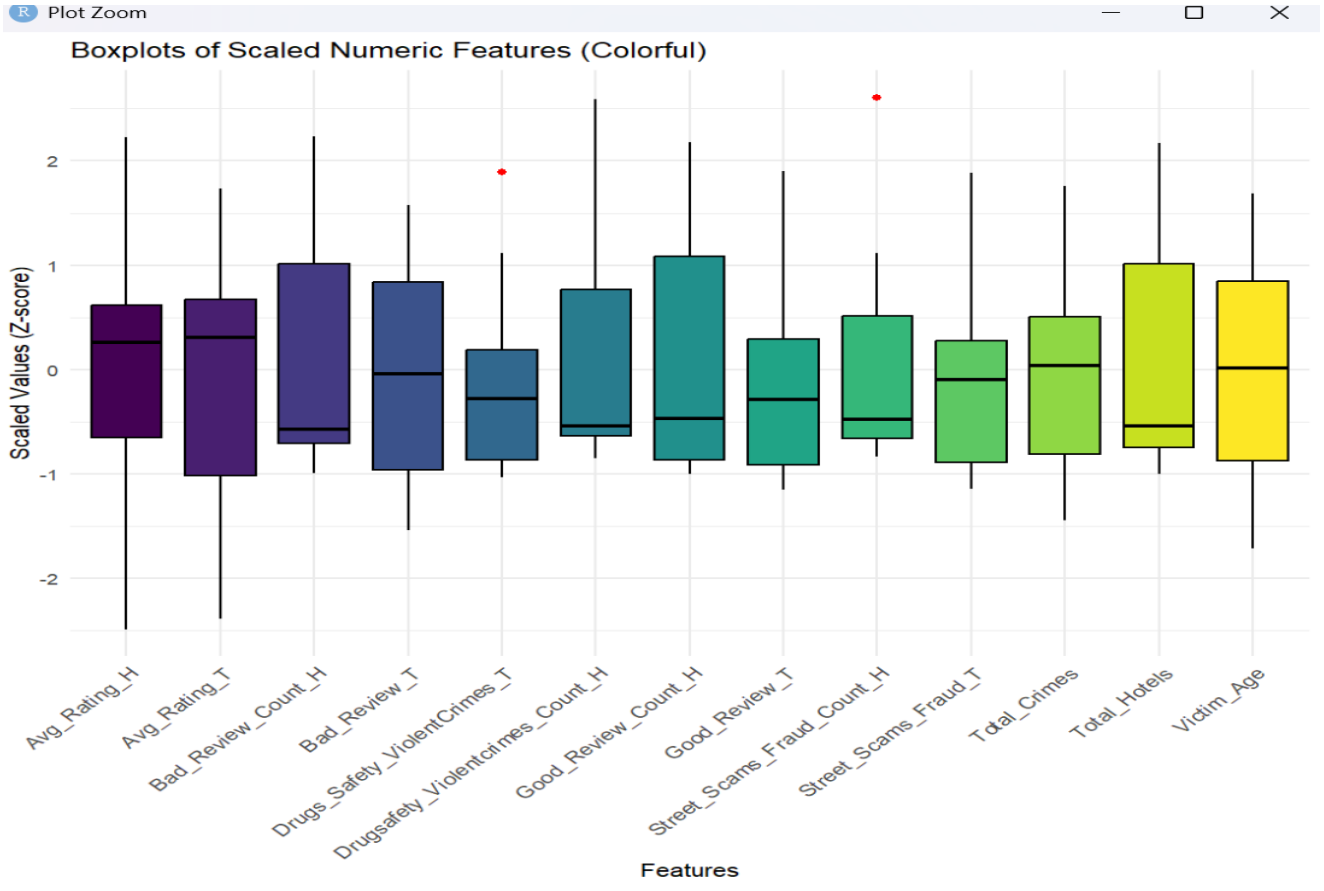

Figure 1. Box Plot of Numerical Variables by Outcome

Figure 1 illustrates the distribution of key numerical variables by safety outcome using scaled box plots. The variables Good Review Count Tourism, Total Hotels, and Street Scams Fraud Tourism have high variance, indicating scams, tourist service density, and attractions are very different in each city. However, the very narrow interquartile ranges of the variables such as Victim Age, Average Rating Hotel and Average Rating Tourism indicate consistent visitor satisfaction and victim demographic activity. Outliers in some of the features such as, Good Review Count Tourism and Street Scams Fraud Tourism, represent locations that are way off the mean regarding reported street scams and visitor reviews. Observe a high degree of variation for each of the crime and safety indicators, indicating that safety issues are highly variable among cities. Such patterns, in addition to the heterogeneous urban profiles by city, provide the rationale for local safety rankings and focused policy design.

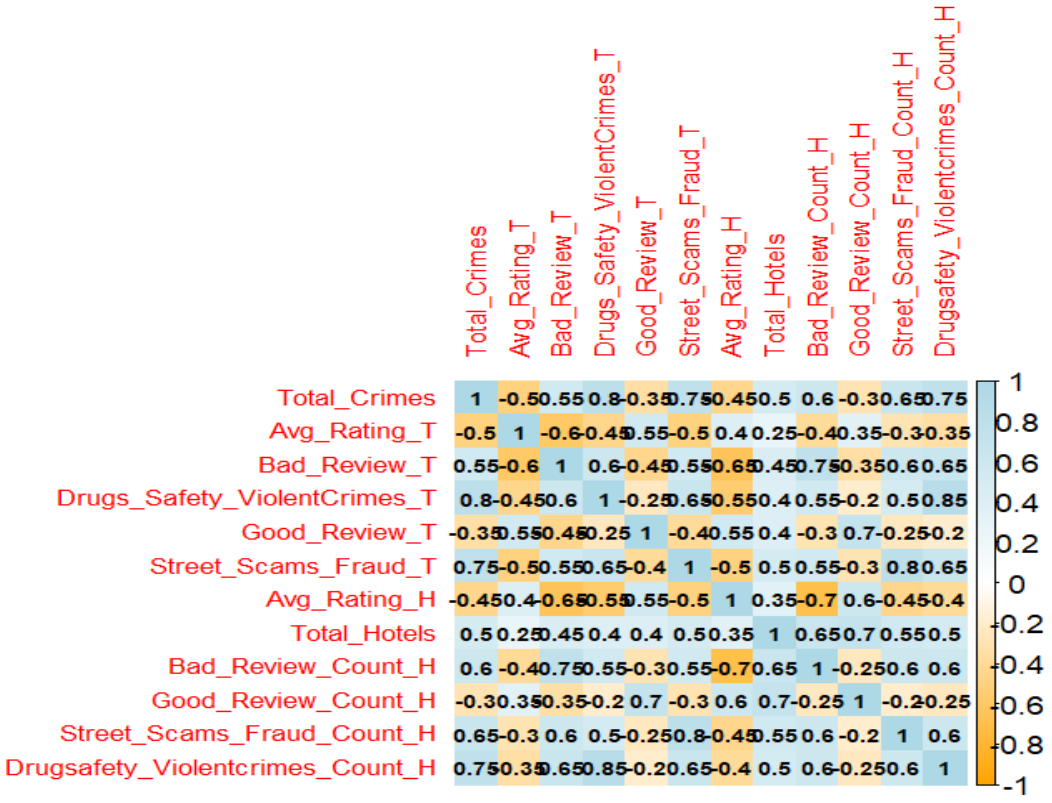

Figure 2. Correlation Heatmap

The correlation heatmap analysis provides valuable information about the relationships between crime statistics, tourist ratings, and hotel data. There was a very high positive correlation between Total Crimes and both Violent Crimes and Scam related crimes in tourism industry with respective coefficients of r = 0.80 and r = 0.75), indicating that areas with high general crime levels are also heavily affected by drug crime and street crime. In addition, Bad Review by tourists had substantial positive correlations with Scam and violent crimes in tourist area (r=0.55, r = 0.60), indicating that adverse tourist reviews often relate to perceived or real safety concerns. Conversely, Average Tourist Ratings was negatively correlated with Total Crimes (r = –0.50) and Bad Review (r = –0.60), which suggests that high crime and negative reviews repress tourist overall satisfaction. These correlations are relevant to learn how crime statistics and safety perceptions impact tourism-related feedback and can be used for informing feature selection in predictive modelling.

5.2 Normality Test

Normality tests indicated that most numeric variables in the data are not normally distributed. Only Victim Age, Average Tourism Rating and Average Accommodation Rating had near-normal distributions with skewness close to zero, kurtosis close to 3, and Shapiro-Wilk p-values greater than 0.05. The rest of the variables exhibited some degree of right-skewness and heavy-tailed distributions features such as Scams and Violent crimes in hotel clusters expressed through high skewness above 1.5, high kurtosis above 4, and strong Shapiro-Wilk test rejections ($p < 0.05$). Visual examination through Q-Q plots and histograms verified these with substantial right-tailed departures for review count and crime related variables. These results suggest that data transformations or non-parametric methods are necessary for formal statistical analysis because parametric tests assume normality.

5.3 Cluster Analysis

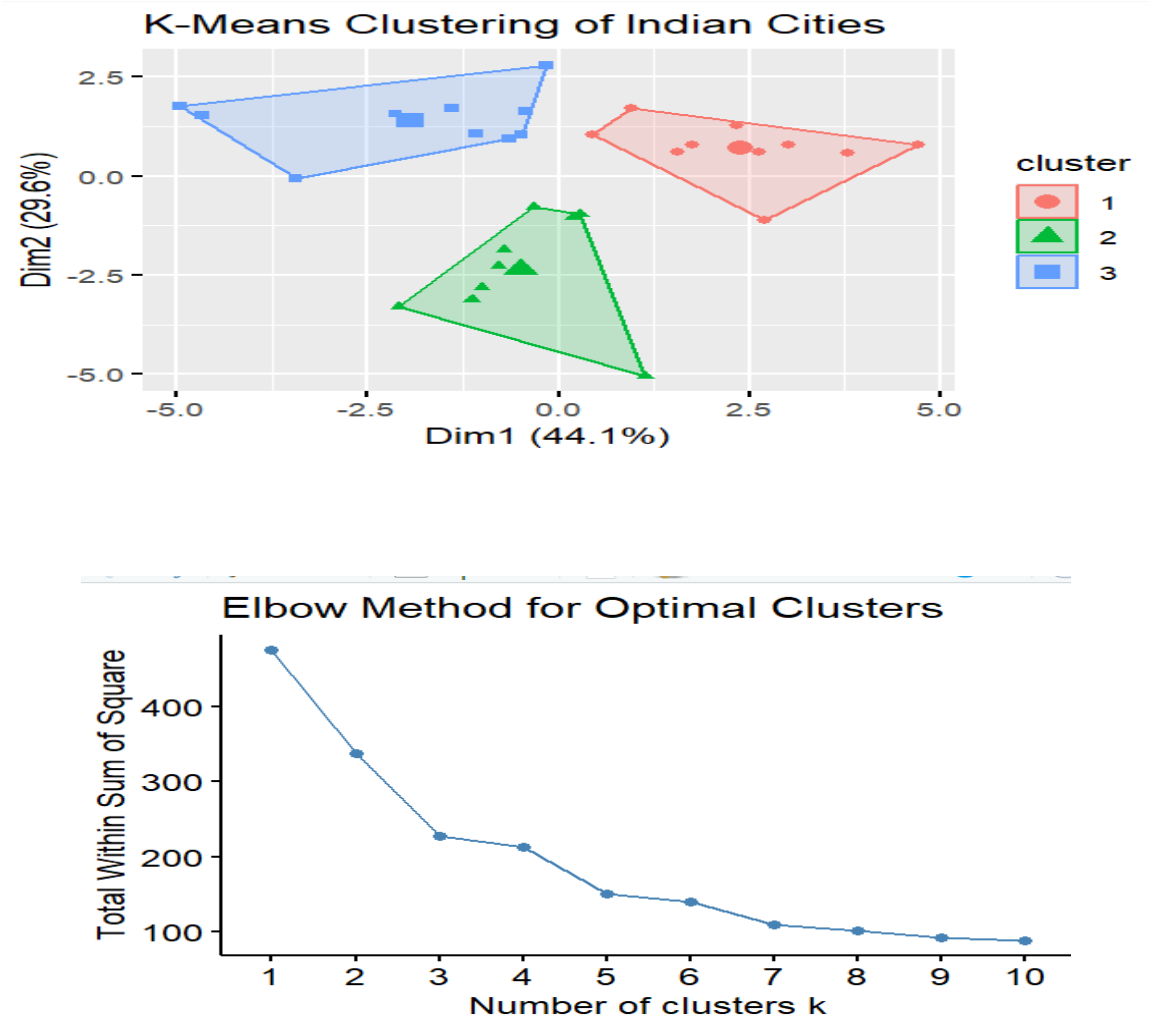

Figure 3. K-Mean Cluster Analysis on Indian Cities

K-men Clustering is performed on standardized, and log transformed crime and review related features of Indian cities. The optimal number of clusters was determined to be three using Elbow method [21]. The first two principal components together capture 73.7% of the total variance, means the plot preserves most of the original data's structure. Cluster 1: predominantly includes cities with moderate crime rates and balanced review scores. Cluster 2: comprises cities characterized by relatively lower crime levels and better safety indicators. Cluster 3: groups' cities with higher reported incidents, lower average ratings and high count of sentiment features. The identified clusters were integrated into Power BI dashboard through dynamic map visuals, allowing users to filter cities by cluster and visualize risk categories geographically using color coded markers. Clustering offers a strong data-driven approach to identifying similar city crime and tourism profiles, which can guide urban development decision-making, police policy in relevant regions to prevent crimes [22] and tourist safety planning.

6. Machine. Learning and Deep Learning Models

6.1 BERT-Based Sentiment Classification Performance

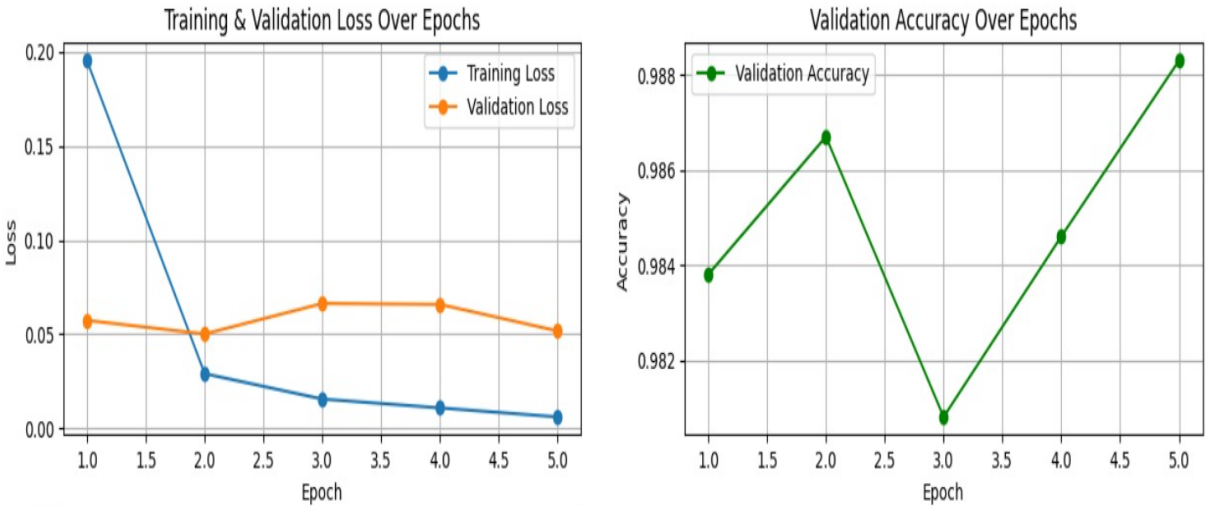

Figure 4. Training and Validation Loss Over Epochs and Validation Accuracy Over Epochs.

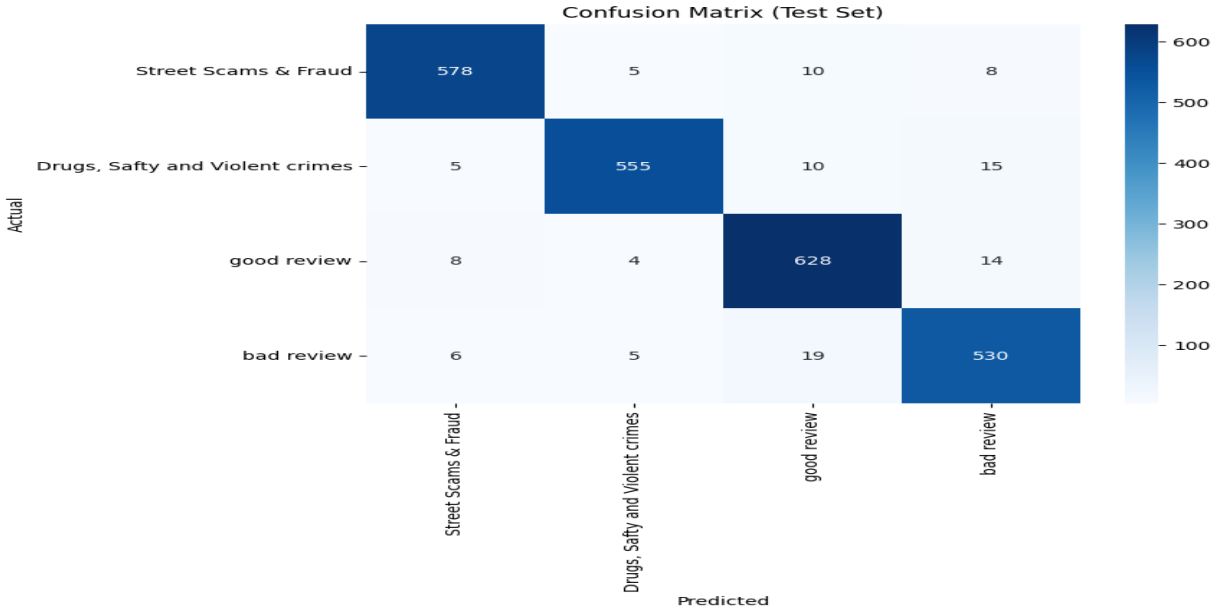

Figure 5. Confusion Matrix of Crime Classification

The fine-tuned BERT model achieved high performance in the multi-class classification of user reviews into four classes. The training plots and performance measures of the BERT model show its better capability in categorizing reviews. The loss plots and accuracy reflect rapid convergence with training accuracy from 95.17% (Epoch 1) to 99.79% (Epoch 5), while validation accuracy reached a plateau at 98.83%, reflecting minimal overfitting. The loss plots also support this, with sharp decrease in training loss from 0.1953 to 0.0060 and stable validation loss 0.05–0.06, supporting robust generalization. The precision-recall balance (0.9883) and F1-score highlight the stability of the model across all four classes. The corresponding confusion matrix(fig5) conforms strong predictive accuracy with minor misclassifications which is common in natural

language tasks. The low rate of entropy, combined with high classification accuracy, reflects a high confidence prediction distribution, rather than random guessing. BERT outperforms lexicon-based methods [23], which limited to binary polarity and does not capture contextual or domain specific semantics through contextual embedding's with evidence based claims [24]. The validation performance consistency despite increased training accuracy highlight's reliability for real-world use.

6.2. Crime Trend Forecasting

In predicting Indian crime trend across risk categories, the XGBoost model consistently achieved the highest predictive accuracy out of three forecasting techniques investigated. The model forecasted the total number of crimes in the country with a MAE of 6.05, MAPE of 29.5 percent and a RMSE of 9.14 respectively. The results significantly surpassed those obtained using traditional approaches. The ETS model reported an RMSE of 11.3 for total crimes, while linear regression obtained the least accurate results with an RMSE of 12.4 lacked the flexibility to capture temporal dynamic effectively.

Although ETS performed reasonably well in forecasting high and medium risk crimes univariate nature restricted the capacity to adjust to latest changes and geographical heterogeneity in crime patterns. The forecast produced by ETS depend on weighted average of historical data, where the weights drop exponentially according to the age of information [25]. The performance of XGBoost can be linked to the capacity to encode categorical factors and integrate lagged values as input features. These capabilities enable the model to generate reliable predictions for multiple risk categories.

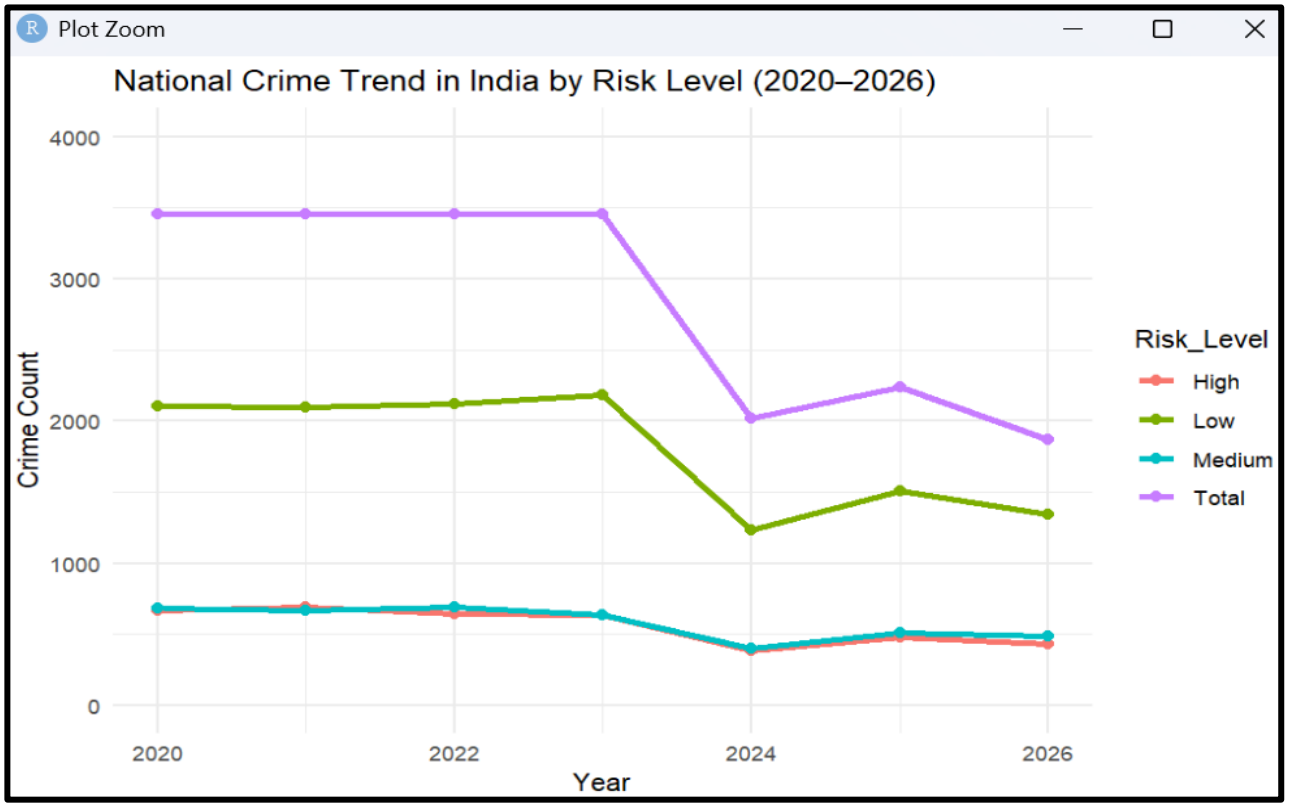

Figure 6. National Crime Trend in India by Risk Level

Figure 6 illustrates the national crime trend from 2020 to 2026. The historical and projected data are clearly aligned. The model captures a substantial decline in total and low risk crimes beginning in 2024 with stabilization observed towards 2026.This modelled crime trend forecasting provide meaningful insights for policymakers and reinforces the methods practical applicability in long term crime forecasting.

6.3 City Safety Score Prediction Results

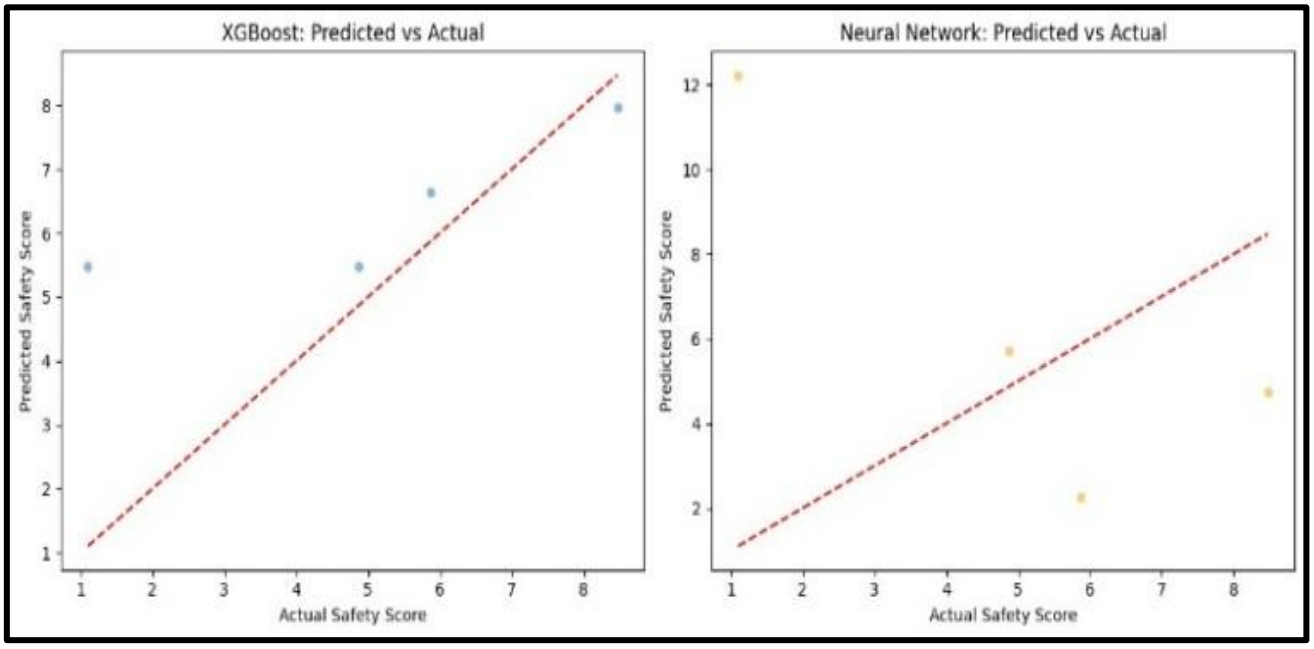

Figure 7. XGBoost and Neural Network Model Evaluation

The comparative analysis of safety score prediction models yield considerably different results for the XGBoost and Neural Network approaches. The XGBoost model demonstrated strong predictive performance with an MSE of 5.09 and R^2 0.27. The result of Residual analysis shows mostly normally distributed errors centred around zero, although slight positive skewness indicating sporadic under prediction of higher values of safety scores. The Neural Network solution performed much worse in all measures. With an MSE of 26.02 and negative R^2 value -2.71, suggesting that the model failed to generalise effectively. Error distribution analysis showed a heavy-tailed behaviour, with numerous large-magnitude residuals pointing towards systematic inability to predict. Training dynamics reflected unstable convergence behaviour, with validation loss reflecting large epoch-to-epoch variance. This difference in performance highlights comparative advantages of tree-based ensemble learning algorithms over deep learning algorithms to structured, medium-dimensional sparse sampling regression problems with sparse training sets. The nature feature selection qualities of XGBoost and insensitivity to noise appear particularly suited to safety score prediction, as opposed to greater parameter complexity within the neural framework against it per available training sample.

6.4 Tourism Safety Dashboard Implementation

The tourism safety dashboard illustrated an integrated platform for analyzing and forecasting tourism related crime patterns. Forecasted crime trend with different risk category shows the upcoming trend of crime pattern in Indian cities. Predicted city-based safety score derived from XGBoost models, identified regions with lower anticipated safety, assisting in risk prioritization. The clustering output groups cities into three distinct categories, illustrated on an interactive map for better geographical interpretation. The overall ratings and sentiment count within crime and safety in both hotels and public spaces provided users insights about hidden information often not shows by traditional tourism websites and lodging platforms. This dashboard provides users with filter options to choose city, crime category, time category, victim characteristics, and year. Through this interactivity, insights remain context-specific and aligned with the user's area of interest. Throughout different sets, the dashboard consistently indicated that crime trends vary by location and time, with identifiable patterns in age and gender distribution, hour of day, and season. especially for scams and safety concerns in both hotels and public spaces. The addition of predictive modelling gave the dashboard a useful forward-looking aspect. The models enable anticipatory decision-making on the part of tourists, authorities, and service providers. The greatest strength is adaptability

through the ability of users to personalize the presentation of information to fit needs, it can act as a decision support tool capable of being suitable for multiple stakeholders. Finally, the dashboard permits safer, data-driven tourist planning by combining statistical crime data, machine learning predictions, and sentiment-driven insights into a unified, accessible interface statistical crime data.

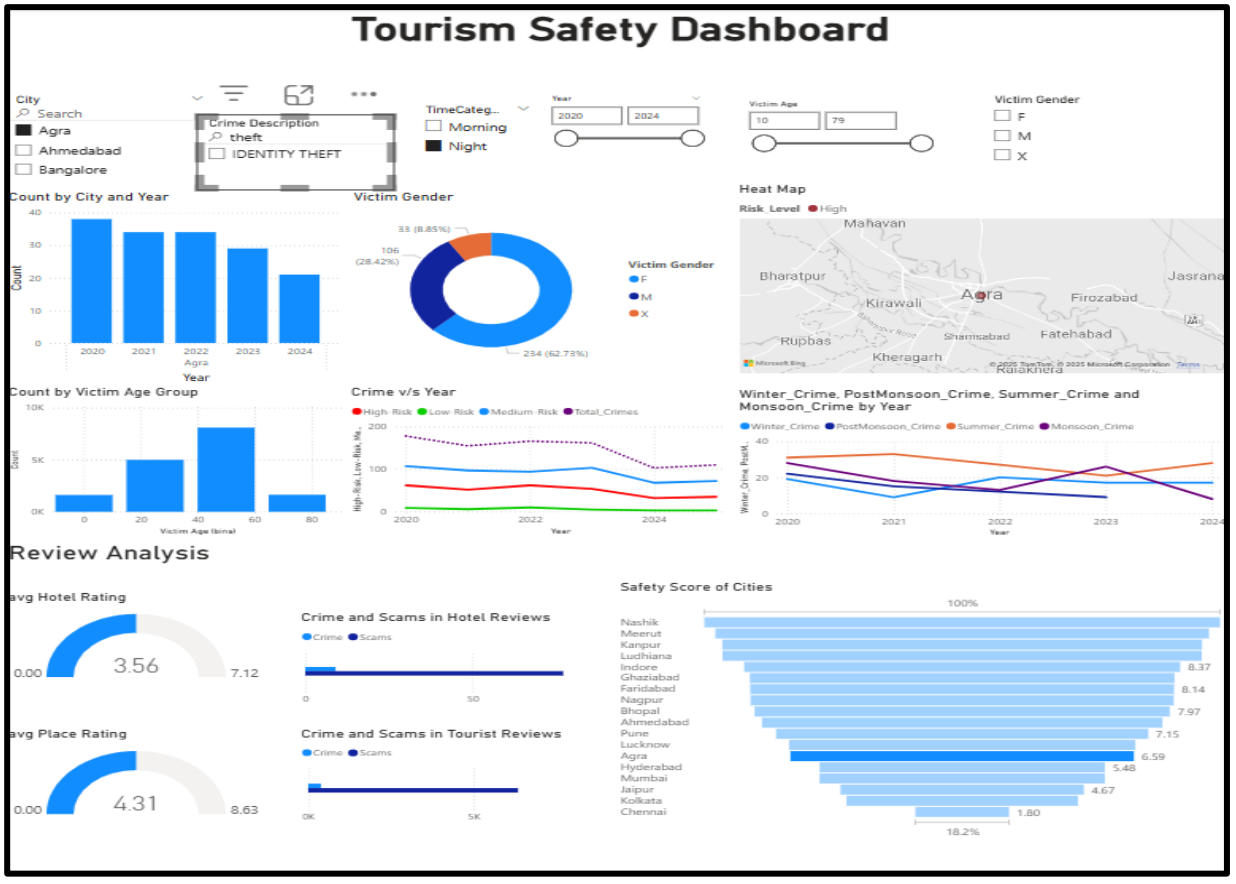

Figure 8. Tourism Safety Dashboard Using Power BI

7. Limitations and Future Enhancements

Despite the robust methodology and comprehensive integration, several limitations exist in the current framework. The crime data are aggregated at the city level, which hide localized crime patterns at the neighborhood level reducing risk detail [26] due to the unavailability of granular data. Secondly, the BERT performance on sentiment classification that is applied only to English reviews restricts generalizability to multilingual situations prevalent in India. The forecasted safety score was derived using weighted equation based on domain intuition and well-informed assumptions which though reasonable but might result in bias. Despite the use of machine learning models to validate and generalize this formula, the approach may benefit from further empirical calibrations or use of ground truth safety labels in future studies. Finally, the dashboard's foundation on sporadic updates of data rather than real-time crime feeds may fail to capture the appearance of new safety threats. Addressing these limitations in future iterations could enhance the model's precision and applicability. This research paradigm can be enriched substantially with the following strategic improvements. The dashboard accuracy and timeliness would be improved by broadening the datasets to include mores sources, such as government tourism boards, international tourism reports and live feeds of police crime. More precise risk evaluation would be obtained by incorporating more geographical data such as data at the district or even neighborhood level would provide better crime hotspots identification. Moreover, incorporating automatic anomaly detection capabilities, predictive alert and user recommendation systems to the dashboard will dramatically enhance decision support and user interactions. To increase the access and impact the dashboard

can be linked to tourism websites such as TripAdvisor and OYO or developed as a mobile application for travel planners. Expanding the NLP platform for multilingual sentiment analysis would extend coverage of diverse tourist opinions. Federated learning can be utilized to develop customized risk evaluation systems to generate customized safety advisories [27] based on user profiles such as traveller group size, gender and age while ensuring data privacy. These enhancements would individually make the system a more responsive, precise, and user-oriented safety system for tourists, and provide policymakers more subtle tools to urban safety planning and hospitality partners with actionable business insights.

8. Conclusion

This research establishes a model to assess crime and tourism connection in Indian cities, providing insights that crime patterns significantly influence tourist experiences and safety perceptions. By Combining crime statistics and tourist sentiment to identify significant spatial and temporal dynamics that influence the appeal of locations. The interactive platform translates raw safety information efficiently into decision-ready information for both travellers and decision-makers. Above all, this research establishes a methodological blueprint for future studies at the intersection of urban security and tourism economics. Clustering techniques enabled effective risk categorization of cities aiding targeted interventions. The flexibility of the framework permits the possibility to extrapolate to other geographical contexts and integrate with new technologies like real-time crime monitoring systems. The research underscores the significance of evidence-based action in tourism security management, while for the hospitality industry, it highlights the growing significance of open safety communication. The results show how machine learning, natural language processing, and visualization tools may be used to analyze tourism safety issues. Higher geographic resolution at the level of the tourist destination, languages, and real-time data sources are possible future enhancements. Lastly, this research advances the field of city tourism resilience, providing both a practical tool for visitors and a conceptual framework for future research into smart tourism and city administration. The model provides a replicable framework for translating publicly available safety data into actionable safer data driven travel planning.

Corresponding author

Adona Kulathinal Josephi
Kulathinal_Joseph-A@ulster.ac.uk

Acknowledgements
NA.

Funding
No funding.

Contributions
A.K.J; M.M; Conceptualization, A.K.J; M.M; Investigation, A.K.J; M.M; Writing (Original Draft), A.K.J; M.M; Writing (Review and Editing) Supervision, A.K.J; M.M; Project Administration.

Ethics declarations
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication
Not applicable.

Competing interests
All authors declare no competing interests.

References

[1] George, R. (2003). Tourists' perceptions of safety and security while visiting Cape Town. *Tourism Management, 24*(5), 575–585. https://doi.org/10.1016/S0261-5177(03)00003-7

[2] Yu, N., & Chen, J. (2022). Design of machine learning algorithm for tourism demand prediction. *Computational and Mathematical Methods in Medicine, 2022*, 1–9. https://doi.org/10.1155/2022/6352381

[3] Shchokin, R., Maziychuk, V., Mykhailyk, O., Kolomiiiets, A., Akifzade, S., & Tymoshenko, Y. (2023). The impact of the crime rate on the hospitality and tourism industry in EU countries. *GeoJournal of Tourism and Geosites, 46*(1), 135–147. https://doi.org/10.30892/gtg.46115-1009

[4] Lisowska, A. (2017). Crime in tourism destinations: Research review. *Turyzm, 27*(1), 31–39. https://doi.org/10.1515/tour-2017-0004

[5] Cohen, E. (2018). Tourism-related crime: Towards a sociology of crime and tourism. *Visions in Leisure and Business, 16*(1). https://scholarworks.bgsu.edu/visions/vol16/iss1/2/

[6] Mawby, R. I., & Vakhitova, Z. I. (2022). Researching the relationship between tourism, crime and security: The tourism industry and the disenfranchised citizens. In *The Handbook of Security* (pp. 581–601). Springer. https://doi.org/10.1007/978-3-030-91735-7_27

[7] Chhabra, J., & Bhattacharjee, M. (2019). Analyzing tourist preferences in India to crime and threat of terrorism. *International Journal of Research in Social Sciences, 9*, 2249–2496.

[8] Hua, N., Li, B., & Zhang, T. (2020). Crime research in hospitality and tourism. *International Journal of Contemporary Hospitality Management, 32*(3), 1299–1323. https://doi.org/10.1108/IJCHM-09-2019-0750

[9] Brunt, P., Mawby, R., & Hambly, Z. (2000). Tourist victimisation and the fear of crime on holiday. *Tourism Management, 21*(4), 417–424. https://doi.org/10.1016/S0261-5177(99)00084-9

[10] Biagi, B., & Detotto, C. (2012). Crime as tourism externality. *Regional Studies, 48*(4), 693–709. https://doi.org/10.1080/00343404.2011.649005

[11] Johnny, L., & Jordan, L. (2007). Tourism and crime in the Caribbean: A case study of St Lucia. *Annals of Leisure Research, 10*(3–4), 475–497. https://doi.org/10.1080/11745398.2007.9686777

[12] Dash, D. P., Dash, A. K., & Parida, Y. (2024). Lethal attraction: Crime as tourism externality—Evidence from Indian states. *Asia Pacific Journal of Tourism Research*, 1–17. https://doi.org/10.1080/10941665.2024.2426196

[13] Tiwari, V., & Omar, A. (2023). The impact of the hotel star rating system on tourists' health safety and risk perceptions. *Journal of Vacation Marketing*. https://doi.org/10.1177/13567667231188880

[14] Alvaro, M., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and review. *SN Computer Science, 4*(4). https://doi.org/10.1007/s42979-023-01828-8

[15] Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. (2024). Research on the application of deep learning-based BERT model in sentiment analysis. *arXiv*. https://arxiv.org/abs/2403.08217

[16] Kumar, R. V. (2021). *Exploratory data analysis using R & RStudio*. https://doi.org/10.13140/RG.2.2.24944.99843

[17] Nelson, L. S. (1998). The Anderson–Darling test for normality. *Journal of Quality Technology, 30*(3), 298–299. https://doi.org/10.1080/00224065.1998.11979858

[18] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

[19] Selvakumar, B., & Lakshmanan, B. (2022). Sentiment analysis on users' reviews using BERT. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2022.03.678

[20] Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Access, 9*, 1–13. https://doi.org/10.1109/ACCESS.2021.3078117

[21] Cui, M. (2020). Introduction to the K-means clustering algorithm based on the elbow method. https://doi.org/10.23977/accaf.2020.010102

[22] Gera, P., & Vohra, R. (n.d.). City crime profiling using cluster analysis.

[23] Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian. *Electronics, 11*(3), 374. https://doi.org/10.3390/electronics11030374

[24] Peters, M., Neumann, M., Zettlemoyer, L., & Yih, W.-T. (2018). Dissecting contextual word embeddings: Architecture and representation. *arXiv*. https://arxiv.org/abs/1808.08949

[25] Athanasopoulos, G., & Weatherburn, D. (2018). *Forecasting male and female inmate numbers: A comparison of ARIMA and ETS modelling results* (Contemporary Issues in Crime and Justice No. 219).

[26] Chainey, S., & Ratcliffe, J. (2005). *GIS and crime mapping*. John Wiley & Sons. https://doi.org/10.1002/9781118685181

[27] Lau, H., Tsang, Y. P., Nakandala, D., & Lee, C. K. M. (2021). Risk quantification in cold chain management: A federated learning-enabled multi-criteria decision-making methodology. *Industrial Management & Data Systems*. https://doi.org/10.1108/IMDS-04-2020-0199

Biographies

Adona Kulathinal Joseph received a master's degree in computer science from Ulster University. Her academic interests include artificial intelligence, data analytics, data visualization, and cybersecurity. She has been actively involved in research projects during her studies and is passionate about exploring the intersection of AI and security. She is passionate about using data-driven approaches to solve practical problems in security and risk analysis. Adona aims to contribute to innovative solutions in technology through continued learning and research. *Email: Kulathinal_Joseph-A@ulster.ac.uk*

Dr. Mahmud Maqsood is a senior faculty member at the School of Computing and a core contributor to the Artificial Intelligence Research Center (AIRC). His interdisciplinary research spans AI risks, cybersecurity, and privacy-preserving technologies, with over 80 publications in top-tier journals and conferences. He has led five funded research projects and holds two US patents in biometric security and NLP encryption. As a certified Web of Science (WoS) Trainer and Senior Fellow of AdvanceHE (UK), he actively mentors researchers and promotes academic excellence. He is also a senior member of IEEE and a frequent speaker at global conferences. *Email: m.mahmud@ulster.ac.uk*