ORIGINAL RESEARCH

Spatio-temporal Crime Analysis and Forecasting on Twitter Data Using Machine Learning Algorithms

Meghashyam Vivek[1] · **Boppuru Rudra Prathap**[1]

Received: 19 October 2022 / Accepted: 30 March 2023 / Published online: 6 May 2023

Abstract
The concept of social media began to gain popularity in the late 1990s and has played a significant role in connecting people across the globe. The constant addition of features to old social media platforms and the creation of new ones have helped amass and retain an extensive user base. Users could now share their views and provide detailed accounts of events from worldwide to reach like-minded people. This led to the popularization of blogging and brought into focus the posts of the commoner. These posts began to be verified and included in mainstream news articles bringing about a revolution in journalism. This research aims to use a social media platform, Twitter, to classify, visualize, and forecast Indian crime tweet data and provide a spatio-temporal view of crime in the country using statistical and machine learning models. The Tweepy Python module's search function and '#crime' query have been used to scrape relevant tweets under geographical constraints, followed by substring-keyword classification using 318 unique crime keywords. The Bokeh and gmaps Python modules create analytical and geospatial visualizations, respectively. Time series forecasting of crime tweet count is performed by comparing the accuracy of Long Short-Term Memory (LSTM), Auto-Regressive Integrated Moving Average (ARIMA), and Seasonal Auto-Regressivee Integrated Moving Average (SARIMA) models to determine the best model.

Keywords Crime analysis · Twitter data · Machine learning algorithms · Forecasting · Crime prevention

Introduction

Crime analysis and forecasting have become an area of focus in criminology. India is a vast country with large regional diversity. Law enforcement agencies would greatly benefit from being able to identify crime hotspots pre-emptively. This would enable these agencies to organize better and distribute their limited resources to take optimal measures to tackle the crime rate. However, the major issue these agencies face is the large-scale inconsistency and fragmentation of crime data. This is where social media comes in.

Social media platforms can serve as a repository of crime data with great geographical specificity as information may be contributed by the residents of a locality. This repository is large and ever-growing, especially in developing nations like India, where great strides are being made to bring internet connectivity to all. Twitter, in particular, has been chosen for its characteristic features. Tweets are largely text-based, allowing easier analysis of crime data through text processing techniques. The 280-character limit on tweets forces users to use keywords that can be identified for data classification. Furthermore, India makes up the world's third-largest Twitter user base, providing an extensive crime-related data set that is publicly accessible. Several regional and national news agencies now have Twitter feeds that further contribute to this data set.

It must be noted that tweet data has its drawbacks like the use of regional languages, incorrect and inconsistent information, and grammatical errors and poor sentence structure. However, the advances made in natural language processing and translation allow us to overcome these drawbacks. Data from official newsfeeds have an advantage over Twitter data in consistency and structure but fail to provide specificity at

This article is part of the topical collection "Cyber Security and Privacy in Communication Networks" guest edited by Rajiv Misra, R K Shyamsunder, Alexiei Dingli, Natalie Denk, Omer Rana, Alexander Pfeiffer, Ashok Patel and Nishtha Kesswani.

✉ Boppuru Rudra Prathap
boppuru.prathap@christuniversity.in

Meghashyam Vivek
meghashyam.vivek@btech.christuniversity.in

[1] Computer Science and Engineering, CHRIST (Deemed to be University), Bangalore, India

a local level. These feeds may be used as a secondary data source to verify the accuracy of Twitter data.

The data scraped from Twitter is first organized and classified using keywords and removing duplicate tweets. The obtained data is then used as a base for visualization techniques like heatmaps and choropleths and forecasting using ARIMA, SARIMA, and LSTM models. These techniques together provide locations with the highest probable occurrence of crime at any given time across the country.

Related Work

In this section, an extensive review of related literature has been made to explicate the nature, importance, trend, and pattern and determinants of spatio-temporal crime. A wide variety of sources have been explored for this purpose, including published articles, journals, theories of crime, available and accessible books, research reports, government reports, Karnataka state police Reports, and various official websites of authorized agencies related to crime. The causes of crime have been the subject matter of much speculation, discussions, research, and debates. There is a large assortment of theories about crime and criminal behavior. Scores of theories related to crime and criminal behavior state crime as a part of human nature. Crime and criminal behavior mainly stem from the psychological, biological, sociological, and economic aspects of human behavior. Various theories explain people's engagement in crime from mental, physical, developmental, economic, social, cultural, and other causes. Theories exploring the causes of crime identified religion, philosophy, politics, economics, and social forces as the main contributors to the ever-increasing crime rate. Several works have been done on crimes, criminals, and different theories of crimes. A few of them are discussed here, which the researcher believes can lend a hand in shaping and explaining the present study.

Cornow et al. [1] proposed a study on the relationship between alcohol retailers and crime in Buffalo, New York, in terms of area and time. The study investigated if a crime was more likely to occur near licensed liquor businesses. Data on licensed alcohol outlets and violent crimes were examined using global and local bivariate space–time k-function techniques from 2005 to 2011. A global bivariate space–time K-function analysis revealed the spatial and temporal distribution of bars and crime. Personal blunders were both collected and dispersed. According to a local survey, outlets selling alcohol and crimes occur simultaneously and in place. Space–time analysis of bars and crime reveals a link. Much time has passed since Louis Thurstone compared different types of crime almost a century ago as stated by Ohyama et al. [2]. Recent research has used methodologies such as the Cambridge Crime Harm Index (CCHI), which assesses the detrimental effects of crime on society and identifies places with high crime harm by assigning a greater weight to the more serious offenses. Although its applicability in Japan, a country with a low rate of violent crime overall and specifically gun violence, is debatable, this index affects urban policy. Catlett et al. [3] provide a prediction approach based on geographical analysis and auto-regressive models. The method predicts urban crime hotspots. This technique will produce a spatio-temporal crime prediction model. The model will use crime hotspots and predictors. Each predictor estimates general crime in its area. The experiment used NYC and Chicago data. According to this review, the system can provide accurate spatial and temporal crime projections over rolling time periods. Hu et al. [4] map and analyze hotspots using a spatial–temporal approach. The proposed paradigm differs in four ways: STKDE incorporates time into predictive hotspot mapping. The best bandwidths are chosen using probability cross-validation. A statistical significance test eliminates false positives in density estimations and the predictive accuracy index (PAI) curve measures predictive hotspot mapping. Anneleen Rummens et al. [5] proposed a predictive analysis for urban research. Invasion, robbery, and battery statistics are collected in 200 m by 200 m grids. The monthly crime rate for 2014 can be estimated using data from the last three years. Monthly forecasts are broken down (day vs. night). The accuracy of a forecast is based on the direct hit rate, the precision, and the prediction index (ratio of direct hit rate versus proportion of total area predicted as high risk). Predictive analysis of crime data at the grid level is used to make functional forecasts. Monthly forecasts that tell the difference between day and night do better than biweekly forecasts, which suggests that the amount of time a forecast covers affects how well it works. Prathap [6] examined 68 crime-related keywords to assist in identifying the type of crime using geographical and temporal information. Keywords are assessed to provide as much information as possible on criminal activities. It is possible to segment criminal activity using news feeds and the Naive Bayes classifier. Mallet extracts terms from many news streams. K-means is utilized to identify crime hotspots. The KDE [23] strategy is utilized while dealing with crime density, and this method has corrected the algorithm's shortcomings. The study uncovered similarities between the ARIMA and crime-predicting models.

Changes in the criminal justice system have been talked about for a long time. Prioritizing crimes based on random models from physics or statistics is a good idea in theory, but it doesn't work well in practice. Data-driven models, especially neural network models, can depict event dynamics. Huge data sets can be mined for information. Spatial–temporal datasets make it hard to learn about crime in a region because of their complexity, intractable correlations, and redundant data. CSAN was made by Qi Wang et al. [7] by putting together variational auto-encoders and context-based sequence generative neural networks. CSAN is better than Conv-LSTM at predicting the number of different types of crimes in a region. Twitter enables the monitoring of spatial and temporal crime statistics on social media. Prathap et al. [8].

Since social media users change rapidly, emotional analysis is an excellent decision-making tool. Twitter is a popular way to obtain news and communicate with others. More than 150 million individuals send 500 million 140-character tweets daily. Twitter is utilized to recommend products depending on the opinions of its users. This paper demonstrates how to examine crime-related tweets from users. The data will demonstrate variations in how people feel about different types of crimes and how they feel about positive or negative crime situations. Crime is the most serious societal problem in emerging countries. Crime has an impact on a country's reputation and way of life. Crime has an economic impact since it necessitates investments in the police force and justice system, increasing the government's financial burden. Law enforcement works to reduce crime. Real-time crime projections can aid in crime reduction. The proposed study by Prathap et al. [9] develops a criminal analytics platform that analyzes newsfeed data to detect crime hotspots. This method assists criminologists in understanding unacknowledged relationships between crime and specific areas. Interactive visualizations aid law enforcement in crime prediction.

Crime analysis using social media data, such as Newsfeeds, Facebook, Twitter, and so on, is a rising area of study for law enforcement agencies worldwide. Data are used to anticipate attacks and arrange reinforcements. Prathap et al. [10] collect and visualize newsfeed data to focus on textual data analytics. By providing crime area coordinates and possible crimes, the research study creates a framework for forecasting 16 types of crime in India and Bangalore. Prathap et al. [11] conducted a research study on criminal activity reports in India and Bangalore. Theft, homicide, alcoholism, assault, etc., are categorized by geographic density and criminal trends, such as time of day, to identify and emphasize national and regional crime. Based on a review of a year's worth of news articles, 68 crime-related terms can be divided into three categories.

Crime clusters are typically reported in locations with a high number of kernels. Time series data can be predicted using the ARIMA model. Using a data mining application, one's different criminal tendencies can be graphically represented. Iqbal H. Sarker describes machine learning methods that can boost an application's intelligence and capabilities [12]. The study investigated the basics of many machine learning algorithms and their relevance to many real-world application areas, including cybersecurity, smart cities, healthcare, e-commerce, and agriculture. The study discusses issues and directions for future research. This project aims to build a technical resource that academic and industrial specialists, practical decision-makers, and others may utilize. Sarker et al. [13] provide a comprehensive overview of "AI-based Modeling" including the principles and capabilities of potential AI techniques that can aid in the development of intelligent and smart systems in real-world application domains, such as business, finance, healthcare, agriculture, smart cities, and more. Our study's research challenges are highlighted. The study contains academics, industry specialists, and decision-makers in real-world scenarios and application areas with a complete introduction to AI-based modeling. Jangada Correia's [14] research recommended various methods for determining public perceptions of cyberterrorism and the need for standardized terminology and framework. The findings of an online poll favor expanding stakeholder diversity to improve terrorism detection and prevention in the UK. Despite general agreement on cyberterrorism, misunderstandings may impair the public's ability to detect and report it. Overall, the literature research and gathering of primary data contribute to developing a cyber terrorism definition consistent with UK legislative definitions and a terrorist activity framework that emphasizes the connections between traditional, cyber-enabled, and cyber-dependent terrorism.

N Kanimozhi et al. [15] presented a study that predicted current crimes using Kaggle open-source crime data. This study assesses the crime that has the greatest impact on a specific place and time period. This study uses machine learning methods such as Naive Bayes to classify criminal patterns more precisely than pre-composed works. Sivanagaleela et al. [16] developed a method based on crime location rather than criminal identification. Initially, the system relied heavily on naive Bayes classification. Data on kidnapping, murder, theft, burglary, cheating, crime against women, and robbery will be clustered using the fuzzy C-Means technique in the current setting. In developing nations like India, crime is rife. Rapid urbanization necessitates constant oversight. Akash Kumar et al. [17] suggested employing KNN to forecast crime rates to avert calamity. It will anticipate a crime's type, date, place, and hour. This data will show local criminal trends, which can help with investigations. It also includes a list of the most serious crimes committed in a specific area. The k-nearest neighbor, machine learning method, is used in the author's work. Wajiha Safat et al. [18] utilized machine learning algorithms, such as logistic regression, support vector machine (SVM), Naive Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), to improve the fit of crime data. Regarding RMSE and MAE, LSTM performed well on both data sets. A data analysis predicts more than 35 categories of crime and an annual decline in crime rates in Chicago and Los Angeles, with February having the lowest crime rate. Chicago's crime rate will continue to grow slowly until decreasing. According to the ARIMA model, the crime rate and the number of offenses in Los Angeles decreased significantly. The results of crime prediction were also seen in the urban cores of both cities. Overall, these results provide Police with a more accurate method than earlier methods for predicting crime, crime hotspots, and future trends. They can be utilized to inform police practice and planning.

Based on the literature, there are very complex spatio-temporal patterns and complicated urban configurations in urban crime. A large number of existing algorithms will not capture all the aspects of the patterns. Therefore, comprehensive techniques are necessary to identify the complex spatio-temporal pattern for analyzing urban crime. Machine learning and Deep learning are good algorithms for capturing spatio-temporal crime patterns and predicting future outcomes.

Motivation and Objective of the Research

India is a fast-developing country with a large population and a young workforce. Urbanization is one of the many by-products of this high-speed development [19, 21]. The degree of urbanization in India has been in a constant ascent from 31.28% in 2011 to 35.39% in 2021. This growth has led to mass migrating of people from rural to urban areas in search of better prospects. Researchers have shown a direct correlation between the concentration of population in urban areas and the increase in crime rate.

To show the consistent increase in crime, we may consider the example of Bangalore city, Karnataka. Bangalore is one of the most prominent metropolitan cities in the country and is considered to be the IT hub of India. As per Karnataka State Police reports [22], cases of thefts recorded were 480 in January 2022 and 725 in August 2022, showing an increase of 151.04% over 7 months. Similarly, a 122.51% increase is seen in the number of Special and Local Laws (SLL) crimes.

The crime rate in India has been consistently increasing since 2018, according to reports by the National Crime Records Bureau (NCRB) [20]. 2020 witnessed a massive surge in the crime rate in response to the imposition of COVID-19 restrictions. This surprised law enforcement agencies, leading to poor mobilization of limited resources.

Growth and prosperity of a nation are largely dependent on the psychological state and well-being of its residents. This, along with the literature review on the use of social media to predict crime, has motivated the use of Twitter data to create a tool that provides a spatio-temporal visualization of criminal activity. Based on this main objective, the following sub-objectives were identified:

1. Development of a dashboard to provide a consolidated view of aggregated data.
2. Identification of crime keywords and categorization of crime tweets.
3. Generation of crime heatmaps, choropleths, and scatter plots.
4. Forecasting of crime tweet count using ARIMA, SARIMA, and LSTM models.

Identifying Problem Based on the Literature

The real-time identification of crime hotspots has become a priority for law enforcement agencies to determine and neutralize increasing threats pre-emptively. However, the vast amount of data collected by these agencies is outdated, inconsistent, and fragmented based on several categories like region, nature of crime, etc. Therefore, a single, consistent source of data collected in real-time that provided an overall view of crime in the country with equal importance to both localized minor events and distributed macroscopic events is required. The use of data from official news sources is promising but lacks the level of coverage provided by social media posts. Therefore, we turn to Twitter, where the required data set is updated with every tweet in real time and is publicly accessible by all. In addition to hotspot prediction, a simple means of visualization is also required. Gerber et al. [23] and Prathap [24] indicating the Kernel Density Estimation to identify crime hotspots in the USA and India, respectively. The literature review shows that social media and police data can predict criminal activity, but none have utilized Twitter data for the same in the Indian context.

List of Crime Keywords Considered

A total of 318 unique crime keywords have been considered to classify crime into 6 categories as shown in Table 1. This model of categorization is modeled after the system used by the Karnataka State Police but is modified for improved understanding of readers without an in-depth understanding of the Indian Penal Code.

Methodology and Implementation

The paper focuses on building a tool that uses Twitter data to identify crime hotspots in India. This section discusses the methodology used to develop the tool, as mentioned earlier. As discussed, the methodology can be broadly divided into classification, visualization, and forecasting components. These components can be further classified into the following seven subsystems:

Table 1 Crime keyword classification

Crime category	Crime keywords
Drug-related crimes	Drug Trafficking, Drug dealing, Dealing drugs, Drug dealer, Alcohol Drinking, Alcohol dealing, Alcohol, Liquor law violation, Liquor, Drug, Narcotics, Heroin, Cocaine, Ganja, Opium, Cannabis, Weed, Overdose, Capsule, Ketamine, Amphetamine
Violent crimes	Gang rape, Rape, Sexual harassment, Sexual assault, Sex offense, Sex abuse, Sexual abuse, Molest, Dishonor, Assault, Fight, Beat, Kick, Punch, Battery, Lash out, Attack, Belt down, Obliterate, Intent to kill, Attempted murder, Murder, Kill, Homicide, Armed robbery, Robbery, Terrorism, Kidnapping, Abduction, Tied, Incapacitate, Ransom, Shot, Gunshot, Shootout, Stab, Harassment, Abuse, Assault, Outrage, Snatch, Put to death, Lynch, Hit and run, Gambling, Hang, Run over, Dacoity, Sex racket, Victim, Goon, Rowdy, Knife, Domestic violence, Dead, Death, Violence, Sexual misconduct, Body, Firing, Fired, Sexual, Hostage, Sex, Injure, Injury, Finger, Slit, Intimidate, Kidnap, Acid, Explosive
Commercial crimes	Official Document Forgery, Currency Forgery, Official Seal Forgery, Official Stamp Forgery, Forgery, Bribery, Bribe, Counterfeit, Cheat, Conned, Impersonator, Impersonate, Fake, Deceptive, Breach of trust, Breach of contract, Breach, Embezzlement, Misappropriate, Fraud, Corrupt, Leak, Abscond, Misconduct, Conspire, Conspiracy, Tax, Excise, Collude, Collusion, Contract, Extort, Lakh, Crore, Gold, Gullible, Chit fund, Chit, Fund, Money, Association, KYC, Bank, Business, Deal, Raid, Copyright, Import, Export
Property crimes	Arson, Motor vehicle theft, Theft, Burglary, Steal, Riot, Violent protest, Protest, Larceny, Barrage, Barrage fire, Fire, Bombardment, Bomb, Explosion, Explode, Shelling, Looting, Trespass, Incendiarism, Shoplifting, Vandalism, Vandalize, Vandals, Encroach, Property, Stole, Smuggle, Thief, Burn, Land dispute, Land, Dispute, Rob, Enter, House, Robbed, Crook, Loot, Robbers, Shop, Burglar, Grand theft auto
Traffic offenses	Speeding, Signal Jump, Jump Signal, Running a Red Light, Reckless, Collision, Collide, High speed, Speed, Fast, Drunk driving, Drink and drive, Drinking and driving, Driving under influence, DUI, Helmet, Accident
Other offenses	Employing Illegal Worker, Illegal worker, Prostitution, Illegal Gambling, Gambling, Begging, Adultery, Homosexuality, Weapons violation, Violation, Weapon, Porn, Video, Child, Public peace violation, Peace, Stalking, Hurt, Dowry, Modesty, Negligence, Suicide, Criminal damage, Harlotry, Whoredom, Espionage, Spy, Pickpocketing, Pilfering, Poaching, Damage, Illegal, Juvenile, Inappropriate, Marriage, Custody, Affair, Student, Hacking, Convict, Prison, Love, Threat, Blackmail, Witness, Accuse, Conflict, Gang, Arrest, Miscreant, Cop, Police, Incident, Criminal, Cyber Crime, Scream, Commit, Seize, Thugs, Uproar, Trap, Hindu, Muslim, BJP, Congress, Romantic, Marry, Politic, Follow, Hunting, Hunt, Tiger, Fur, Skin, Husband, Wife, Follow, Girlfriend, Cyber, Crime, Suspect, Abase, Warrant, Social media, YouTube, Facebook, Instagram, Twitter, False news, Law, Minor, Girl, Group, Chain snatch, Chain, Tribal, Caste, Adulterate, Hijack, Hate, Schedule, Tribe, Atrocities, Atrocity, Abet, Conceal, Habit, Repeat, Major minerals, Minor minerals, Mineral, Mining, Mine, Vehicle, Car, Bike

1. Data Aggregation
2. Data Cleaning
3. Location Identification
4. Classification
5. Visualization
6. Forecasting
7. Dashboard

Figure 1 shows the proposed framework.

Data Aggregation

Data aggregation refers to scraping appropriate tweets from Twitter using the search function of the Tweepy module. The Tweepy module is used to interact with the Twitter API. The hashtag '#crime' is used to query crime-related tweets

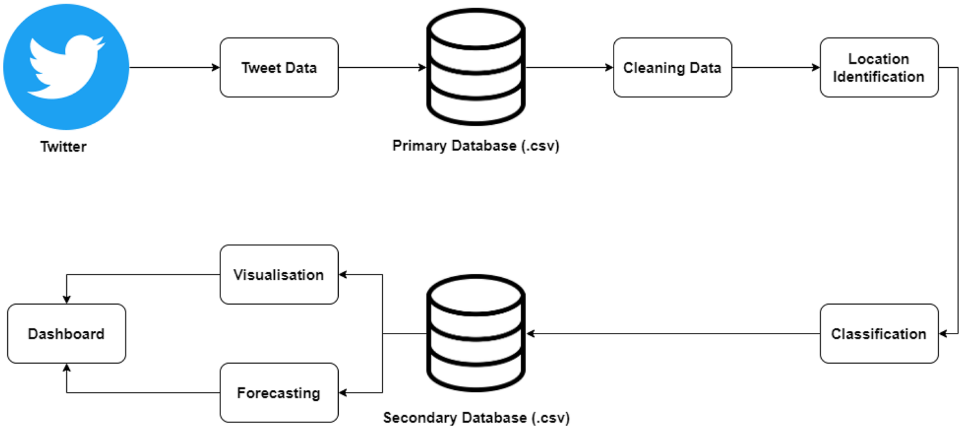


Fig. 1 Proposed framework

and coordinates isolating tweets from and around the Indian subcontinent. Many retrieved tweets are written in various regional languages and translated into English for better analysis. The tweets are then stored in a.csv file.

Data Cleaning

Data cleaning helps improve the integrity of the data set by removing duplicate tweets. Each tweet is assigned a hashcode, and duplicate hashcodes are removed. Tweet hashtags are extracted for each tweet. Tweets with only '#crime' are said to have 'No Hashtags'. The API provides the timestamp of tweet creation in UTC, which is converted to the corresponding time in IST. A URL pointing to each tweet on the Twitter platform is generated and stored to confirm the authenticity of the data set.

Location Identification

Location identification involves the identification of the geographical location of crime by searching tweet attributes in a prioritized manner. First, the name of a city is searched for in the tweet hashtags, followed by the tweet text and eventually the tweet geolocation. A city once identified is coupled with the corresponding state and stored. The second round of searches is carried out on state names in the same manner. A state once identified is coupled with its capital city and stored. Finally, the word 'India' is searched for in the same fashion. Identification of the word maps the city to Delhi. All tweets still left without a geographical location are considered to belong to neighboring countries and removed. The latitudes and the longitudes of identified cities are stored.

Classification

Classification refers to the categorization of tweets based on the type of crime. Crime keyword is identified using a substring-keyword search on the tweet hashtags followed by the tweet text. The keyword is then referenced against the crime keyword database to determine the crime type. Data from the above subsystems are stored in the secondary database.

Visualization

Visualization provides a comprehensible diagrammatic representation of the data stored in the secondary database. The applied visualization techniques can be classified into analytical and geospatial. Geospatial visualization includes a heatmap for crime density representation, a choropleth for state-wise crime distribution and a scatter plot to pinpoint crime location. Analytical visualization includes bar graphs, pie charts, clustered bar graphs, and line graphs to represent quantifiable measures of collected data.

Forecasting

Forecasting involves a comparative analysis of ARIMA, SARIMA, and LSTM models to determine the most accurate forecasting model for crime tweet count.

ARIMA

ARMA models were developed for stationary time series. However, a new class of models was introduced by integrating a phase for the removal of non-stationarity in a time series. These models are known as ARIMA models and developed by Box and Jenkins which is a stochastic process. ARIMA is described as a three-stage iterative model which includes time series identification, estimation, and verification. The Box–Jenkins approach mainly uses the integration filter, AR filter, and MA filter. The integration filter produces a filtered (differenced) series from observed data. The AR filter generates an intermediate series which is further processed by the MA filter which results in random white noise.

The formula for an ARIMA(p,d,q) model is given by:

$$\begin{aligned}\mathrm{ARIMA(p,d,q)} &= (n) = \mu + \theta 1(n-1) + \cdots + \theta p(n-p) \\ &\quad + \theta(n) + \theta 1(n-1) + \cdots + \theta q(n-q)\end{aligned} \tag{1}$$

In Eq. (1), μ is the mean of the series, $\theta 1$, …, θp are the autoregression coefficients, $\theta 1$, …, θq are the moving average coefficients. $f(n)$ is the forecasted value at time n n is the current time step $n-1$, $n-2$, …, $n-p$ are previous time steps.

This equation can be broken down into three components: Autoregression (AR) term: the autoregression term represents the linear dependence between an observation and a number of lagged observations. It is represented by the sum of $\theta 1(n-1)+\cdots+\theta p(n-p)$. Differences (I) term: the differences term represents the amount of differencing applied to make the time series stationary. It is represented by d in the ARIMA(p,d,q) formula.

Moving average (MA) term: The moving average term represents the error or residual at a certain point modeled as a linear function of the previous errors or residuals. It is represented by the sum of $\theta 1(n-1)+\cdots+\theta q(n-q)$. In summary, the ARIMA model predicts the next value in a time series as a weighted sum of past observations and past forecast errors, with weights determined by the values of p, d, and q.

SARIMA

SARIMA (Seasonal Autoregressive Integrated Moving Average) is a statistical model that is used to analyze and forecast time series data with a seasonal component. It is an extension of the ARIMA model, which takes into account the presence of seasonality in the data.

The formula for a SARIMA(p, d, q) (P, D, Q)m model is given by: SARIMA(p, d, q) (P, D, Q)m

$$\begin{aligned}(n) = {} & \mu + \theta 1(n-1) + \cdots + \theta p(n-p) \\ & + \theta(n) + \theta 1(n-\mathrm{m}) + \cdots + \theta q(n-q\mathrm{m}) \\ & + \gamma 1(n-1-\mathrm{mD}) + \cdots + \gamma P(n-P\mathrm{m}-\mathrm{mD})\end{aligned} \tag{2}$$

In Eq. (2), μ is the mean of the series, $\theta 1$, …, θp are the autoregression coefficients, $\theta 1$, …, θq are the moving average coefficients $\gamma 1$, …, γP are the seasonal autoregression coefficients, $f(n)$ is the forecasted value at time n, n is the current time step, $n-1, n-2, \ldots, n-p$ are previous non-seasonal time steps. $n-m, n-m-1, \ldots, n-m-$q are previous seasonal time steps, m is the number of seasonal periods per cycle

D is the order of seasonal differencing and d is the order of non-seasonal differencing.

In summary, the SARIMA model predicts the next value in a time series as a weighted sum of past observations, past forecast errors, and past seasonal errors, with weights determined by the values of p, d, q, P, D, Q, and m. It has a form similar to the ARIMA model, but with added terms of seasonal autoregression ($\gamma 1$, …, γP) and seasonal differencing (D).

LSTM

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) architecture that is capable of learning long-term dependencies in data. An RNN is a neural network architecture that is designed to process sequential data, such as text, speech, or time series data. The LSTM architecture was designed to overcome the problem of vanishing gradients that occur in traditional RNNs when the input sequence is very long.

In an LSTM network, each neuron, or "memory cell," has three gates: an input gate, a forget gate, and an output gate. These gates are used to control the flow of information into and out of the cell, and to allow the LSTM network to learn when to remember or forget information. The gates are controlled by weights that are learned during training, and the values of these weights determine how much information is allowed to flow through the gates at each time step.

Implementation of the Process

The various subsystems mentioned in the methodology section are implemented using Python programs. The Tweepy Python module is used to communicate with the Twitter API.

Data Aggregation

To extract tweets from Twitter, a valid API key with appropriate access level authorization and the API key secret is required. API key authentication request is sent using the 0Auth2AppHandler function in Tweepy. Once Twitter successfully authenticates the API key, the data aggregation process may begin.

Scraping is performed using the search function of the Tweepy Python module. The search query '#crime' along with geocoding '20.5937 (latitude), 78.9629 (longitude), 3000 km (the radius for tweet extraction)' are passed as parameters. This geocode encompasses all of India and small parts of its neighboring countries. The search_tweets query returns pages of data traversed using the tweepy cursor to extract tweet attributes. Tweet text is translated to English using the googletrans python module. Any URLs contained in the tweet text are eliminated before translation using simple substring elimination with 'https' as a key. The aggregated data is written onto a pandas data frame and appended to the primary CSV file.

Data Cleaning

Hashcodes are generated using the standard Python hashing library. Duplicate hashcodes are removed using the removeduplicates() function of the pandas module. Time zone conversion is performed using the dateutil Python package.

Location Identification

The presence of geographic keywords referenced from the locations database is found in tweet attributes using the find() function. find() returns a value of − 1 if the keyword is not found.

Classification

Crime keywords referenced from the keyword database are searched for in the tweet attributes using the find() function. find() returns a value of − 1 if a keyword is not found.

Visualization

Geospatial heatmaps and scatterplots are generated using the gmaps Python package while the interactive choropleth is generated using the bokeh module and a shape (.SHP) file of India's administrative divisions. Analytical visualizations are generated using the matplotlib Python package.

Forecasting

The SARIMA model has been implemented using the SARIMAX function from statsmodels.tsa.statespace package in

Python. Autocorrelation plot (ACF) and Partial Autocorrelation plot (PACF) have been implemented using the acf and pacf functions from statsmodels.graphics.tsaplots package in Python. The ARIMA model has been implemented using the ARIMA function from the statsmodels.tsa.arima.model. ARIMA parameters are found using the auto_arima function from the pmdarima package. To train the LSTM model, the crime counts are scaled to a value between 0 and 1 using the MinMaxScaler function from the sklearn. preprocessing package. The time series is then formatted for the training process using the TimeSeriesGenerator function of the keras.preprocessing.sequence module. The LSTM layer is added to the model with 100 neurons and RELU activation. The dense layer is added to the model and the model is compiled using the Adam optimizer and Mean Squared Error as a loss. The LSTM model can be found in keras.layers as LSTM. The model is trained for 50 epochs while saving only the best model using ModelCheckpoints.

Dashboard

Built using HTML, CSS, and Javascript to display interactive plots and other analytics in an organized form.

Figure 2 shows a detailed framework of the program implementation.

Results and Discussion

State-wise Crime Distribution Using Choropleth

Figure 3 displays the absolute crime count determined by crime tweets and does not consider population. Therefore, populous states like Uttar Pradesh tend to have a higher crime count. The five states with the highest crime count in decreasing order are as follows: Delhi (2377), Maharashtra (2275), Uttar Pradesh (1670), Tamil Nadu (1252), and Gujarat (1121). A 2022 report shows Delhi has the country's highest crime rate, with a crime index of 59.58. The figures generated from tweet data show a mismatch with NCRB reports. This is because not all crimes are tweeted about. The accuracy of tweet data concerning real-world data is expected to improve over time with the addition of new users. However, the relative state-wise accuracy of data can be used to determine the state-wise distribution of Twitter users. For example, in Delhi, the crime tweet data is similar to real-world crime data, it can be assumed to have a higher percentage share of Twitter users. In contrast to this, states like Arunachal Pradesh where there is a large disparity between crime tweet data and real-world crime data can be assumed to have few Twitter users. States along the border of the country, like Ladakh, Sikkim, and Arunachal Pradesh seem to have fewer Twitter users and crime-related tweets. This may be due to the following factors—geopolitical tension, lack of internet infrastructure, or low population density.

Crime Density Detection Using Heatmaps

Seven heatmaps are generated in total, 6 of which display the geospatial density of each crime category and one displays the total geospatial crime density, including all types of crime. Figure 4 shows the different heatmaps generated.

Figure 4 shows that the density distribution for each crime category is different as various factors affect their distribution. Criminal activity appears denser in metropolitan cities. It can be inferred from the above plots that violent crimes make up the majority of the total crime count due to the similarities in density distribution. Violent crimes appear more concentrated in North India, with a decreasing density as we move South. Delhi, Uttar Pradesh, and Bihar belts appear to have the country's highest concentration of criminal activity. Delhi appears to have a high concentration of criminal activity under all categories. Drug-related crimes are sparsely distributed, with Goa, Delhi, and Mumbai being hotspots. Commercial crimes are concentrated around metropolitan and port cities, which are seats of commerce. Traffic offenses are sparsely distributed with concentrations along the western and eastern ghats. Property crimes are concentrated in cities like Mumbai, Delhi-NCR, Bangalore, Chennai, Hyderabad, Pune, and Kolkata which are home to some of the most expensive localities to buy real estate in the entire country. Other offenses are sparsely scattered and concentrated around metropolitan cities without any distinct patterns.

Pinpointing Crime Locations Using Scatter Plots

Heatmaps provide a great visualization of the crime density, but a scatter plot is required to study each event individually. Figure 5 represents the scatter plot for the crime-related tweets collected thus far.

Parts of Telangana and Odisha appear not to have any pointers over them. This can be verified with the corresponding heatmap showing zero crime density. A clustered scatter plot produces a heatmap. The view of the map at the default zoom level is not very comprehensive. Zooming into a particular area of interest and clicking the marker provides more comprehensive results as shown in Fig. 6,

Similar to heatmaps, scatter plots may also be filtered by crime type. Figure 7 shows a scatter plot of only drug-related crimes.

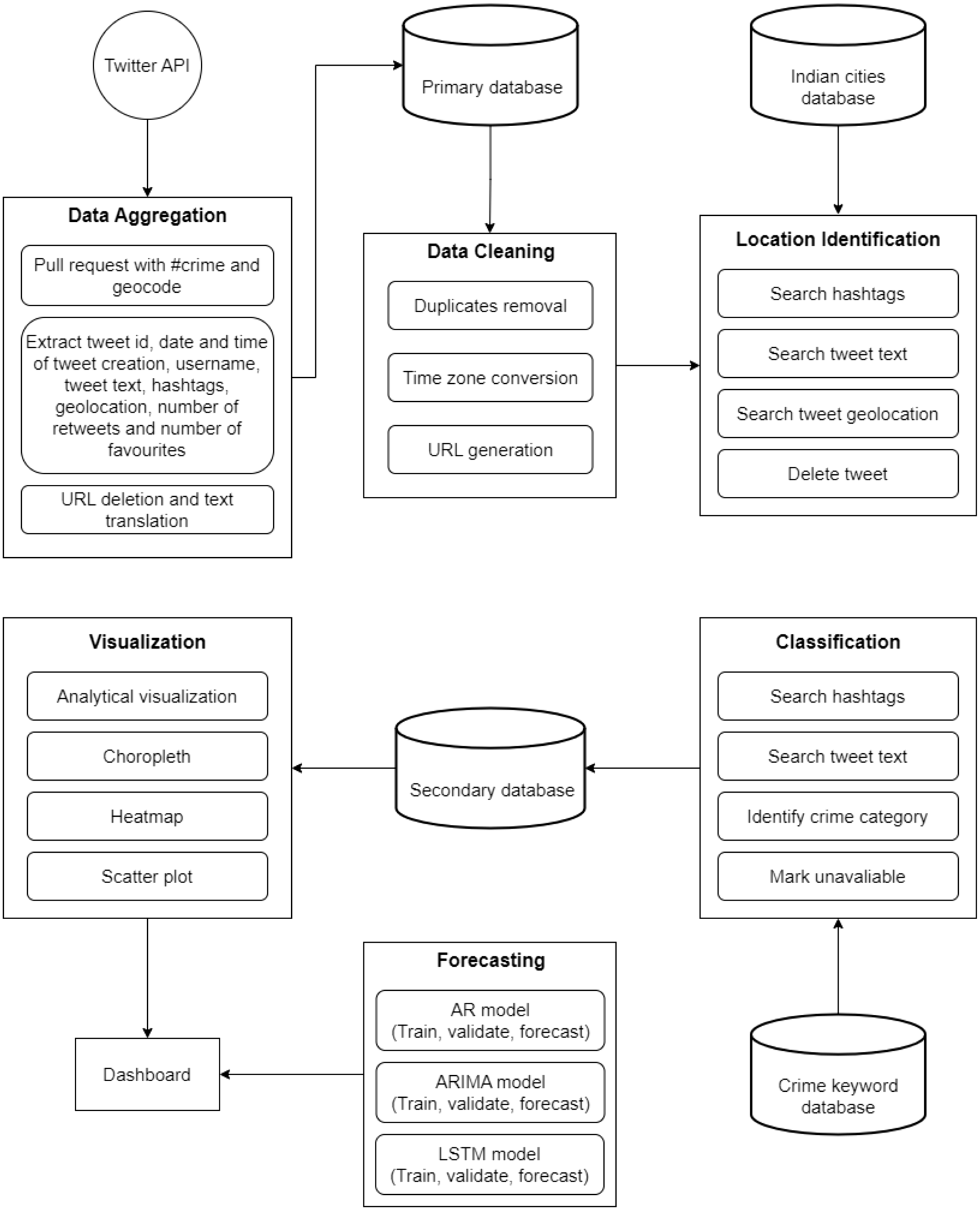


Fig. 2 Detailed framework

Analytics

Percentage Share of Crime Categories

Figure 8 shows each crime category as a percentage of the total crime based on tweet data.

As of 2022, India has recorded a crime index of 44.57 against an average count of 44.74, according to World Crime Index Reports. Here, Violent crimes are categorized as actions that intentionally cause bodily harm. The other offenses category has a higher share as it includes any crime that does not fit under the other five categories. Twitter data

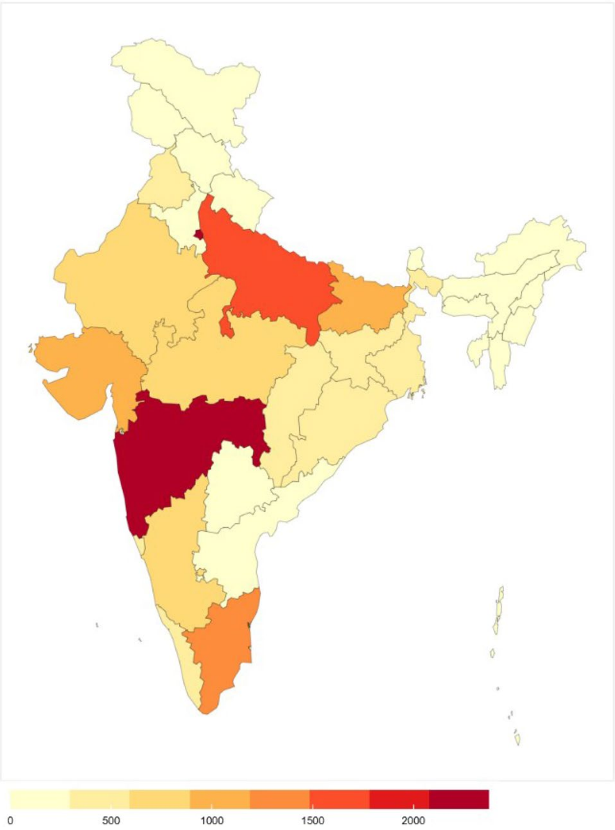


Fig. 3 State-wise crime distribution using choropleth

might be an efficient way of tracking violent crimes. The percentage of traffic offenses is much higher, according to NCB reports. This implies that this category of tweets gets less traction on social media and is not covered by major news channels.

Crime Keyword Analytics

Analysis run on the keywords can help identify the level of awareness the public has toward the criminal activity around them. Figure 9 displays the top ten crime keywords with the highest hit rate.

6 of the top 10 keywords with the highest hits are from the violent crime category. This shows that tweet crime data has a high percentage of violent crimes about real-life crime reports. Crime, Police, and Arrest are general keywords used when no other event-specific keyword is found and classified as other offenses. However, the presence of police as the keyword with the 5th highest number of hits indicates that the public is quite extensively using Twitter to bring the attention of law enforcement agencies to specific issues. Law enforcement agencies could consider tweets as a factor while determining resource deployment. Alternately, this also indicates the possibility of using crime tweet data to analyze the effectiveness of law enforcement agencies. It can also be said that these words are of common knowledge to most people and have specific translations from various regional languages.

Crime Keywords with the Highest Social Impact

It is possible to determine the crime keywords that Twitter users focus on by integrating the number of retweets associated with that particular tweet. A retweet is a feature of Twitter that enables one user to repost another user's tweet, thereby sharing it with his followers and increasing the number of user interactions with the original post. Figure 10 shows the average number of retweets for a tweet with a particular keyword.

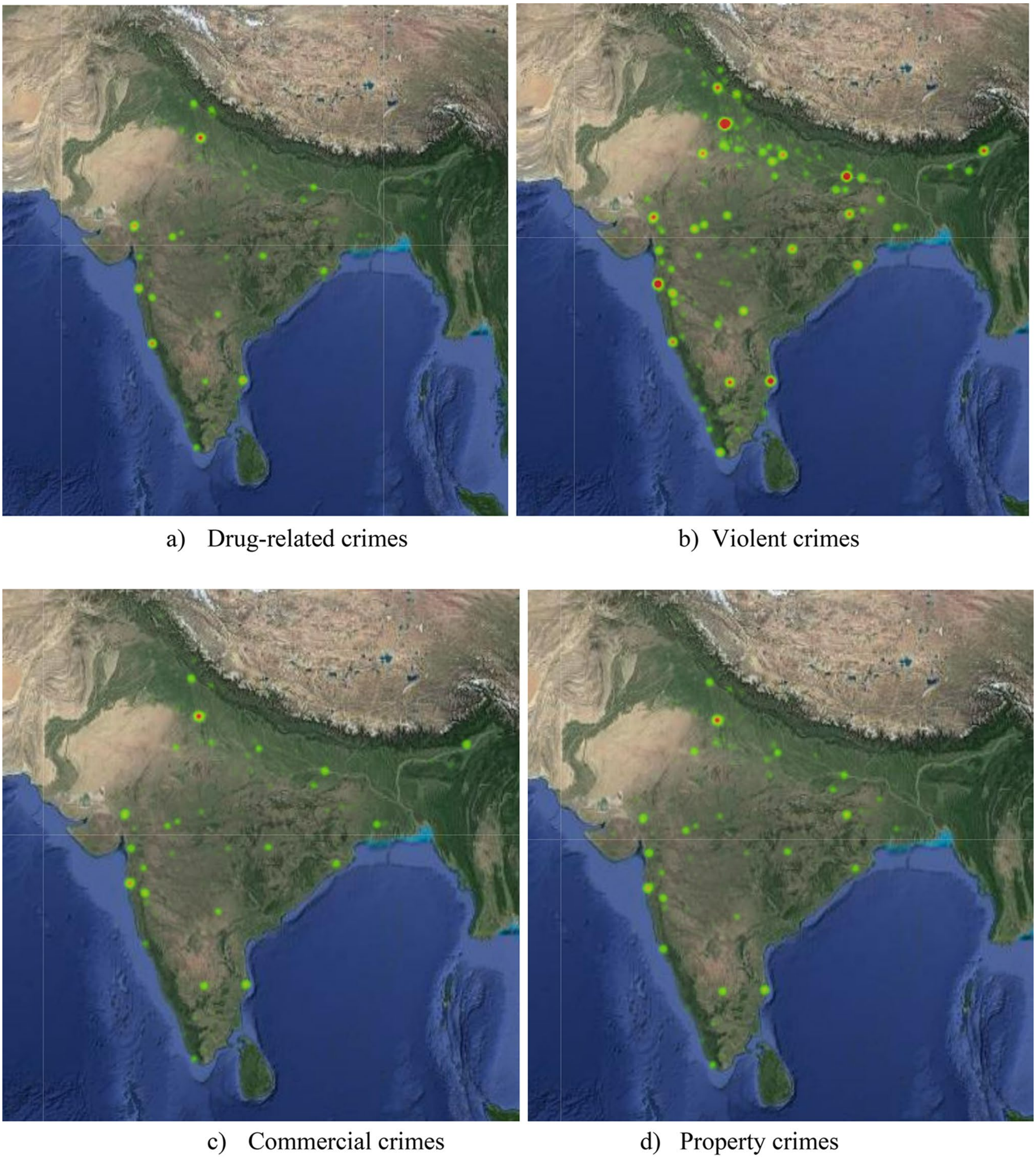
The keyword Trap tops the list with 56.75 retweets on average. This could be to raise awareness against conmen and fraudulent schemes. Burn may be used to indicate both arson as well as the use of fire to inflict bodily harm. Both these cases are serious offenses and retweeting helps raise awareness against them. The keyword Girl consistently repeats in most tweets conveying sexual crimes against minors and deeply impacts society. This data could be tracked to identify repeat offenders and help people identify and avoid areas with poor safety for children. Twitter being a platform for the heated exchange of political ideologies leads to the appearance of the keyword Congress. Keywords that appear on this as well as the previous chart indicate crimes that are most talked about on Twitter and also have a deep impact on society. In this case, the keyword is found to be Shot.

Crime in Metropolitan Cities

Rapid urbanization has made cities hotspots of crime. The clustered bar graph in Fig. 11 represents the counts of each of the 6 crime categories in 6 major metropolitan cities. The crime rate is the highest in Delhi. This follows a 2022 report ranking Delhi as the city with the highest crime rate, a population of 18,980,000 and a crime index of 59.58. Kolkata has the second lowest overall crime count with no drug crimes reported on Twitter. This follows the 2022 NCB report stating Kolkata is the safest Indian city to live in. It must be noted that the count on this chart is also proportional to the Twitter user base in each city.

States with the Highest Crime Rates

The state-wise distribution of crime as seen in Fig. 3 is a representation of the absolute crime count. The crime rate for each state can be calculated by taking into account the population using the formula,


a) Drug-related crimes

b) Violent crimes

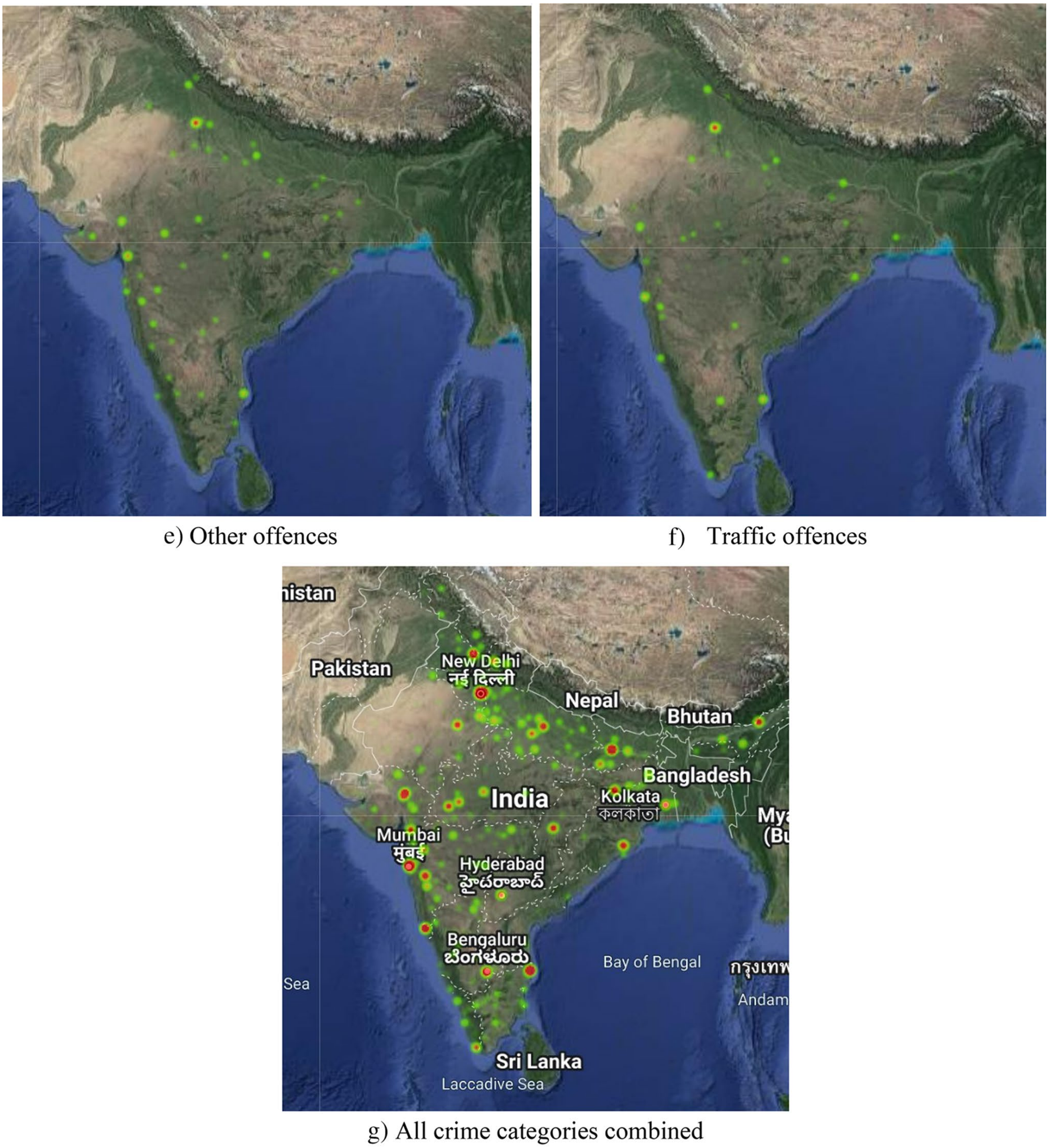
c) Commercial crimes

d) Property crimes

Fig. 4 Crime density detection using heatmaps. **a** Drug-related crimes, **b** Violent crimes, **c** Commercial crimes, **d** Property crimes, **e** other offenses, **f** Traffic offenses, **g** All crime categories combined

$$\text{Crime rate} = (\text{Total number of crimes}/\text{Total population of the state}) \times 100,000$$

Table 2 shows 10 states with the highest crime rates. These rates are updated in real time and change with the addition of new data.



e) Other offences

f) Traffic offences

g) All crime categories combined

Fig. 4 (continued)

These crime rates are calculated on the basis of crime tweet data for a period of 2 months ranging from 18/08/2022 to 16/10/2022. Delhi, Goa, Gujarat, Tamil Nadu, Maharashtra, and Chandigarh are part of the top 10 states with the highest crime rates as per NCRB data. This shows a 60% match with crime data obtained from Twitter. The remaining states on this list may have a relatively larger number of Twitter users leading to a discrepancy with NCRB data.

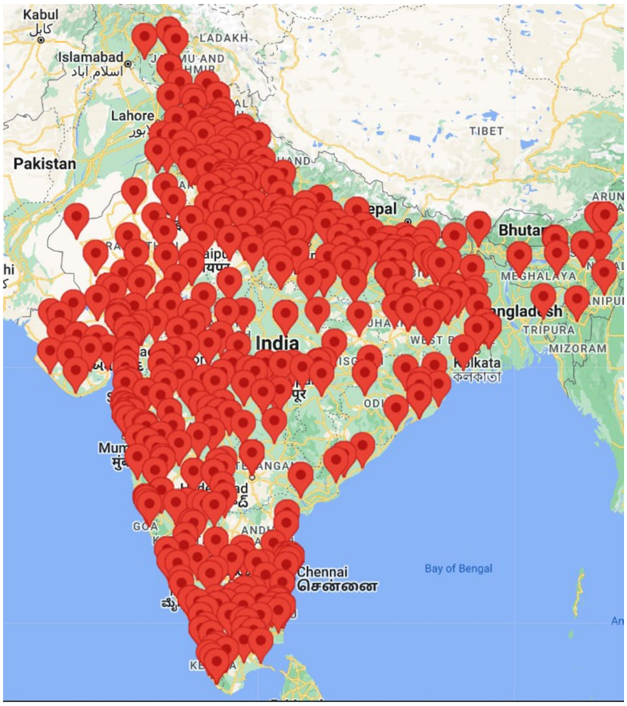


Fig. 5 Scatter plot with standard zoom level

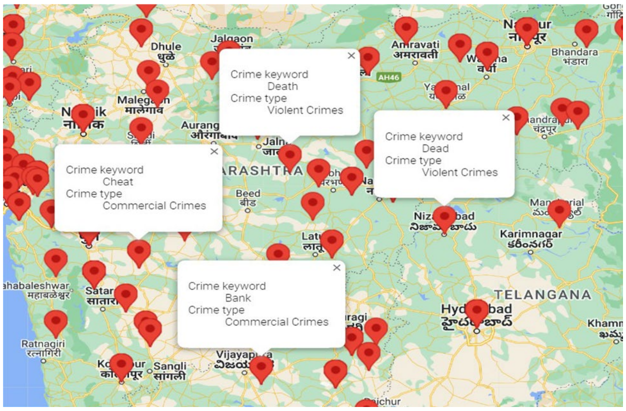


Fig. 6 Pinpointing crimes through scatter plots

Time Series Analysis

The real-time data used in this paper has been collected consistently from the 18th of August, 2022 to the 16th of October, 2022 for approximately 2 months. This section contains a basic analysis of the daily and hourly crime count plots. Note that these counts represent the number of crime tweets recorded and not the actual number of crimes but are treated as analogous in subsequent parts of this section.

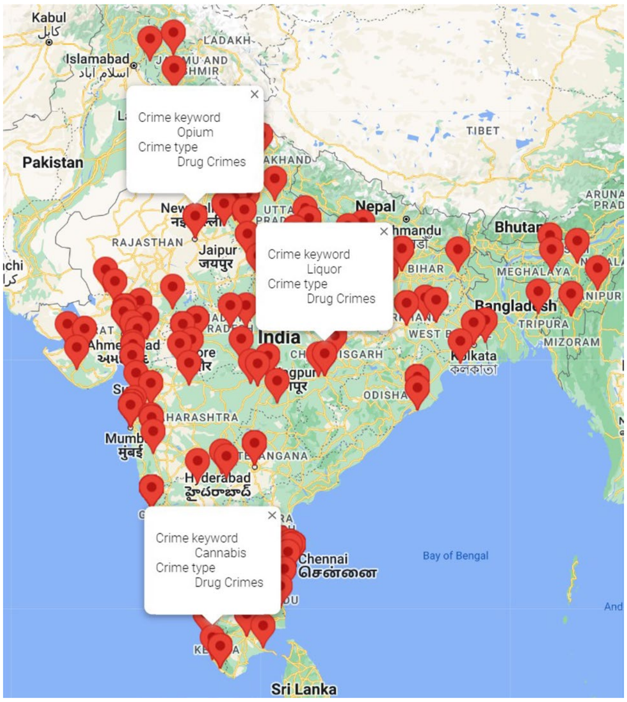


Fig. 7 Scatter plot of Drug-related crimes

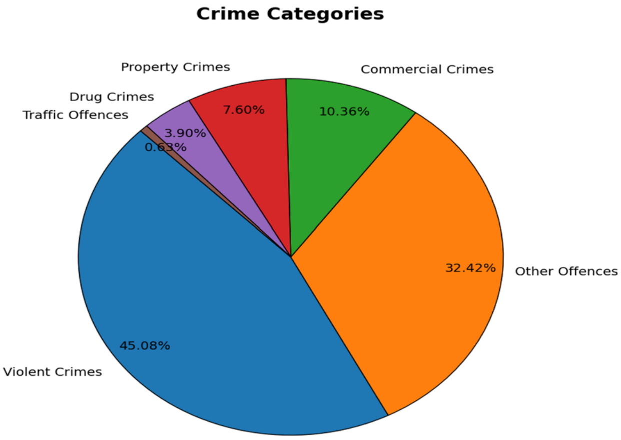


Fig. 8 Percentage shares of crime categories

Daily Plot of Crime Tweet Count

Figure 12 represents the crime count as a daily time series. It can be seen that some days have higher crime counts than others. There is a sharp fall in crime rates on weekends especially Sundays (for example 21/08/2022, 23/08/2022). The number of crimes seems to increase during festivals (for example 18/08/2022—Krishna Janmashtami, 31/08/2022—Ganesh Chaturthi, 08/09/2022—Onam).

There also appears to be a weak correlation between rainfall and crime count. This can be better analyzed with the collection of more data. In this case, the daily time series contains only 60 entries which are insufficient to create accurate forecasts using statistical models Therefore, an hourly time series is developed.

Hourly Plot of Crime Tweet Count

Figure 13 provides an hourly representation of the time series. It is seen that the least number of crimes is recorded during the early morning from 2:00 to 5:00 and the highest number of crimes is recorded during the late afternoon from 14:00 to 17:00. This time series representation is suitable for forecasting. The ARIMA model works best on

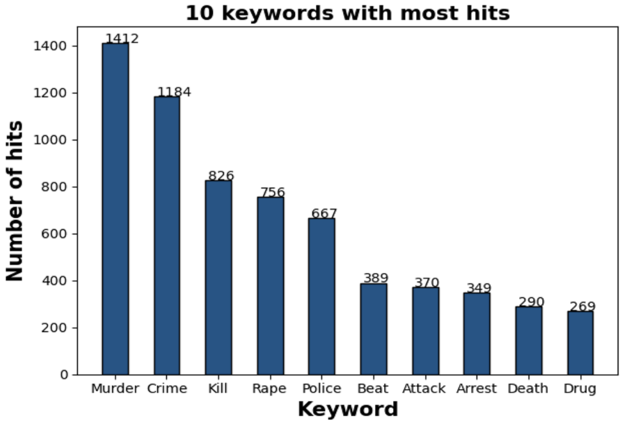
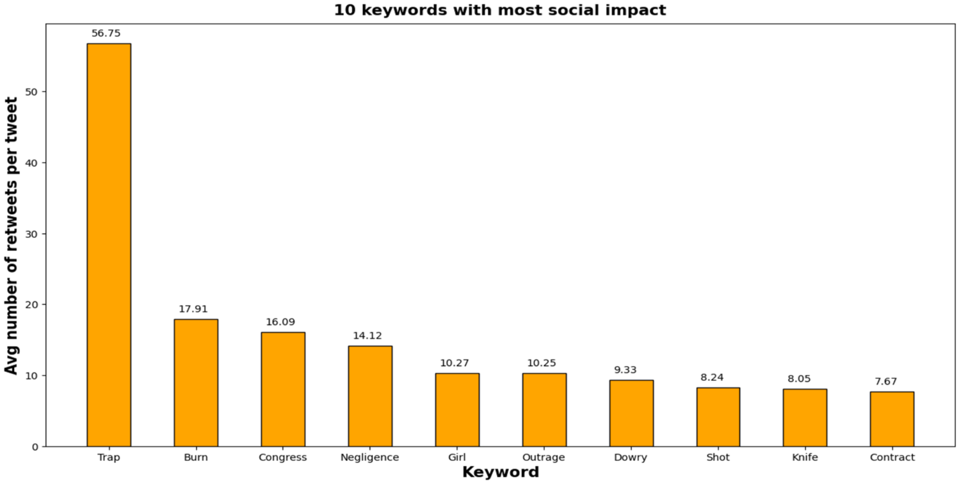


Fig. 9 Most used crime keywords

Fig. 10 Social impact of crime keywords



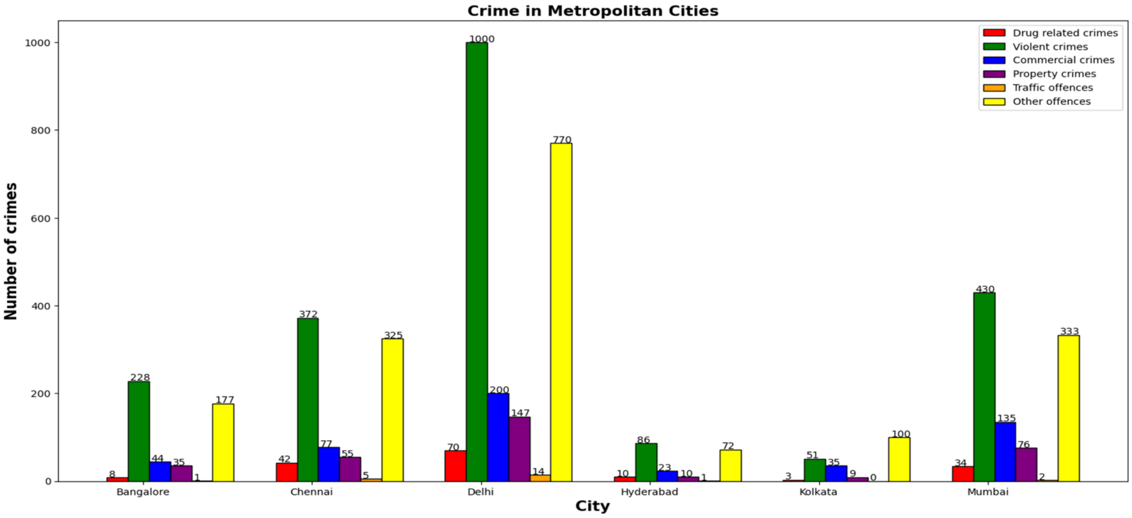


Fig. 11 Crime in metropolitan cities

Table 2 10 States with the highest crime rates

State	Population	Crime count	Crime rate
Goa	15,21,992	377	24.77016962
Delhi	1,93,01,096	2377	12.31536282
Chandigarh	11,58,040	92	7.944457877
Maharashtra	12,49,04,071	2275	1.821397799
Puducherry	16,46,050	28	1.701041888
Uttarakhand	1,17,00,099	194	1.658105628
Gujarat	7,04,00,153	1121	1.592326085
Tamil Nadu	8,36,97,770	1252	1.495858253
Kerala	3,46,98,876	495	1.426559177
Meghalaya	37,72,103	42	1.113437252

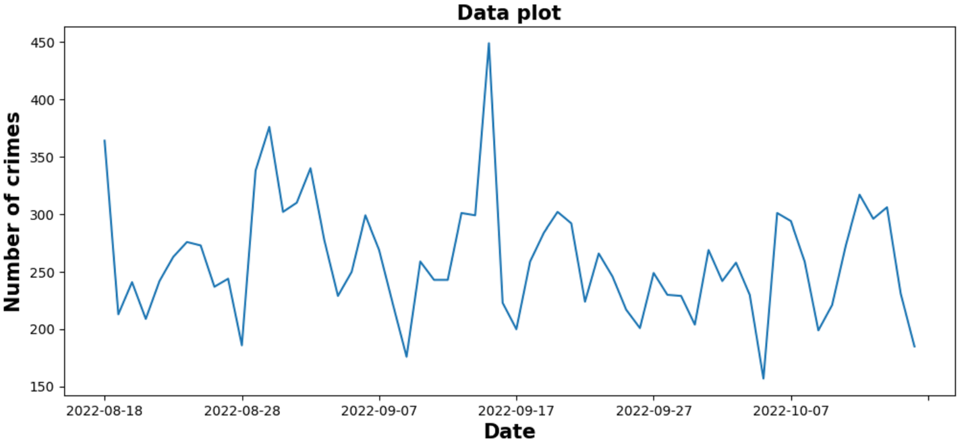
stationary data. A data set that does not have a clear trend can be said to be stationary. This can be identified using the Augmented Dicky–Fuller Test. The test is implemented using the adfuller function from the statsmodels.tsa.stattools package in Python. Figure 14 shows the results obtained from the Augmented Dicky–Fuller Test.

The value of significance to us is the P-Value. The lower the P-Value, the more stationary is the data. A P-Value of under 0.05 can be considered to be stationary. Therefore, the crime count data set is stationary.

Time Series Forecasting

The process of analyzing past and present data to quantifiably predict future data is considered to be time series forecasting. Figure 15 shows the Autocorrelation and Partial Autocorrelation plots used to identify the order of the SARIMA and ARIMA models.

Fig. 12 Daily time series



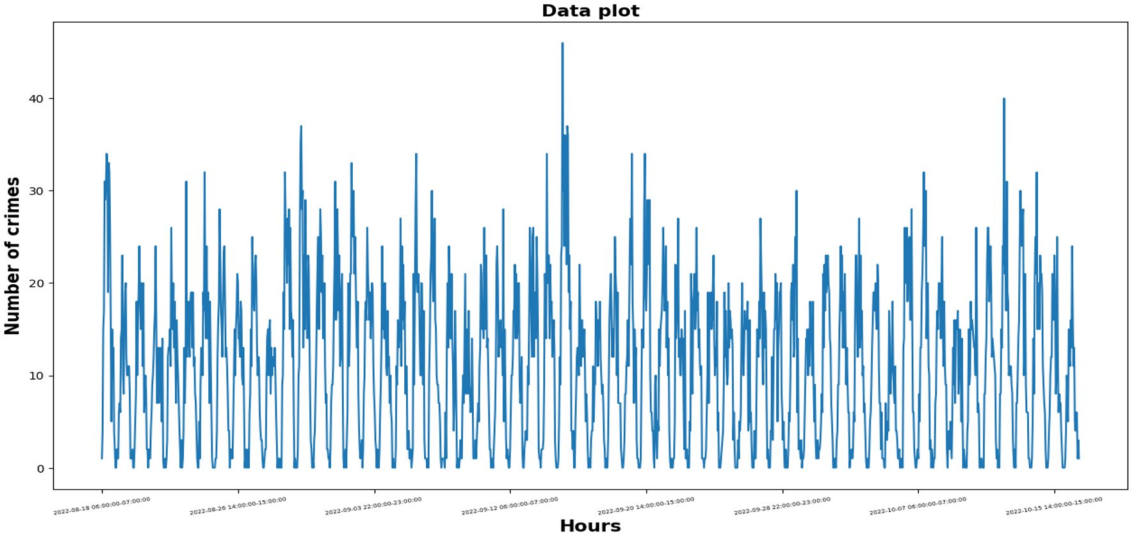


Fig. 13 Hourly time series

The autocorrelation function gives the correlation between a time series and its lags. The partial autocorrelation function gives the same correlation but after removing the relations explained by previous lags. The order of the ARIMA model as generated by the autoarima function is (2, 0, 5) and that of the SARIMA model is (0, 1, 2)x(2, 1, 1, 24) due to daily seasonality of data.

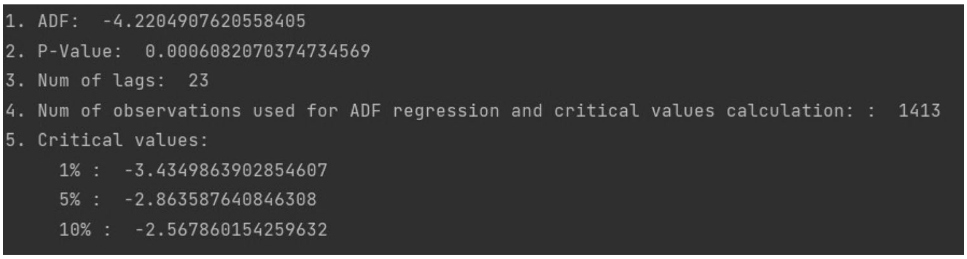
Figures 16, 17, and 18 show the validation plots generated for each of the three models. These plots compare forecasted data with real-time data to determine the accuracy of each model. The last 24 h of training set data is used for validation.

Table 3 shows the Root Mean Squared Errors (RMSE) and Mean Absolute Error (MAE) generated from the validation plots for each of the three forecasting models.

From Fig. 19, the ARIMA model is found to have the least RMSE of the three models. Therefore, the ARIMA model is the best-suited model for time series forecasting of crime tweet count. Figure 20 compares the 24-h forecast plots generated by the three models.

Table 4 shows the hourly forecast of crime provided by the ARIMA model for the period 17/10/2022–18/10/2022.

Fig. 14 Augmented Dicky–Fuller test



```
1. ADF:  -4.2204907620558405
2. P-Value:  0.0006082070374734569
3. Num of lags:  23
4. Num of observations used for ADF regression and critical values calculation: :  1413
5. Critical values:
     1% :  -3.4349863902854607
     5% :  -2.863587640846308
     10% :  -2.567860154259632
```

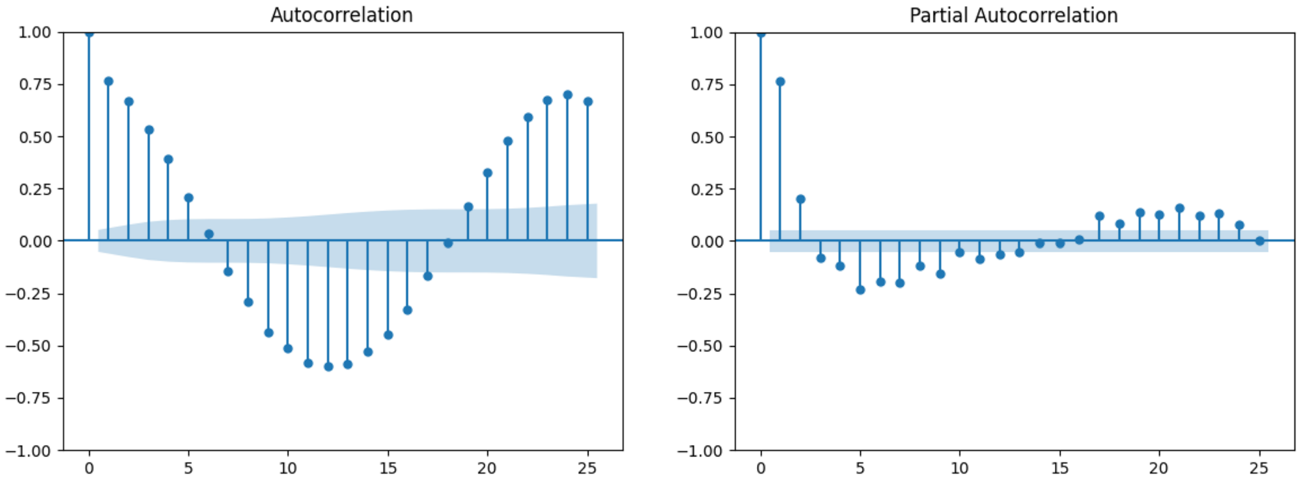
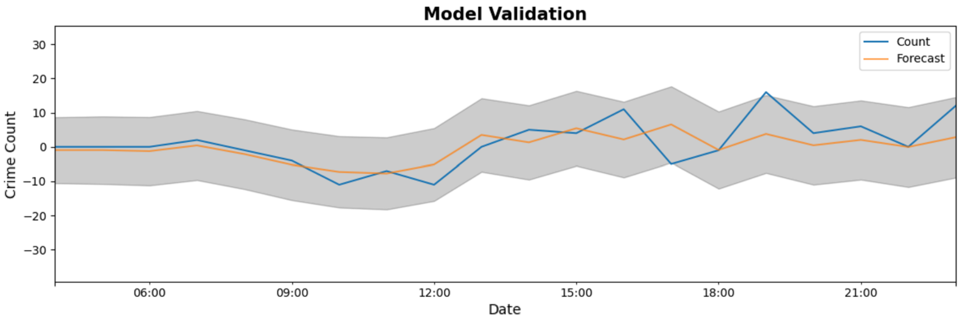


Fig. 15 ACF and PACF plots of aggregated data

Fig. 16 Validation plot for SARIMA model



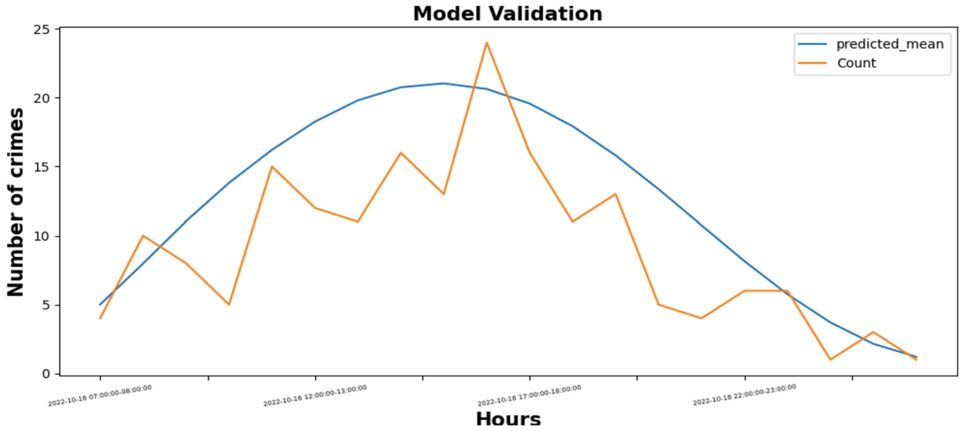


Fig. 17 Validation plot for ARIMA model

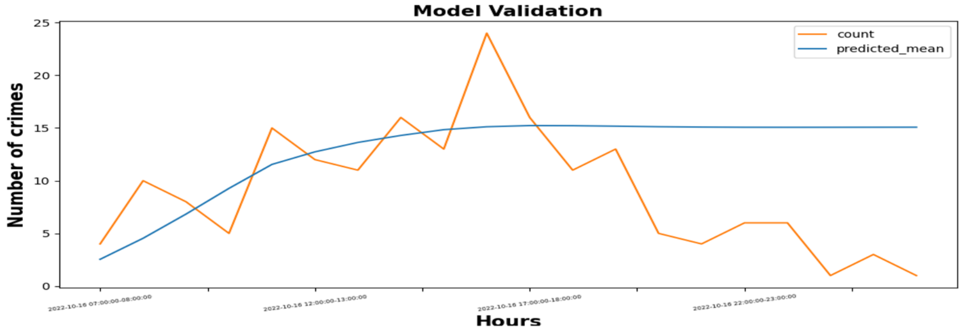


Fig. 18 Validation plot for LSTM model

Table 3 Model errors

Model	RMSE	MAE
ARIMA	3.283	2.005
SARIMA	5.287	3.772
LSTM	7.781	6.285

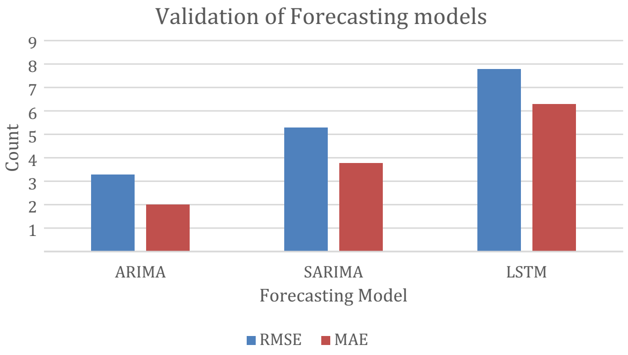


Fig. 19 Comparision of forecasting models

Verification with Karnataka State Police (KSP) Data

The keywords were restructured to match the crime categories of the Karnataka State Police. Certain categories that are not of common knowledge or public importance were ignored leaving behind the following categories—Spl & Local laws, Theft, Murder, Cases of Hurt, 107 Cr.P.C., Cybercrime, NDPS Cases, Cheating, Burglary—Night, Riots, POCSO, Rape, Robbery, Cr. Br. of Trust, Dacoity. The crime data scraped from Twitter along with the corresponding KSP data for the month of September 2022 are listed in Tables 5 and 6.

From Figs. 21 and 22, the number of tweets does not match the number of crimes directly as not every crime is tweeted about. Therefore, the twitter data are validated against KSP data by comparing the order of crime categories based on the number of crimes when arranged from the highest to lowest. As seen from Tables 5 and 6, there is a mismatch for the categories of Murder and Rape, giving a match of 86.67%. This shows that Twitter is an ideal source of data for crime forecasting (Table 6).

Suitability of the ARIMA Model

The ARIMA model can handle data sets with trends but cannot handle seasonality in data. The ADF test in Sect. "Hourly plot of crime tweet count" shows a P-value of 0.006 showing that the data is stationary that is with low enough trend and seasonality to apply the ARIMA model. However, on

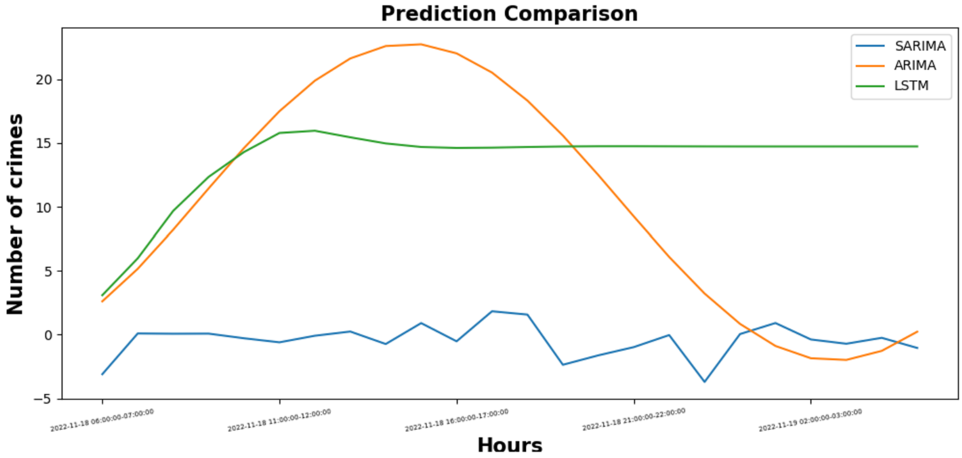


Fig. 20 Comparison of SARIMA, ARIMA, and LSTM forecasts

Table 4 ARIMA forecast

Time	Crime count
2022–10-17 03:00:00–04:00:00	1.014640865
2022–10-17 04:00:00–05:00:00	1.558521878
2022–10-17 05:00:00–06:00:00	2.58573119
2022–10-17 06:00:00–07:00:00	4.281474358
2022–10-17 07:00:00–08:00:00	6.411219367
2022–10-17 08:00:00–09:00:00	8.842567881
2022–10-17 09:00:00–10:00:00	11.40985838
2022–10-17 10:00:00–11:00:00	13.93816941
2022–10-17 11:00:00–12:00:00	16.25523778
2022–10-17 12:00:00–13:00:00	18.2031955
2022–10-17 13:00:00–14:00:00	19.64932577
2022–10-17 14:00:00–15:00:00	20.49510508
2022–10-17 15:00:00–16:00:00	20.68291559
2022–10-17 16:00:00–17:00:00	20.19997023
2022–10-17 17:00:00–18:00:00	19.07918337
2022–10-17 18:00:00–19:00:00	17.39692752
2022–10-17 19:00:00–20:00:00	15.26782924
2022–10-17 20:00:00–21:00:00	12.83695856
2022–10-17 21:00:00–22:00:00	10.26994447
2022–10-17 22:00:00–23:00:00	7.741689586
2022–10-17 23:00:00–00:00:00	5.424453273
2022–10-18 00:00:00–01:00:00	3.47611496
2022–10-18 01:00:00–02:00:00	2.029417403
2022–10-18 02:00:00–03:00:00	1.182922747

verification with the KPSS test, P-value of 0.01 was found indicating a seasonal trend in contradiction to the ADF test. Additive seasonal decomposition is applied on the data set to obtain the trending, seasonal, and residual components of the data set using the seasonal_decompose() function in the statsmodels.tsa.seasonal module in Python. Fig. 23 shows the trend data, Fig. 24 shows the seasonal module data, and Fig. 25 shows the residual data of the crime incidents.

Table 5 Crime data as per Twitter

Sl. no.	Crime category	Count
1	Spl and Local laws	82
2	Theft	60
3	Murder	58
4	Cases of Hurt	57
5	107 Cr. P. C	47
6	Cybercrime	34
7	NDPS Cases	33
8	Cheating	31
9	Burglary–N	28
10	Riots	26
11	POCSO	20
12	Rape	19
13	Robbery	17
14	Cr. Br. Of Trust	14
15	Dacoity	3

Table 6 Crime data as per KSP

Sl. no.	Crime category	Count
1	Spl and Local laws	4210
2	Theft	1652
3	Cases of Hurt	1203
4	107 Cr. P. C	1185
5	Cybercrime	1073
6	NDPS Cases	609
7	Cheating	542
8	Burglary–N	368
9	Riots	285
10	POCSO	224
11	Murder	111
12	Robbery	104
13	Rape	48
14	Cr. Br. Of Trust	29
15	Dacoity	13

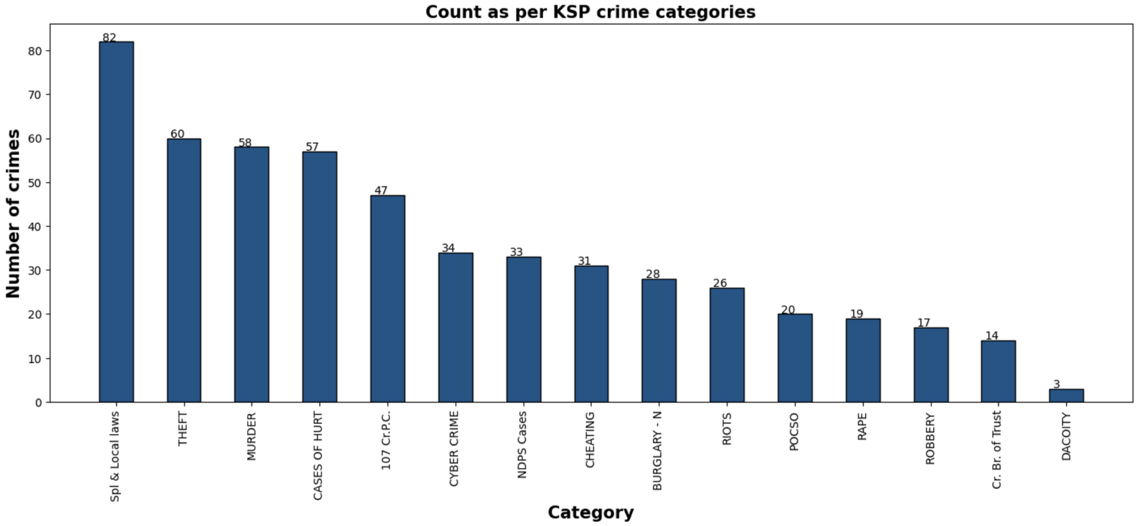


Fig. 21 Twitter crime data reorganized as per KSP categories

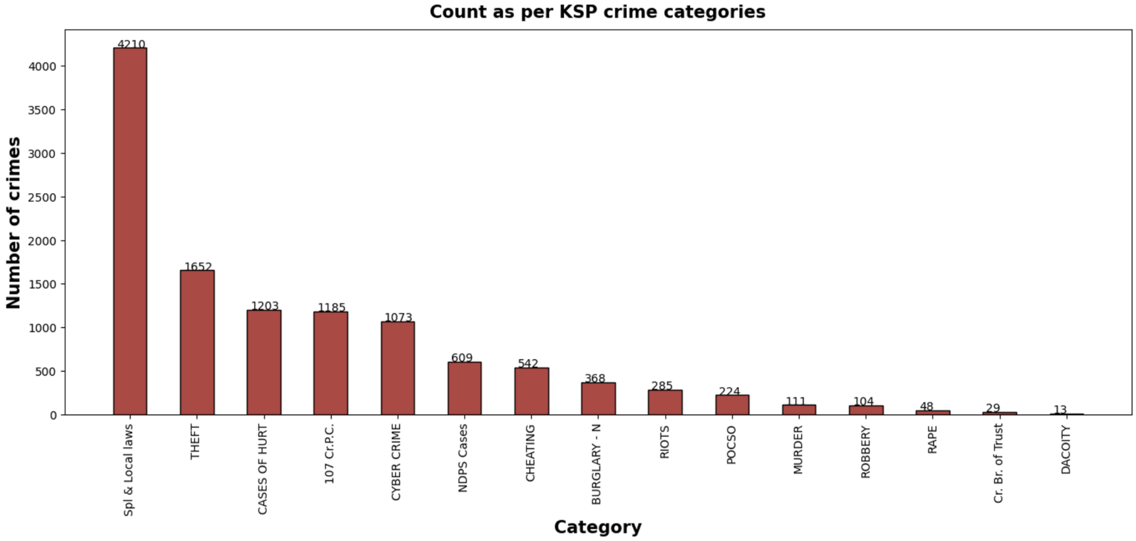


Fig. 22 KSP crime data

Validation plot for ARIMA on residual data. As seen from Fig. 26, the plot the residual data has neither seasonality nor trend and is completely stationary. ARIMA (2, 0, 2) was trained on this data to produce the following validation plot.

Figure 26 shows the model validation of residual values of RMSE of 3.135 and an MAE of 1.998, which is approximately equal to the RMSE and MAE of the ARIMA model when applied to the original data set as seen in Table 3. Therefore, the limited seasonality of the original data set does not greatly impact the accuracy of the ARIMA model. The higher accuracy of the ARIMA model over the SARIMA model is due to the method of order identification. The order of ARIMA is found using the auto_arima function while that of the SARIMA model is calculated manually. Thus, auto_arima provides a better fit model for the training data set.

Fig. 23 Trend data

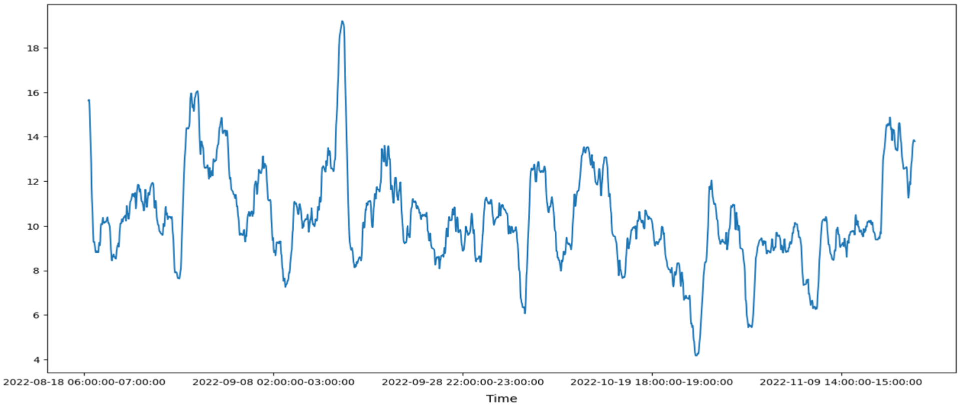
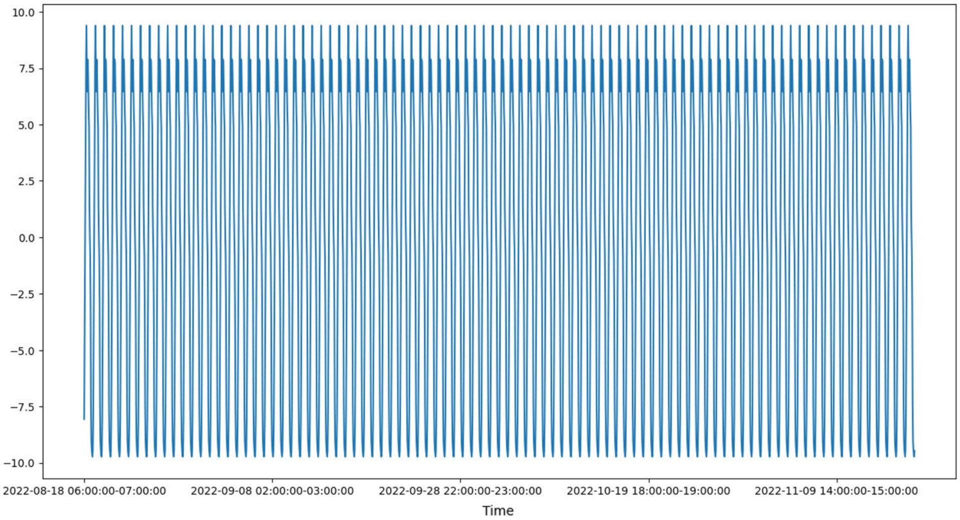


Fig. 24 Seasonality of data



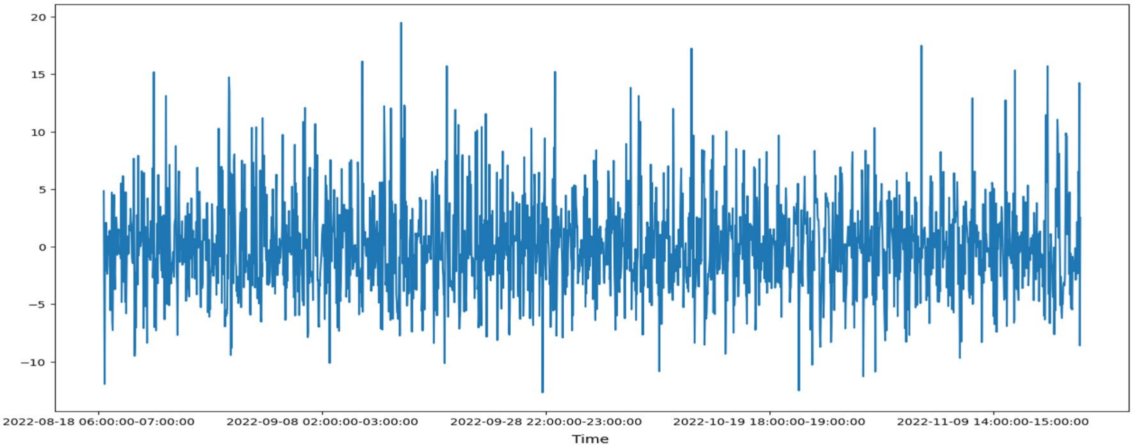


Fig. 25 Residual data

Conclusion

In this research, 2 months of relevant crime tweets scraped from Twitter were cleaned and classified into six categories based on 318 unique crime keywords. Analytical and geospatial visualization techniques were applied to the 15,601 tweets collected to produce valuable insights. The comparison of ARIMA, SARIMA, and

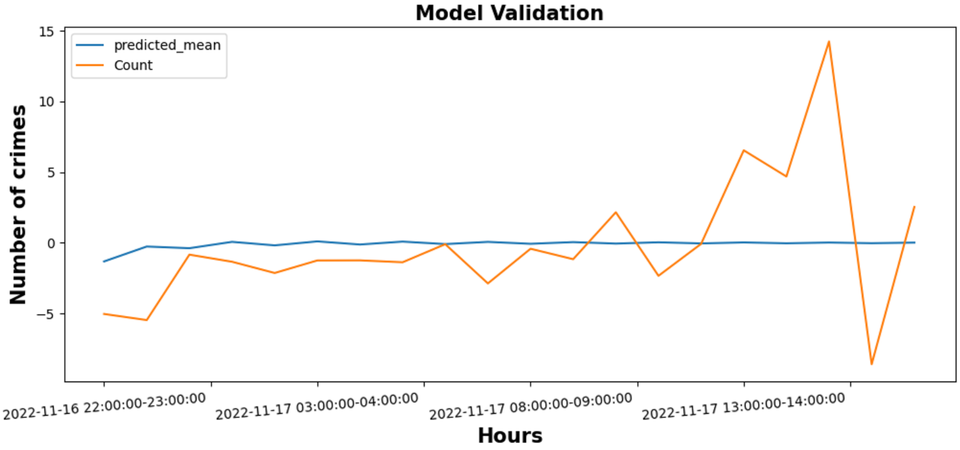


Fig. 26 Model Validation for residual data

LSTM models showed that ARIMA is most suitable for time series forecasting of crime tweet count. Adding new users over time and increasing median user experience will further improve the quality of the crime tweet data set. The outcome of this study shows that Twitter is a viable option for analyzing and forecasting crime. Twitter data may also be used in combination with other sources of data as a means of verification. This research focuses only on univariate forecasting techniques and uses only Twitter as a source. There is a vast scope for using advanced forecasting techniques and multiple social media platforms for crime analysis and forecasting in the field of criminology. Thus, law enforcement agencies can use the developed system to optimize resource distribution, thereby improving crime response rates.

Funding The authors would like to confirm that there was no funding received for the above work.

Data availability The author collects the Dataset used in this research which will be provided on request.

Declarations

Conflict of interest I know of no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome. As the corresponding author, I confirm that the manuscript has been read and approved for submission by the named author.

References

1. Conrow L, Aldstadt J, Mendoza NS. A spatio-temporal analysis of on-premises alcohol outlets and violent crime events in Buffalo, NY. Appl Geogr. 2015;58:198–205.
2. Ohyama T, et al. Investigating crime harm index in the low and downward crime contexts: a spatio-temporal analysis of the Japanese Crime Harm Index. Cities. 2022;130:103922.
3. Catlett C, et al. Spatio-temporal crime predictions in smart cities: a data-driven approach and experiments. Pervas Mob Comput. 2019;53:62–74.
4. Hu Y, et al. A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. Appl Geogr. 2018;99:89–97.
5. Rummens A, Hardyns W, Pauwels L. The use of predictive analysis in spatiotemporal crime forecasting: building and testing a model in an urban context. Appl Geogr. 2017;86:255–61.
6. Prathap BR. Geospatial crime analysis and forecasting with machine learning techniques. In: Artificial Intelligence and Machine Learning for EDGE Computing. Academic Press, pp 87–102. 2022
7. Wang Q, et al. CSAN: a neural network benchmark model for crime forecasting in spatio-temporal scale. Knowl Based Syst. 2020;189:105120.
8. Prathap BR, Ramesha K. Twitter sentiment for analyzing different types of crimes. In: 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT). IEEE. 2018
9. Prathap BR, Ramesha K. Geospatial crime analysis to determine crime density using Kernel density estimation for the Indian context. J Comput Theor Nanosci. 2020;171:74–86.
10. Boppuru PR, Ramesha K. Geo-spatial crime analysis using newsfeed data in Indian context. IJWLTT. 2019;14(4):49–64. https://doi.org/10.4018/IJWLTT.2019100103.
11. Boppuru PR, Ramesha K. Spatio-temporal crime analysis using KDE and ARIMA models in the Indian context. Int J Digit Crime Foren (IJDCF). 2020;12(4):1–19. https://doi.org/10.4018/IJDCF.2020100101.
12. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2:160. https://doi.org/10.1007/s42979-021-00592-x.
13. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci. 2022;3:158. https://doi.org/10.1007/s42979-022-01043-x.
14. Jangada Correia V. An explorative study into the importance of defining and classifying cyber terrorism in the UK. SN Comput Sci. 2022;3:84. https://doi.org/10.1007/s42979-021-00962-5.
15. Kanimozhi N, Keerthana NV, Pavithra GS, Ranjitha G, Yuvarani S. CRIME type and occurrence prediction using machine learning algorithm. Int Conf Artif Intell Smart Syst (ICAIS). 2021;2021:266–73. https://doi.org/10.1109/ICAIS50930.2021.9395953.
16. Sivanagaleela B, Rajesh S. Crime analysis and prediction using fuzzy C-means algorithm. In: 2019 3rd International Conference

on Trends in Electronics and Informatics (ICOEI), pp. 595–599. 2019. https://doi.org/10.1109/ICOEI.2019.8862691.

17. Kumar A, Verma A, Shinde G, Sukhdeve Y, Lal N. Crime prediction using k-nearest neighboring algorithm. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–4. 2020. https://doi.org/10.1109/ic-ETITE47903.2020.155
18. Safat W, Asghar S, Gillani SA. Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access. 2021;9:70080–94. https://doi.org/10.1109/ACCESS.2021.3078117.
19. Wikipedia. Crime in India. Wikimedia Foundation. Last modified September 14, 2022. https://en.wikipedia.org/wiki/Crime_in_India.
20. O'Neill A. India—Urbanization 2021. Statista, July 29, 2022. https://www.statista.com/statistics/271312/urbanization-in-india/.
21. Malik AA. Urbanization and crime: a relational analysis. J Hum Soc Sci. 2016;21:68–9.
22. Gadagpolice. "Monthly Crime Review." Monthly Crime Review - Karnataka State Police. Accessed October 17, 2022. https://ksp.karnataka.gov.in/new-page/Monthly%20Crime%20Review/en.
23. Gerber MS. Predicting crime using Twitter and kernel density estimation. Decis Sup Syst. 2014;61:115–25.
24. Prathap BR. Geo-spatial crime density attribution using optimized machine learning algorithms. Int j inf tecnol. 2023;15:1167–78. https://doi.org/10.1007/s41870-023-01160-7.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.