"THE IMPACT OF SOCIAL MEDIA ON CRIMINAL INVESTIGATIONS AND TRIALS IN INDIA: A STUDY ON THE OPPORTUNITIES AND CHALLENGES"

A DISSERTATION TO BE SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF DEGREE OF MASTER OF LAWS

SUBMITTED BY
MEGHNA RAGHAV
ROLL NO – 1220997021
LL. M. (CRIMINAL LAW)

UNDER THE GUIDANCE OF
DR. SUDHIR AWASTHI
DEAN
SCHOOL OF LEGAL STUDIES

SESSION 2022-23

DECLARATION

Title of Project Report "**THE IMPACT OF SOCIAL MEDIA ON CRIMINAL INVESTIGATIONS AND TRIALS IN INDIA: A STUDY ON THE OPPORTUNITIES AND CHALLENGES"**

I understand what plagiarism is and am aware of the University's policy in this regard. MEGHNA RAGHAV

I declare that

(a) The work submitted by me in partial fulfilment of the requirement for the award of degree **LLM** Assessment in this **DISSERTATION** is my own, it has not previously been presented for another assessment.

(b) I declare that this **DISSERTATION** is my original work. Wherever work from other source has been used, all debts (for words, data, arguments and ideas) have been appropriately acknowledged.

(c) I have not used this work previously produced by another student or any other person to submit it as my own.

(d) I have not permitted, and will not permit, anybody to copy my work with the purpose of passing it off as his or her own work.

(e) The work conforms to the guidelines for layout, content and style as set out in the Regulations and Guidelines.

Date: NAME MEGHNA RAGHAV

UNIVERSITY ROLL No.1220997021

LL.M. (2022-23)

CERTIFICATE

This is to certify that the research work entitled **"THE IMPACT OF SOCIAL MEDIA ON CRIMINAL INVESTIGATIONS AND TRIALS IN INDIA: A STUDY ON THE OPPORTUNITIES AND CHALLENGES**" is the work done by a student of Babu Banarsi Das University, Lucknow , under my guidance and supervision for the partial fulfillment of the requirement for the Degree of (LLM) in Babu Banarsi Das University Lucknow, Uttar Pradesh. According to the best of my knowledge, he/she has fulfilled all the necessary requirements prescribed under the University Guideline with regard to the submission of this dissertation.

I wish him/her success in life.

Date : DR. SUDHIR AWASTHI

(DEAN)

Babu Banarasi Das University

ACKNOWLEDGMENT

I would like to take this opportunity to express my heartfelt gratitude and appreciation to all those who have contributed to the completion of my dissertation study and related research. This significant milestone would not have been possible without the support and encouragement of several remarkable individuals, and I would like to acknowledge them with deep appreciation.

First and foremost, I extend my sincere thanks to the Almighty God for granting me the strength, knowledge, and guidance throughout this journey. The blessings and divine intervention have been instrumental in shaping my thoughts and providing the perseverance to overcome challenges along the way.

I would like to extend my gratitude to my esteemed Dean, Mr. Sudhir Awasthi. His guidance, mentorship, and valuable insights have been crucial in shaping the direction of my research. His vast knowledge and expertise in the field have immensely contributed to the quality of my work. I am grateful for his unwavering support and encouragement throughout this journey.

I am incredibly indebted to my family for their unwavering love, encouragement, and understanding. Their constant support, both emotionally and financially, has been invaluable in helping me pursue this research endeavour. I am grateful for their sacrifices, patience, and belief in my abilities.

I would also like to express my heartfelt appreciation to my friends for their continuous

motivation and inspiration. Their intellectual discussions, insightful suggestions, and moral support have been instrumental in refining my ideas and boosting my confidence. Their belief in my potential has been a constant source of motivation.

Finally, I extend my thanks to all the teachers, professors, and colleagues who have provided me with valuable feedback, shared their knowledge, and assisted me in various aspects of my research. Their expertise and constructive criticism have been instrumental in refining my work and expanding my understanding.

In conclusion, I am immensely grateful to God, my family, friends, and my esteemed Dean, Mr. Sudhir Awasthi, along with all those who have contributed to my dissertation study and related research. Their unwavering support, guidance, and belief in my abilities have been the pillars of strength that have enabled me to complete this academic endeavour successfully.

Thank you all from the bottom of my heart.

MEGHNA RAGHAV

UNIVERSITY ROLL No.1220997021

LL.M. (2022-23)

LIST OF ABBREVIATIONS

Abbreviations	Full Form Of Abbreviations
KDE	Kernel Density Estimation
RTI	Right to Information
RSS	Rich Site Summary and Really Simple Syndication
GIS	Geographical information system
RDF	Resource Description Framework
XML	eXtensible Markup Language
FIR	First Information Report
RFID	Radio Frequency Identification
GPS	Global Positioning System
ARIMA	Autoregressive Moving Average
GDpatterns	Geospatial Discriminative patterns
NLP	Natural Language Processing
CDM	Concentration Driven Model
SVM	Support Vector Machine
GUI	Graphical User Interface
API	Application Program Interface
NCRB	National Crime Records Bureau
CCTNS	Crime and Criminal Tracking Network System
TIS	Talash Information System
CIPA	Common Integrated Police Application

PCA	Principal Component Analysis
LDA	Latent Dirichlet Allocation
CDR	Call Detail Record
COSMOS	Community-oriented Online Social Media Observatory
ESRC	Economic and Social Research Council
KSP	Karnataka State Police
VADER	Valence Aware Dictionary for sEntiment Reasoning
EDA	Exploratory data analysis
IDA	Initial Data Analysis
CSV	Comma Separated Values
PDF	Probability Density Function
ACF	Auto Correlation Function
ACVF	Auto Correlation and Auto Co-Variance

LIST OF CASES

- AG v Shivkumar Yadhav & Anrs, CrlA. 2015. AK Gopalan v Noordeen, AIR 1969 2 SCC 734.
- Anita Whitney v California, 274 US 357 (1927).
- Ankur Chandra Pradan v Union of India, (1996) 6 SCC 354. Arnold v Emperor, 1914 PC 116.
- Attorney General v English, (1983) 1 AC 116.
- Attorney General v Guardian Newspaper, (1990) 1 AC 109. Attorney General v Times Newspaper Ltd, (1973) 2 AIIER 54. Austin v Keefe, 402 US 495, (1971).
- Australian Securities and Investment Commission v Rich, (2001) 51 NSWLR 643. Bachan Singh v State of Punjab, (1980) 2 SCC 684.
- Balhine Ramakrishna Reddy v State of Maharashtra, (1925) SCR 425. Bennett Colemans v Union of India, AIR 1973 SC 106.
- Bijoyananda v Bala Kush, AIR 1953 Ori 249. Bridge v California, 314 US 252 1941.
- Brij Bushan v State of Delhi, AIR 1950 SC 129.
- Chintamani Rao v State of Madya Pradesh, AIR 1951 SC 118. Court on its own Motion v State, (2008) 146 DLT 149.
- Court on its Own Motion v The Publisher, Times of India, Civil Writ Petition Number 7160 of (2013).
- Dagenais v Canadian Broadcasting Corporation, (1994) SCR 835. DC Saxena v Hon'ble The Chief Justice of India, (1996) 5 SCC 216.
- Delhi Judicial Service Association Tiz Hazari v State of Gujarat & Anrs, AIR 1991 SCW 2419.
- EM Shankaran Namboodiripad v T. Narayana Nambiar, (1970) 2 SCC 325.
- Express Newspaper v Union of India, AIR 1958 SC 578. Goodwin v UK, (1996) 2 EHRR 123.
- Govind v State of Madya Pradesh, AIR 1975 SC 1378. Govind Shahi v State of Uttar Pradesh,

AIR 1968 SC 1513. Gujarat Water Supply v Unique erectors, AIR 1989 SC 973. Hamdard Dawakhana v Union of India, 2 SCR 671 (1960). Hussainara Khatoon v State of Bihar, AIR 1979 1369.

- Indian Express Newspaper Bombay Pvt Ltd v Union of India, AIR 1986 SC 515. In Re Harijai Singh & Anrs, (1996) 6 SCC 466.
- In Re S.Mulgaokar, AIR 1978 SC 727.
- In Re Subrahmanyam, AIR 1953 Mad 422.
- In Re Vinay Chandra Mishra, (1995) SC 2348. JR Prasad v Prashant Bushan,(2001) 6 SCC 735. KA Abbas v Union of India, AIR 1971 481.
- Kartongen Kemi Och Forvaltning AB v State through CBI,(2004) 72 DRJ 693. Kehar Singh v State AIR 1988 SC 1883.
- Kharak Sing v State of Rajasthan, AIR 1963 1295. Kishore Singh v State of Rajasthan, AIR 1981 625. Leo Roy Fray v R.Prasad, AIR 1958 P&H 377.
- LIC v Manubhai D Shah, AIR 1992 3 SCC. Lowell v Griffin, (1939) 444 US.
- Mahesh Bhatt v Union of India, 2008 (147) DLT 561. Maneka Gandhi v Union of India, AIR 1978 SC 597. Manu Sharma v State of Delhi, (1010) 6 SCC 1.
- Mirror v Superior Court, 3 Cal. 2d 309.
- Mohammed Ajmal Amir Kasab v State of Maharashtra, AIR 2012 9 SCC 1.
- Mother Diary Food & Processing Ltd v Zee Telefilms, AIR 2005 Delhi 195.
- M.P. Lohia v State of West Bengal, (2005) 2 SCC 686.
- National Legal Services Authority vs. Union of India, AIR 2014 SC 1863. Near v Minnesota, 283 US 697 (1931).
- Nebraska Press Association v Stuart, 427 US 593 (1976). Neelam Katara v UOI , 2003 DLH 84.

- Neeraj Sridhar Mirajkar v State of Maharashtra, AIR 1967 SC 1.
- N. Ram vs Siby Mathew And Anr, 2000 CriLJ 3118. New York Times Co. v US, 403 US 703.
- New York Times Sullivian, 376 US 254 (1964).
- Papansam Labour Union v Madura Courts Ltd, AIR 1995 2200. Pennekamp v Florida, 328 US 331 (1946).
- Perspective Publications v State of Maharashtra, AIR 1971 SC 221.
- P.N. Krishna Lal v Government of Kerala, 2009 CriLJ 2974. Prabhu Dutt v Union of India, AIR 1982 SC 6.
- Rajendra Sail v Madhya Pradesh High Court Bar Association, (2005) 6 SCC 109. Rakeysh Omprakash Mehra vs. Governmen. of NCT of Delhi, 2013 (197) DLT 413. Ram Dayal v State of Uttar Pradesh, AIR 1978 SC 921.
- Rao Harnarain v Gumori Ram, AIR 1958 P& H 273.
- Reliance Petrochemicals v Proprietors of Indian Express, AIR 1989 SC 190. Rex v BS Nayyar, AIR 1950 551.
- RK Anand v Registrar, 8 SCC 106(DEL.2009). RK Garg v SA Azad, AIR 1967 AII 37.
- Romesh Thappar v State of Madras, AIR 1950 SC 124. R Rajagopal v State of Tamil Nadu, AIR 1995 264.
- Russell v Russell, (1976) 134 CLR 495.
- R v Oaks, (1986) 26 DLR 20.
- Sahara India Real Estate V Securities Exchange Board of India, (2012) 10 SCC 603. Saibal Kumar v B.K. Sen, AIR 1961 S.C 633.
- Sakal Papers v Union of India, AIR 1962 SC 305. Scott v Scott, (1913) AC 417.
- Sharad Birdhichand v State of Maharashtra, AIR 1984 1622. Sheela Barse v UOI, AIR 1986

1773.

- Sheppard v Maxwell, 384 US 333 1966.
- Smt. Padmawathi Devi v R.K. Karanjia, AIR 1963 MP 61.
- S.P Gupta v Union of India, AIR 1982 SC 149. State v Biswanath Mohapatra, AIR 1955 Ori 169.
- State v Editor, Printer Publisher of Mathrubhumi, 1954 CriLJ 926. State of Maharashtra v Rajendra Jawanmal Gandhi, (1997) 8 SCC 386. State v Simpson, No. BA 097211 (Cal. Super.Ct filed July 22 1994).
- Stroble v California, 343 US 181 1952.
- Subhash Chandra v S.M. Agarwal, (1984) CriLJ 481 DEL Sunil Batra v Delhi Administration, AIR 1980 1579.
- Surendra Mohanty v State of Orissa, Crl App No 107(56),
- The Sunday Times v UK, (1979), Series A No. 30, 14 EHRR 229.
- Thornhill v Alabama, (1940) 310 US 88. Virendra v State of Punjab, AIR 1957 896. Visakha v State of Rajasthan, AIR 1997 SC 3011.
- Y.V Hanumantha Rao v KR Pattabhiram and Anr, AIR 1975 SC 1821. Zahira Habibulla Sheikh v State of Gujarat, (2004) 4 SCC 158.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

BACKGROUND INFORMATION

People increasingly use social media for sharing their ideas, views, and news associated with the incidents in the geographical area they are interested in. Crime prediction is a well-known research area for improving public safety. Studies show that social media data can use crime analytics and forecasting. Rapid advancements in technology have supported the analysis of big data. However, crime is increasing due to urban congestion. National security is an important goal for any nation. Countries have invested in criminology studies to understand distinct criminal characteristics and data mining utilized in this application. Crime analysis is performed by law enforcement agencies to analyze the tendencies and patterns in crime and take the systematic review.

There are various methods for analyzing crime. This research is mainly identifying crime related to spatial location and time. There is quite a lot of crime data mining approaches available corresponding to clustering systems, sequential pattern mining, association rule mining, string assessment, and classification. Social media data can use for analysis and prediction of crime.

Crime analysis plays a significant role in resolving criminal cases. Rapid advancements in technology have enabled faster investigation of crimes. Population growth and unregulated migration play a substantial role in crimes occurring in cities. Law enforcement organizations and Intelligence organizations collect large amounts of crime data for predicting future occurrences. Since the quantity of data is high, manual methods of data analysis are not fruitful. Therefore crime analysis is one of the critical problems that remain to solve for law enforcement agencies.

Unlike developed countries like the United Kingdom, the United States of America, etc., India does not share crime-related data in the online domain. Also, crime

recording and analysis have done manually, which makes interpretation of the information very difficult. Crime analysis has done for several reasons, such as identifying crime patterns, trends, etc. to maximize the utilization of limited resources available to police agencies. There are various data mining techniques, such as classification, clustering, pattern mining, etc.

In addition to the crime data available to the police, social media data can have used as a potential source for identifying hotspots for criminal activity. With this data doubling every two years, data analytics can help in predicting human behaviours (Witten, Frank and Hall, (2011).

This research work showcases a crime analytics system that makes use of social media data, i.e. Newsfeeds (RSS feeds). Using crime analysis, crime events that have similar characteristics such as frequency, time, location and victim patterns, etc. can be identified. Spatial and temporal information used in social media data can have used to create a profile of social misconduct. The insights gained from social media data have used for taking corrective actions. Anomalies can be detected and used to avoid social malfunctions utilizing a highly efficient data pipeline (Manoochehri, 2014).

OVERVIEW

Crime analysis using data is the emerging discipline in criminology. Law enforcement agencies are focusing on methods that enable them to predict future attacks, and This allows them to utilize their limited resources effectively. The significant challenges faced by these agencies are the complexities involved in the processing of large volumes of data. The complexity of crime data and a variety of geographical diversity have made crime analysis difficult. Researchers are focusing their efforts on data mining algorithms that can extract meaningful information from crime data.

There are various sources of crime data. In this research used news feed data about

crime in total India and on specific to Bangalore region. Social media data has richdata about user emotions on a particular topic. The main advantage of this kind of data is that it has precise spatial and temporal coordinates. These spatial-temporal data can have used for crime prediction, which had processed with the help of linguistic analysis and topical statistical modelling. Social media data can be used as auxiliary sources in addition to traditional data sources to increase the accuracy of the prediction. However, there are also limitations to using social media data. Tweets have inconsistent information, on the fly word invention, misspellings, symbol use and syntactic structures that cannot be handled by computational algorithms. Even though the content has personalized content in real-time, it is hard to process. Newsfeed has used as the primary source of data in this paper. The data needs to be processed with automated text analysis, smart segregation and filtering methods.

The spatiotemporal analysis provides insights about the situational awareness of local events, enables understanding of the severity, consequences, and the time-evolving nature of the crime. The spatiotemporal analysis has been done based on volume-based importance. In this case, the messages are extracted from news feed data and then filtered and sorted by space and time. A major challenge in using news feed data is that critical information has mostly obscured by high volumes of inconsistent, incomplete and inaccurate data.

WHAT IS CRIME

Crime is a multifaceted concept that had explained in both legal and non-legal sense. From a legal perspective, it has shown that certain activities are not acceptable in particular geographic areas. The breach of the criminal laws that aim at protecting the rights, property, and lives of citizens present within those areas is defined as a crime. The crimes in which the criminal justice system is involved include breach of state legislation and territory that provides for offences related to property (theft and damage), persons (sexual assault and murder) and regulations (traffic rules

violations), etc. There is commonwealth legislation related to matters like commerce and trade, exports and imports, defence, external affairs, and taxation.

The non-legal perspective of crime indicates it as an act that violates the socially and morally accepted rules of human behaviour. The definition of crime in this context depends on the moral principles that underpin it, and hence, it can change when the tenets change. The examples of behaviours that have decriminalized include attempted suicide, prostitution, abortion, and homosexual intercourse. There are activities such as credit card fraud and tax evasion that have become criminalized over time. It is necessary to understand the contradiction between these two viewpoints. The data collected by law enforcement agencies had based on the legal definition of crime. Data collected on victimization had found on the individual interpretation of crime. This disparity affects the crime definition at extreme ends.

TYPES OF CRIME ANALYSIS

There are six different types of crime analysis, as follows. This analysis has some common characteristics of crime analysis, but the review and its use cases are mixed. Unlike law enforcement analysts do these activities. The policing analysis is of various types. These analyses are done by crime analysts to give a complete picture of the crime.

Tactical Crime Analysis: Tactical crime analysis has defined as the analysis of the day to day activities that looks for patterns, series, hotspots, sprees, and hot dots that affect the jurisdiction. This analysis also studies specific crime-related information such as a point of entry, method of entry, type of victim, suspect actions and type of weapons used, etc. This analysis makes use of date, location, time of the crime, and kind of location. There are field information used, such as persons with tattoos, criminal trespass warnings, suspicious activity calls for getting service, scars, or marks. It is possible to gain the insights needed for the crime analysis. This analysis is

used for (i) day today, (ii) for patterns, series, hotspots, sprees, and (iii) used for administration and deployment.

Strategic Crime Analysis: This crime analysis analyzes the socio-economic and spatial factors affecting crime in a particular area. This study helps in problem solving and understanding of the long term patterns of the activity. Also, this study supports the officials to prepare procedures and responses in reducing crime. This analysis used for (i) Find out unusual activity levels by location or time and (ii) Forecast potential crime events or concentration.

Administrative/Academic Crime Analysis: The study of crime information along with socio-demographic and spatial factors to identify the long term "patterns" inactivity, to evaluate and research the procedures and responses, and to help in problem-solving.This analysis used for (i) Statistical summaries or reports for public, commanders and grant funding and (ii) Implications on policy beyond the law enforcement agencies.

Operations Analysis: Operations analysis analyzes the policing practices related to patrol allocation and resource utilization. The examples include analysis of overtime worked by police officers in a year. Strategic crime analysis and operational analysis and help the patrol officers to use their resources efficiently. This study used for (i) Assessing needs (population of data, calls for service and demographics) and (ii) Generate projections for resource allocation and deployment.

Intelligence Analysis: Intelligence analysis is the analysis of criminal enterprises and organizations, the key players involved in the process, and their linkages. This analysis enables the prosecution and investigation units within the police force, also Using intelligence analysis, it is possible to assist police officials in identifying the networks and apprehend the individuals responsible for the crime. This study can help in preventing further criminal activity. This analysis used for (i) The linkage between

crime organizations and enterprises and (ii) Relate elements including agencies, companies, times, people, days, places and to crimes.

Investigative Analysis: In the case of investigative analysis, the information from the crime scene has utilized. It includes forensic and psychological analysis used in crime locations. This study helps in catching arsonists and criminals and serial killers. The primary purpose of the criminal investigative study is to create patterns in serial crimes that cross state, city, national, or even international boundaries by linking evidence and behaviour among and within incidents to gain an understanding of the offender and solve cases. This analysis used for (i)Psychological, forensic, and Crime scene information and (ii)Link serial crimes or related events.

PHASES OF CRIME ANALYSIS AND MAPPING

Crime has considered as one of the biggest threats to the development of a country. The understanding of crime behaviours had limited until the advent of big data. Crime generally occurs in clusters. This study has direct implications on the crime prevention strategies of law enforcement agencies. Criminal activities are on the rise in major cities such as Bangalore. This study had attributed to the population density, urban immigration, and the existence of slums in the city area. By understanding the factors that drive criminal behaviour in spatial and temporal terms, law enforcement agencies can identify crime clusters and take appropriate action.

Research conducted by Algahtany, Kumar, Barclay and Khormi (2017) shows that crime happens due to conditions called crime generators. The first success of crime encourages the criminal to conduct the activity repeatedly within their surroundings. Criminals create a safety zone around their surroundings and then gradually expand the area. They do serial offending and repeat victimization. It has found that nearly 68% of the crime happens in the same area. The consistent occurrence of crime in a particular area creates a crime hotspot. This study leads to many social and economic

consequences for the crime clusters, e.g., depreciating house prices, increased fear, etc. The discovery of crime clusters requires focused and prompt police action. By observing the possible stimulants of crime, the risk can be identified.

There are certain attractors for a crime, such as drug abuse, weather patterns, and specific land uses. Crime is seasonal. It has identified that cold season triggers violent crime, and hot weather triggers nonviolent crime. There are also complex associations between crime and weather. According to Biswas and Maiti (2016), crime peaks at nights, on weekends and during holidays**.**

CRIME DETECTION

Crime detection is the method used for identifying general and specific crime patterns, trends, and series in a continuous and timely manner to maximize the use of the limited police resources to prevent crime regionally, locally, and nationally. This study enables the crime-fighting agencies to be more proactive in their approach towards crime.

A study performed by (Lalit Kumar et al., 2017) examines the relationship between crime and places in Saudi Arabia. It uses geographic information systems to identify and visualize the spatial distributions of regional and national crime rates in Saudi Arabia. The crimes that had classified include assault, murder, theft, alcohol, and drug crimes over ten years, i.e., from 2003-2012. The role of "place" in crime analysis has become increasingly important in the area of environmental criminology. The spatial distribution of crime reflects the different organizational structures within the community. The focus of ecological criminologists includes spatial analysis rather than criminogenic causes such as developmental, biological and social characteristics of an offender (Bogomolov et al., 2015).

CRIME ANALYSIS

In crime analysis, domestic or international crime data has collected. This activity involves a considerable amount of data. Therefore manual techniques for analyzing these data with high variation have resulted in low productivity and ineffective utilization of human resources. This study has a significant problem in the intelligence agencies and law enforcement agencies. There are different types of crime data mining techniques available such as association rule mining, clustering techniques, classification, sequential pattern mining, and string comparison.

Crime data mining is the use of data mining techniques for the analysis of crime. There are various subcategories of crime, depending on the criteria. They include drug offences, traffic violations, theft, fraud, sex crimes, cybercrimes, arson, and violent crimes. Each category has different definitions in both local and national law enforcement agencies. Other definitions include categories such as organized crime, corporate crime, and a war crime, etc. defined by IPTC (international press telecommunications council).

There are different types of crime data mining present, including clustering techniques, association rule mining, entity extraction, classification, and sequential pattern mining. Different crime types had analyzed using automation techniques. However, there is no unified framework for applying these techniques to crime types.

GIS IN CRIME ANALYSIS

Geographical information systems and crime mapping are considered as an essential apparatus for crime detection by police agencies. The availability of geographic data sources and technology development enables police departments to use GIS and crime mapping. These are devices for understanding the reasons for crime occurrences. This study allows law enforcement agencies to take action to prevent crime proactively.

With the help of GIS, historical data have evaluated, and the importance of place explained in accurately assessing crime problems.

According to Brantingham and Brantingham (1991), location is the most critical aspect of the crime. The geographical information of the crime and place can help in the prevention program by identifying the characteristics of criminals. This study allows criminologists to get a better perspective of environmental reasons regarding crime incidents. Analysts can identify areas with high crime concentration. It is a well-known fact that crime is location specific and had not uniformly distributed. There are locations with high crime density and those with low crime density within the same city. Most researchers are interested in the high criminality of places such as towns or cities rather than the individuals. A GIS helps the researchers in understanding the factors that support crime.

CRIME MAPPING

Crime mapping had implemented through software. This method has the unique capacity to overlay, in digital map layers, different data sources, such as police calls for service, arrest reports, crime reports, the location of specific sites, and even citizen complaints to use them for analysis. Olligschlaeger (1997) notes that police are using GIS for analyzing large amounts of information to understand crime and its root causes. Spatial analysis, data management, and data visualization are essential components in crime mapping and GIS. These systems help in connecting criminal activities and their characteristics with geolocation due to the combined insights of these three components. There are other overlay functions, such as population characteristics and location of crime events. Currently, GIS and crime mapping provide a superior solution compared to other methods.

Other benefits of GIS and crime mapping include the use of the system in answering fundamental and essential questions about crime and their policing strategy. This

study enables patrolling officers to assess the specific places that need to patrolled and which areas they need to increase surveillance to prevent crimes.

Additionally, mapping helps in gaining insights about crime and criminal behaviour and delivering it to the public (Saddler, 1999). By giving specific information about crime components to the public, crime mapping and GIS provide an opportunity for the community to help police departments and participate in the prevention of crime in their neighbourhoods. In this sense, it connects police and neighbourhoods.

One vital benefit of crime mapping and GIS is preparing police departments for future crime problems in their jurisdictions by giving opportunities to the researchers to foresee future crime patterns. Consequently, crime mapping and GIS can use for deciding the proper place for new facilities according to future crime problems. As the usage of crime mapping and GIS increases day by day, law enforcement departments will use it as a functional device to find out the right place for setting up substations, headquarters, new police checkpoints, and other facilities.

CHAPTER 2

REVIEW OF LITERATURE

INTRODUCTION TO CRIME ANALYSIS

Most countries are rapidly urbanizing to improve the social conditions of their citizens. According to a survey from the United Nations, 60% of the developing world and 83% of the developed world will become urbanized. This study shows that the urban population will exceed the present world population. The correlation between the inequality of urbanites and urbanization has been studied extensively by researchers. Research suggests that cities with disparities will have a high frequency of crime. For example, a record from 1980 to 2000 showed that crime increases from 2500 to 3000 for every 100,000 people. Some researchers have demonstrated that public safety is a significant factor in the sustainable development of urban cities and the quality of citizen's life.

Researchers have shown that safety is the most fundamental aspect of a citizen. It is essential for the psychological and physical needs of the residents. Sustainable urban development depends on well-planned security systems that are available citywide. It should also be community-based, gender-sensitive. A comprehensive and integrated urban crime prevention and urban safety strategies had required for large cities (Algahtany, Kumar, Barclay & M. Khormi, 2017). The traditional crime research makes use of current demographic data. These include the wealth gap, income level, and education level, religious and ethnic differences. This demographic data is inadequate when it comes to an understanding of the complexity and dynamics of crime. One of the main reasons is that demographic features are stable for an extended period. However, crime locality is dynamic, and it cannot indicate the dynamic nature of the community accurately. The second reason is that the majority of cities have

similar demographic profiles, and hence, it is not possible to accurately represent the differences between these communities.

Recently law enforcement agencies have been implementing fine-grained data collection systems and mechanisms. This study has enabled them to collect numerous crime-related data and record them (Chainey & Radcliffe, 2018). This data gives helpful context information regarding crime in urban areas. For example, there are useful environmental factors available with human mobility, such as the function of regional or residential stability. It can significantly affect criminal activities, as found out by ecological criminology. Also, weather information and other meteorological data have found to influence urban crimes. Therefore big data from urban areas have fine-grained and rich contextual data about where and when data is collected. This varied information helps the researcher to understand the various influential factors and evolution of crime and also to understand crime from different perspectives (Mowafy, Rezk & El-bakry, 2018).

Therefore big data from urban areas opens up numerous opportunities for conducting advanced investigations on crime (Ohlan, 2019). This study also helps in the formation and testing of various criminal theories developed using criminology to understand the different varieties of criminal phenomena. For example, the methods from environmental criminologies, such as rational choice theory and routine activity theory, suggest that time and space play a significant role in understanding criminal activities (Perera, Sajeewa, Wijewardane & Wijayasiri, 2015). Ecological factors related to human mobility also play a role in criminal activities. Approaches in social disorganization, such as social criminology, show that there is a direct link between neighbourhood ecological characteristics and crime rates. Culture conflict theory shows that crime occurs due to a clash of values between different communities (Zhang & Wu, 2011). This study can happen due to the difference of opinion in what is acceptable or proper. With the help of these theories, it is possible to understand the

mechanism by which crime occurs. These theories can help in bridging the gap between what authorities have as data and what can be inferred about urban crimes.

However, the biggest problem with the urban data is that it is noisy, dynamic, heterogeneous, and large scale. Hence, more effective and efficient computational solutions required for the same. Various research has been ongoing on big urban data with several computational approaches proposed. This study has resulted in the formation of numerous computational models that integrates crime patterns and criminal theories. Research about data mining of crime information is essential to have an overview of urban crime (Bogomolov et al., 2019). The research has been going on in social criminal theories, and environmental theories from criminology, urban crime patterns, important computational crime tasks with representative algorithms.

INTRODUCTION TO SPATIOTEMPORAL CRIME ANALYSIS

There are various ways in which the data distribution of crime can occur across time and space. The traditional methods used for visualization are hotspot maps. It makes use of Kernel Density Estimation (KDE) methods for fitting historical crime record data with two-dimensional spatial probability density function (Shiode & Shiode, 2014). Hotspot maps utilized as they help the analyst in plotting the data and create visualizations that are intuitive and better identify the areas that have high crime concentrations.

SPATIAL AND TEMPORAL PATTERN ANALYSIS METHODS

It suggested by the theorists of criminology that crime has highly correlated to location and time. Big data from urban areas give valuable information on crime from spatial and temporal perspectives. This study has spurred research on Spatio-temporal pattern analysis. This analysis is a procedure that gains understanding from Spatio-temporal related sources and enables crime analysts to analyze the data. In practice,

perception depends on different environmental conditions (Chae, Thom, Bosch, Jang & Maciejewski, 2018). Different types of Spatio-temporal analysis techniques used for acquiring the appropriate information.

a) SPATIAL PATTERN ANALYSIS

In an urban area, crimes are can't randomly or evenly be distributed. In general, crime occurs more in dense regions and less in other areas (Zhao & Tang, 2018). Spatial pattern analysis aims to understand the aggregation of crime hotspots in a city. In addition to it, crimes correlated to the environmental contexts. This section discusses an introduction to spatial factor analysis and crime hotspots in detail.

Crime hotspot has defined as a geographical location that has a large number of criminal activities. It is also a location that has a higher risk of victimization. On the other hand, there are a number of hotspots that have a lesser density of crime. In general, hotspot analysis makes use of spatial clustering to find spatial patterns (Chainey, Tompson & Uhlig, 2016).

Spatial factor analysis helps in identifying the main spatial factors of crime. The central hypothesis of spatial analysis shows that crime should be correlated to environmental contexts. Many criminal theories support this hypothesis. For example, routine activity theory shows that the three elements, a suitable target, motivated offender and victim, and the nonavailability of a guardian is necessary to converge in space and time for a crime to occur.

b) TEMPORAL PATTERN ANALYSIS

Crime analysis using temporal patterns is complicated because the time can be divided in various ways, such as years, weeks, months, seasons, etc. (Yoon & Shin, 2018). In general, the temporal analysis identifies the temporal patterns in the crime data. The various types of temporal pattern analysis are as follows.

- Crime tendency has defined as the percentage contribution of a crime type in

a particular time and region. For example, property crime rates are less in the USA

- Crime periodicity has defined as the period in which the crime repeats. E.g., some crimes have seasonality
- Similarity search of crime targets uses the similarity between crimes to identify the right one
- In sequential behaviour analysis, the subsequent behaviour of the offender is analyzed before or after committing the crime. E.g., a burglar performs a robbery after buying drugs.

c) SPATIO TEMPORAL PATTERN ANALYSIS

Spatio-temporal pattern analysis of crime helps in understanding the geographic and time-related crime data. Spatio-temporal pattern analysis of crime aims to gain understanding from time and geo-related crime data. The main challenge is to find out the patterns from the interplay between time, space, and crime. Crime patterns correlated with location and time (Clancey, Kent, Lyons & Westcott, 2017). In this subsection, discusses the main spatial and temporal patterns of urban crime.

- Earthquake-like pattern: It can be shown that an earthquake will happen in an area that is near to that of the original one, From the concentrating patterns (Zhang, 2015). This kind of phenomenon is also available in crime formation, such as thieves attacking nearby areas over a time period. This study shows that seismology approaches, such as the self-exciting point process, can be used for processing urban crime.

- Spatio-temporal hotspot: Gang violence and other crimes occur in concentrated space and time (Jiang, Yang & Li, 2018). A spatiotemporal hotspot is a geographical location that has a timestamp in which a large number of crimes occur. This study will include temporal patterns in the

geographic region.

- Spatiotemporal correlations: this shows the Spatio-temporal correlations. The intra region temporal correlation created for an urban region. (i) It is most likely to have similar crime numbers for two-time slots (Kedia, 2019). If the difference between the time slots is high, it can increase the crime difference. The inter-region spatial correlation is present in urban areas (i) two regions that are near will have similar crime numbers (ii) the crime difference will increase with an increase in spatial distance between the regions.

CRIME PREDICTION USING SPATIOTEMPORAL DATA

Crime prediction defined in the theory that future crimes happen near to that of past crimes.This can be done with the help of hotspot maps (Zahra, 2018). The spatiotemporal clustering of urban crimes can be modelled using self-exciting point process models. But there are a number of limitations. For example, it is not possible to create models that are independent of historical data (Khalid, Wang, Shakeel & Nan, 2019). The solution cannot be scaled to different geographical areas. Social media data cannot be used for information extraction.

The first two limitations of the hotspot mapping method are addressed by correlating feature space with that of data. This uses historical and spatial variables for predicting crime. The use of spatial information also helps in forecasting crimes in geographical areas without historical data.

Prediction using social media

Predictive modelling based on social media data is an area with solutions in the detection or prediction of earthquakes, disease outbreaks, commercial performance of films, etc. The spatial resolution of crime data is different in this prediction modelling. These research areas have a spatial resolution that encompasses an entire city with one prediction. However, there is a block to block difference in city crime. According to (Wang et al., 2004), tweets from news agencies used for predicting

crime. However, the data is available only for vehicle hit and run, home break-in, and enter crimes. However, the research did not have GPS information present in the tweets to provide geolocation information.

The research paper aims to use GPS information and news feed data to identify the clusters. Such details are categorized on the basis of the types of crimes such as rape, fraud, theft, vehicle theft, assault, sex offence, prostitution, shootout, kidnapping, homicide, murder, pickpocket, injury, terrorism, explosion bomb, etc. a number of statistical languages processing methods are used for extracting analysis from news feeds.

TECHNIQUES FOR CRIME PREDICTION

Big data from urban areas are typically noisy, large-scale, heterogeneous, and dynamic (Martina & Ralphs, 2014). It calls for effective and efficient computational solutions. Therefore, a number of computational solutions are proposed to improve crime research in urban areas. This segment will discuss important computational tasks with suitable algorithms for the prediction of urban criminal activity.

Crime Rate Prediction

Crime rate prediction predicts the future crime rate in a particular region (MBURU, 2017). In this subsection, we divide the crime prediction models based on data sources such as crime data, environmental context data, and social media data etc.

Prediction Based on Crime Data

Some research papers have proposed crime prediction nearly thirty days ahead in smaller areas such as police precincts. The techniques used by the police compared with that of univariate time series models for predictive accuracy. According to a fixed effect regression model with a 100% prediction error, the average number of offences must be greater than 30 to gain less than 20% prediction error (Eck &

Weisburd, 2018). It's also known that the most precise model for predicting precinct-level crime is Holt exponential smoothing. An autoregressive integrated moving average (ARIMA) used for predicting property crimes that may occur near the future. The ARIMA model makes use of 50 weeks of property crime data to predict crimes one week ahead. ARIMA model has higher prediction precision and fitting compared to exponential smoothing.

Some papers have experimented with a four order tensor predicting crime. The tensor encodes the latitude, time, longitude, and other information related to crime. The tensor then processes the data sparsity because the orders are in the lower dimension (McClendon & Meghanathan, 2015). Besides, the tensor maintains the geometric structure properly. The empirical discriminative tensor analysis algorithm leverages the tensor framework to reduce empirical risk and acquire adequate discriminative information at the same time. The new approach proposed by (Behrens & Robert-Nicoud., 2014) to selecting characteristics and buildings that use spatial and temporal patterns as predictive methods. This Spatiotemporal pattern has multi-dimensional features that can build on top of the regional crime cluster distribution at various levels. Then the cluster-confidence-rate-boosting framework is utilized to combine the global crime patterns into local Spatio-temporal patterns. This study has then deployed for crime prediction.

A point process model is used to analyze the temporal patterns of dynamics present in the violence for predicting urban crime. The crime rate divided into the sum of the self-exciting component and Poisson background rate in which the crime stimulates the growth in the process rate. Specifically, every crime data that is produced by the process will create a number of offspring crimes based on the Poisson distribution (Zhao & Tang, 2018). The background rate is usually assigned initially for crimes. The crimes predicted with the help of self-exciting point process models. They make use of a nonparametric evaluation strategy to get a clarity of the spatial-temporal

triggering function and the temporal tendencies in the burglary background rate. In particular, self-exciting effects and background intensity estimation can use for evaluating spatial heterogeneity.

Prediction Based on Environmental Context Data

Periodicity and the tendency of the crime rate had identified using a routine activity method for predicting crime (X. Zhao and J. Tang 2018). Specifically, it assumed that homes that far given opportunities to commit a crime. This study can yield a high value of crime rates. This assumption helps in understanding crime rate tendencies in the USA from 1947- 1974. This study also a consequence of various social changes, such as single-parent families and labour force involvement. The crime prediction made using seasonal crime patterns. The crime data evaluated on monthly, quarterly, and annual data exhibits substantial evidence that temperature is a significant factor in criminal activities. Seasonal variations play a role in the crime (Bowers & Newton, 2018). The main explanation is that during higher temperatures, individuals to spend more time away from home. This research is close to standard crime theories and to increase crime. The results show that temperature is the main reason to be considered for describing the quarter to quarter variation of urban crime.

Prediction Based on Social Media Data

Twitter posts contain both rich context and event-based information that can leverage for the prediction of criminal incidents. The framework has two components. The Spatio-temporal generalized additive model is the first component. It makes use of feature-based methods for predicting future crimes at a given time and location (S. Gerber, 2017). The second component gets textual information from the semantic role labelling based latent Dirichlet allocation. In addition to this, there is a new feature selection approach that helps in designing essential features. The proposal includes crime prediction through twitter information. This method contains an

intelligent semantic analysis of twitter posts and uses latent Dirichlet allocation for dimensionality reduction. Mathematical topic modelling and linguistic analysis are specific to twitter introduced to identify discussion topics on twitter related to urban areas (Corso, Leroy & Alsusdais, 2015). These topics used to create a crime prediction model. The researchers show that the use of twitter information increases the accuracy of crime prediction when compared with kernel density estimation. Various performance bottlenecks need to take into account for using twitter data in the decision support system. Forecasting crime tendencies is done using Twitter content. Historical data extracted using a twitter sampling approach for solving the missing data problem over a period of time (S. Marzan, C. Baculo & de Dios Bulos, 2017). The experiments showed the relationship between crime tendency and twitter content. Besides, there are some crime types such as burglary that have closer relationships with twitter content than others.

Crime Hotspot Detection

Crime hotspot mapping and detection is a spatial mapping technique that focuses on identifying the concentration of crime events in an urban area (Ristea, Kounadi, & Leitner, 2018). In this subsection discussues the classification of different crime hotspot detection methods based on the techniques used, i.e., (1) KDE – based techniques which is a non-parametric method that calculates the probability density function of the crime (2) Reaction-Diffusionon based techniques – a mathematical framework that is based on reaction-diffusion partial differential equations to understand the dynamic nature of crime hotspots and (3) other techniques including grid thematic mapping, geographic boundary thematic mapping hotspot optimization tools and spatial ellipses.

KDE based Techniques

Interpolation techniques, bandwidth, and grid cell size used to analyze the prospective crime hotspot detection methods. In particular, this works explains the science behind the KDE hotspot maps quality based on the variations in the user-defined settings that included in the interpolation. The analytical technique was used to evaluate the impact on different crime types such as robbery, assault, and burglary (Marzan & C. Baculo, 2018). The KDE hotspot with low resolution converted to a new map with contour lines. The output is a hotspot map that has smooth boundaries, has faster generation speed, and equally accurate to KDE hotspot maps with smaller cell sizes. The new maps are more representative of hotspots of the real world than the original KDE maps. There are many helpful suggestions for setting the KDE parameters such as search radius (bandwidth) and grid cell size. These proposed for hotspot detection tasks (Zhang & Shixiong, 2009).

Reaction-Diffusion based Techniques

The dynamics and formation of crime hotspots are studied using a computational framework with reaction-diffusion partial differential equations. This framework represents how criminals interact with victims and move in a given area. According to the analysis, crime hotspots due to recurring crimes diffuse locally and not over a period of time (Walther & Kaisser, 2017). The crime hotspots are detected using nonlinear analysis with the help of the reaction-diffusion system. The researchers use amplitude equations that modify the starting process of crime hotspots with the help of the perturbation method. Subcritical hotspots are identified that come from trans-critical or sub-critical bifurcations of the geometry. The rigorous detection of the hotspot is based on the reaction-diffusion based approach (Zhao & Tang, 2018). The presence of steady states is identified with multiple spikes with two distinct patterns. (1) Multiple spikes have the same amplitude and (2) Multiple spikes that have

different amplitudes. They make use of a strategy that is based on Liapunov-Schmidt reduction and convert it into a quasilinear crime hotspot model.

Other Techniques

The spatial distributions of urban crimes modelled using geographic boundary thematic mapping. This study requires very little knowledge of interpretation and can quickly generate hotspot maps. The government defines the boundary regions in an arbitrary way, such as in a police precinct. Grid thematic mapping method is used to process the situation of different shapes and sizes of regions such as police precincts (Catlett, Cesario, Talia & Vinci, 2019). The crimes in the hotspot map focus on these regions that shaded with corresponding crime numbers inside them. In this case, the grids with the same shape and size drawn on an urban map covering the study area. Therefore all the regions in the map will be comparable and have uniform dimensions. This study helps in easy and quick identification of crime hotspots. Hotspot detection software, such as spatial ellipses, are used to identify hotspots within a study area. It identifies the aggregation of hot clusters and then creates a "standard deviation ellipse" to each one of them. The crime clusters are ranked using the ellipses based on their properties and sizes. A hotspot optimization tool is utilized to increase the detection of hotspots by optimizing the boundary based on the spatial patterns of the crime (Cheng, Gui & Wu, 2019). The mix of values of different variables is indicated by means of a pattern that can distinguish normal regions and hotspots from the spatial perspective known as Geospatial Discriminative patterns. The proposed model identifies the GDpatterns and detects the crime hotspots automatically at the same time.

Base papers explanation

The relationship between urban environment, people dynamics, and crime studied with the help of urban activist Jane Jacobs. She underlines the importance of natural

surveillance, i.e., the presence of high diversity and visitor density is essential for ensuring the safety of an area and reduce crime. Criminologists investigate the crime concentrations at micro levels in geography. Data drove and place centric approaches are used to determine whether the geographic area can become a crime scene.

Another interesting area of research is the news content analysis. In this case, semi-structured information present in news articles can use to do data mining, sentiment analysis and predicting issues of the society, etc. interesting pattern analysis used in the news content analysis (Das & Das, 2019). NER (Named Entity Recognition) used for the extraction of location data and identification of a crime location. By combining event similarity, temporal proximity, and temporal relationships, the relationships between various events can be analyzed.

(Spengler et al., 2009) have created a web-based tool that can analyze necessary information from unstructured data about the method of operation of the various drug cartels. These methods have identified multiple insights about the Mexican drug cartel, methods for selling drugs and how they adapt, etc. Arulanandam, Savarimuthu, and Purvis have used the conditional random field to extract crime information from newspaper articles. NLP is used by other researchers to improve the efficiency of the system.

Research is done by (Bao et al., 2013) makes use of the Concentration Driven Model (CDM) for analyzing the crime locations. Centro based models used for macroscopic prediction of places of future crime. In this case, the keywords taken from the news body that represents the articles (Agarwal, Nagpal, and Sehgal, 2013).

2.6. Data sources for crime analysis

Newsfeeds: (Jayaweera et al., 2015) has researched by analyzing newspaper articles such as Daily Mirror, The Island, etc. A crawler collects the crime news. The SVM

classifier used for classification. Duplicates removed using data preprocessing. The results are plotted using the web GUI after post-processing. Research done by Piek Vossen uses a multilingual model for understanding information such as what, when, who, etc. from news articles. The results matched with other analytics platforms such as Yahoo Finance, Google trends, Google finance and Reuters (Morshed & Jayaraman, 2019). The main advantage is that present news can merge with historical news for getting a complete timeline of events.

(Roya Hassanian-Esfahani et al., 2016) has used different techniques for news retrieval, visualization, and analysis. (Khmael Rakm Rahem et al., 2014) have done investigations on drug crime. The information such as different types of drugs sold, the nationality of drug dealers, it's quality, the price at which it sold, the location of drugs, etc. are extracted from news articles. Shiju Sathyadevan has utilized used a Naive Bayes Classifier algorithm for identifying crime information related to murder, thievery, group assault, vandalism, burglary, sexual assault, etc. pattern identification made using Apriori calculation. Prediction made using a choice tree (Malathi, and Baboo, 2011). (Xifan Zheng et al., 2011) and created a retrieval system called (i-JEN) and analyzed Malaysian news. The results are collected using various classification and clustering algorithms. (Chung-Hsien Yu et al., 2011) analyzed 22 years of crime data from the New York Times. Techniques such as clustering and entropy used for predicting the future occurrence of mob attacks, sickness episodes, etc. visualizations then created from them.

Twitter: Matthew S. Gerber used Twitter data to predict 25 different types of crime based on twitter feeds coming from the Chicago area. He used official twitter streaming API and defined coordinates for getting location-specific tweets of users. Prediction models created using KDE (Kernel Density Estimation), and the results are shown in visualizations. The source code released as open-source as an "asymmetric threat tracker." (Anthony J. Corso et al., 2015) have used NLP techniques. They have

created a predictive GIS artefact that consumes noisy knowledge from social media, preprocessed government knowledge, and subject knowledge. The correlation between domain-specific activities and social media news proven with this research (Oliveira & Menezes, 2019).

(Maximilian Walther and Michael Kaiser 2013) have researched geospatial event detection for extracting location data from tweets. The real-world events identified using machine learning. Statistical methods are used by (Tony H. Grubesic 2006) for identifying robbery, burglary and assault crime patterns using real-time Twitter data. The spatial trajectories of social media users are investigated by (M. Wang and M. S. Gerber 2015). They have created a correlation matrix for crime occurrence in the USA. This data is taken from twitter feeds. The geoparsing analysis is used by (Nikhil Dhavase and M. Bagade 2016) to identify geospatial data using newsfeeds automatically. Location specification such as street address and building names used to determine the event location. Situational and behavioral data are used by (T. Mantoro et al., 2014) for correlating crime with other information. Automated linguistics analysis and spatiality reduction, i.e. linear modelling prediction and Dirichlet, are used by (Wang X, Gerber M.S et al., 2015) for creating a criminal incident prediction system based on twitter information. The system is tested on the prediction of future crimes and found to be performing better than the baseline model used in the same offence. (Vukosi Marivate and Nyalleng Moorosi 2015) analyzed concerns related to privacy while mining social media data. The research identifies the current gaps in the regulations related to privacy protection challenges, possession of private data and legal protection of non-public information (Wang & E. Brown, 2015). The moral problems associated with new information creation also discussed.

Instagram: (Ke Xie et al., 2014) performed a time series analysis with the location tag data in Instagram photos. They can identify irregular signals in specific areas. The actual location identified using a classifier. A total of 905746 photos are collected

from Instagram using geographical boundaries. The candidate event classification for time series and spatial data prediction done using Gaussian process regression.

Mobile data: (Bruno Lepri and Andrey Bogomolov 2015) made use of anonymized mobile data with demographic information for predicting crime occurrences in London. 63.57% accuracy was achieved using the Random forest algorithm applied to demographic and mobile data. The efficiency has increased to 69.56% while adding census data. 70% accuracy is available for prediction of crime hotspots if human behavioural data also added. Six mobile applications have studied 1. Community Alert. 2. Crime Watch Mobile 3. Community against Crime 4. Malaysia Crime 5. MyDistress 6. Enforce Crime Map. The following parameters used for evaluating the apps: map viewing, crime list, reports of authority, sharing of incidents, tutorial, sharing data in media and system support. These apps can use for alerting the community about crime.

Teddy Mantoro develops the crime assistance application for helping victims and police. The victim can use the app to send information such as location, time with the use of Google API. The alert is sent to the nearest police station automatically. The mobile app integrates Google map API to identify crime locations. An upgraded version with location-aware capabilities expected for public release (Ahmad, Syal & Tinna, 2019).

Police data: (Christoffer Gahlin and Erik Johanson 2015) analyzed the performance and accuracy of the Kernel Density Estimation algorithm with small datasets. They also analyzed the data required for developing a reliable hotspot. They have selected three geographic locations, Stockholm, Sweden, and Gothenburg, and studied them for over a year.

(Shoaib Khalid et al., 2015) used location and time analysis on crime data taken from the Faisalabad police department in Pakistan. The research conducted in four steps.

The first step is Data collection, Graphical positioning survey, Zone separation, and digitization of data and final interview of police officials. The second step is crime report geocoding, locating on the map, density analysis of crime and network. In step three, a COMSTAT model called CRIMEGEOGRFIX developed (B.M. & D, 2019). The final step is Operational Analysis, observation of Strategic Plans, Revised Duty Rosters and New locations of Police Checkpoints. The following crimes identified from model 1. Theft of bike and car in parking places 2. Burglary in posh city areas 3. Chain snatching in areas that lack street lights.

(Raju K. Gopal and Arunima S. Kumar 2015) analyze the data mining systems used in crime investigations from the Indian perspective. The crime analysis has three frameworks 1. Regional crime analysis program 2. 1. Data mining framework for crime pattern identification. 3. Tucson police department's Narcotics network in the Indian initiatives. Based on this framework, NCRB (National Crime Records Bureau) has created the following systems 1. Crime and Criminal Tracking Network System (CCTNS), 2. Talash Information System (TIS) and 3. Common Integrated Police Application (CIPA).

(Ehsan Khoramshahi and Somayeh Nezami 2016) have analyzed crime data of drugs from the country Iran. Crime distribution is studied based on various variables. The distance between the town and the closest police station modelled using crime distribution (Baumbach, Sharm, Ahmed & Dengel, 2019). The police knowledge of southern Khorasan province is analyzed using Geographically Weighted Regression algorithm.

(Mohammad A. Tayebi et al., 2014) used spatial crime analysis for studying crime hotspots. They have used a probabilistic model for spatial analysis called Crime Tracker. (Omowunmi E. Isafiade and Antoine B. Bagula 2013) used data in the African Republic nation for analyzing the crime. The algorithm used is City safe algorithmic rule and mistreatment FP-Growth algorithmic rule. Crime hotspots are

identified with the help of visual imaging techniques. This has helped public safety organizations and law enforcement agencies to implement effective interventions (Ardis Hanson, 2014).

(Jianping Wu, Zhanhong Wang, Bailang Yu 2011) used data from Shangai town for developing hotspot maps. There are two different types of crimes identified. PCA (Principal Component Analysis) used for identifying the 18 crime distribution indicators. (M. Shiblee, S. B. Changalasetty, O. A. Khalid, M. Alalyan, L. S. Thota and A. F. Fathima 2016) used the National Crime Records Bureau (NCRB) dataset to Indian administration. It also used K-Means clustering to analyze the data. (Veena Karjigi and H.P. Kavya 2014) The author used keyword data coming from telephonic calls stored by security companies. Total data collected from 39 phones.

Hardware-related: (Chaolun Xia et al., 2014) used social media data using a tool called CityBeat. It has camera data along with geotagging for understanding the time and area of the crime. The information characterized using the SVM algorithm.

Dataset: (Amit Kumar Manjhvar and Nidhi Tomar 2106) made use of K-means clustering to identify new crime groups in the data. They have used ruff information collection for testing.

(Guiyun Zhou et al., 2015) filtered crime and non-crime locations using web applications. Czech Republic data also used for this.

(H. Jiří, I. Igor 2016) makes use of OLAP for implantation of a multidimensional information base. The author has analyzed wellbeing, joblessness, populace, and properties, etc. It analyzed with time and space measurements.

(J. Azeez et al. 2015) extracted crime designs in various databases using algorithms. The algorithms have good quality. A spatial crime tree with H-table used for analysis. The data used as Ruff dataset.

(Xifan Zheng, Yang Cao, and Zhiyu 2011) Ma developed a model that predicts serial crime using location, geographic profiling, and time information. The results identify the likelihood density of prediction. Alice Hutchings used models for analyzing criminal purpose website takedowns such as achievement of money mules, faux websites, phishing websites, counterfeit and illicit product sales, child offence content and malware dissemination. Detailed discussions were done with people who have done website takedowns (Bunch, Murray, Gao & Hunt, 2019). Others include companies affected by criminals, businesses that offer specialized services, UK social control, and repair, UN answer takedown requests, and suppliers.

(Ubon Thongsatapornwatana 2016) using data mining methods such as classification, association, and clustering algorithms for gaining insights about crime data from websites, databases, and sensors.

Organizations: (Qiang Zhang and Pingmei Yuan 2016) utilized LDA-KNN for hotspot mapping and fleeting spatial qualities. Dimensionality diminution and KNN are used as the investigation methods. The National Natural Science establishment of China has supported this work. Data from Nanchang city is used for identifying cases such as robbery, burglary, etc. Important festivals such as classification techniques such as KNN and dimensionality diminishment strategy are used as part of the Straight Discriminant investigation. This work is done with the support from the National Natural Science establishment of China. The research has taken 2014-15 data from the Nanchang city zone for identifying crime occurrences such as burglary, robbery, etc. Important occasions such as Valentine's Day, Chinese Spring Festival, Lantern Festival, May Day, Tomb Sweeping Festival, Double Seventh Festival, National Day, Mid-Autumn Festival, New Year's Day and Christmas are considered for crime prediction (Catlett, Cesario, Talia & Vinci, 2018).

(Mehmet Sait Vural et al. 2013) suggests that a realistic data set is not used in unattended approaches to crime analysis. The model uses GIS for calculating

population characteristics in real life. The results are combined into a GIS map for visualization.

Email:(Mugdha Sharma 2014) has created an improved Decision Tree algorithm for identifying suspicious messages application called “Z-Crime” is used for identifying suspicious criminal activities. It also proposes a prevention plan. The main information sources are emails.

Telephone Phone Call information: (M. Hanumanthappa, T. V. S. Kumar, and M. Kumar 2015) have implemented a call detail record (CDR) for identifying criminal suspects. The CDR database is analyzed with visualization techniques for creating insights.

Software’s related to crime analysis and prediction: Cortana solutions have created My Neighbourhood for getting information about what is happening and who is nearby etc. The application used by law enforcement agencies. The user can know what is happening in the neighbourhood. This is sponsored by the local police agencies. (Daya Sagar, Krishna, Kumar, Teja & Babu, 2019).

Social Media and Data Mining, Community-oriented Online Social Media Observatory (COSMOS), a universe is a software that reduces methodical hindrances in the investigation of online networking. It is developed by the Economic and Social Research Council (ESRC) using Big Data elements such as social, political, PC, factual, wellbeing, and scientific researchers to identify hypothetical, methodological, specialized, and observational measurements of social networking sites for analyzing social situations (Wang B, Dong H., et al.).

NewsReader stores volumes of information from online news. The project used a large amount of information to get millions of field archives. These include life stories and friend’s databases (Fast et al., 2019). The framework concentrates information about who has done what, on which occasion by utilizing labels of the

news feed and storing in an organized database. It can handle diverse dialects such as Italian, Spanish, English, and Dutch.

Malaysia Crime is a website that gives crime reports in specific areas in Malaysia. Visitors can identify the crime density in a particular area. The project is open-source and makes use of twitter data. Open Engines have 2500 clients all over the world that give crime reports for specific places on the map.

Table 2.1 Summary on Crime analysis and prediction Data Sources

Source of data	Performed researchers
News feeds	Jayaweera et al.,(2015),Piek Vossen et al.,(2014) ,Richard Fletcher et al.,(2018),Roya Hassanian-Esfahani et al., (2016),Khmael Rakm Rahem et al., (2014) ,Shiju Sathyadevan et al., (2014) ,Xifan Zheng et al., (2011),Chung-Hsien Yu et al.,(2011), Richard Fletcher et al., (2018)
Twitter	Matthew S. Gerber et.al., (2014),Leitner M et al., (2018),Michael Kaiser et al., (2013),Grubesic et al., (2008),Wang et al., (2014) ,Bagade et al., (2016),T. Mantoro et al., (2014),Xiaofeng Wang et al., (2015) ,Nyalleng Moorosi et al., (2015), Xiangyu Zhao et al., (2018).
Mobile	Andrey Bogomolov et al., (2015),Izyana Ariffin et al., (2014),Teddy Mantoro et al., (2014),
Instagram	Ke Xie et al., (2013),Izyana Ariffin et al., (2014)
Police Record	Christoffer Gahlin et al., (2015),Shoaib Khalid et al., (2015),Arunima S. Kumar et al., (2015),Somayeh Nezami et al., (2016),Mohammad A. Tayebi et al.,(2014) ,Zhanhong Wang et al., (2012), Y. Bédard et al., (1999), Fotheringham A S(2002).
Ruff datasets	Nidhi Tomar (2016),Guiyun Zhou (2012),Priyanka Das et al.,(2019),H. Jiří, I. Igor (2016),Zhiyu Ma (2012),Alice Hutchings et al.,(2016), Thongsatapornwatana (2016),Rahu Garg et al.,(2018), Babakura(2014),

	Mehmet Sait Vural (2013), Ahishakiye et al(2017), Zahra, S. A. (2018).
Organizations	Qiang Zhang et al.,(2016)
Emails	Mugdha Sharma et al., (2014)
Telephone call data	M. Kumar et al.,(2015)

Table 2.2 Summary of Algorithms used for Crime ananlysis and prediction.

Source of data	Used Algorithms
News feeds	SVM ,Naive Bayes Classifier, Entropy Clustering.
Twitter	KDE, Risk Terrain Modeling, Naïve Bayes ,Multilayer perception, Pruned C4.5, decision tree, ZeroR, K-means, GWR,ARIMA.
Mobile	Random forest algorithm
Instagram	SVM, Gaussian Processes Regression(Time series prediction).
Police Record	KDE, Geographically Weighted Regression, Naive Bayes Classifier, Decision Tree.
Ruff datasets	Ant Colony Optimization (ACO) ,algorithm to improve the K-Means algorithm , Decision Tree, KDE, KNN(K Nearest Neighbor)
Organizations	LDA((Linear Discriminant Analysis)-KNN(K Nearest Neighbor)
Emails	Decision Tree Algorithm
Telephone call data	Naive Bayes Classifier

Table 2.1 Show the Researchers involved in the Crime analysis with respect to different social meida and other data sources of crime data investigations.Table 2.2 shows the Diffent algorithms used for crime analysis and prediction.

CRITICAL REVIEW AND GAP ANALYSIS BASED ON LITERATURE

A study on crime patterns shows that the increase in crime relates to slow down in economic activity at all levels, both national and regional. In the early part of the 20th century, criminologists have focused on the relationship between criminals and the sociological factors of the neighbourhood. These factors include poverty index, exposure to specific peer groups, characteristics of the neighbourhood, etc. (Ahishakiye, Omulo & Taremwa, 2017). Studies have devoted attention to the relationship between criminal behaviour dynamics and the place-centric perspective. Research by (Wang et al., 2004) has implemented a people-centric approach called Series Finder, a machine learning algorithm for extracting patterns from crime data using gender profiling. The empirical knowledge of criminal behaviour is combined with place and time information to identify patterns.

Studies from Ratcliffe have explored the constraints of spatiotemporal constraints in crime. Quantitative tools from physics, mathematics and signal processing are needed to analyze spatial and temporal patterns in crime data, and it is necessary to have quantitative tools from physics, mathematics, and signal processing. (Toole et al., 2011) have identified the presence of complex multiscale relationships of crime data with both space and time. (Buczak and Gifford 2010) derived a finite set of rules with the help of fuzzy association rule mining on demographic information data. (Eck et al.,1995) have extracted useful insights using kernel density estimation. (Mohler et al.,2011) have used a self-exciting point process model for modelling crime data. However, the main problem with this method is that they cannot use in areas for which no data is available since they rely on the historical information of the crime (Garg, Malik & Raj, 2018).

There are also various types of clustering methods used in data analysis, such as hierarchical, agglomerative, divisive, constraint-based clustering. The number of wanted clusters is updated and optimized in the K-means algorithm. The hierarchical clustering method seeks to segment similar data by using techniques such as Euclidean distance, Hamming distance, etc. Expectation maximization is a popular method used to segment the incomplete data into clusters based on probability. To improve the accuracy of these clustering algorithms, a segmented multiple metric similarity measures (SMMSM) proposed to identify suspected crime.

However, there are issues such as incremental updating mined patterns (Garg, Malik & Raj, 2014). To solve this, there are various nature-inspired algorithms like the Genetic algorithm, Bees algorithm, Gravitational search algorithm. There has been considerable research in pattern identification through textual content with a particular focus on news content.

CHAPTER SUMMARY

In this section discusses about some insights gained from the data and future directions.

More Crime Patterns

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. Deep learning is a good algorithm for capturing Spatio-temporal patterns and predicting future outcomes. It has been used in a number of applications, such as air quality prediction and traffic flow forecasting. The promising application is to have novel models in deep learning to extract features from the complex Spatio-temporal patterns for improving the analysis of urban crimes (Chen, Chau, Qin & Chung, 2014).

More Advanced Techniques

Urban crime is hard to understand the phenomenon with a dynamic interplay between time, space, and other factors such as urban configuration, Environment, and economics. Crime analysis is done with the help of advanced techniques. For example, urban crime with dynamic nature is better modeled with the help of deep reinforcement learning models. This can continuously update its algorithm with every interaction with the environment. There are various sources such as point of interest (PoI), meteorological information, and human mobility data that can impact urban crime. The various sources are analyzed linearly (M. Leigh, J. Dunnett & M. Jackson, 2016). However, it fails to map the nonlinear subordinations and connections equally. Therefore advanced features are used to incorporate features from the different sources of analysis. In addition to this, there should be an automatic way of feature extraction from different resources. Handcrafted features are not sufficient to understand the complex spatiotemporal patterns. This requires the use of end to end frameworks that combine computational tasks and feature extraction for urban crimes.

More Computational Tasks

There are various developments in big data technologies that can aid in urban crime analysis. It provides a unique and unprecedented opportunity to design sophisticated models for tackling the practical policing tasks in the real world. For example, New York City has a stop-question and frisk program for the prevention of crime. It detains temporarily, questions, and searches citizens for contraband and weapons (Nasridinov, Ihm & Park, 2019). But this policy has generated controversy of racism and has failed to reduce crime such as robbery and burglary. Therefore there needs to be concentrated efforts for preventing crime that minimize the rights infringements of citizens and increase deterrent value at the same time. Also, a number of urban tasks need to be jointly planned for a smarter and safer city such as health, education,

economic development, urban planning, policy, employment, immigration, justice, integration, and poverty, etc.

Urban Simulation

Strategies of policing need to be re-evaluated before they are implemented. This will save deployment costs and prevent any negative impacts from occurring on urban safety (Nasridinov, Ihm & Park, 2019). Therefore an urban environment simulation is required for the visualization and offline evaluation of the new patrolling strategies. Consequently, the use of urban environment simulation can be used by the police department and researchers to gain insights, perform investigations and create policing strategy for boosting the communities and integrating the relationships between economy, crimes, and urban configurations.

Some part of this chapter are published in following journals and conferences:

1. Boppuru Rudra Prathap, Ramesha K, "A Pragmatic Study on Heuristic Algorithms for Prediction and Analysis of Crime Using Social Media Data", ***Journal of Advanced Research in Dynamical and Control Systems***,Volume 10,Issue no 2,pp-30-36 April 2019.(**Scopus Indexed Journal**).
2. Boppuru Rudra Prathap**,** Ramesha K, "An Overview of Heuristic Based Crime Prediction and AnalysisUsing Social Media Data", ***Journal of Emerging Technologies and Innovative Research***, Volume 5,Issue no 12,pp-582-590 Dec-2018.(**UGC Journal No:63975**).
3. Boppuru Rudra Prathap and Ramesha K, "A Literature Survey on Heuristic Algorithms to Predict Crime Using Social Media Data", ***International conference on Sustainable Advanced Computing (ICSAC)*** held at CHRIST (Deemed to be University) Mar 2nd to 3rd 2018, Bengaluru, Karnataka, India.

CHAPTER 3

SPATIO-TEMPORAL CRIME ANALYSISUSING NAÏVE BAYES AND K-MEANS CLUSTERING

INTRODUCTION

India is a rapidly urbanizing country in the world. United Nations predicted that about 86% of the developed world and 68% of the developing world would be urbanized by 2050. This statistic implies the entire urban populace, later on, will be more than that of the present total populace. A large number of rural people are migrating to city centres. This statistics has resulted in inequality between urbanites (McClendon & Meghanathan, 2015). The wealth gap is a brooding ground for rising crime. For example, the crime rates have increased from 2300 to 3000 for every 12000 residents according to data from 1980 to 2000. Analysts have demonstrated that there is an intimate connection between the practical improvement of urban areas and the personal satisfaction of urban residents. The primary psychological and physical need for urban dwellers is safety. For a city to have sustainable development, there is a need for urban crime prevention methods that are well planned, community-based, gender-sensitive and have extensive city coverage (Perera, Sajeewa & Wijewardane, 2019).

Conventional demographic data such as a socio-economic profile of a population, including income level, ethnicity, education, religion, and wealth gaps have been using in traditional urban crime research. However, it has discovered that the statistic information isn't adequate to comprehend the unpredictability and elements of urban crime. The reason is that the demographic features of a population do not change over time. It does not capture the dynamics that occur in a particular community (Walther & Kaisser, 2013). Also, the demographic features of different cities are similar,

making it challenging to analyze the differences in various communities.

The Internet has turned into the essential vehicle for expending and moving data on a vast scale. The massive development of new technologies has enabled fine-grained data collection, and this had empowered the chronicle of urban crime data. It has created a wide variety of information that can use for many analytics purposes.Newsfeed data contains spatiotemporal information apart from the information about the incidents. It also has context information about urban crime. This spatiotemporal information can come from social media, RFID, GPS, mobile phones, and smart meters (Wang & Brown, 2012). This research is looking for new ways to better understand criminal behaviour using data. Urban areas require effective safety measures to ensure that the security of the population is guaranteed.

Data such as human mobility in a particular area can help in identifying residential and regional stability. Research from environmental criminology states that this can affect criminal activities in a specific region (Reddy, Saini & Mahajan, 2019). There is also a correlation between urban crimes and weather patterns. This news feed data has fine-grained and rich context information about where and when a crime has occurred. With this data, it is conceivable to comprehend the development of crime and the investigation of crime in alternate points of view (Zhao and Tang, 2018). Therefore this research convinced that urban news feed data has unprecedented opportunities to predict urban crime and improve the safety of the citizens.

This research was mainly focusing on India and Bangalore crime data. We got the motivation after reading the literature review about the various types of social media data that have used for crime prediction. Then we decided to consider the newsfeed data. We have narrowed down to Bangalore, Karnataka, as it is one of the most populous metro cities in India. It is furthermore a standout amongst the other ten urban networks with high incidents of crime. To identify the crime rates, we have used the government website http://ksp.gov.in/. Based on the motivation, we decided

which area to consider, content to take, sources, etc.

This thesis uses Naïve Bayes theory and K-means clustering for solving crime prediction problem. The crime prediction problem has defined as finding the most likely location prone to crime. This method makes use of historical data of crime in a particular location and time. The incident level crime data from newsfeeds had given 9+as the crime dataset in which the location, incident date, criminal ID, crime type extracted. The Spatio-temporal analysis is done on the crime data to identify the prospective crime occurrences.

Data Collection From News Feeds

RSS news feed data are rich in both location and context for prediction of crime incidents (Behrens & Robert-Nicoud, 2014). This method has two steps. The first step has a Spatio-temporal model that uses feature-based extractions for predicting future crime rates in a particular location. The second step involves the extraction of textual information through semantic role labelling. The essential features in the news feed extracted using linguistic analysis and mathematical topic analysis. The addition of news feed data to traditional crime data sources increases crime prediction accuracy. This method can be extended to form a decision support system. A sampling approach used to handle the missing data over time. Some crime types seem to have a close relationship with the internet and social media data (Sun & Du, 2019). Data collected with the help of scrapping algorithms from news feeds available on the Internet. The Hindu is the primary source of news feed data. The data available in various formats such as RSS, RDF, and Atom have converted into XML for processing. The textual analysis was done to extract the crime-related information such as crime location and time.

Text Pre-Processing

Text pre-processing is the method in which the news feed information converted into a machine-readable format for further processing. Machine Learning for Language Toolkit (MALLET) used for extracting features from text and classify them into various clusters. It is a sophisticated natural language processing tool for topic analysis, document classification, information extraction, and machine learning. It consists of multiple algorithms that remain useful in doing the classification. The text is then classified based on features. Since news feeds have natural language data that has messy information (Tunley, Button, Shepherd & Blackbourn, 2019). It is full of noisy data like spelling mistakes, dis-influences or unexpected foreign text, etc. It also faces structural problems when it comes to an understanding the meaning. There are two challenges – content less material and inconsistency of form. These two challenges addressed with the help of stop word removal and lemmatization. These solutions should be specific to a problem and an application. Text pre-processing having the following steps.

Lemmatization

Most word content in the English language can take a number of forms. Sometimes they have grammatical context, and sometimes they are meaningful. If there is a change in form with the same meaning, then it is said to be related derivationally from one form to another. A change in the form that correlates with the grammatical context is called inflected. Adjectives, nouns, and verbs are inflected in English.

Lemmatization is used to standardize the derivationally related forms to the base form and inflectional alternates (VK, 2019). Stemming is the method of slicing or altering characters with the suffix. This renders every form with a derived state that is unambiguous and non-inflected. An example of lemmatization is shown in the following Table 3.1.

Table 3.1 Lemmatization Example

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

Stemming

The most popular method of stemming is to use Porter Stemmer. The stemmer reduces the irrelevant punctuation into a consistent form. The Porter Stemmer retains some forms. It relies on a consistent form of the base and matches the normal suffix patterns. An example of Stemming is shown in the following Table 3.2.

Table 3.2 Stemming Example

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Replacement

Lemmatization techniques have different shortcomings that may require the use of techniques such as replacement techniques. This technique requires a curated list of English words and the relationship between them. The common resource is Wordnet. Wordnet can be imported directly and worked on. There will be a wrapper in the nltk. The replacement method produces output that can be interpreted easily (Yadav, Timbadia, Yadav, Vishwakarma & Yadav, 2017). It handles irregular forms and takes the words that exist in the database as the reference. However, in practice, out of vocabulary items or neologisms can occur in a large body of the text. Wordnet is limited in a number of ways in returning a usable output.

When to lemmatize?

Lemmatization is used in various steps, but it is not appropriate for all applications. Some applications, such as Topic modelling, benefit from lemmatization. Topic modelling is based on content words distribution. The identification of these words depends on a word string match. This is achieved by form lemmatization so that the variants occur consistently across documents (Malec & Schienle, 2013). LDA and TD-IDF are the beneficiaries of lemmatization. This method is also important in training word vectors since the word counts would be disrupted by irrelevant inflexion such as present tense inflexion or simple plural inflexion.

The general rule for lemmatizing is as follows: if the performance is not improved with lemmatization, then it is not used. This is a conservative approach and is the default method if there is no significant gain in performance. For example, VADER, a popular sentiment analysis method, has various ratings that vary with the word form. Therefore the input should not be lemmatized or stemmed. Applying lemmatization regardless of the problem, will not give significant results. In some cases, it can impede the success of the model.

Stop words

Contentless material, for, eg. ‘The’ and ‘of’ are present in natural language. These are called stop words, and they are essential because they are used for grammatical relationships of content materials. Most NLP libraries have a list of stop words, but there is no proper definition of stop words. The common ways of identifying stop words include selecting the most frequent words in the body of the text. Many NLP applications eliminate the stop words as they leverage the statistical profile of the input for success. Contentless and irrelevant words that occur frequently are considered as noise. This is true in topic modelling. But in some cases, removing stop words are problematic (Ahmad, Uddin & Goparaju, 2018).

Table 3.3 Stop words example

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

any resources and packages provide lists of stop words or methodologies for removing them. The method for eliminating stop words is simple. After processing, the sentence lacks all the instances of 'a,' 'is,' 'against,' 'the' and 'of.' Example for Stops word identification removing shows in the following table.

CRIMES MAPPED

In this Research identified the different characteristics of crime data such as Types of crimes (68 Crime keywords-6 Subclasses), Newsfeeds, and frequency of crime, Geographic Locations, and Temporal facts. We focus on utilizing the newsfeed data effectively to predict crime.

Table 3.4 List of Crime Keywords Considered for Research

Category of crime	**Crime Keyword**
Drug-Related Crimes	Drug Trafficking, Drug dealing, Drugs smuggling, Narcotics, drugs, and alcohol.
Violent Crimes	Rape, Murder, Terrorism, Kidnapping, Assault, Sexual Harassment, Sexual assault, Homicide, Gunshot, Intentional Killing peoples, Shootout, Gang-rape, Attempt to murder, Sexual abuse, Putting to death.

Commercial Crimes	Official Document Forgery, Currency Forgery, Official Seal Forgery, Official Stamp Forgery, Bribery, Counterfeiting, Cheating
Property Crimes	Arson, Motor vehicle theft, Theft, Burglary, Robbery, Riots, Criminal breach of trust, Stealing, Barrage fire, Bombardment, Electric battery, Shelling, Looting,
	Embezzlement, Trespass, Incendiarism, Shoplifting, Vandalism
Traffic Offences	Speeding, Signal Jump, Running a Red Light, drunk, and drive.
Other Offences	Prostitution, Illegal Gambling, Adultery, Homosexuality, Weapons violation, Offense involving children, Public peace violation, Stalking, Cheating, Hurt, Counterfeiting, Dowry deaths, Outrage her modesty, Causing death by negligence, Suicide, Criminal damage.

Classification of Crimes Using Naïve Bayes Algorithm

Naïve Bayes algorithm is a classification algorithm for multi-class classification problems and two-class (binary). The technique is easier to implement in the case of binary or categorical input values. The technique is called Naïve Bayes because the probability calculation is simplified to make the evaluation tractable. The algorithm does not calculate the value of each attribute individually. The algorithm is conditionally independent, given the target value and calculated as P(d1|h) * P(d2|H) and so on. The assumption here is that the attributes do not interact, which is very unlikely in real data. However, this approach works well on data in which this

assumption is not applicable (Gupta, Sabitha, Choudhury & Bansal, 2018). This research proposed model used Machine Learning for Language Toolkit (MALLET) for extracting features from text and classify them into various clusters. MALLET consists of the Naïve Bayes model, which classified the 68 different types of Crime keywords into 6 main classes.

Representations in Naïve Bayes Models

The Naïve Bayes is represented by means of probabilities. The model is built using a list of probabilities. These include Conditional probabilities, the conditional probabilities of each input value is calculated based on the class value and Class probabilities - the probabilities of each class in the training dataset.

Learning Naïve Bayes model from data

The method for learning a Naïve Bayes model from the training data is fast. This is because the method requires the calculation of only the probability of each class and that of given different input (x) values (Lo, 2018). There is no need for coefficients in the optimization procedures.

Calculation of class probabilities

The class probabilities are the frequency of instances that are available with each class partitioned by the all outnumber of cases. For instance, in a parallel grouping, the likelihood of a case having a place with class 1 would be determined as:

$$P(class=1) = count(class=1)/(count(class=0) + count(class=1))$$

In the easiest case, each class would have a likelihood of 0.5 or half for a twofold grouping issue with a similar number of occasions in each class.

Figuring of contingent probabilities

The contingent probabilities are the recurrence of each characteristic incentive for a given class worth separated by the recurrence of cases with that class esteem.

Mathematical model:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (3.1)$$

Crime hotspot identification using the K-Means algorithm

K-means algorithm is one of the least complex solo learning calculations that aides in taking care of the clustering issues. The procedure has an easy and simple way for the classification of a given dataset into a number of clusters (approx.. k clusters) fixed apriori. The objective is to define k centres for each of the clusters (R. Hipp, Bates, Lichman & Smyth, 2018). These centers should be placed in such a way that different location gives different results. It should be balanced so that they are far away from one another. The following stage is to take each point in the given dataset and associate it to the nearest centre. At the point when there is no pending point, the primary stage is finished. At this point, the k-means centroids are re-calculated as the barycenter of the clusters coming from the previous step. Once we have the k-new centroids, there has to be a new binding between the nearest new centre and those same data points. In this way, a loop is generated. As a result of this loop, the k centres change their location step by step until there are no changes done, or in other words, the centres do not change their position. Finally, this algorithm aims at minimizing an objective function known as the squared error function given by:

$$c(v) = \sum_{i=1}^{d} \sum_{j=1}^{i} \left(\left\| x_i - v_j \right\| \right)^2 \quad (3.2)$$

Where $'||x_i - v_j||'$ is the Euclidean distance between x_i and v_j.

'd_i' is the number of data points in i^{th} cluster.

'd' is the number of cluster centres.

PROPOSED METHODOLOGY FRAMEWORK

The proposed methodology shows seven stages of work as shown in figure 3.1. The first two levels show the architecture for collecting crime information from RSS feeds of periods 2017 Jan to 2018 Jan (27782 News related to crime).

65 types of crime-related keywords are classified and showed in 6 subcategories of classes. RSS, known as Rich Site summary, is a website format that enables users to get up-to-date information about the website in a standard way. The third level is called the standard Preprocessing algorithm that is used for cleaning data and changed into the data frame. The preprocessing is done using Lemmatizing, Stop words, Stemming and POS (Remove noisy data and cleaning process) using tools such as R language toolkits like diplyr & plyr. The processed data (.rdat) is stored using a standard database. A query language is used for retrieving data. The output is stored in .csv format.

The fourth level enables data to be cleaned and preprocessed and converted into hash code. The details and headline of the parent agency are converted to hash code for reducing data redundancy. The fifth level is preprocessing that has all information such as crime occurrences and location. The process repeats until enough information is available. Then 16 types of crimes such as Drunkenness, Assault, Burglary, Fraud, Arson, Hurt, Molested, Gambling, Suicide, Robbery, Kill, Murder, Vandalism, Harassment, Warrants, and Trespass.

The algorithm has the synonyms of the 16 types of crime with standard dictionary names. It uses the stemming and lemmatization process, which enables identification of familiar dictionary words such as the ARIMA time series model and Kernel density estimation. The seventh level shows crime visualization based on place and crime type. The detailed Proposed framework consists of 3 Phases which explained in figure 3.1 and Section 3.3.1 to 3.3.3

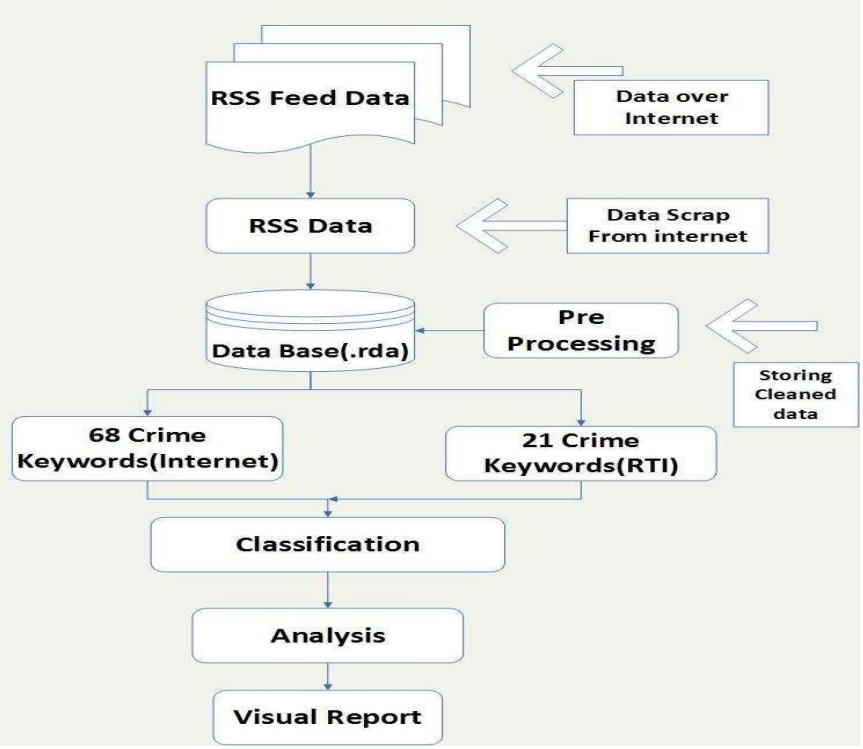


Figure 3.1 Proposed Framework

DATA MINING, CLEANING, AND EXPLORATORY DATA ANALYTICS

Exploratory data analysis (EDA) is a technique in statistics that is used for the data set analysis to summarize the patterns and main characteristics with the help of visual methods. EDA can be accomplished with or without the help of a statistical model. It helps the crime analyst to see what the data can tell beyond the hypothesis testing and formal modelling process. This method encourages the statisticians to explore the data in various ways and possibly formulate the hypothesis that can lead to new experiments and data collection. EDA is different from that of IDA (Initial Data Analysis). In IDA, the assumptions are tested using hypothesis and model fitting, transforming the variables and handling missing values as required. IDA is a subset of exploratory data analysis.

In this first layer of data analytics, the RSS feeds are mined from different sources, and duplicate feeds are removed with the help of hash code. The feeds are then stored

in a database. An algorithm is used to remove the duplicates, and new information is merged with the old data. Using a scraping algorithm, textual information related to crime type and location are extracted and stored in a database. An XML parser is used to retrieve the crime-related text from the news feeds that are in XML format. The preprocessing of the news feeds is done using tools such as POS, Lemmatizing, Stop words and stemming. The noisy data is removed and cleaned using tools like plyr and diplyr. The output data is then stored in XML format in a local database.

Data analysis is defined as the procedure for analyzing the data with the help of techniques to interpret the results of the procedures. This makes the analysis more precise, easy, and more accurate. Mathematical analysis is done with the help of statistical methods. There are several statistical computing packages that are developed to speed up the analysis. There are specific programming languages such as R and S-Plus that have high-quality dynamic visualization capabilities. This allows the statisticians to identify trends, outliers and patterns in data that merit further study.

There are two other concepts in statistical theory: robust statistics and nonparametric statistics. Both methods reduce the sensitivity of the statistical inferences to errors in the formulation of statistical models. These methods make use of a five-number summary of numerical data, i.e., the median, the extremes (minimum and maximum), and the quartiles. These are functions of the empirical distribution, unlike standard deviation and mean. The median and quartiles are more robust to heavy-tailed or skewed distributions compared to traditional summaries (standard deviation and mean). The software packages use resampling statistics, which are robust and nonparametric. Robust statistics, exploratory data analysis, nonparametric statistics and the use of statistical programming languages facilitate the statisticians' work on engineering and scientific problems.

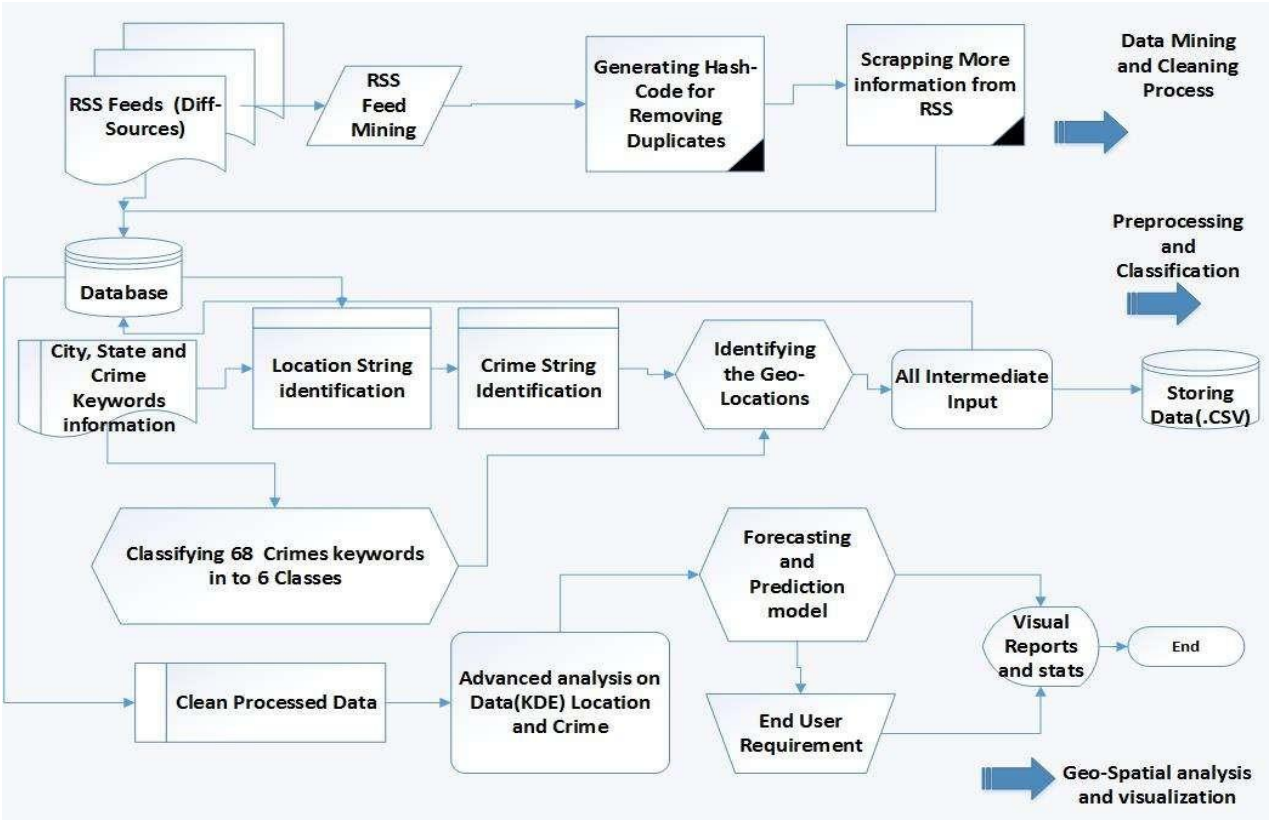


Figure 3.2 Detailed Proposed Framework

Preprocessing and Classification

The data pipeline begins with the collection of data and end with the communication of results. The process is a difficult one. There are various steps involved, such as data preprocessing. There are a number of sub-steps for data preprocessing, and the necessary steps depend on the nature of data, type of data file and different value types, etc. Mallet package is used for extracting the keywords from the newsfeeds.

Data preprocessing is defined as the data mining technique in which raw data is transformed into a human-understandable format. Real-world data is often inconsistent, incomplete and lacking in certain trends or behaviors and contains many errors. Data preprocessing helps in resolving the issues. Data preprocessing is a step that prepares data for further processing. This step is used in database-driven

applications such as rules-based applications and customer relationship management. Data preprocessing is an important step in machine learning.

From the news feed data, specific feature-based information such as city, state information, and the type of crime are extracted. The data is preprocessed by identifying the location string in the news feeds. More specifically, crime-related text strings are identified in the news feeds. Then crime data associated with Bangalore is identified and separated from the overall data. All the intermediate outputs are converted to excel form and stored in a database. Digest package has been used for duplicate detection with the help of hash code. The collected data is periodically checked for adequacy. A minimum of 3000 trial and error procedures are done to review the data. Ggmap is used for locating hotspots in the data. Then the classification model is prepared for preprocessing the data. Both XML and CSV format is used for extracting the data.

a) Importance of Data Pre-processing

Machine learning problems such as classification rely on data preprocessing to handle missing values, correct outliers, normalize and scale data, or feature engineering. If the preprocessing is not done, then it can give false positives and false negatives. The machine learning model is just a piece of code. It requires training data to work properly. Therefore the training model needs to have the right data to provide the right predictions. Issues such as missing values, inappropriate values, etc. can lead to misleading answers/predictions for the unknowns.

b) Getting Started with Data Pre-processing

Data preprocessing includes instance selection, cleaning, normalization, the transformation of data, extraction of features and selection, etc. The output of

data preprocessing is the final training set. There are well-known algorithms for each step of the data preprocessing.

Data visualization is an essential part of the process. When the size of the dataset increases, it becomes difficult to understand insights using excel spreadsheets. It is easier to understand the data when it has visualization. Data visualization is also considered equivalent to visual communication. It involves the study of data and the creation of a visual representation of the same. Data visualization makes use of plots, information graphics, statistical graphics, and other tools. The numerical data available can be represented as lines, dots or bars to communicate the message quantitatively. Data visualization makes use of different plots and graphs to visualize complex data to simplify the discovery of data patterns.

c) Importance of Visualization

CSV data can be difficult to process for getting insights. Human brains can easily process information using graphs or charts to visualize complex data rather than using reports or spreadsheets. Concepts are conveyed easily using data visualization. Data visualization helps in understanding the outliers that impact the machine learning model. It also helps us understand the parameters that impact the results. Visualization can be used both before modelling and after modelling. It can help in identifying the different clusters in the dataset, which is difficult to identify compared to simple visualizations. Visualization is very much useful in the EDA (Exploratory Data Analysis) phase to understand patterns in the data.

3.3.3 Geospatial analysis and visualization

A geospatial visualization is a number of tools and methods that help in geospatial analysis using interactive visualizations. Geospatial visualization emphasizes the

construction of knowledge over information transmission or knowledge storage. To perform this operation, geo-visualization tools communicate geospatial information in various ways, including data exploration, human understanding, and decision-making processes.

Static maps which are traditionally used have a limited exploration capability. The underlying geographical information is linked to graphical representations. Geovisualization and GIS enable more interactive maps, including the ability to explore the different layers of the map, to modify the visual appearance of the map, to zoom in or out, etc. on a computer display. The method of Geovisualization has a number of cartographic practices and technologies that make use of the ability of the advanced microprocessors to create changes to maps in real-time. This allows the users to adjust the mapped data on the fly.

The crime patterns in urban areas are not randomly or evenly distributed. The typical pattern is that the crime occurs rather dense in some regions of a city and sparse in other areas (Bowers & Newton, 2018). With the help of spatial pattern analysis, it is possible to identify the hotspots, i.e., the area in which there is a high aggregation of crime. Also, the environmental context plays a vital role in the occurrence of crime. The definitions of spatial pattern analysis are as follows: crime hotspots are defined as the geographic locations in which criminal activities repeatedly occur (Marzan & C. Baculo, 2018). Individuals have a higher risk of victimization in these areas compared to other places. Cold spots are locations in which there is decidedly less criminal activity. Spatial clustering is used to identify the hotspot patterns in crime. The spatial analysis correlates with environmental contexts.

Real-world scenarios and ‘what-if’ scenarios are analyzed using geospatial visualization. There are two domains – the private domain that is used by professionals for geo-visualization and public domain that is used for “visual

thinking" for the public. This requires collaboration between the professionals and the public.

Geo-visualization tools are used by planners to model the public's policy concerns and environmental scenarios. They use 3D photorealistic representations of urban redevelopment. The dynamic computer simulations show the pollution diffusion for the next few years. The use of the internet by the public has impacted collaborative planning causing the public to participate increasingly. This reduces the time for controversial planning.

Tools like plyr and diplyr are used for cleaning the processed data. This is done to ensure that the collected data does not have any outliers. The data is analyzed multiple times, and the latitude and longitude are processed. A model is built to analyze the crime. In some cases, the data are missing. These missing values are inserted in case of the location not available. The final results are shown in a hotspot visualization on the map of India and Bangalore. The data analytics system also has the option to add filters to visualize the crime hotspots in different ways.

CASE STUDY-1

The geospatial visualization and analysis are used to identify patterns in the crime rates in India. Urban crime has become one of the vital problems for modern cities due to immigration and population growth. Law enforcement agencies collect a vast amount of data to model and predict crime. However, they lack real-time information about the crime. Social media is a data source that has the potential to model crime in real-time and predict it. Based on research studies, spatial and temporal data gathered from social media can be utilized for prediction & analysis of crime. Different social media sources can be used for Spatio-temporal crime analysis like news feeds, Twitter, Facebook, Sample data sets, Police data, etc.

The study says that the crime rate is increasing day by day, which demands Spatio-Temporal visualization techniques such as hotspots detections, Density identification, and Forecasting for better Crime investigations. We have identified the different characteristics of crime data such as Types of crimes (68 Crime keywords-6 Subclasses), Newsfeeds, Frequency of crime, Geographic Locations, and Temporal facts. We focus on utilizing the newsfeed data effectively to predict crime.

GEOSPATIAL ANALYSIS OF CRIME – INDIA

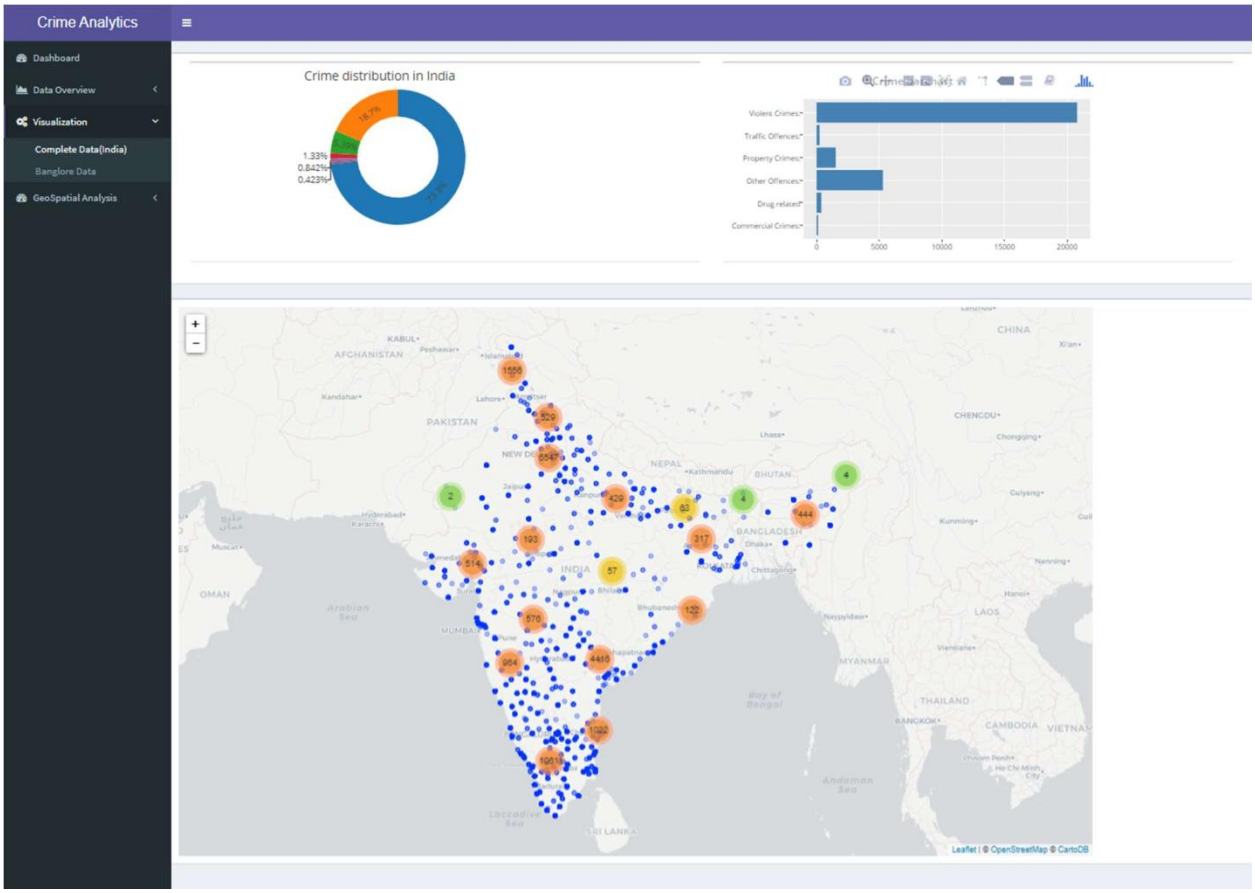


Figure 3.3 Geo-Spatial crime hot spots in India using KNN Algorithm-India

Figure 3 Shows the Geo-Spatial Crime identifications in the Indian context with hotspot identification using KNN Algorithm.

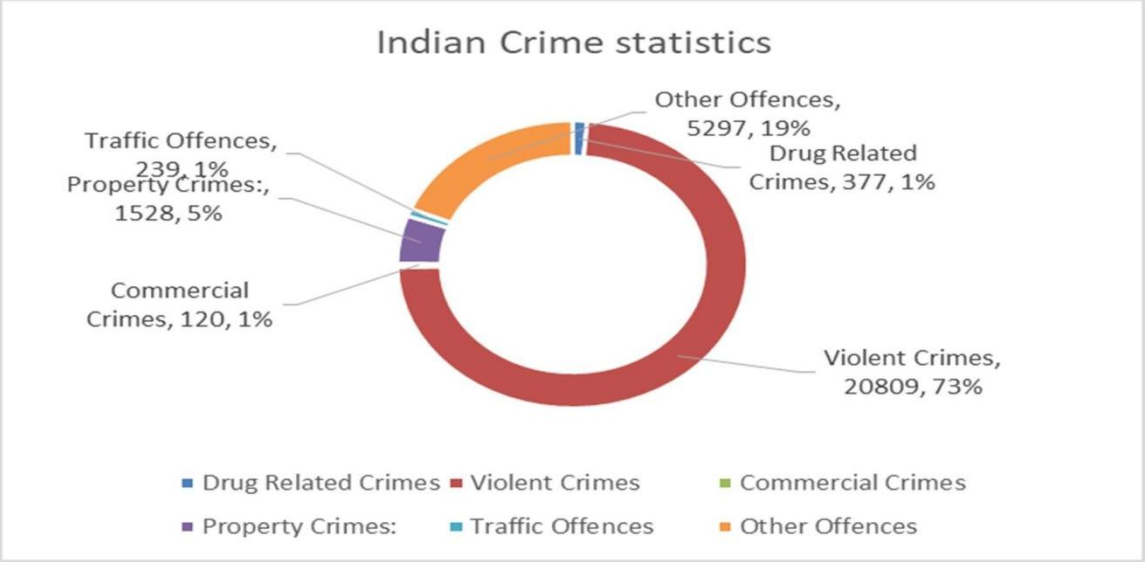


Figure 3.4 Crime Statistics-India

From Figure.4 data analysis, it is found that most of the crimes happening in India are violent crimes constituting 73% of the crime incidents. Property crimes are in second place, with 5% of incidences. Other offenses are at 19%. This analysis has been done with the help of newsfeed data from the Hindu newspaper.

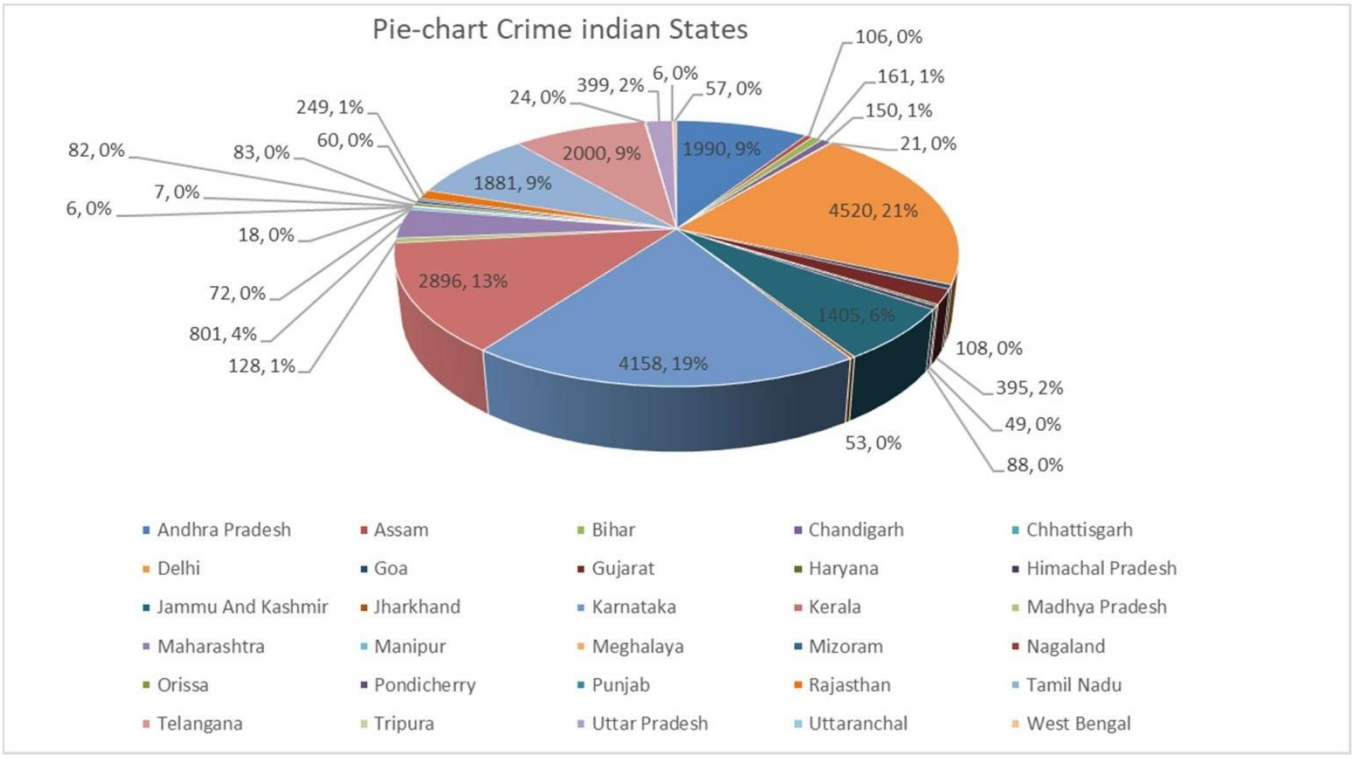


Figure 3.5 State-wise crime analysis

From Figure5. The analysis was done based on location state-wise shows that Andhra Pradesh and Delhi are the top states for crime incidents. This shows that a high density of crime is present in urban areas compared to rural areas in India.

GEOSPATIAL ANALYSIS OF CRIME – BENGALURU

Figure.6 gives the overall picture of the crime rates in Bangalore city with Geo-location wise. Figure.7 shows the analysis of various crimes that occur in the Geolocation wise. It has been found that Violent-crimes are more reported crime in the city.

(a) (b)

Figure 3.6(a) and (b). Geo Spatial crime hot spots in India using KNN Algorithm-Bengaluru

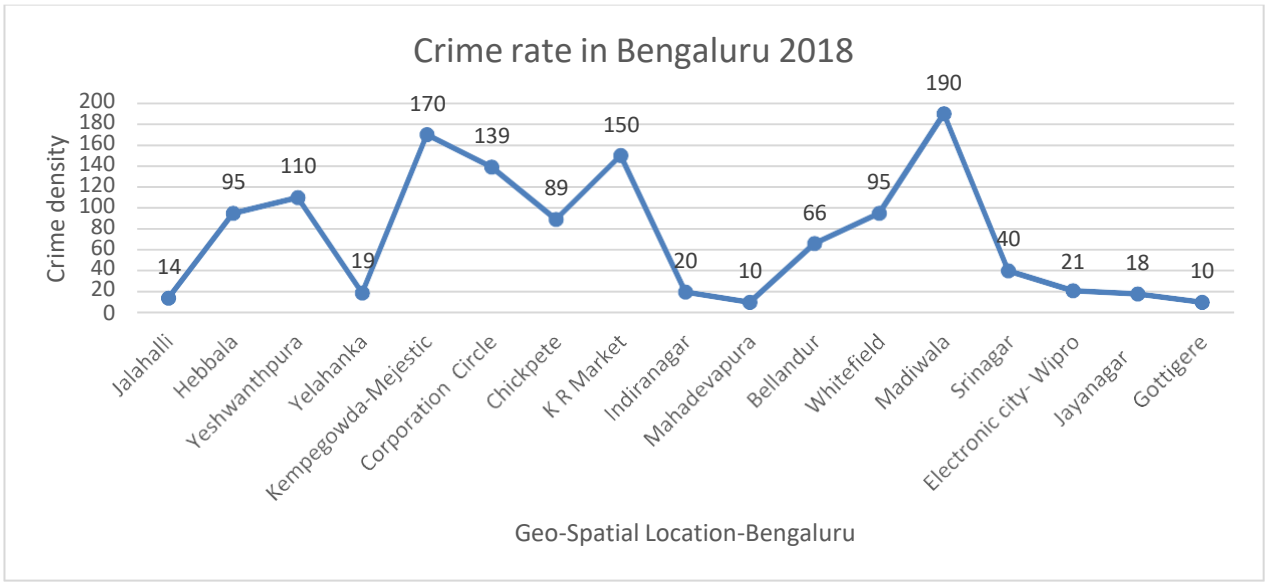


Figure. 3.7 Geo-Spatial Crime Density-Bangalore

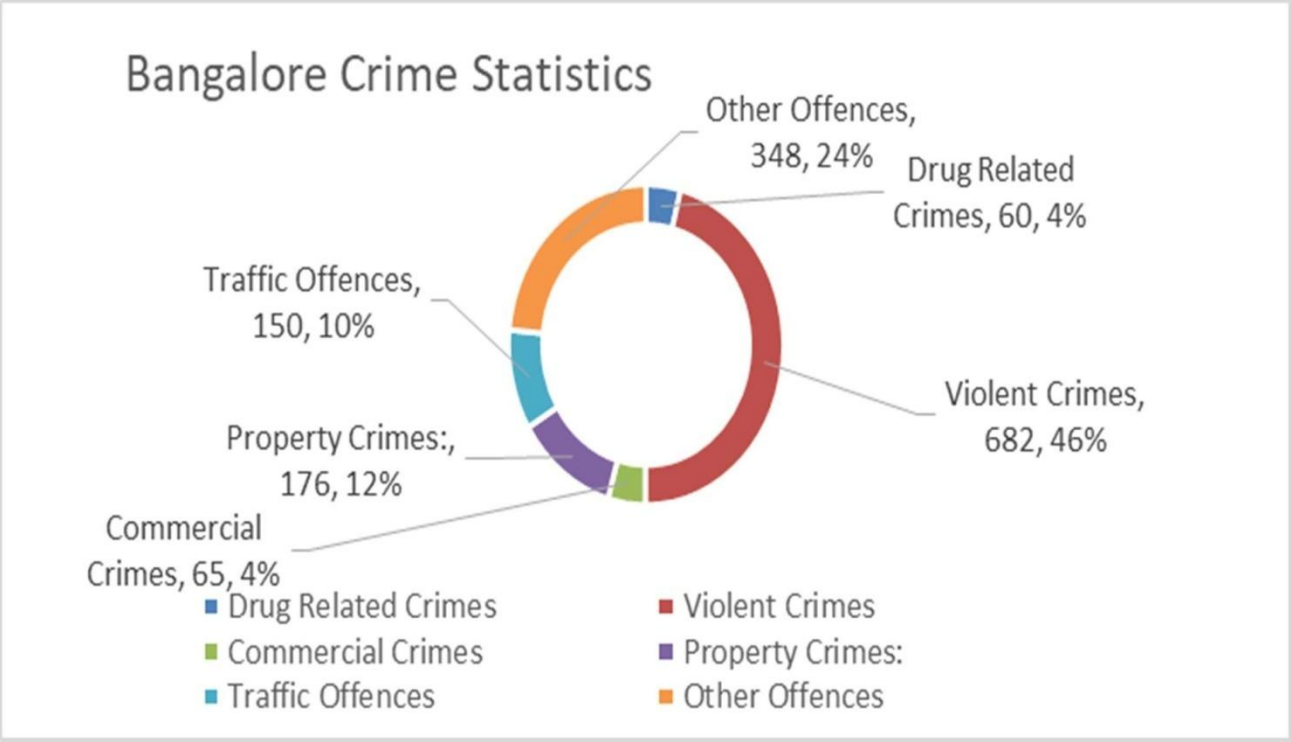


Figure.3.8 Crime Statistics-Bangalore

From Figure.6, 7, and Figure.8 analysis says that violent crime is the top crime in Bangalore city with 46%. Property crimes constitute another 12% in the city. We can gain these insights with the help of Hindu news feeds. This analysis can help police officials to plan their patrolling activities.

GEOSPATIAL ANALYSIS OF CRIME – BENGALURU (Crime branch data)

By filing RTI (Right to Information)(Refer-Appendix-B), we Received and compared the data obtained from the Bengaluru crime branch with news feeds crime data. We can attain a good correlation between the two. This will help in modelling the crime forecasting, which we tackle in the latter part of the thesis.

Table 3.5 Crime statistics (Crime branch data)

Crime head	Crime Branch data
Theft	10966
Assault	3269

Cheating	3199
Burglary	1813
Kidnapping And Abduction	1046
Robbery	986
Molestation	976
Drugs	354
Riots	320
Murder	282
Suicide	172
Dowry Deaths	48
Counterfeiting	10
Total	23441

Chapter Summary

This chapter gives the introduction to the methodology used to collect and clean the crime data from various internet sources. The underlying theory behind the data collection and cleaning are explained in this chapter. Naïve Bayes classification algorithm is used for the classification of the crime into different classes. Mallet package is used for extracting the keywords from the newsfeeds. K-means algorithm is used to identify the hotspots in the crime locations. We have taken the case study of crime analysis in India and Bangalore to implement Geospatial analysis methodologies. The primary reason for that is that crime rates in Bangalore has been rising more compared to that of other cities in India due to urbanization.

Some part of this chapter are published in following journals :

1. Boppuru Rudra Prathap, Ramesha K., "Geo-Spatial Crime Analysis Using Newsfeed Data in Indian Context**",** ***International Journal of Web-Based Learning and Teaching Technologies*** *(IJWLTT),* vol.14,no.4,pp.49-64,2019. doi:10.4018/IJWLTT.2019100103 **(Scopus Indexed Journal)**

CHAPTER 4

CRIME DENSITY IDENTIFICATIONUSING KERNEL DENSITY ESTIMATION

INTRODUCTION

Crime density analysis is one of the fundamental concepts in the pattern analysis of crime. A range of methods can be applied for the analysis right from the underlying summary to advanced spatial analysis. Kernel density estimation is a type of pattern analysis used for crime analysis (Agarwal, Nagpal & Sehgal, 2019). While identifying the crime patterns, hotspots identification should do as useful and accurate as possible. Kernel density estimation methods are used for visualization and analysis of spatial data with the aim of understanding and predicting future event patterns. These methods have several applications, such as damage and risk assessment analysis, road accidents, and emergency planning for rescue services. KDE maps widely used in crime analysis. KDE is very useful in the detection of hotspots utilizing a series of estimations made over a grid that placed on the complete point pattern. These estimations show the crime intensity at a particular location. It also identifies the lows and highs of the point pattern densities. The only role of the user is to identify the appropriate bandwidth for an estimation, which is an important activity. If the bandwidth is set too large, the right information may get lost. If the bandwidth is too small, local data will influence the result more. The process can be eased with the help of a bandwidth slider tool that outputs the preprocessed KDE maps along with the specified bandwidth. In this way, the influence of the bandwidth can be identified in the algorithm. This study also helps in determining the appropriate bandwidth for the problem visually. Even the other issue with KDE is the classification of the kernel density output roster (Clancey, Kent, Lyons & Westcott, 2017). The objective of the classification is to perform an approximation of the original surface as closely as possible, along with the preservation of the characteristicpatterns of the phenomenon.

Kernel density estimation for detection of crime density

Based on the literature, Kernel density estimation is used for calculating the density of point features in every output raster cell. It can be used as a common point feature tool. The algorithm used by the tool fits smoothly with the curved surface at each point. The surface value is large at the point location and reduces when the distance from the point increases. The value is zero at the bandwidth (search radius) distance from the origin point (Dağlar & Argun, 2016). The tool calculates the bandwidth of the input dataset by default. There is a linear unit of projection of the output spatial reference that forms the basis of the search radius. The cell output size determines the output raster that needs to be created. This is the specific value of the environment. If the environment is not determined, the height or width of the point features is defined as the cell size.

PROPOSED METHODOLOGY

Crime hotspot detection is the spatial mapping technique that identifies the concentration of various crimes in the urban area (Berestycki, Wei & Winter, 2014). The most widely used crime hotspot detection method is the Kernel Density Estimation method. KDE is a non-parametric method of estimation in which the probability density of crimes is calculated. KDE uses grid cell size, interpolation methods, and bandwidth to identify the precision of kernel density. The interpolation process has various user-defined settings, thus increasing the quality of KDE hotspots (Wang & Luo, 2018). This analytical technique used for multiple types of crime, such as burglary, robbery, and assault, etc. The KDE hotspot with low resolution converted into one with contour lines. The hotspot generated with smooth boundaries; hence, the generation speed increased. Parameters such as bandwidth and grid cell size are specified for generating hotspots.

Crime density identification using KDE

Kernel Density Estimation (KDE) is the research method used for estimating the data. There are various terminologies used in KDE as follows.

a) Bangalore Crime Density –Case study

Bangalore divided into several geographical areas such as blocks, council wards, local areas etc. For such regions, we compute a measure of crime intensity such as total no of crimes, number crimes in each group, crime density related to land area, population etc. For example, consider an area such as City Market which has an area of say 1600 k square meters with a population 500 k assume that the crime numbers for the Six groups for the calendar year 2017-18 shown in following Table 4.1.

Table 4.1 Sample Bangalore Crime statistics

Crime Group No	Crime Group Name	Crime count
1.	Drug-Related Crimes	89
2.	Violent Crimes	180
3.	Commercial Crimes	42
4.	Property Crimes	50
5.	Traffic Offences	58
6.	Other Offences	35

If we consider Voilent crime, the crime density related to land area is Cd_{area}= # Crimes/ area= = 180/1600 = 0.1125 per K Sq.meter.$Cd_{population}$ = # Crimes/ Population =180/500 =0.36 Per K Sq. meter The following Table 4.2 shows the both Cd_{area} and $Cd_{population}$ for all the six groups of crimes.

Based on Table 4.2 Area based method provided an effective approach due to the following reasons:

- Bangalore is a city, and hence almost every space is occupied. So the number of crimes per area provides a better approach to the transaction.

- Population gives only the people living in the area. However, Bangalore is a city people live in different locations but commit the crime at this location under study.

In our research, we consider the only area-based crime density approach based on the above case study. As the crimes are events happening in the city, every crime is associated with a few key attributes

- Date and time of occurrence.
- Location of occurrence represented by longitude and latitude.

Table 4.2 Geographical-Bangalore Violent crime density based on area and population

S.no	**Crime group**	**Geographical area name**	**#Crime count**	**Cd_{area}**	**$Cd_{population}$**
1.	Violent Crimes	City Market	180	0.1125	0.36
2.	Violent Crimes	Corporation Circle	139	0.1158	0.46
3.	Violent Crimes	Kempegowda-Mejestic	170	0.130	0.56
4.	Violent Crimes	Madiwala	190	0.19	0.6785
5.	Violent Crimes	Whitefield	95	0.05375	0.23
6.	Violent Crimes	Yeshwanthpura	110	0.0611	0.22

Crimes occur in pointed location once or multiple times at different times. For example consider the occurrence shown in Figure 4.1 In the area marked as a circle, the total number of crime occurrence is 1+3+1+2+20+8+15 = 50. The radius of the circle is 0.7.

Area of circle = $\pi r^2 = \pi$ x 0.7 x0.7 = 1.5394

So, crime density = 50 / 1.5394 = 32.48.

The density correspondence to the center of circle of the whole area of the circle depending up on how we want to analyze the data.

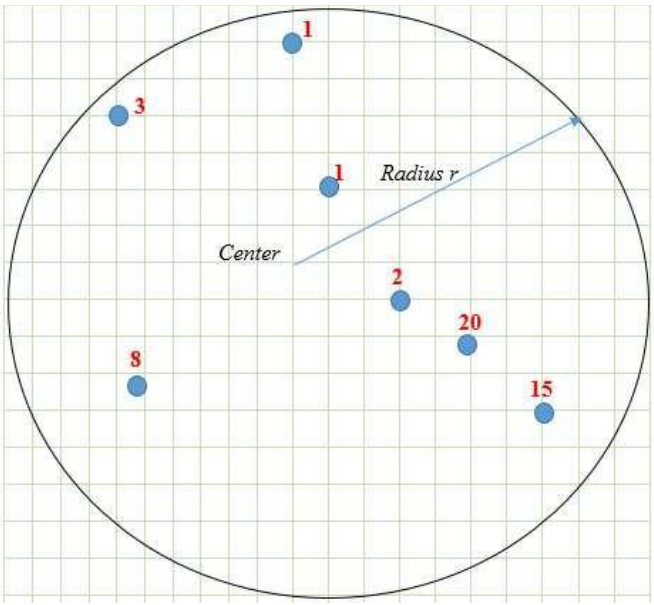


Figure 4.1 Sample crime occurrence hotspots-Scernario-1

The same density assumed for the whole area will mislead the study as there are two spots at the location right with 20 and 15 occurrences and hence the density has to be assumed to be more in that location.

The following dimensions are to be seriously looked into while analyzing the data.

1. Though crime occurrence at point, the same crime may also happen at a neighbouring point of the grid. So instead of a discrete set of data, we need a continuous probability density function for each point of the grid.

2. The same crime density cannot be considered for all the grid points because there is a point with 20 frequency with another one point.

b) Probability density function

Probability theory deals with quantities that have a random distribution. The probability density function (PDF) is defined as the probability of a random value fitting into a range of values in the function (Ezra, Crucecia, Ivan, Ayoo & Olwedo, 2019). In this methodology, the integral of the value's density is identified. The

resultant value gives the probability of the new random value. The probability of the random number is given by the area covered by the density function.

$$p(a < X < b) = \int_{a}^{b} (d)dx \quad a < b \qquad (4.1)$$

Equation-(4.1) shows the relationship between crime type 'X,' crime density 'd,' and the bandwidth of the crime from a to b.

Bandwidth: Bandwidth is a smoothing parameter that denotes the width of the sample in KDE (Jiang, Yang & Li, 2018). It determines the search radius of the function. The function can be over or under smoothed. Bandwidth can be estimated based on thumb rules. Perfect computational solutions cannot be applied.

Grid size: The grid cell size defines the resolution of the KDE algorithm. Large grid sizes lead to the low-quality hotspot and low visualization. The right grid size is identified by the standard deviation of latitude and longitude. Grid size is denoted by a and b.

c) Kernel density estimation Existing

Kernel density estimation is used to identify the crime hotspot of a city easily. There is a spatial mapping of the various parts of the city, such as K.R Market in Bengaluru. The strength of the hotspot is measured by counting the events in a particular area (Jin-Cheon, Hao, Yong, Mia Hao & Mani Kandan, 2019). KDE aims at smoothing the discrete values and providing a continuous probability density function.

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \qquad (4.2)$$

In equation (2) *f(x)* -Crime desity Estimation, *h*- Maximum distance covered,x_i-specific crime density,*k*-kernel function.

Consider the crime occurrence data given in Figure 4.2 h = r = 0.7, n= 7.

i	1	2	3	4	5	6	7
x_i	1	3	1	2	20	8	15

Figure 4.2 Sample crime occurrence

Assume that *k(a)* is non-negative kernel function. For any value x from 0 to a large integer

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) = \frac{1}{7(0.7)}\left[k\left(\frac{x-1}{0.7}\right) + k\left(\frac{x-3}{0.7}\right) + k\left(\frac{x-1}{0.7}\right) + k\left(\frac{x-2}{0.7}\right) + k\left(\frac{x-20}{0.7}\right) + k\left(\frac{x-8}{0.7}\right) + k\left(\frac{x-15}{0.7}\right)\right]$$

We have following serious observations evaluation of $f(x)$ is dependent of the center. For example there are two points with occurrences 20 and 15 from far center in Figure4.1 . The fact that they are far way has no impact on the computation. For example Figure 4.3 we present two scenarios. Both of them have the same KDE.

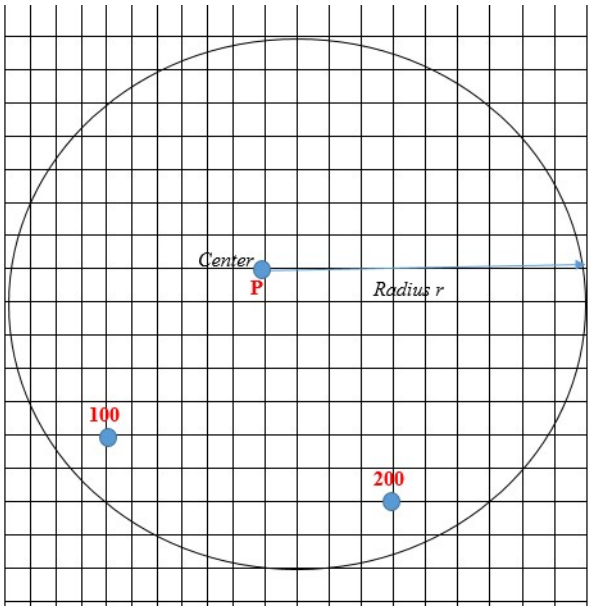


Figure 4.3(a)

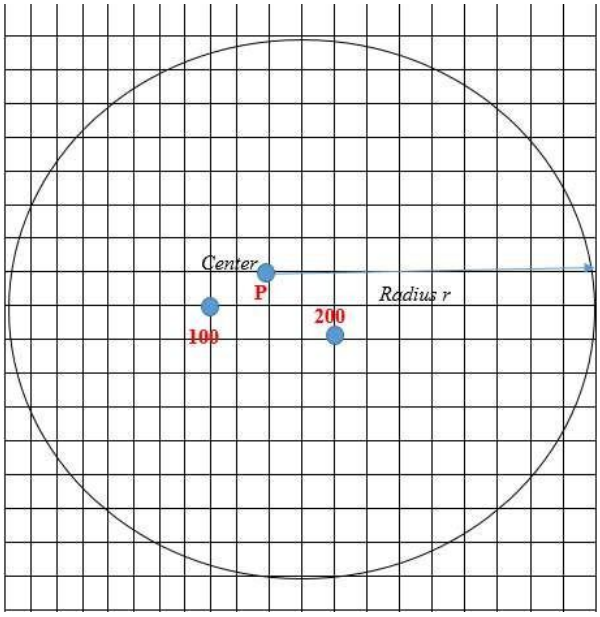


Figure 4.3 (b)

In Figure 4.3 (a) there are two crime spots with 100 and 200 frequency far away from the center.In Figure 4.3 (b) the two point of occurance with 100 and 200 frequency are very near to the center. However in both the cases they equally contribute in the KDE function with

$$f(x) = \frac{1}{nh}\left[k\left(\frac{x-100}{0.7}\right) + k\left(\frac{x-200}{0.7}\right)\right]$$

The above observation motivated us to build a new kernel density estimation function. This new KDE has the following characterstics.

1. *Smoothing:* the function $f(x)$ is a smooth continuous function for all non-negative real values of *x.* As *x* denotes the mean crime occurrence in the area with center *p* and radius *h*, *x* can be any non-negative real number.

2. *Weighted contribution of point occurrence:* In the circle assume that there are n points of occurrences denoted by p_1, p_2, p_3…….p_n with number of crimes x_1,x_2,x_3.....x_n Assume that the geographical distance between *p* and p_i is h_i.

   h_i will contribute to the computation of *f(x)* inversely. The crime count x_i happened at far away location contributes less to the calculation of *f(x)* and vice versa. If h_i is high, x_i and h_i contribution to the calculation of *f(x)* is less. This means that occurrence of crime at a far location contribution less to *f(x).*

$$f(x) \propto \frac{1}{h_i} \qquad (4.3)$$

3. *Kernel function: f(x)* is primarily driven by the kernel estimation function *k.* we consider a non-negative kernel function for evaluating *f(x).* As the value of *x* is non-negative the following are possible choices of the kernel function.

a) Gaussian function: The Gaussian function defined as

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Gaussian function supports for even negative values of *x*. However for the crime estimate problem, *x* is always non-negative.

b) Logistic function:The logistic function is defined as

$$k(u) = \frac{1}{e^{u} + 2 + e^{-u}} \qquad (4.4)$$

c) Sigmoid function:The sigmoid function is defined as

$$k(u) = \frac{2}{\pi} \frac{1}{e^{u} + e^{-u}} \qquad (4.5)$$

d) Bandwidth selection

In KDE estimation, the bandwidth selection plays a major role. (Jones.M.C and J.S.J Shelther,. 1996) have given summary of approach.

Problems identified in the existing system

1. **Identification of more potential Crime Geo locations.**

The potential Geo locations of the crime area cannot be identified with the help of the existing Kernel density estimation method. The KDE algorithm can be applied only for a small number of destinations. This causes coverage issues for the data. Example of a case study with respect to Bangalore crime data showed in Figure 4.3 (a). From figure it is identified that identification of Specific geographical location is difficult.

2. **Visualization of specific crime Geo locations**

The results of the existing KDE algorithm cannot be visualized for specific crime Geo-location. This is because the complete bandwidth ‘h’ is considered. Figure 4.3 (a)

explains the visualization of Specific geo-location visualization is difficult because of bandwidth selection.

Proposed Kernel Density Estimation Model

Consider an area represented by the latitude and longitude. Let *p* be a point at which we want to construct the weighted KDE based up on the crime data around point *p*. let *h* be the radius of the circle with *p* as the radius. *h* is called the bandwidth as discussed already. Consider the crime data on points in the circle let $p_1,p_2, p_3.....p_n$ be the n points within the circle. Let $h_1,h_2, h_3....h_n$ be the geographical distance of points $p_1,p_2,p_3....p_n$ from the center *p*. let $x_1,x_2,x_3...x_n$ be the crime numbers at $p_1,p_2,p_3....p_n$.

The weighted KDE *f(x)* for the point *p* with bandwidth *h* is defined as

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{\frac{h_i^2}{h_{mean}}}\right) \quad \text{---------- (4.6)}$$

In Equation 4.4 *h*-Maximum distance covered, *n*-Total number of crimes, *x*-Specific crime density, x_i-Crimes mean, *k*- Kernel function, h_i- Specific geographic Distance, *f(x)*-Crime Density estimation, Where $x_l,x_2,.....$ x_n is the randomly selected newsfeed data sample, kernel function is depicted by $k(),(x - x_i)$ gives the distance between the event x_i and the estimated points '*h*' is bandwidth,

$h_{mean} = meanof\{h_1, h_2, h_3 .....h_n\}$,i.e $h_{mean} = \frac{h_1 + h_2 + h_3.....h_n}{n}$.

Table 4.3 Sample Crime occurances

Index i	1	2	3	4	5	6	7
Point p	P_1	P_2	P_3	P_4	P_5	P_6	P_7
x_i	1	33	1	2	20	8	15
h_i	1.4	4.1	4	2.3	3.4	3.4	5.9

Consider the crime occurrence data given in Figure-1. $h = 0.7, n=7, h_{mean}=3.5$.

$$f(x) = \frac{h_{mean}}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h_i^2}\right)$$

$$= \frac{3.5}{7(0.7)} \left[k\left(\frac{(1.4)^2}{x\ 20}\right) + k\left(\frac{(4.1)^2}{x\ 8}\right) + k\left(\frac{(4)^2}{x\ 15}\right) + k\left(\frac{(2.3)^2}{\quad}\right) \right]$$

$$= 0.71429 \left[k\left(\frac{(3.4)^2}{1.96}\right) + k\left(\frac{(3.4)^2}{16.81}\right) + k\left(\frac{(5.9)^2}{16}\right) + k\left(\frac{5.29}{\quad}\right) \right]$$

$$k\left(\frac{11.56}{\quad}\right) + k\left(\frac{11.56}{\quad}\right) + k\left(\frac{34.81}{\quad}\right)$$

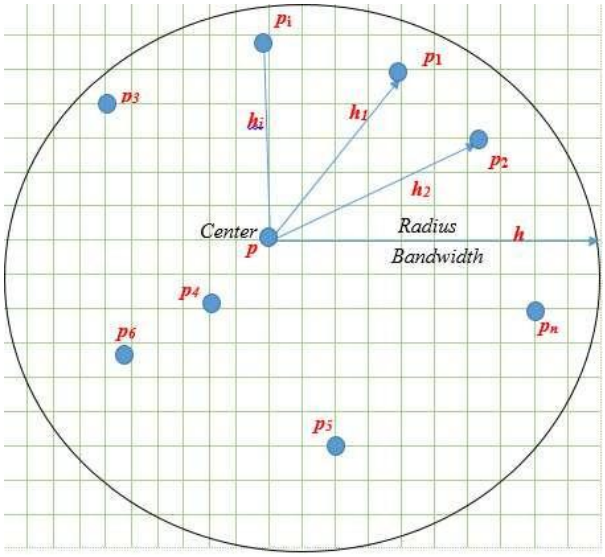


Figure 4.4 Sample crime occurrence hotspots-Scenario 4

Example of a case study with respect to Bangalore crime data showed in Figure 4.5 (a) and Figure 4.5(b). From Figure 4.5(b) it is identified that identification of Specific geographical location becomes easy for individual crime and its identified visualization made it easy to understand to the user.

From the formula, it is found that bandwidth influences KDE. The point density change is smooth when '*h*' increases, and the change is rough when '*h*' decreases. Kernel density estimation is derived from the moving window and is represented by the point process smooth intensity. In this Research, we had identified 6 classes of crimes mapped and identified the density of crimes using KDE.

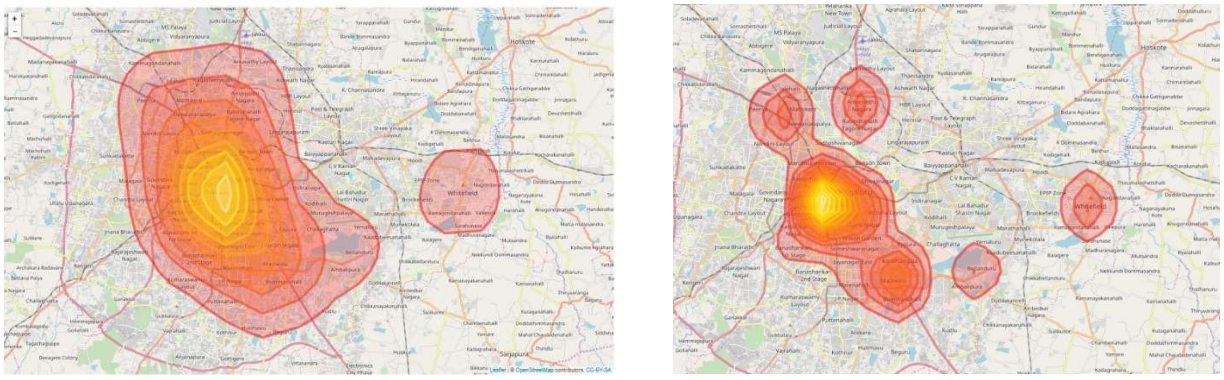
The working procedure of the proposed algorithm is as follows

Algorithm: Geo Spatial crime Density using Kernel Density Estimation

Input: Crime news gathered from a newsfeeds in particular time period.

Output: Density Points in geospatial maps.

1. Gather crime related news data from newsfeeds.
2. Preprocess the data gathered from news feeds.
    a) Remove the html tags, advertisements
    b) Remove the white spaces present in the data set.
    c) Remove the start and stop words from the data set.
    d) Apply the porter stemmer algorithm to remove the stemming words.
3. Identify the Crime and location form the news feed.
4. Find the kernel density function k of features f in spatial point for every *x*
5. $$f(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h_i^2}\right) h_{mean}$$
6. End for
7. Return the Density of crime with respect to spatial and temporal characteristics.



(a) **(b)**

Figure 4.5 Sample output for Existing and Proposed model (a).Existing KDE Result, **(b)** Proposed KDE Result

Figure 4.3 shows the results of the existing KDE model (Figure 4.5(a)), wherein the exact location or the density of the crime is not clear. We can see a hotspot in the whole area. On the other hand, if we know the result of the proposed KDE model (Figure 4.5 (b)), the variation with the value of '*h*' gives us the more accurate and exact location of the crime. Even it shows which area has more crime density as compared to others.

CASE STUDY -2
Crime Density Analysis – India News Feed Data

a) Crime density overall

Figure 4.6 shows the crime density analysis of the crime data from India. It is found that Karnataka, Delhi, and Andhra Pradesh are the top states for violent crimes. There are also other crimes such as theft, burglary, etc. that are more prevalent in Kerala. The reason is because of the dense population in these areas.

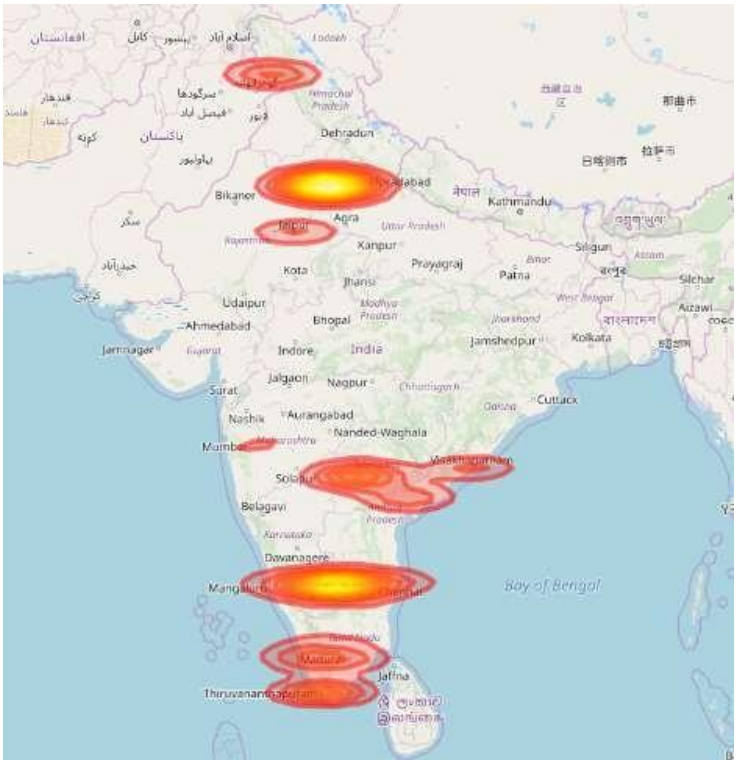


Figure 4.6 Crime density identification for all 6 crime classes- India

The overall density of all six crimes categories is 23945. Overall crime density is calculated with a total of 68 types of crime in 6 classes like Drug-Related Crimes, Violent Crimes, Commercial Crimes, Property Crimes, Traffic Offences, and Other Offences. The following sections explain the detailed result of individual crime density identification.

b) Crime density –Violent Crimes:

Figure 4.7 Crime density identification for Violent crime - India

Figure 4.7 shows the crime density analysis of the violent crime data from India. The Violent Crime class includes 15 different types of crime keywords such as Rape, Murder, Terrorism, Kidnapping, Assault, Sexual Harassment, Sexual assault, Homicide, Gunshot, Intentional Killing peoples, Shootout, Gang-rape, Attempt to murder, Sexual abuse, Putting to death. The total density of violent crimes identified

as 14120 and also violent crimes identified as the highest crime in India. Figure 4.7 found that Karnataka, Delhi, and Andhra Pradesh are the top states for violent crimes. Having data visualizations for individual crimes can help the police in analyzing the root causes easily.

c) Crime Density-Drug related

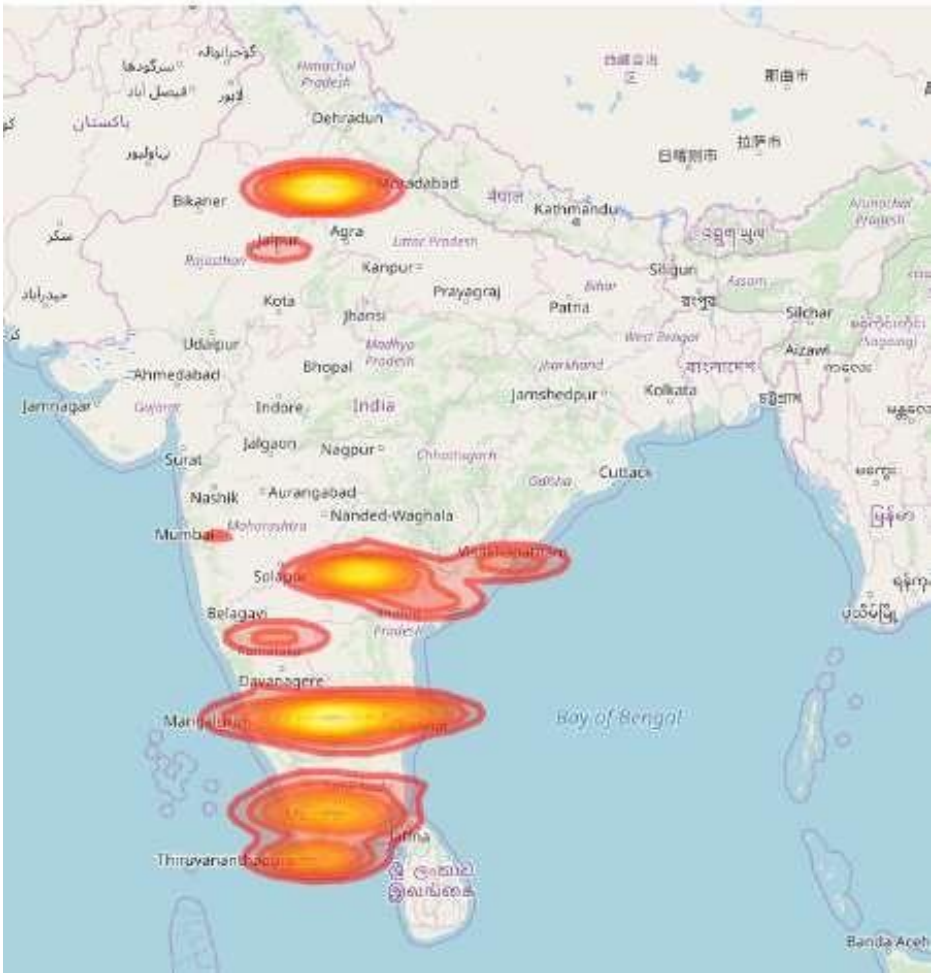


Figure 4.8 Crime density identification for Drug-related crime - India

Figure 4.8 shows the crime density analysis of the drug-related crime data from India. The Drug-related crime class includes 5 different types of crime keywords such as Drug Trafficking, Drug dealing, Drugs smuggling, Narcotics, drugs, and alcohol. It is found that the southern part of India is very much affected due to this crime. The total density of drug-related crimes identified as 2030. The most affected states are Tamil

Nadu, Karnataka, and Andhra Pradesh. Having data visualizations for individual crimes can help the police in analyzing the root causes easily. The crime density can be varied, and location-specific data can be obtained easily.

d) Crime density – Property related

Figure 4.9 Crime density identification for Property related Crimes - India

Figure 4.9 shows the crime density analysis of the property-related crime data from India. Property-related crime class includes 16 different types of crime keywords such as Arson, Motor vehicle theft, Theft, Burglary, Robbery, Riots, Criminal breach of trust, Stealing, Barrage fire, Bombardment, Electric battery, Shelling, Looting, Embezzlement, Trespass, Incendiarism, Shoplifting, Vandalism. The total density of

property-related crimes identified as 3880. It is found that the southern part of India is very much affected due to this crime. The most affected states are Tamil Nadu, Karnataka, and Andhra Pradesh. The crime density can be varied, and location-specific data can be obtained easily.

e) Crime density – Other offenses

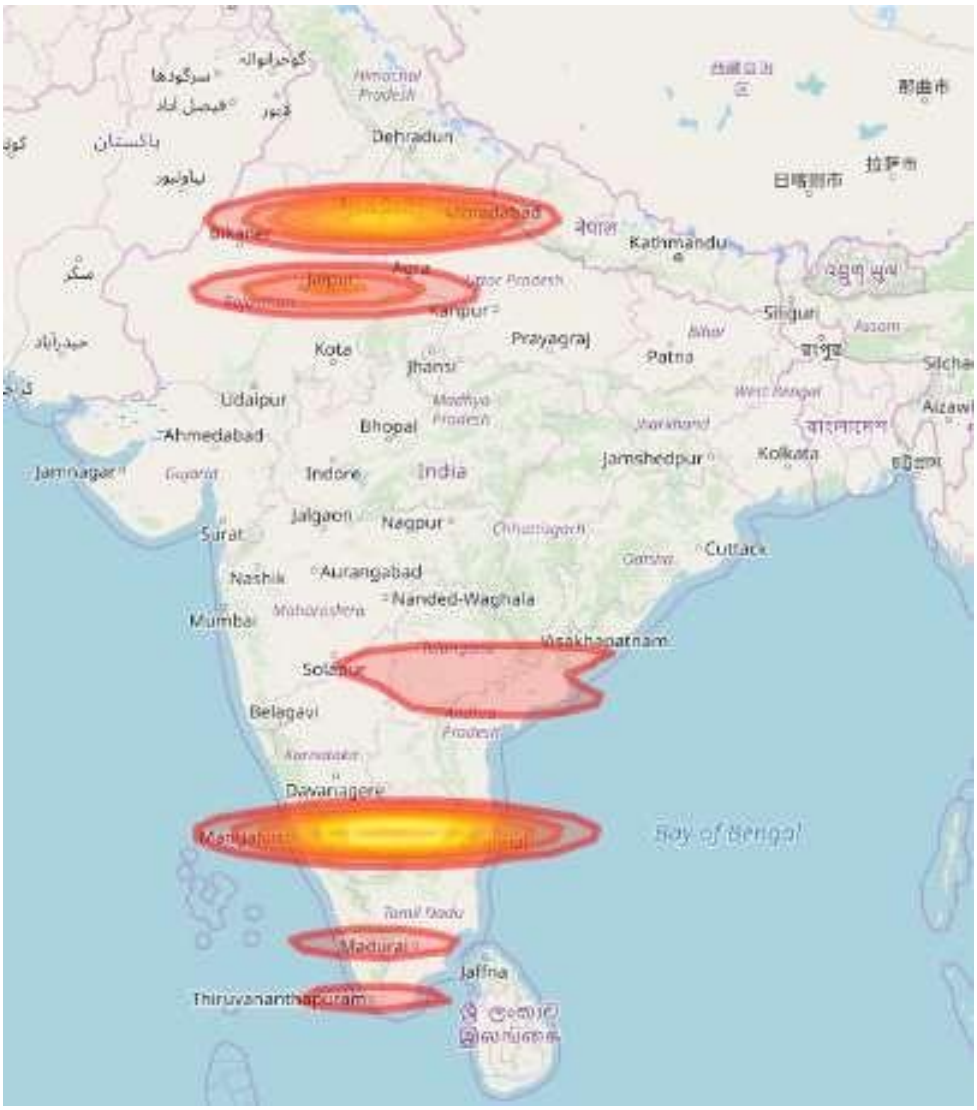


Figure 4.10 Crime density identification for Other Offenses - India

Figure 4.10 shows the crime density analysis of the other offenses data from India. Other offenses crime class includes 17 different types of crime keywords such as Prostitution, Illegal Gambling, Adultery, Homosexuality, Weapons violation, Offense involving children, Public peace violation, Stalking, Cheating, Hurt, Counterfeiting,

Dowry deaths, Outrage her modesty, Causing death by negligence, Suicide, Criminal damage. The total density of other offenses identified as 895. It is found that the southern part of India is very much affected due to this crime. The most affected states are Tamil Nadu, Karnataka, and Andhra Pradesh. In the northern part of India, Delhi and North Rajasthan are the most affected.

f) Crime density- Traffic offenses

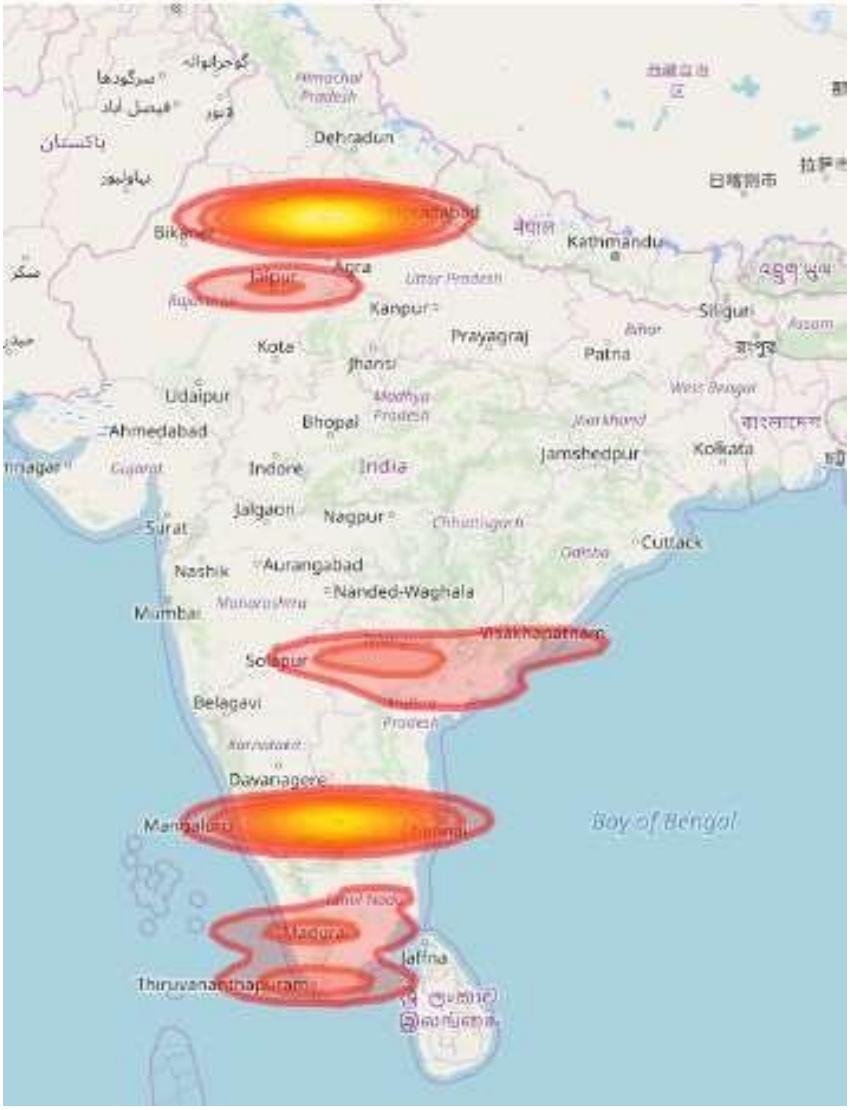


Figure 4.11 Crime density identification for Traffic offenses - India

Figure 4.11 shows the crime density analysis of the Traffic offenses data from India. The Traffic offenses include 4 types of crime keywords such as Speeding, Signal

Jump, Running a Red Light, drink, and drive. The total density of traffic offenses identified as 2320. It is found that the southern part of India is very much affected due to this crime. The most affected states are Tamil Nadu, Karnataka, and Andhra Pradesh. In the northern part of India, Delhi and North Rajasthan are the most affected.

g) Crime density- Commercial Crimes

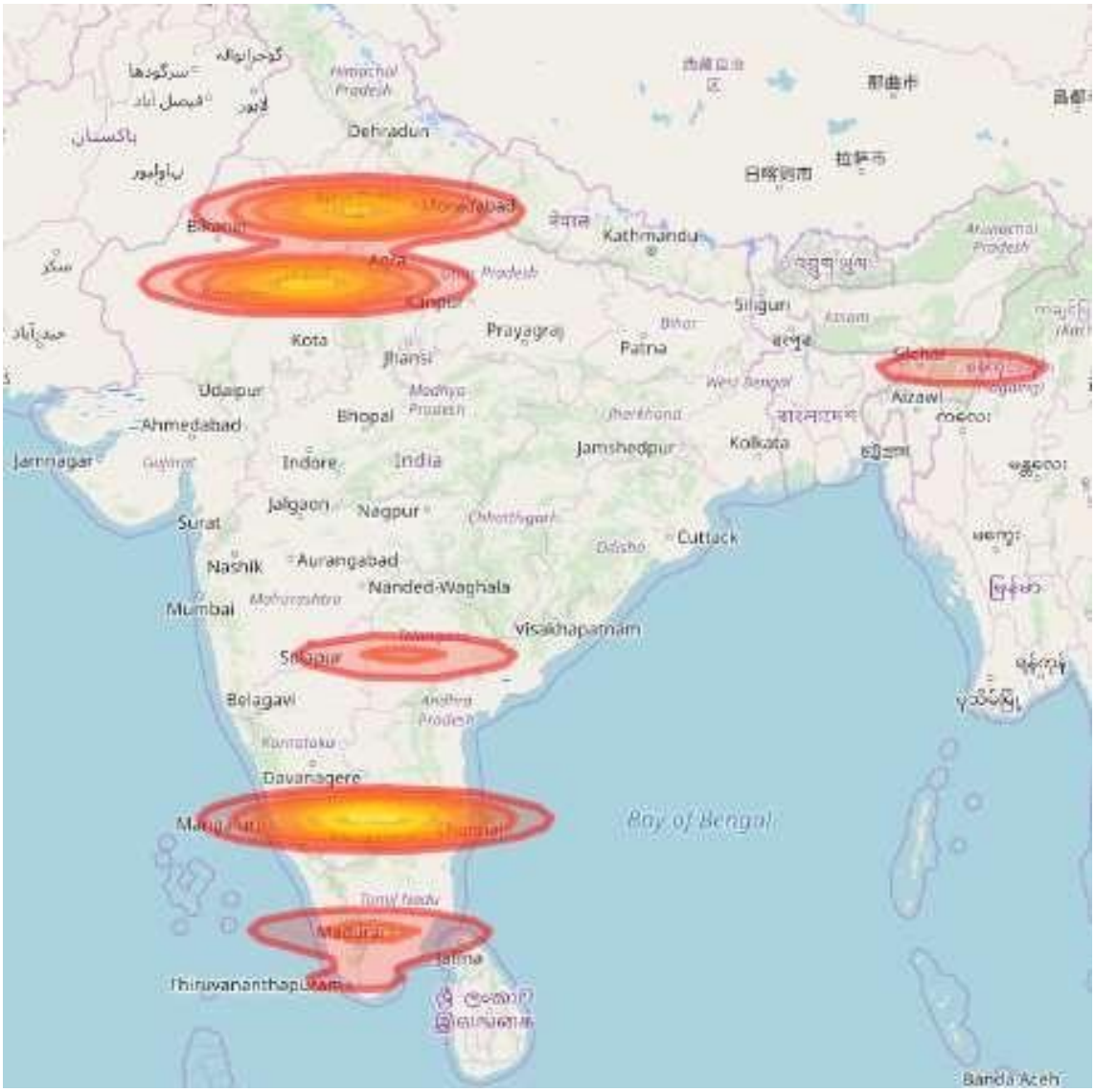


Figure 4.12 Crime density identification for Commercial Crimes - India

Figure 4.12 shows the crime density analysis of the commercial crime data from India. Commercial crime class includes the 7 types of crime keywords such as Official

Document Forgery, Currency Forgery, Official Seal Forgery, Official Stamp Forgery, Bribery, Counterfeiting, Cheating. The total density of Commercial crimes identified as 780. Figure 4.8 found that the northern part of India is very much affected due to this crime. The most affected states are Tamil Nadu, Karnataka, and Andhra Pradesh. In the northern part of India, Delhi and North Rajasthan are the most affected.

Crime Density Analysis – Bengaluru News Feed Data

a) Crime density Bangalore-Overall crimes

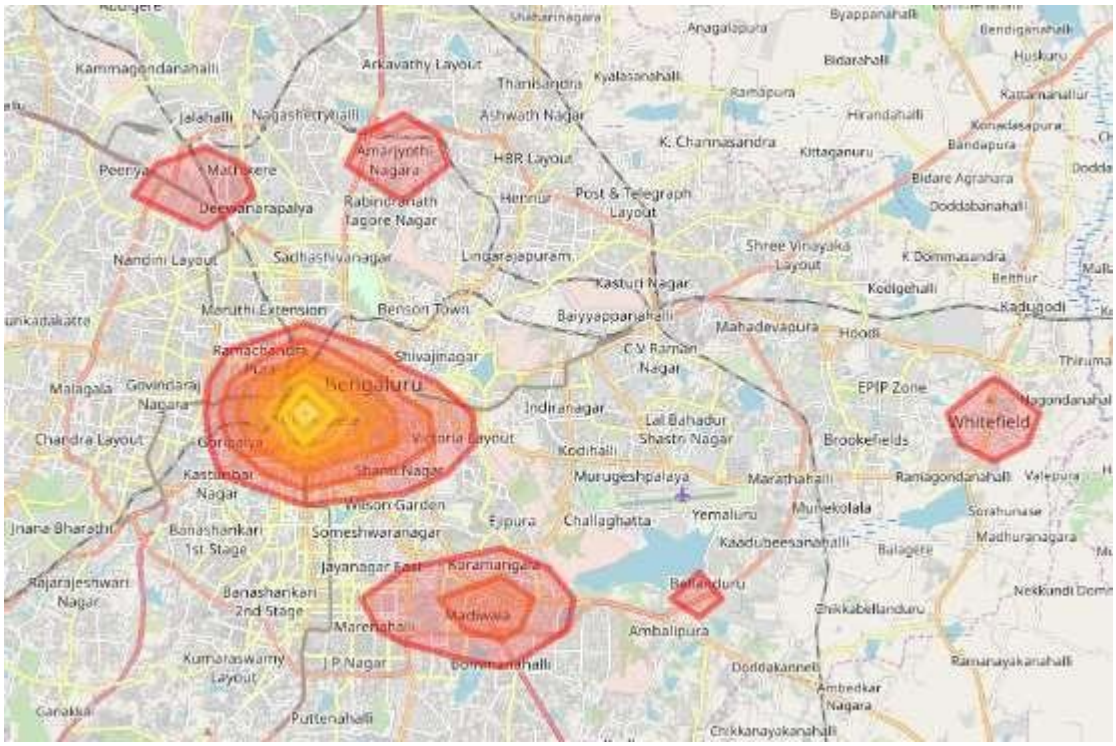


Figure 4.13 Crime density identification for all 6 crime classes - Bengaluru

Figure 4.13 shows the crime density for all 6 crime classes such as Drug-Related Crimes, Violent Crimes, Commercial Crimes, Property Crimes, Traffic Offences, and Other Offences in the context of Bangalore city. The total density of all 6 crimes identified as 1544. Figure 4.13 shows that the central region of Bengaluru more on specific Geographic locations like Yeshwanthpura, Kempegowda Majestic, Corporation Circle, K R Market, Chickpete, etc.. are most affected by crime. This is probably because the population density is higher in these Geographical areas.

b) Crime density Bangalore-Traffic offenses

Figure 4.14 shows the traffic offense crime density of Bangalore city crime data. The Traffic offenses include 4 types of crime keywords such as Speeding, Signal Jump, Running a Red Light, drink, and drive. The total density of Traffic offenses identified as 148. Figure 4.14 shows that the central region of Bangalore, more specific Geographic locations like Whitefield, NH48, Airforce Yellanka, NH48, BETL, etc. are most affected by crime. There are also pockets of high crime density areas on the outskirts of the city. It is also observed more traffic offenses are happening at highways like NH44, BETL, NH48, etc

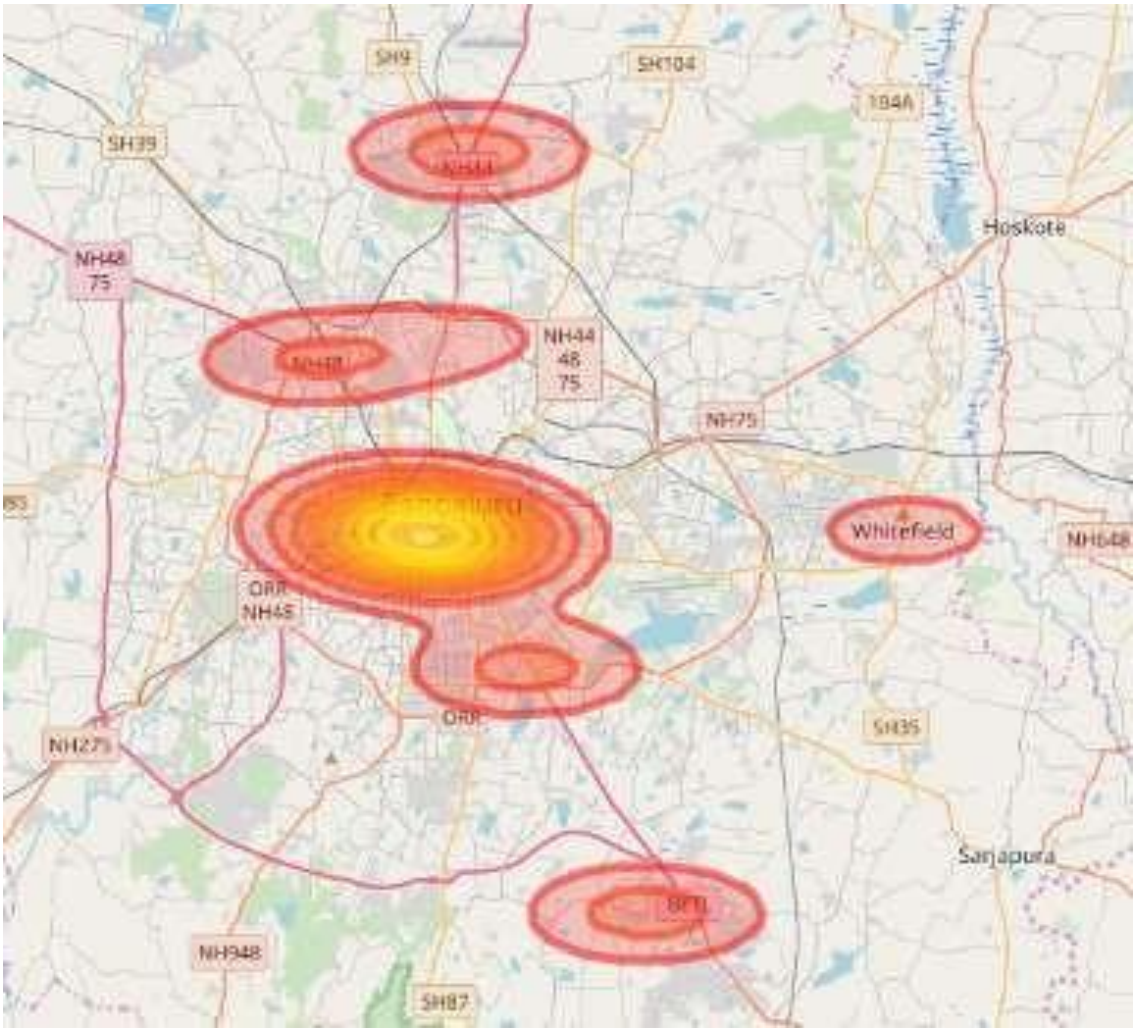


Figure 4.14 Crime density identification for Traffic offenses - Bengaluru

c) Crime density Bangalore-Property Related crimes

Figure 4.15 shows the property crime density of Bangalore city crime data. Property-related crime class includes 16 different types of crime keywords such as Arson, Motor vehicle theft, Theft, Burglary, Robbery, Riots, Criminal breach of trust, Stealing, Barrage fire, Bombardment, Electric battery, Shelling, Looting, Embezzlement, Trespass, Incendiarism, Shoplifting, Vandalism. The total density of Property related crimes identified as 178. Figure 4.15 shows that the central region of Bangalore more on specific Geographic locations like Chikpet, Peenya, Mathikere, Madiwala, Koramangala, Whitefield, etc. are most affected by crime. This is probably because the diversity of population density is higher in these areas.

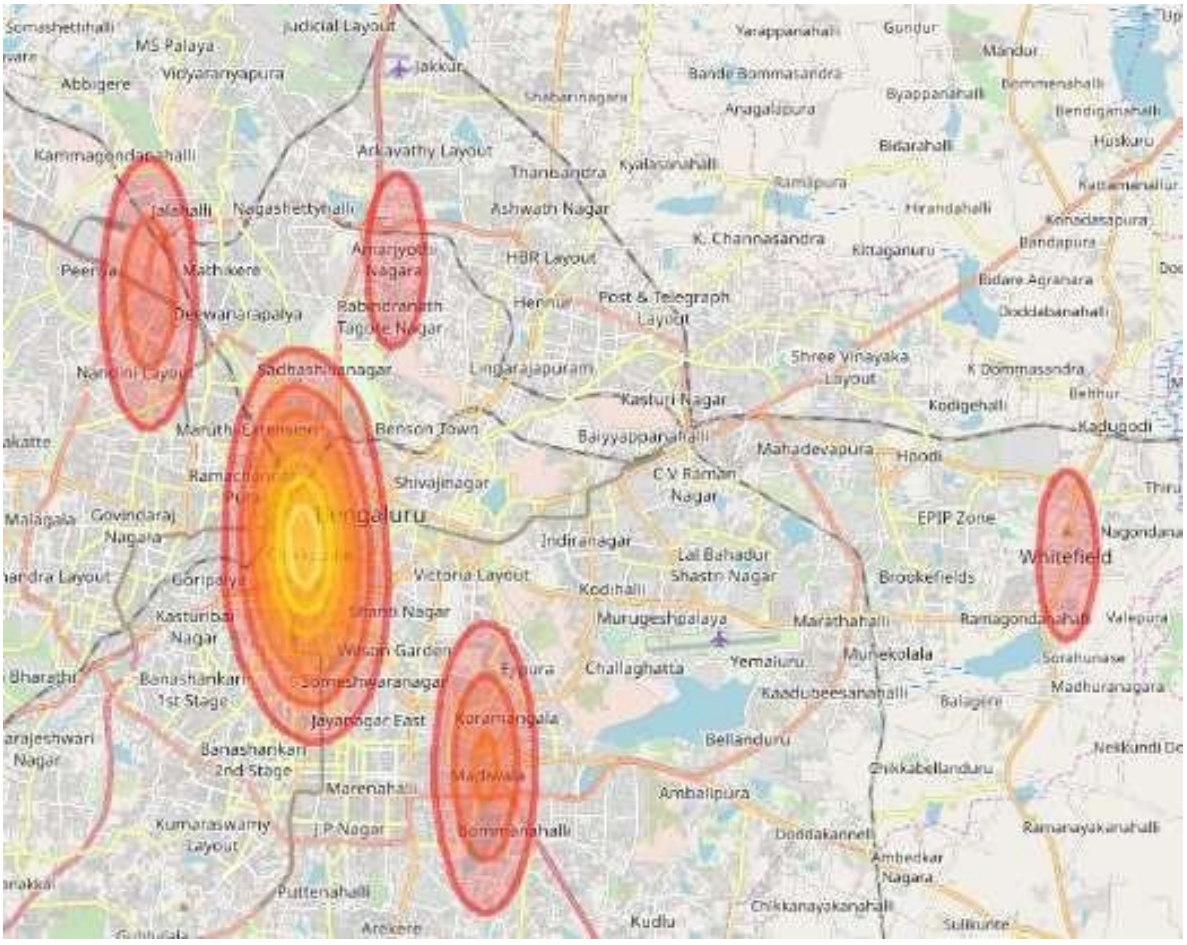


Figure 4.15 Crime density identification for Property related crimes - Bengaluru

d) Crime density Bangalore-Drug-related Crimes

Figure 4.16 shows the drug crime density of Bangalore city crime data. The Drug-related crime class includes 5 different types of crime keywords such as Drug Trafficking, Drug dealing, Drugs smuggling, Narcotics, drugs, and alcohol. The total density of Drug-related crimes identified as 53. Figure 4.16 shows that the central region of Bengaluru more on specific Geographic locations like Marenahalli, Whitefield, K R Market, Madiwala, Amar Jyothi Nagar, Majestic, etc. are most affected by crime. Geographic population density influences the occurrence of drug crimes in the region.

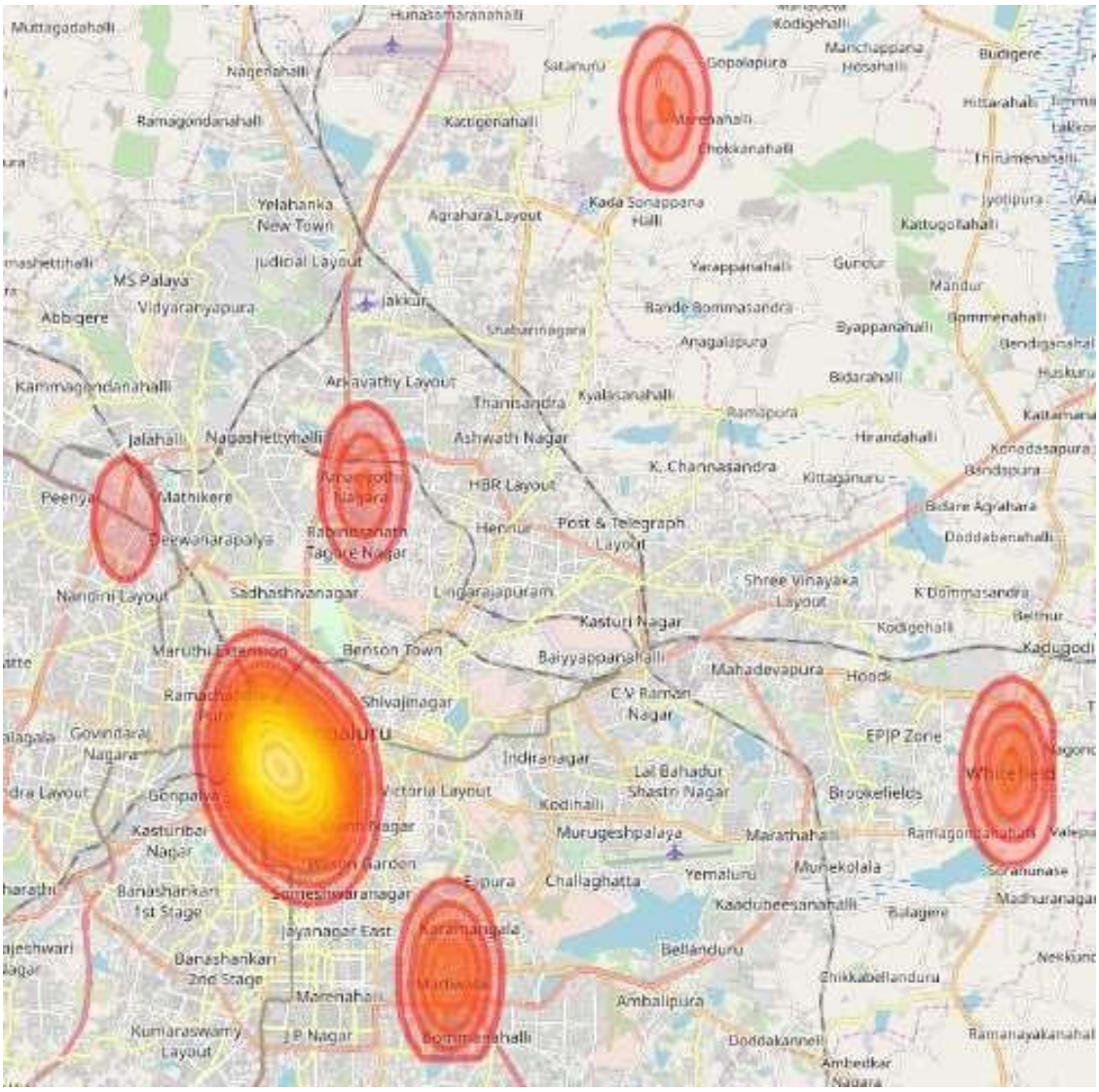


Figure 4.16 Crime density identification for Drug-related crimes – Bengaluru

e) Crime density Bangalore -Commercial Crimes

Figure 4.17 shows the commercial crime density of Bangalore city crime data. Commercial crime class includes the 7 types of crime keywords such as Official Document Forgery, Currency Forgery, Official Seal Forgery, Official Stamp Forgery, Bribery, Counterfeiting, and Cheating. The total density of Commercial crimes identified as 36. Figure 4.17 shows that the central region of Bangalore more on specific Geographic locations like Amar Jyothinagar, Peenaya, Madiwala, Marathahalli, K R Market, etc. are most affected by crime. There are also areas in the eastern region of Bangalore that are affected by this crime.

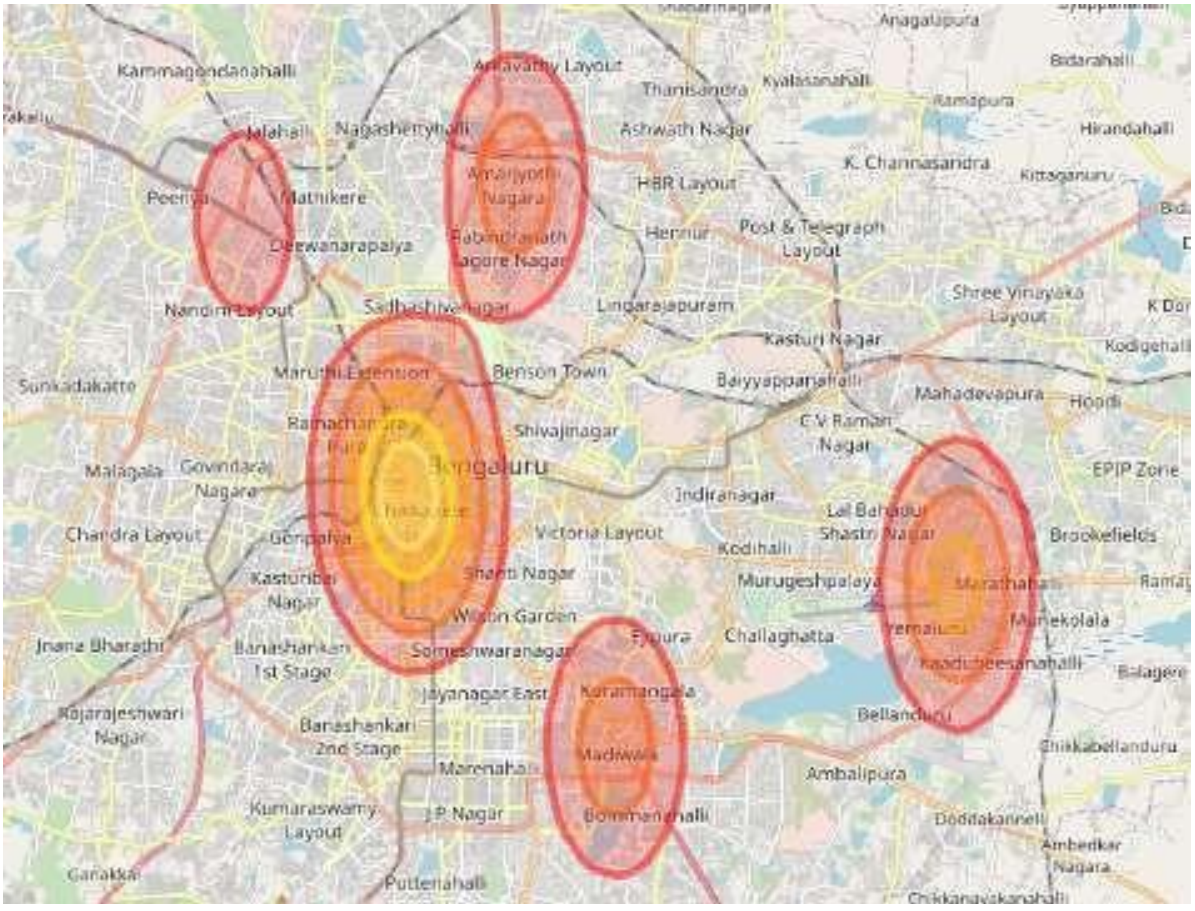


Figure 4.17 Crime density identification for Commercial crimes - Bengaluru

f) Crime density Bangalore-Violent Crimes:

Figure 4.18 shows the violent crime density of Bangalore city crime data. The Violent Crime class includes 15 different types of crime keywords such as Rape, Murder, Terrorism, Kidnapping, Assault, Sexual Harassment, Sexual assault, Homicide, Gunshot, Intentional Killing peoples, Shootout, Gang-rape, Attempt to murder, Sexual abuse, Putting to death. The total density of Violent crimes identified as 662 and also violent crimes is the highest crime rate in the Bangalore context. Figure 4.18 shows that the central region of Bangalore more on specific Geographic locations like Corporation circle, K R Market, Madiwala, Marathahalli, Whitefield, Sivaji Nagar, etc. are most affected by crime. There are also areas in the eastern region of Bangalore that are affected by this crime.

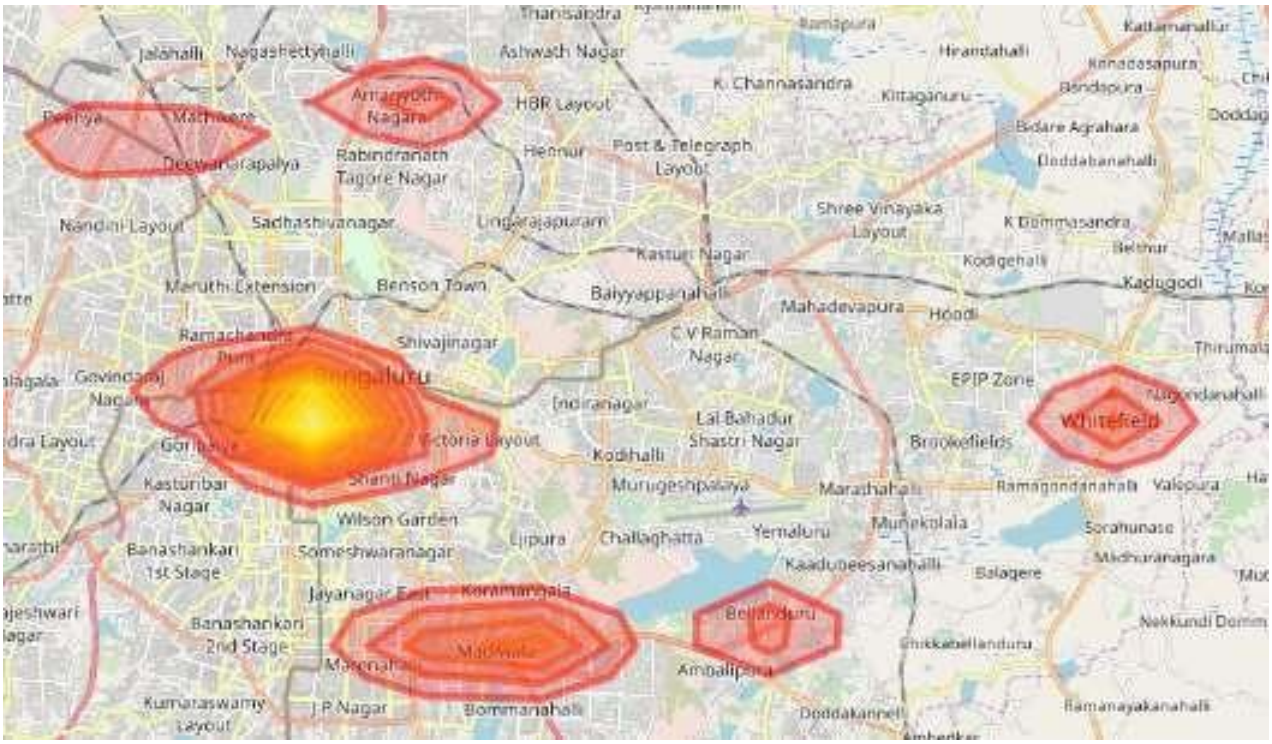


Figure 4.18 Crime density identification for violent crimes - Bengaluru

Crime Density Analysis – Bengaluru Crime Branch Data

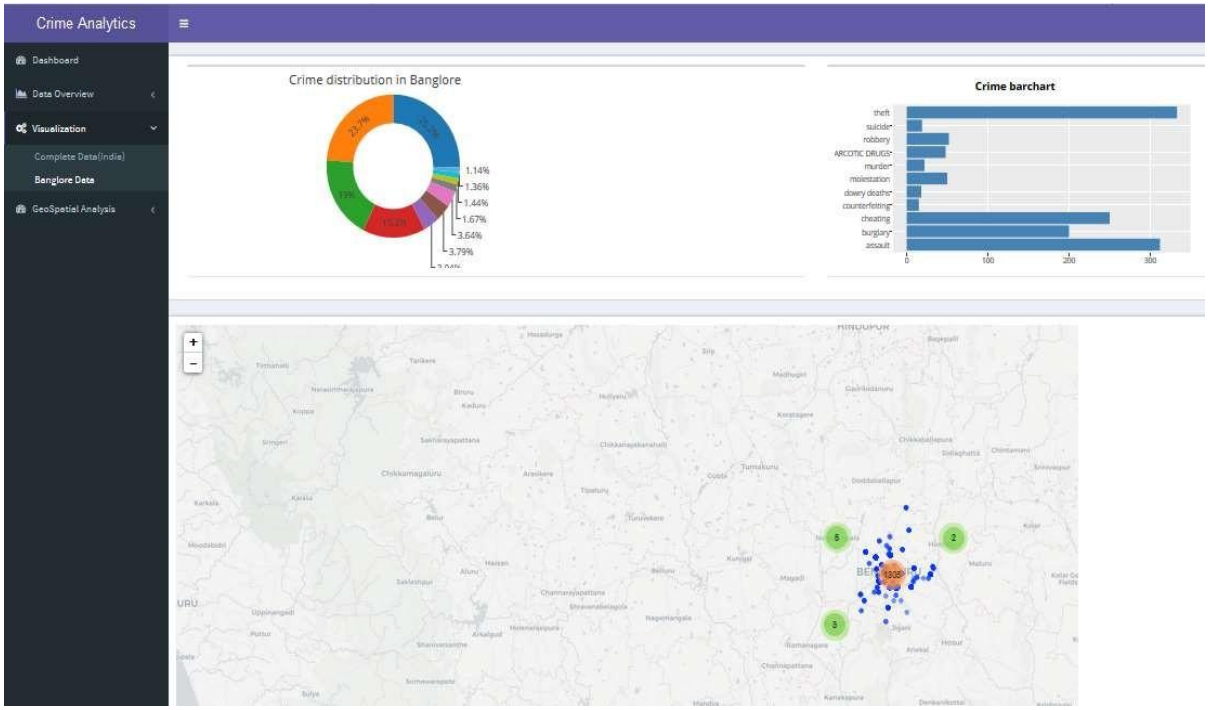


Figure 4.19 Crime density analysis Bengaluru crime branch data

The crime density obtained through newsfeed data is compared with Bangalore crime branch data. We had identified a significant correlation between these two data, both overall and by crime type. This study shows that the KDE model proposed by us is accurate. We have received the crime branch data through available local police records by applying RTI [Appendix-B]. The first major challenge we faced is the collection of crime branch data from police authorities. The second major challenge we identified is newsfeeds update. Content from newspapers such as The Hindu, Times of India does not have continuity of the crime. There is also not complete information about the crime incident such as the mode of crime, time of the crime, etc. Table 4.4 shows the validation of the proposed model crime density with official data-RTI crime density. It is identified that Crime density count is not matching with official data-RTI(Appendix-B) Crime density count. Still, the highest density sequence is matching only popular news gets published in the contemporary newspapers.

Table 4.4 Validation of Prosed crime density with RTI Data

Crime head	Actual count from work	Percentage of Crime head with actual work	RTI Crime headcount	Percentage of RTI Crime head
Theft	333	19.78609626	10966	53.01426154
Fraud	250	14.85442662	3199	15.46531303
Burglary	300	17.82531194	1373	6.63766014
Assault	312	18.53832442	1050	5.076142132
Kidnap	50	2.970885324	1046	5.056804448
Robbery	150	8.912655971	986	4.766739183
Molested	86	5.109922757	976	4.718394972
Drunkenness	59	3.505644682	354	1.711385062
Riots	20	1.18835413	271	1.310128112
Murder	59	3.505644682	234	1.131254532
Suicide	19	1.128936423	172	0.831520425
Dowry	30	1.782531194	48	0.232052212
Counter Feiting	15	0.891265597	10	0.048344211
Total	1683		20685	

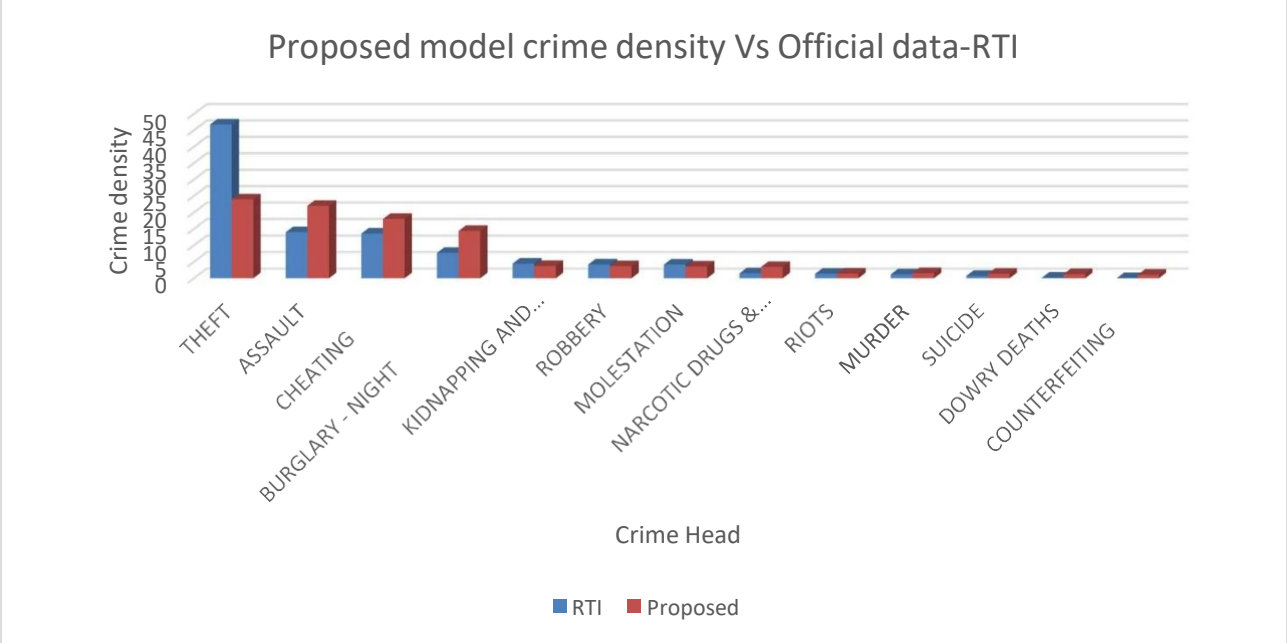


Figure 4.20 Validation of Prosed crime density with RTI Data-Graph-1

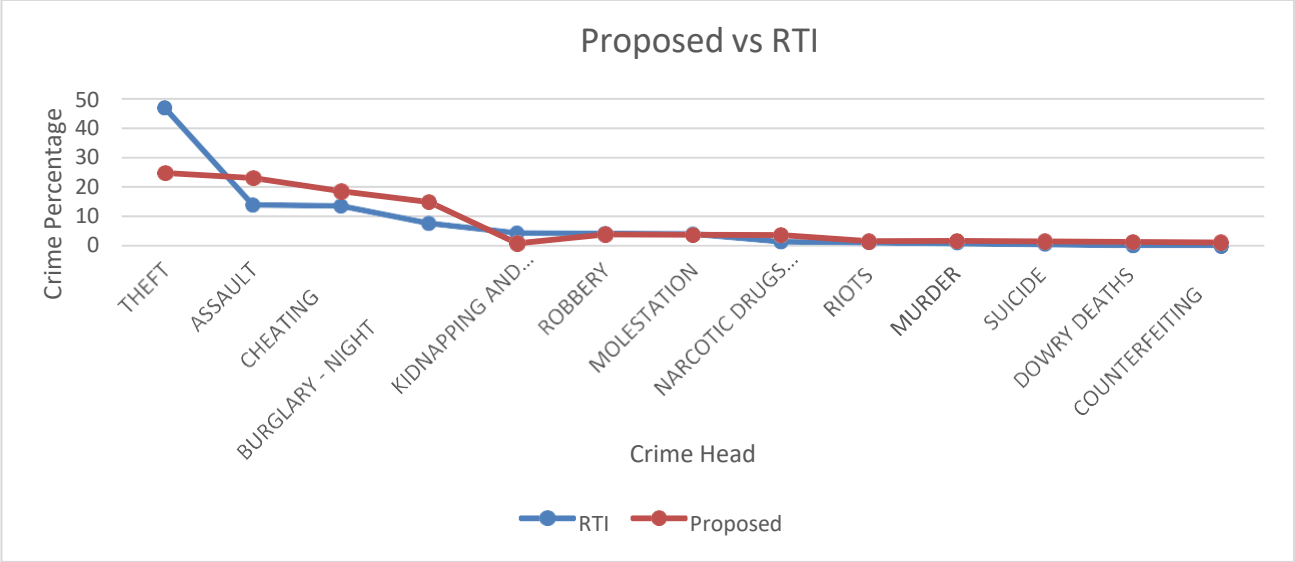


Figure 4.21 Validation of Prosed crime density with RTI Data Graph-2

We have analyzed the crime count for a particular crime and found a good correlation between the proposed model and RTI(Appendix-B) data. Figure 4.20 & Figure 4.21 shows the correlation between the proposed model and official data-RTI. The correlation is the highest for the theft category of crime in both cases. The data from RTI and news feed data are analyzed for matching words for the crime. We also classified minor crimes in the theft category, such as burglaries, snapping chains, etc. Certain words of crime are kept in line with the RTI data. RTI data have verified the info from our algorithm. The frequency of crime matches with that of newsfeed data we have used. Table 4.5 shows the crime density within Bangalore geographical location. The geographical crime density is matched with official data-RTI (Appendix-B).

The location information is obtained from the scrapping of newspaper articles. From the proposed research analysis, it is found that the Kempegowda-Majestic area has a higher number of crimes followed by the corporation circle and K R Market(Appendix-B). We have found a direct correlation between crime rates and population density in the Bangalore region. The crime density also depends on the mixture of the immigrant populations from other states in a particular region. When

diversity is more, crime rates are too high. There is also a good correlation between crime density from RTI data and geographical crime density.

Table 4.5 Geographical Crime density -Bangalore

Geo Graphical Location(Bangalore)	Crime Density
Jalahalli	14
Hebbala	95
Yeshwanthpura	110
Yelahanka	19
Kempegowda-Majestic	170
Corporation Circle	139
Chickpete	89
K R Market	150
Indiranagar	20
Mahadevapura	10
Bellandur	66
Whitefield	95
Madiwala	190
Srinagar	40
Electronic city- Wipro	21
Jayanagar	18
Gottigere	10

CHAPTER SUMMARY

This chapter elaborates on the Kernel density method utilized in the analysis of crime data. The crime density is measured for India and then Bangalore. It has found thatthe crime density estimated through news feed data, and the one determined through Bangalore crime data are correlated well. The news feed analysis model can be usedto understand the crime occurrences in the city.

Some part of this chapter are published in following journals :

1. Boppuru Rudra Prathap, Ramesha K, "Geospatial crime analysis to determine crime density using KDE for the Indian context", ***Journal of Computational and Theoretical Nanoscience (JCTN)***. (**In Press**) **Scopus indexed Journal**.

CHAPTER 5

TIME SERIES ANALYSIS AND FORECASTING USING ARIMA MODEL

INTRODUCTION

Time series forecasting is a method in which data collected regarding a particular event, and a model is generated to represent the underlying relationship (Chen & Yuan, 2008). The model is then used to forecast the future values of the event through time series extrapolation. This method is useful in estimating future behaviuor when there is no real correlation identified. Autoregressive integrated moving average (ARIMA) model is the most widely used time series models. In the ARIMA model, future values are the linear projections of the past value. The application of nonlinear forecasting is minimal.

ARIMA is used to capture even the complex relationships since it can take error terms and observations of the lagged terms too. These models are based on regressing a variable on past values (Hiropoulos & Porter, 2014). The essence of the ARIMA model is that past time points of time series data can impact current and future time points. ARIMA models use this concept to forecast current as well as future values. ARIMA uses several lagged observations of time series to predict observations. A specific weight is applied to each of the past terms, and the weights can vary based on how recent they are.

AR(x) means x lagged error terms are going to be used in the ARIMA model. ARIMA relies on Auto Regression. Autoregression is a process of regressing a variable on past values of itself (McClendon & Meghanathan, 2015). Autocorrelations gradually decay and estimate the degree to which white noise characterizes a series of data.

The work by Box and Jenkins has developed a practical method to implement

ARIMA models. This method works in three iterative steps. It includes model identification as the first step, parameter estimation as the second step, and diagnostic checking as the third step. Model identification ensures that the time series generated will have auto correlational properties (Nau, 2017). The data is transformed into the model identification step to make the stationary time series. Once the approximate model is developed, parameter estimation is done to reduce the overall amount of errors. The model adequacy is then checked with the help of diagnostic checking. This ensures that the model's future predictions fit with the historical data. This three-step iterative process is performed multiple times to identify the right model fit. The final selected model can then be used for prediction purposes (Irvin-Erickson & La Vigne, 2015).

Auto Regressive Integrated Moving Average model.

$$X(t) = A(1) * X(t-1) + E(t) \qquad (5.3)$$

Equation (3), X(t) is the time series that is being investigated, A(1) is the order 1 autoregressive parameter, X (t-1) is the time series which has by lagged 1 period, E(t) is the model's error term.

METHODOLOGY

Time series forecasting is a method in which a future behaviour is predicted based on past data of a particular event. This is done by understanding the underlying relationship between data. The relationship model is utilized to forecast future behaviour in the form of a time series. Autoregressive integrated moving average (ARIMA) model is the most commonly used time series models. ARIMA model is favoured over other models because it has Incorporated Box - Jenkins methodology and statistical properties (Skilling & Rogers, 2019). ARIMA is a linear time series forecasting tool that helps in predicting future occurrences. The data is populated on a linear time scale, and the ARIMA forecasting model is used to extrapolate the data. ARIMA model identifies the correlation between crime data over the last year and

predicts the future value of the crime data. This research considered the data from November 2017 to July 2018 as the data before that has incorrect information.

ANALYTICAL MODEL

Autocorrelation refers to how correlated a time series is with its past values, whereas the ACF is the plot used to see the correlation between the points, up to and including the lag unit. In ACF, the correlation coefficient is in the x-axis, whereas, the number of lags is shown in the y-axis.

Suppose that $\{X_t\}$ is a stationary time series. Its mean is $\mu = E[X_t]$. Its autocovariance function is

$$\gamma(h) = Cov(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)] \quad (5.4)$$

The autocorrelation function (ACF) of $\{X_t\}$ was defined similarly to that of the function $\rho(h)$ whose value at lag h is:

$$p(h) = \frac{\gamma(h)}{\gamma(0)} \quad (5.5)$$

The ACF and ACVF give a good measure of the degree of dependence in the values of time series during different times (Zhang, 2019). This plays a major role in predicting the future values of the series with the help of the past and the present values (Charpentier & Gallic, 2014).

Linear differential equations define ARIMA processes with constant coefficients.

CASE STUDY-3

This section discusses about the case study on crime prediction with respect to different time periods on Indian and Bangalore context. ARIMA model was used for the prediction of the crime rate.

CRIME FORECASTING – INDIA

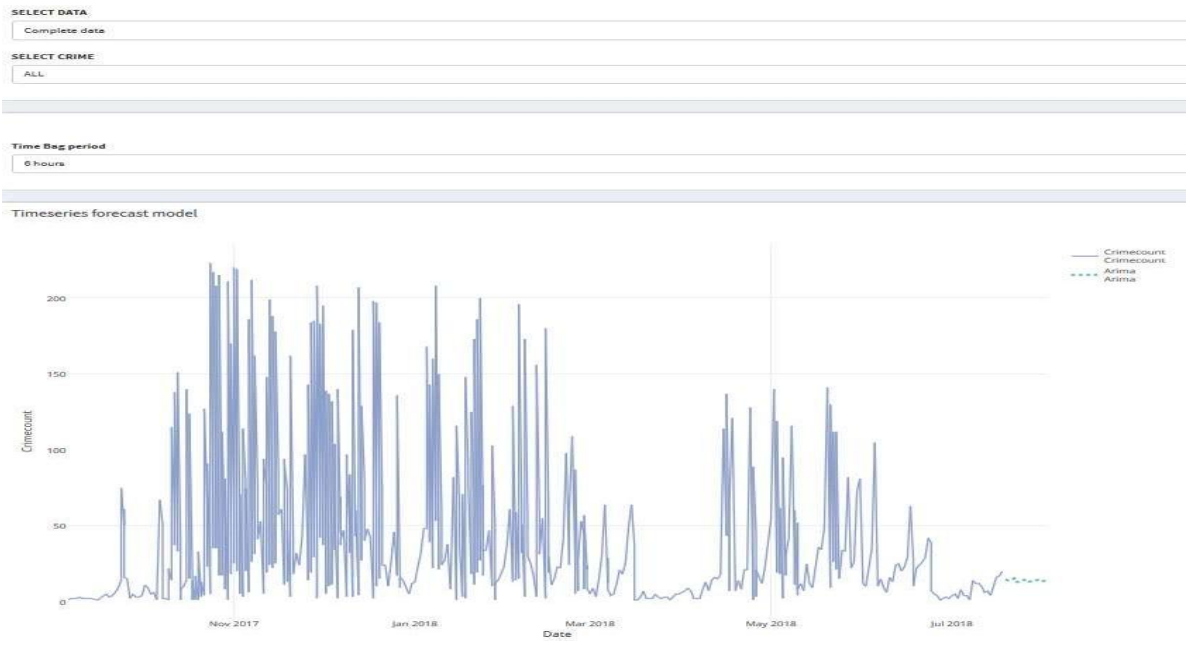


Figure 5.1 Crime forecasting analysis India- 6 Hours

Figure 5.1 shows the time series forecasting of crime occurrences in India for 6 hours. The graph represents 2 types of lines dotted and thick line. The dotted line represents the prediction of crime with respect to time in the Indian context. Thick line represents the crimes identified in a specific period. It is found that the crime occurrences are more in the March time period. This research method and data is verified with RTI information from the crime branch. There it is found to be a close correlation in the predictive model. The crime prediction count value is 10 crimes in every 6 hours of the timeline.

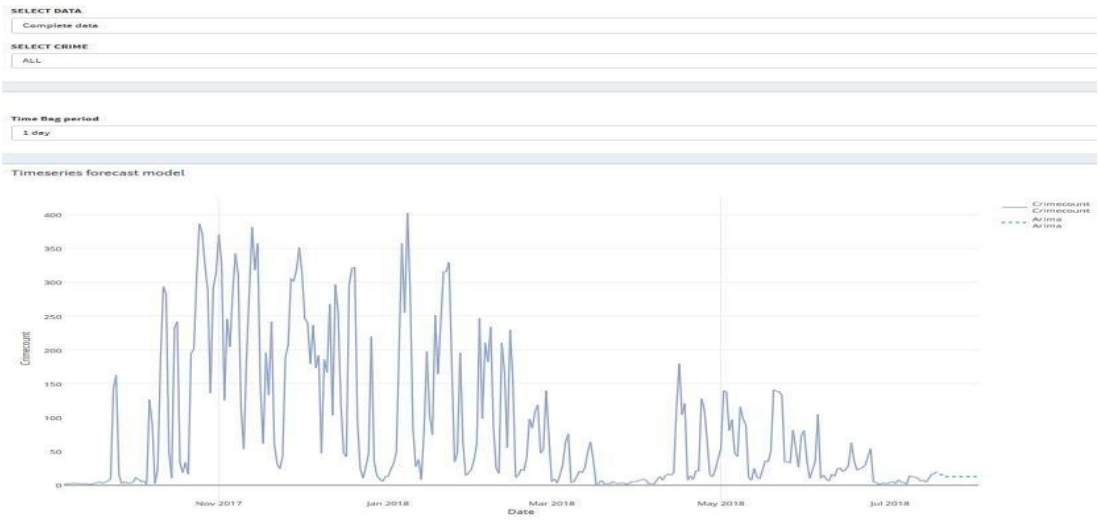


Figure 5.2 Crime forecasting analysis India- 1 Day

Figure 5.2 shows the time series forecasting of crime occurrences in India for the whole day. This has taken into account all the crime data. The crime count value predicted for this data is 5.

Figure 5.3 shows the crime prediction rate for all crimes occurring in India for three days. The ARIMA model value for this is 6.

Figure 5.4 shows the time series forecasting of crime occurrences in India for 1 Week time line. It is found that the crime occurrences are going to be more during March. The crime count value is 6 crimes for a week's timeline.

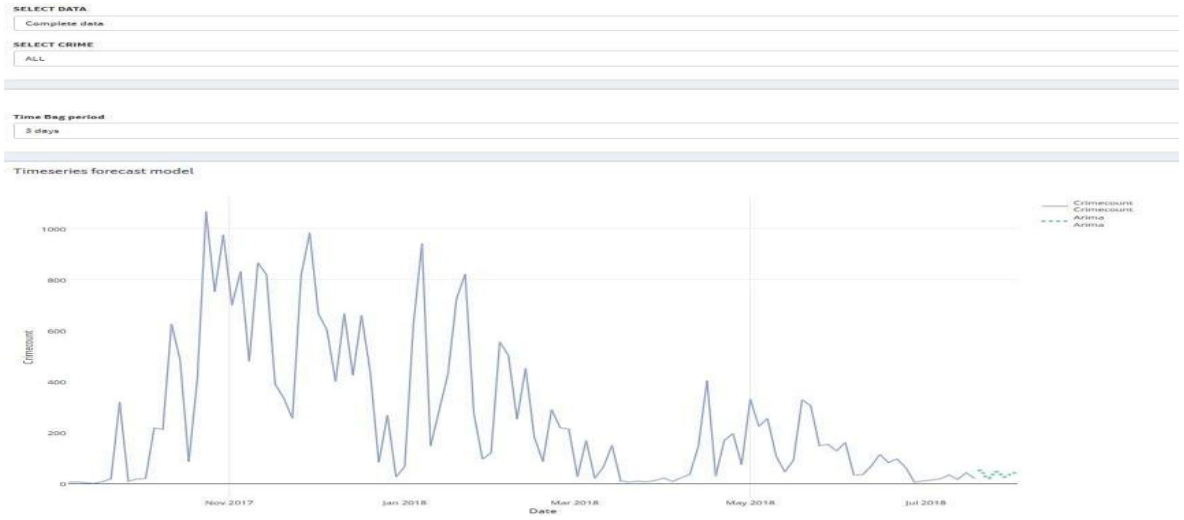


Figure 5.3 Crime forecasting analysis India- 3 Days

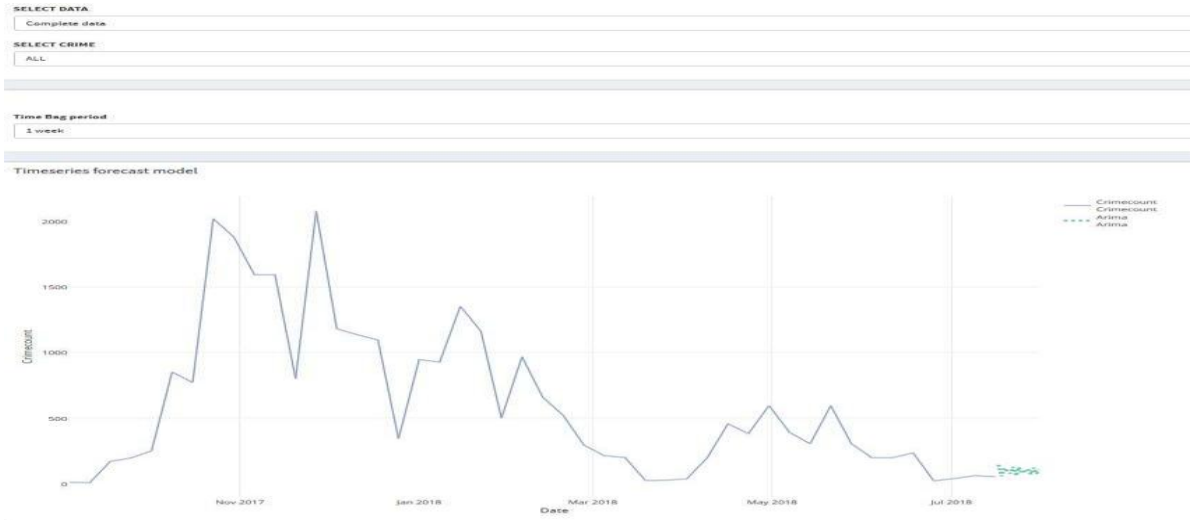


Figure 5.4 Crime forecasting analysis India- 1Week

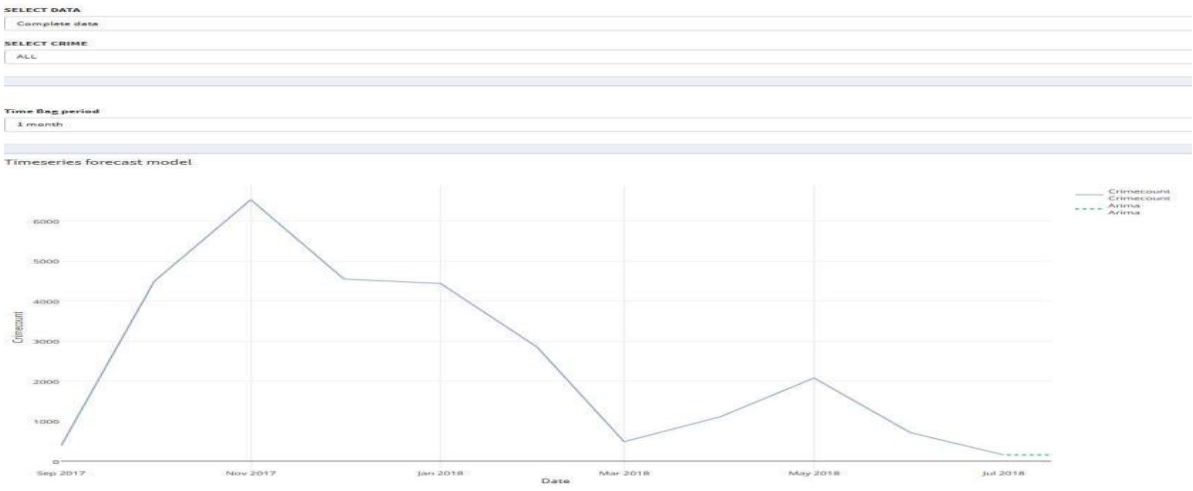


Figure 5.5 Crime forecasting analysis India- 1 Month

Figure 5.5 shows the time series forecasting for all crimes occurring in India for three days. The ARIMA model value for this is 25.

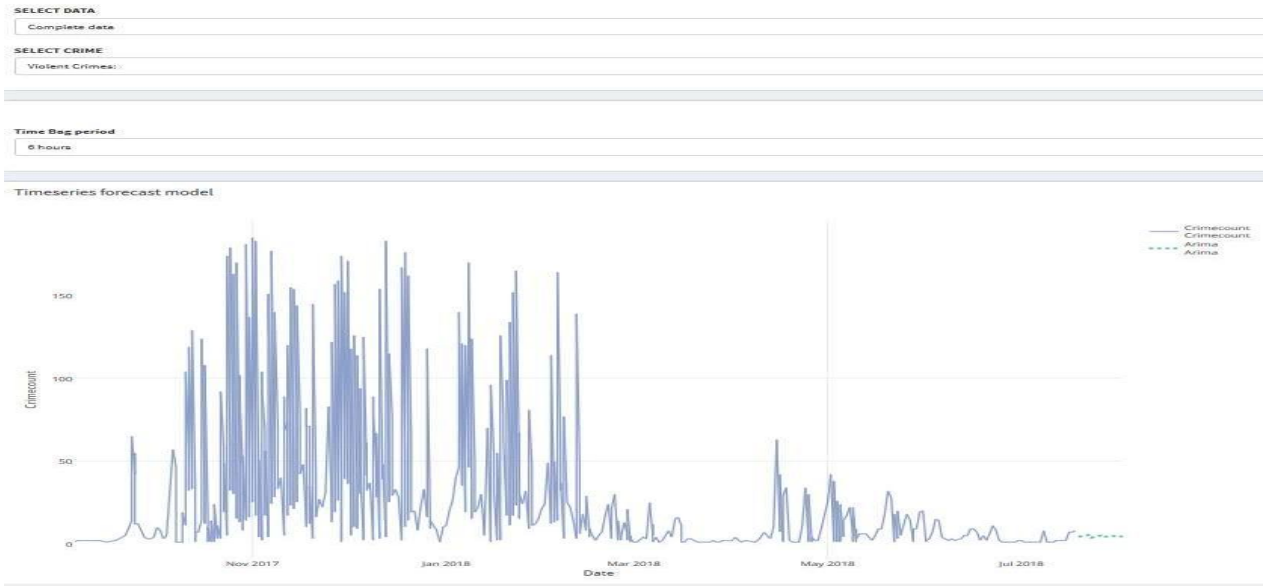


Figure 5.6 Violent Crime forecasting analysis India- 6 Hours

Figure 5.6 shows the crime prediction rate for violent crimes occurring in India for 6 hours. The ARIMA model value for this is 5.

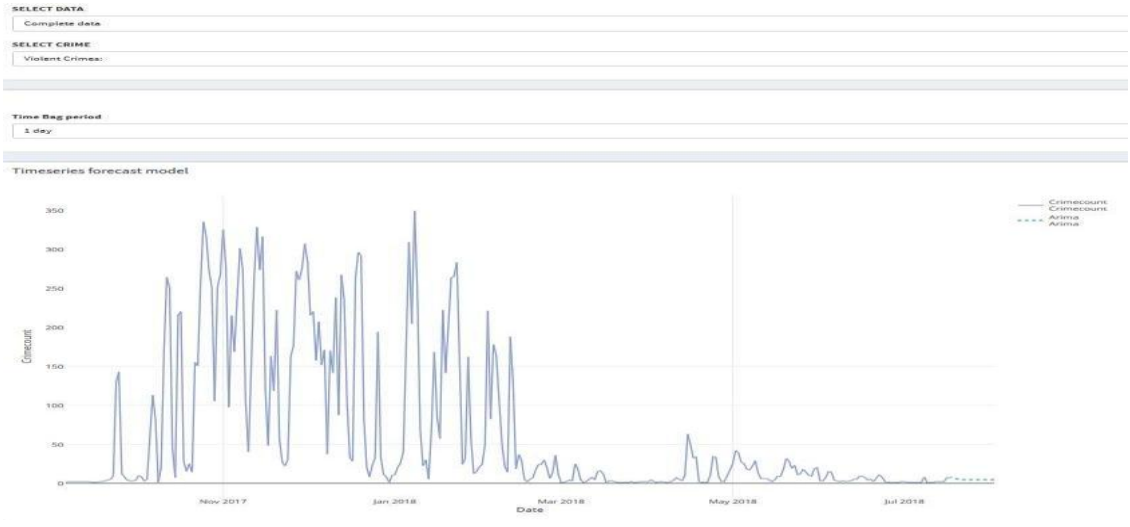


Figure 5.7 Violent Crime forecasting analysis India- 1 Day

Figure 5.7 shows the crime prediction model for India for one-day time bag. The crime count value of violent crime predicted for this data is 10.

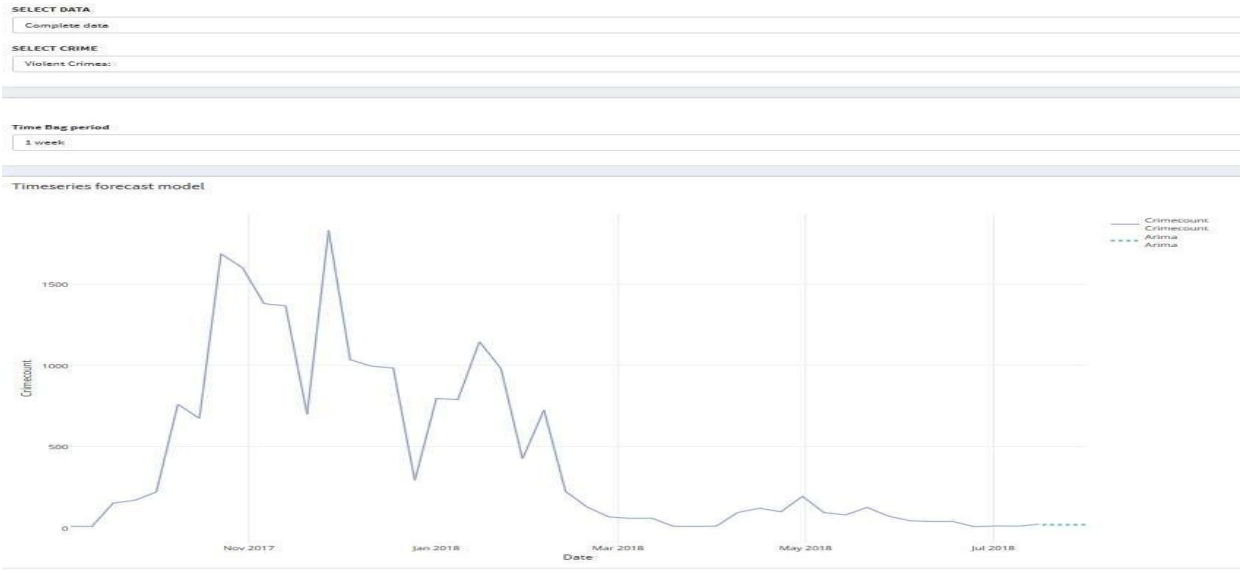


Figure 5.8 Violent Crime forecasting analysis India- 1 Week

Figure 5.8 shows the prediction of violent crimes occurring in India for 1 week. This accounts for the 100 crimes as the prediction output in a week period time bag.

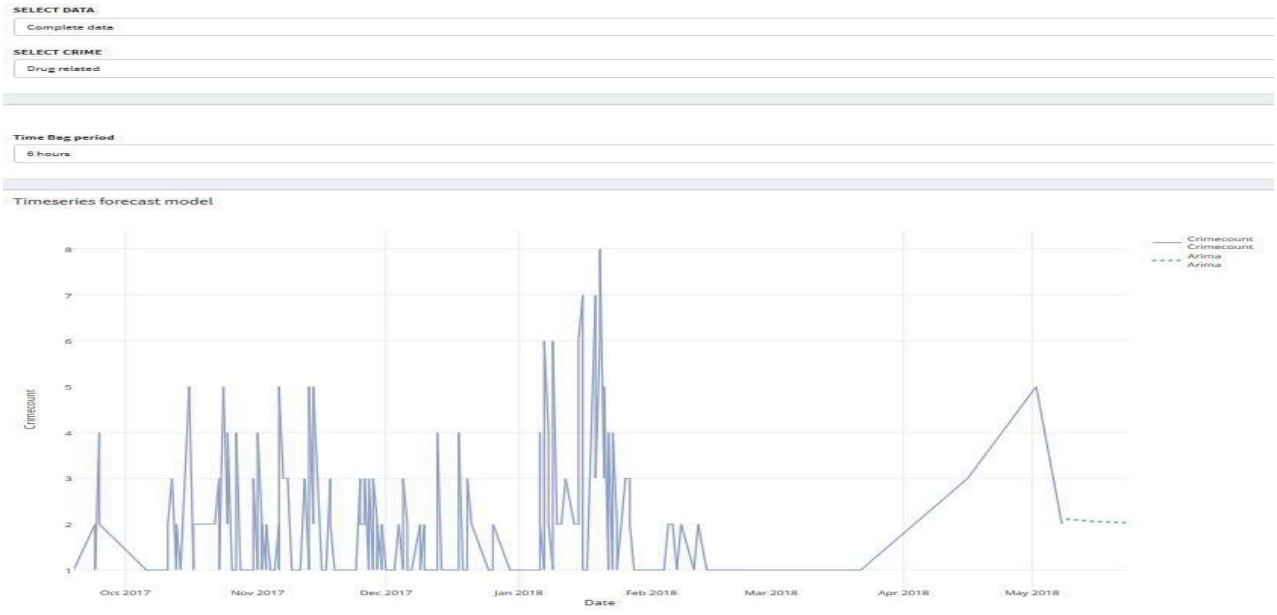


Figure 5.9 Drug-related Crime forecasting analysis India- 6 Hours

Figure 5.9 shows the crime prediction rate for drug-related crimes occurring in India for 6 hours. The forecasting value using the ARIMA model is 3 means in every 6 hours possible 3 drug-related crimes are happening in India.

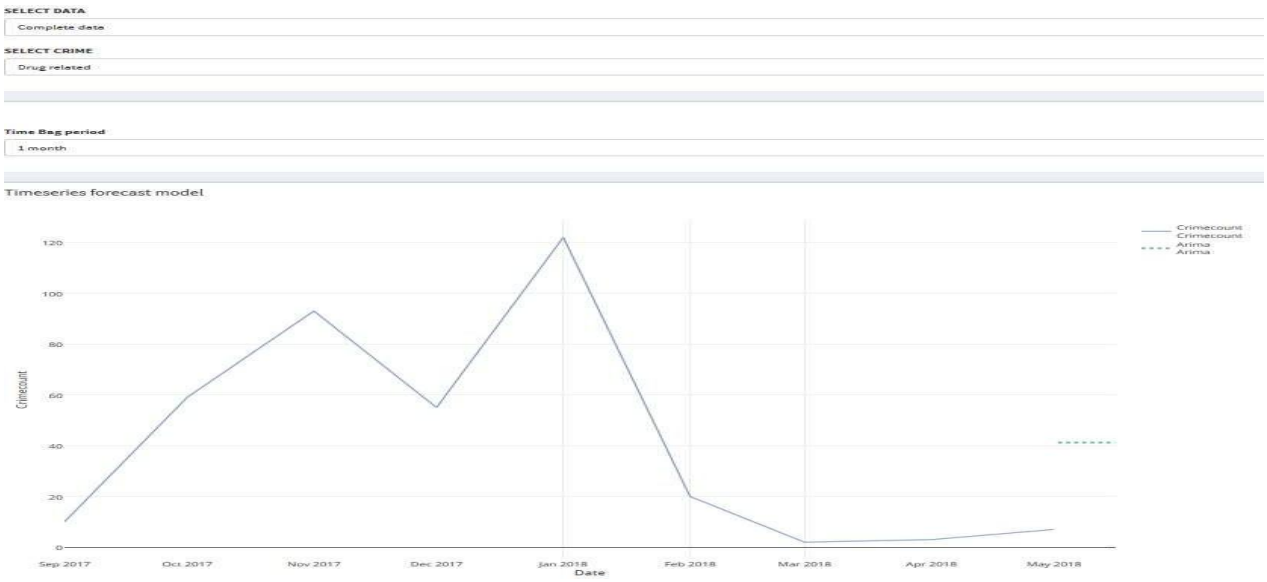


Figure 5.10 Drug-related Crime forecasting analysis India- 1 Month

The figure shows the Drug-related crime prediction model for India's entire Month. This has taken into account all the crime data. The Drug-related crime count value predicted for this data is 5. This shows that there is a possibility of a minimum 5 drug-related crimes that may possibly in one month of time bag.

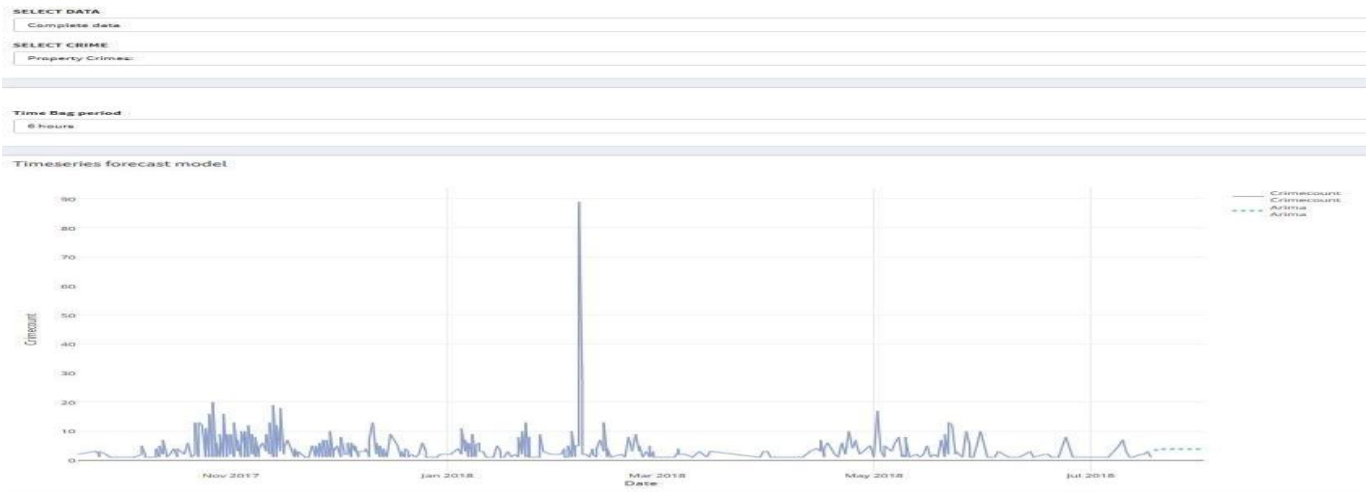


Figure 5.11 Property related Crime forecasting analysis India- 6 Hours

Figure 5.11 shows the time series forecasting of crime occurrences in India for 6 hours. It is found that the crime occurrences are going to be more during March. The crime count value is 4 crimes for 6 hours timeline.

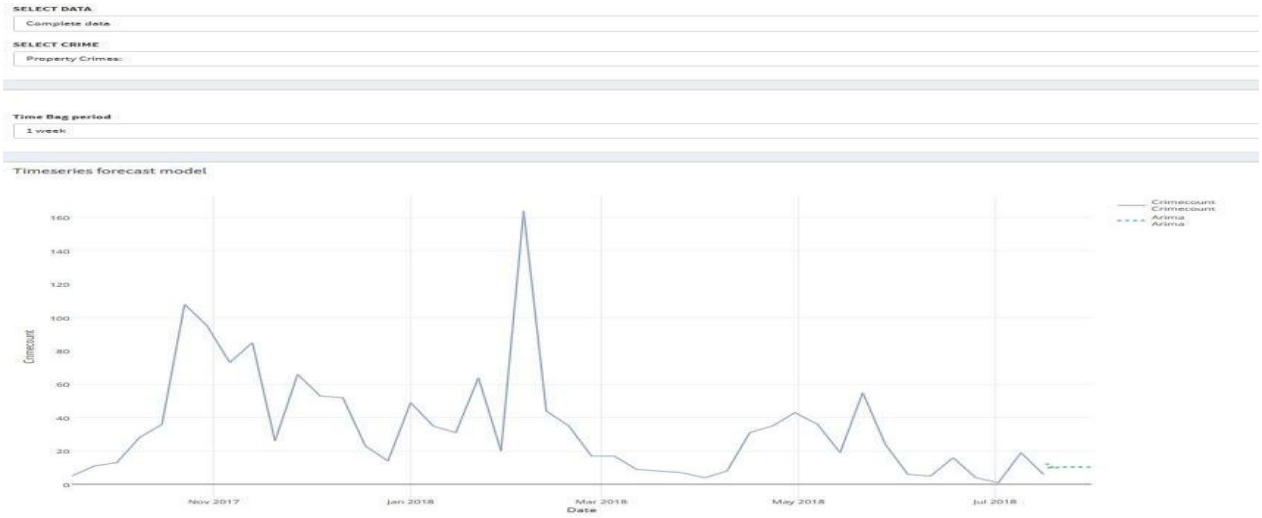


Figure 5.12 Property related Crime forecasting analysis India- 2 Weeks

Figure 5.12 shows the Property related crime prediction model for India for 2 weeks. This has taken into account all the crime data. The crime count value predicted for this data is 5.

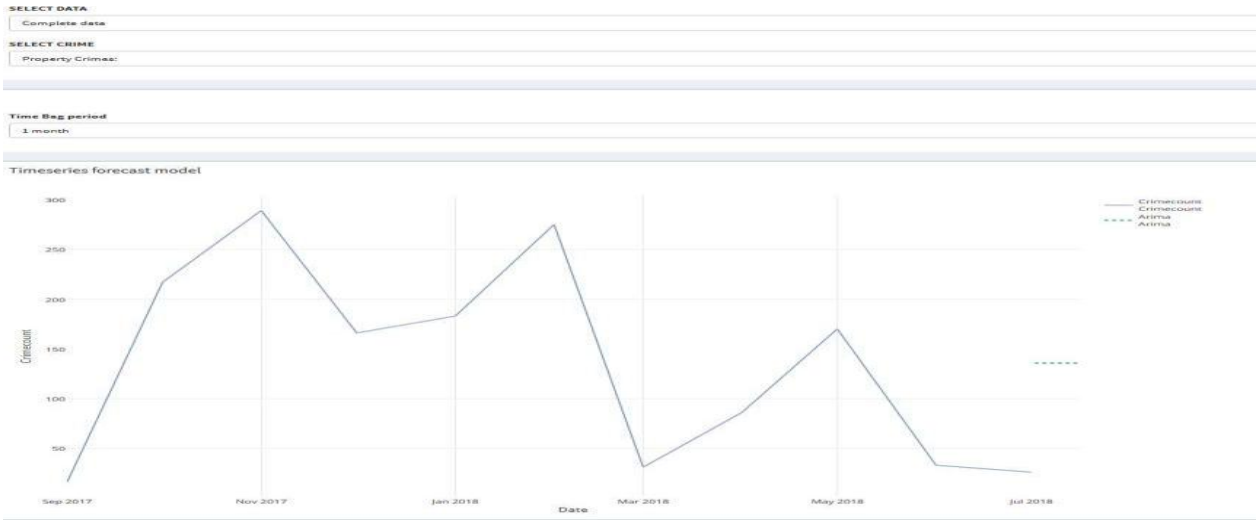


Figure 5.13 Property related Crime forecasting analysis India- 1 Month

Figure 5.13 shows the time series forecasting of Property related crime occurrences in India for 1 month. It is found that the crime occurrences will be more during March. The crime count value is 6 crimes for a month's timeline.

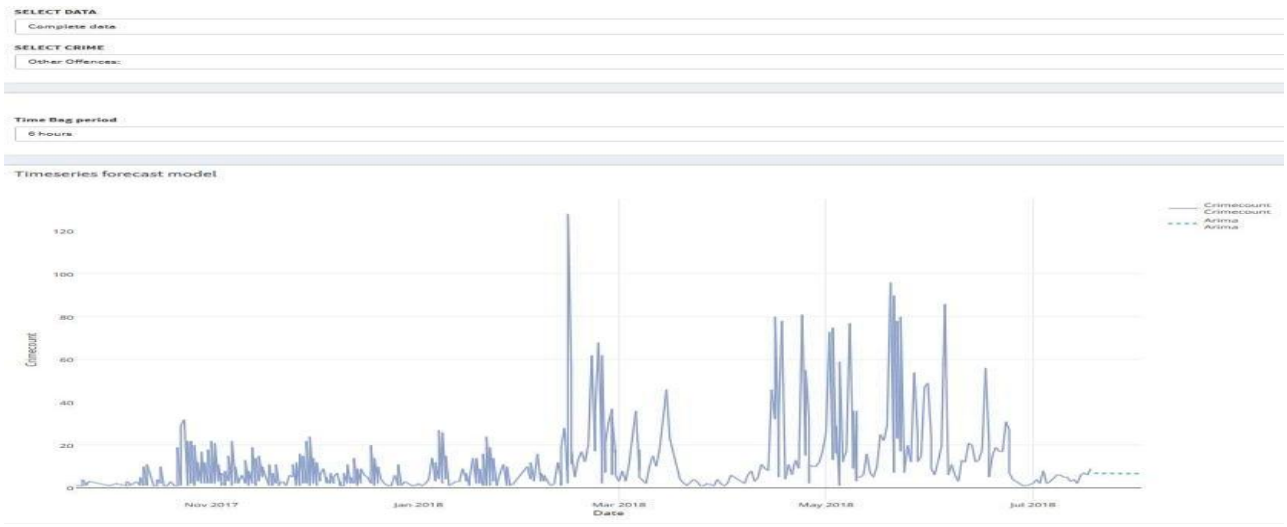


Figure 5.14 Other offenses forecasting analysis India- 6 Hours

Figure 5.14 shows the time series forecasting of other crime occurrences in India for 6 hours' time bag. It is found that crime occurrences are going to be more in March. The crime count value is 10 crimes for a month's timeline.

CRIME FORECASTING –BENGALURU

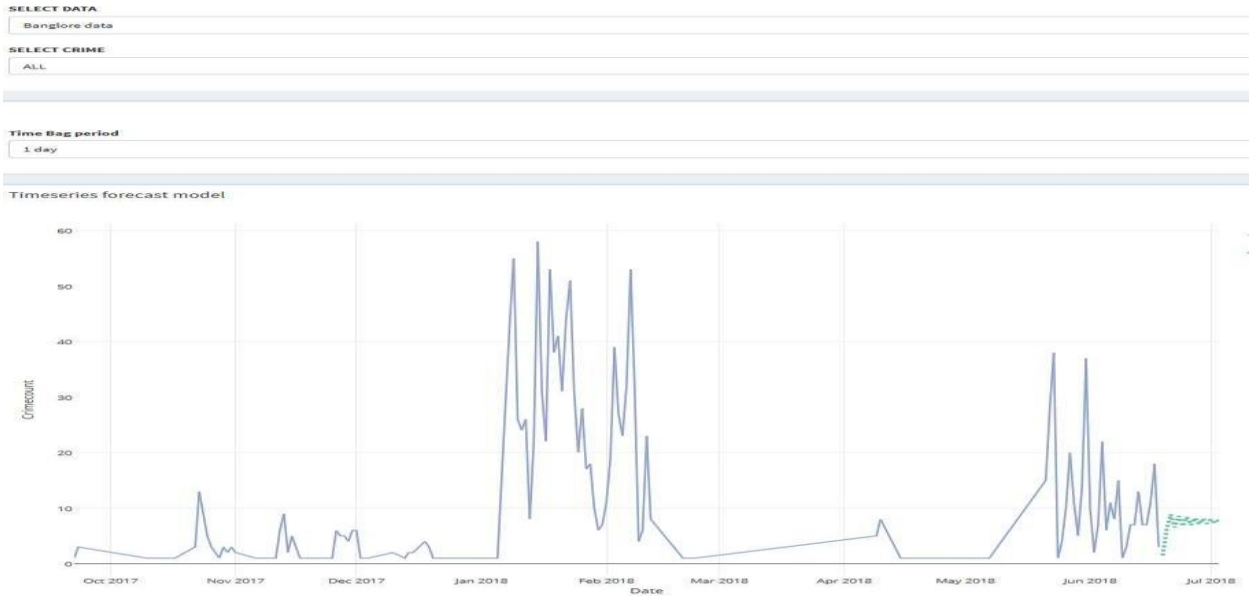


Figure 5.15 Crime forecasting analysis Bengaluru- 1Day

Figure 5.15 shows the time series forecasting of crime occurrences in Bengaluru. It is found that crime occurrences are going to be more in the January-February period. We verified this data with that of the RTI information from the crime branch. There is found to be a close correlation in the predictive model.

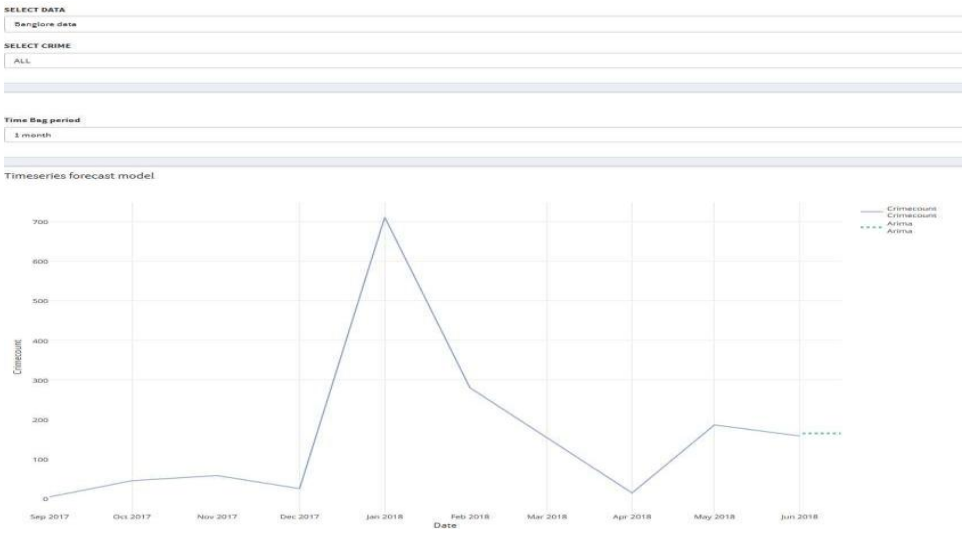


Figure 5.16 Crime forecasting analysis Bengaluru- 1Month

Figure 5.16 shows the time series forecasting of crime occurrences in Bangalore for 1 month. It is found that the crime occurrences are going to be more during March. The crime count value is 200 crimes for a month's timeline.

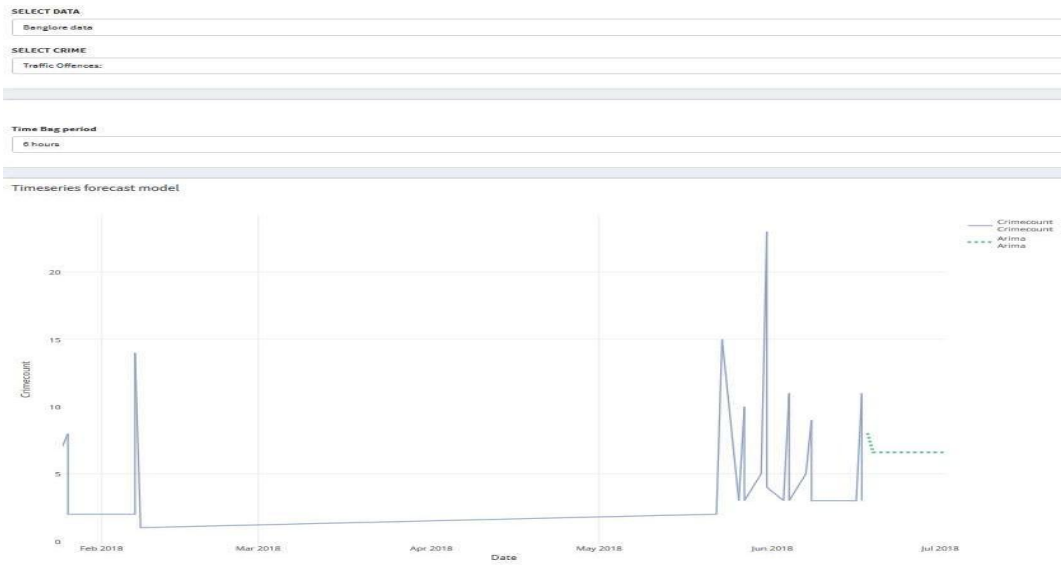


Figure 5.17 Traffic offenses forecasting analysis Bengaluru- 6 Hours

Figure 5.17 shows the time series forecasting of traffic crime occurrences in Bangalore for 6 hours month. It is found that the crime occurrences are going to be more during March. The crime count value is 10 crimes for a month's timeline.

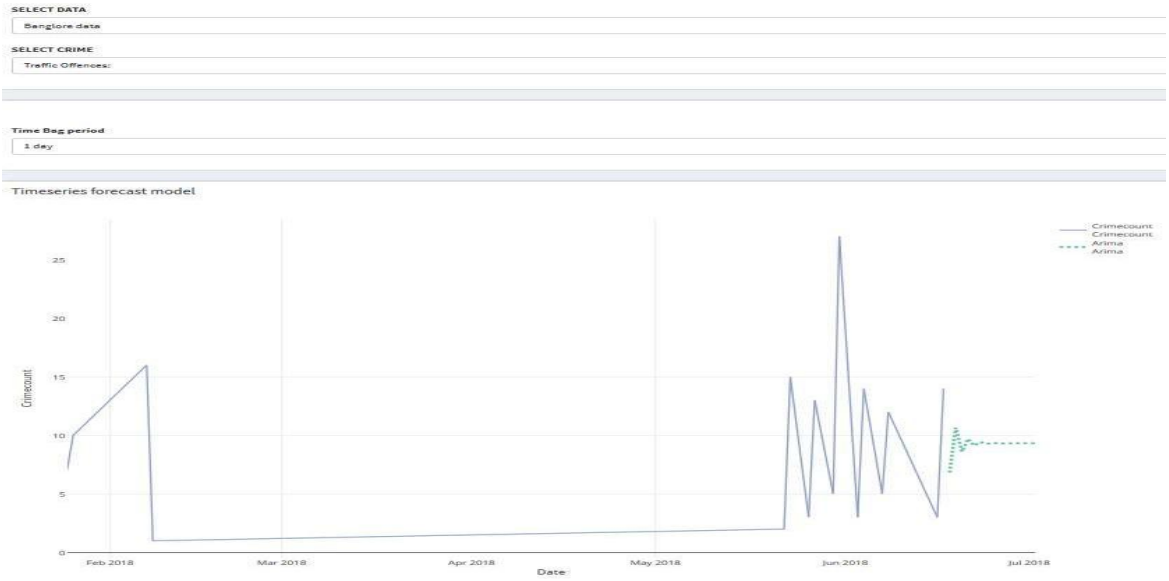


Figure 5.18 Traffic offenses forecasting analysis Bengaluru- 1 Day

Figure 5.18 shows the time series forecasting of Traffic offences occurrences in Bengaluru. It is found that the crime occurrences are going to be more in the January-February time period. The outcome of research is verified with RTI information from the crime branch. There is found to be a close correlation in the predictive model.

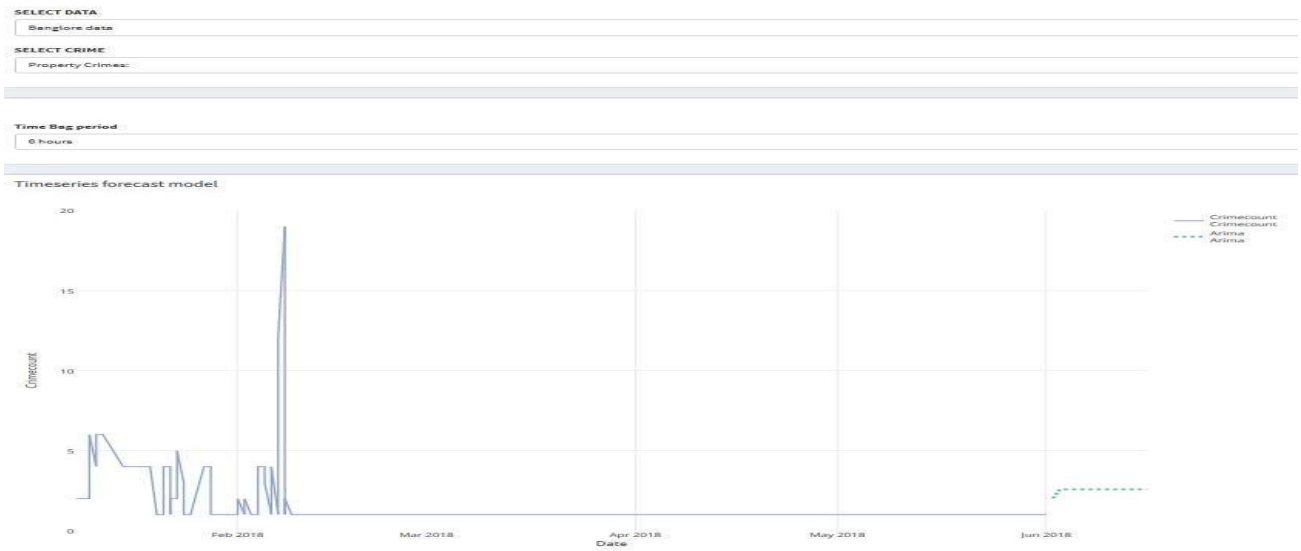


Figure 5.19 Property crime forecasting analysis Bengaluru- 6 Hours

Figure 5.19 shows the time series forecasting of Property crime occurrences in Bangalore for 6 hours month. It is found that crime occurrences are going to be more in March. The crime count value is 10 crimes for a month's timeline.

Figure 5.20 Property crime forecasting analysis Bengaluru- 1 Day

Figure 5.20 shows the time series forecasting of Property crime occurrences in Bengaluru. It is found that crime occurrences are going to be more in the January-February period. The future crimes are predicted at 3 crimes at the same timeline.

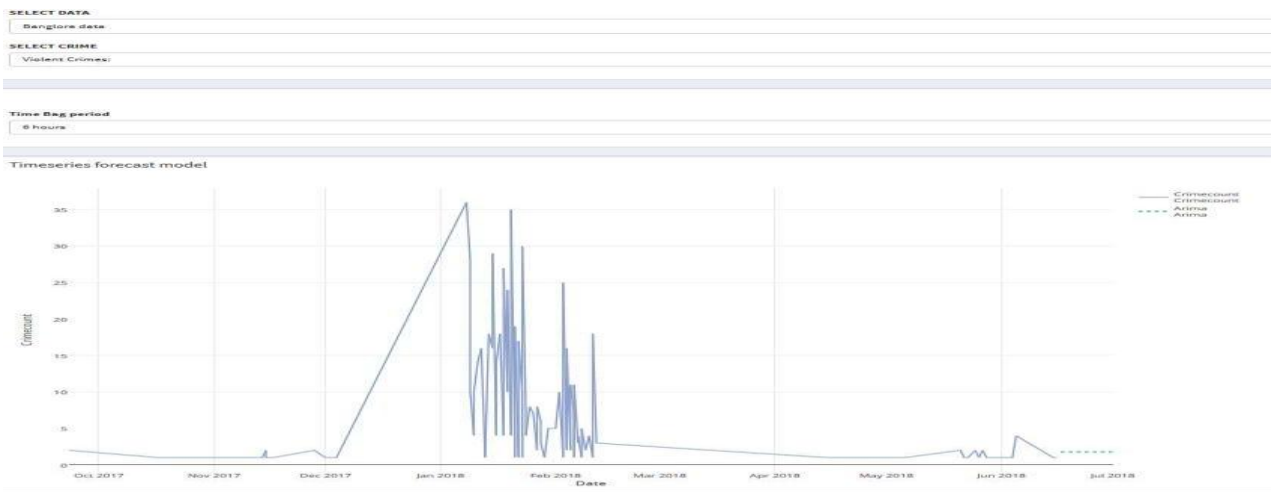


Figure 5.21 Violent crime forecasting analysis Bengaluru- 6 Hours

Figure 5.21 shows the time series forecasting of violent crime occurrences in Bangalore for 6 hours month. It is found that crime occurrences are going to be more in March. The crime count value is 10 crimes for a month's timeline.

Figure 5.22 shows the time series forecasting of Violent crime occurrences in Bengaluru. It is found that crime occurrences are going to be more in the January-February period.

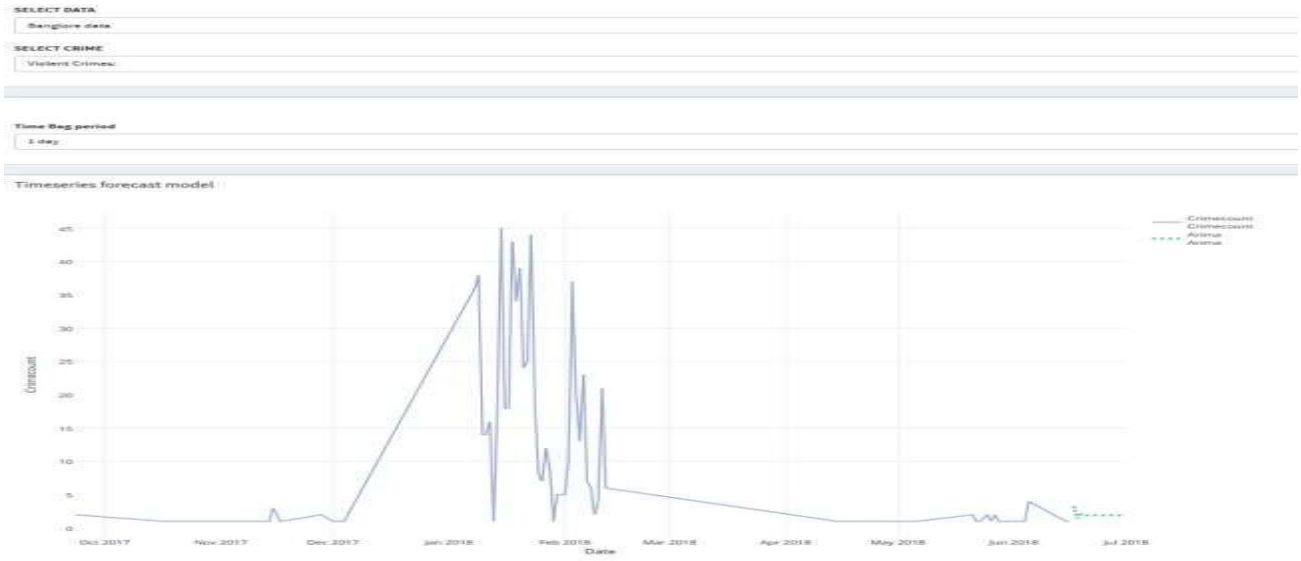


Figure 5.22 Violent crime forecasting analysis Bengaluru- 1 Day

Outcome of Researched is verified with that of the RTI information from the crime branch. There is found to be a close correlation in the predictive model. The future crimes are predicted at 3 crimes at the same timeline.

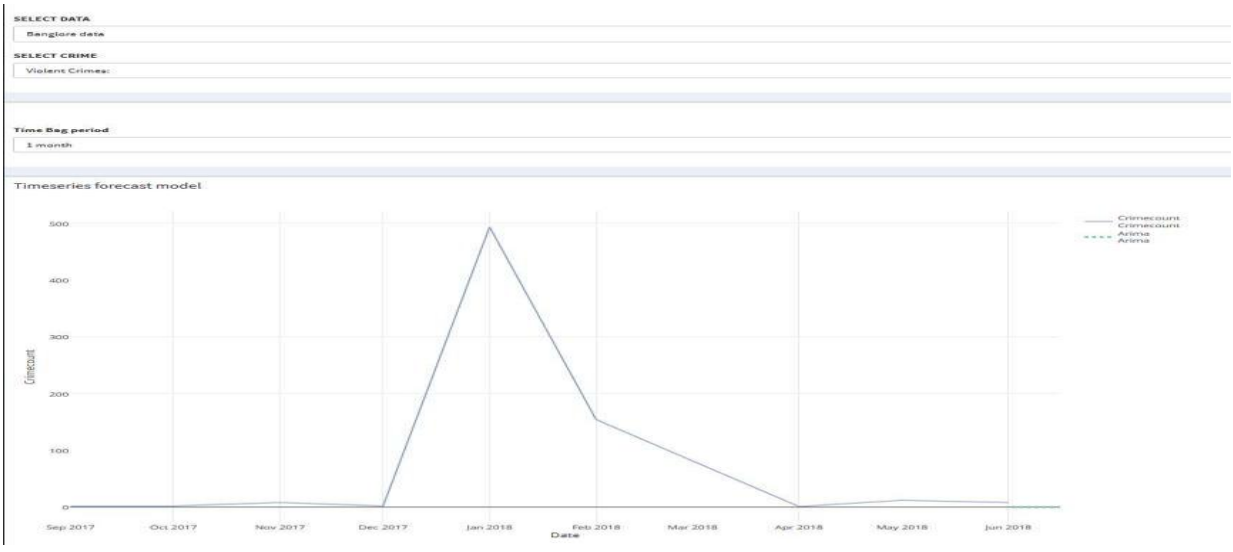


Figure 5.23 Violent crime forecasting analysis Bengaluru- 1 Month

Figure 5.223shows the time series forecasting of crime occurrences in Bengaluru. It is found that crime occurrences are going to be more in the January-February period. Outcome of research is verified with that of the RTI information from the crime branch. There is found to be a close correlation in the predictive model. The future crimes are predicted at 3 crimes at the same timeline.

Validation of Data with RTI –Percentage

Table 5.1 Validation of Proposed method vs Crime branch data

CRIME HEAD	PROPOSED MODEL CRIME	RTI CRIME
THEFT	46.78	24.7
ASSAULT	14	23
CHEATING	13.6	18.5
BURGLARY - NIGHT	7.734	14.8
KIDNAPPING AND ABDUCTION	4.462	0.74
ROBBERY	4.206	3.78
MOLESTATION	4.164	3.71
NARCOTIC DRUGS & PSHYCOTROPIC SUBSTANCES	1.51	3.56
RIOTS	1.365	1.48
MURDER	1.203	1.63
SUICIDE	0.734	1.41
DOWRY DEATHS	0.205	1.34
COUNTERFEITING	0.043	1.11

Table5.1 & Figure 5.24 shows the validation of research work with crime branch data. Research gathered the Bengaluru crime count using RTI(Right to information act), and outcome of research validated Proposed model count (newsfeed data) with that of RTI data that is collected from police agencies. Outcome of reseach found good matching between both data, as shown in the graph. The crime data for robbery, molestation, murder, etc. match with that of RTI data.

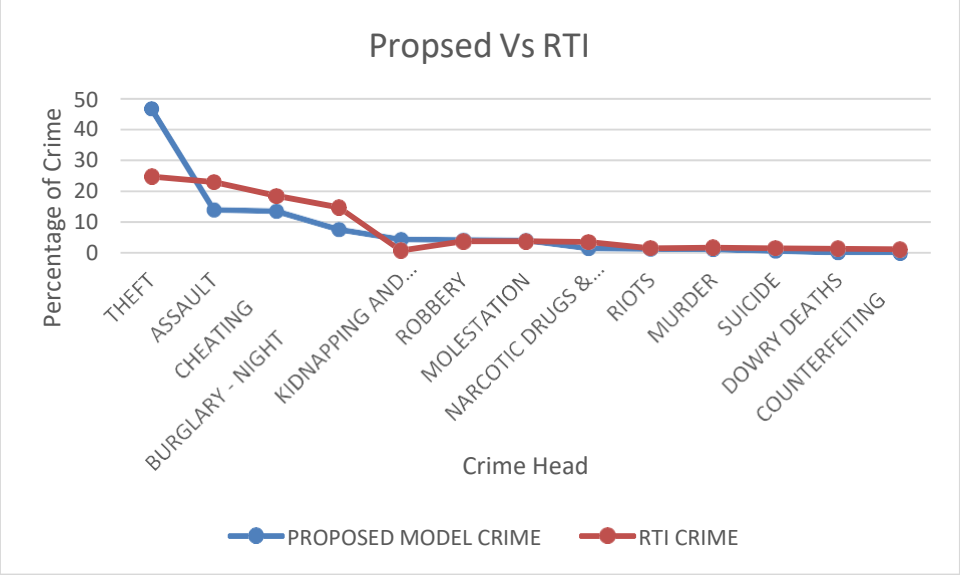


Figure 5.24 Validation of Proposed method vs Crime branch data

Comparison of work with ARIMA model

Table 5.2 Validation of Proposed method vs. ARIMA model

S.No	Day	Density Estimation (Proposed Method)	ARIMA Forecast Value	Percentage Matching
1	366	15	14	93.3
2	367	14	8.99	64.2
3	368	18	12.22	67.8
4	369	14	12.25	87.5
5	370	17	14	82.3
6	371	16	11	68.7
7	372	12	10	83.3
8	373	15	17	86.6
9	374	14	11	78.5
10	375	15	11	73.3
11	376	12	12	100
12	377	15	10	66.6
13	378	14	11	78.5
14	379	9	5	55.5
15	380	8	6	75

The accuracy of the proposed model is calculated by using the following formula

Accuracy= (Average (100-ABS (ARIMA Forecast model)*100/Proposed Model

Outcome of research is compared proposed research methodology with that of the traditional ARIMA model and have reached an average accuracy percentage of 77.49%. The accuracy has been higher on certain days, reaching up to even 100%. Figure5.24 shows the Validation of the Research work-ARIMA Model. Our model, based on news feed data, has performed with accuracy comparable to that of the traditional ARIMA model.

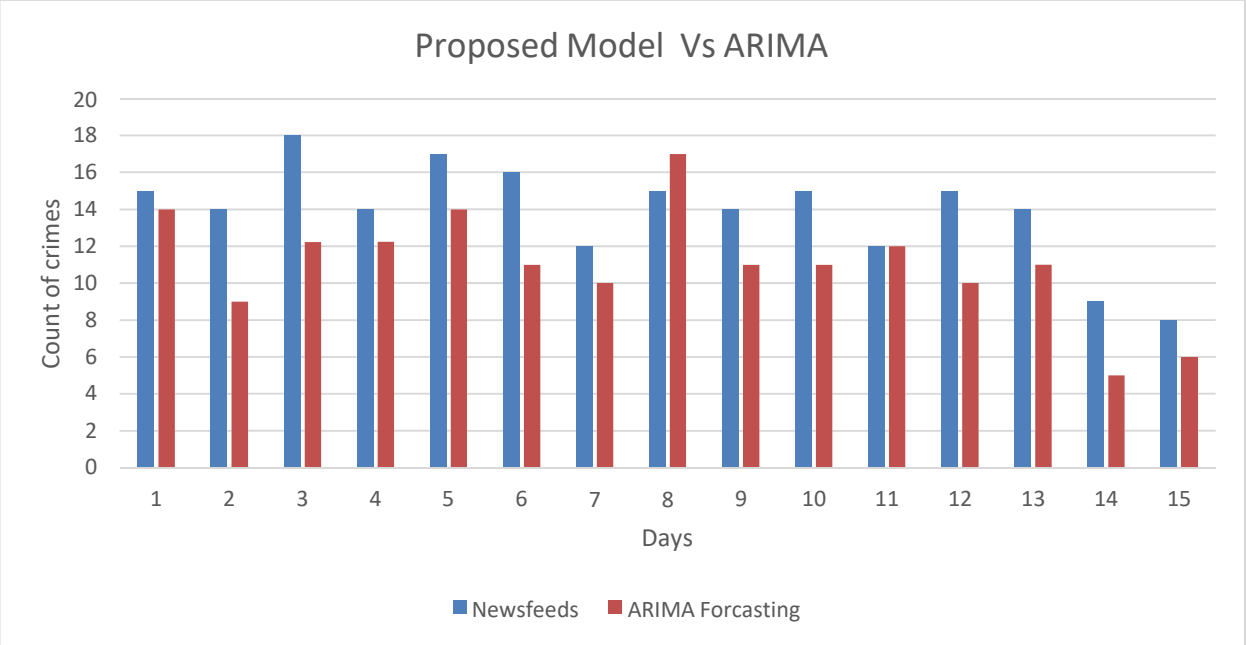


Figure 5.25 Validation of Research work-ARIMA Model

CHAPTER SUMMARY

The research work proposes a framework for crime analysis using Newsfeed data, mainly concentrates on the hotspot detection, Density identification of individual crime, Forecasting of crime using Time series analysis, in Indian and Bangalore context. This chapter elaborates on the time series forecasting implemented using ARIMA models. It is found that there is a close correlation between the predictive

model and the actual crime data. The proposed model predicts the possible crimes for the time span of 6 hours, one day, three Days, one Week, one Month to help the crime authority to take preventive measures well in advance. Our results have shown that the news feed data can be used for extracting spatiotemporal information about the prediction performance of 68 types of crime. The outcome of research achieved a prediction accuracy of 77.49% with our crime prediction models and also outcome of research had validated with proposed crime prediction model with that of the ARIMA model and found equivalent prediction performance.

Some part of this chapter are published in following journals :

1. Boppuru Rudra Prathap, Ramesha K, "Spatio-Temporal Crime Analysis Using KDE and ARIMA Models In Indian Context" ***International Journal of Digital Crime and Forensics (IJDCF)***. **(In Press) Scopus & Web of Sciences Indexed Journal.**

CHAPTER 6

SUMMARY AND CONCULSIONS

SUMMARY

This chapter serves as a summary of the research work carried out. The need for Spatiotemporal crime analysis using news feed data in Indian and Bengaluru context has identified. The literature review revealed a lack of sound in crime analytics in the context of Bangalore. Based on the literature, it has determined that there is a requirement of Framework for crime analysis using social media(Newsfeed data) concerning spatial and temporal data. And also study says that the crime rate is increasing day by day, which demands Spatio-Temporal visualization techniques such as hotspots detections, Density identification and Forecasting for better Crime investigations. This research has given a brief introduction of the conceptual underpinning of a crime prediction system which includes data mining, spatial and temporal analysis, time series prediction, etc. with the help of the system research have proposed a method to identify crime hotspots in the area and predict crime using time series forecasting.

This research also discusses the methodology used to collect and clean the crime data from various internet sources. The underlying theory behind the data collection and cleaning are explained. Naïve Bayes classification algorithm is used for the classification of the crime into different classes. Mallet package is used for extracting the keywords from the newsfeeds. K-means algorithm is used to identify the hotspots in the crime locations. For research purpose considered the case study of crime analysis in India and Bangalore to implement Geospatial analysis methodologies. The primary reason for that is that crime rates in Bangalore has been rising more compared to that of other cities in India due to urbanization.

This research also elaborates on the Kernel density method utilized in the analysis of

crime data. The crime density is measured for India and then Bangalore. It is found that the crime density estimated through news feed data, and the one estimated through Bangalore crime data are correlated well. The news feed analysis model can be used to understand the crime occurrences in the city. The proposed approach has overcome the challenges in the existing KDE algorithm. It also estimates the crime density at different geographical locations such as Madivala, K R Market, etc. which matches RTI data.

This research also elaborates on the time series forecasting implemented using ARIMA models. It is found that there is a close correlation between the predictive model and the actual crime data. The proposed model predicts the possible crimes for the time span of 6 hours, one day, three Days, one Week, one Month to help the crime authority to take preventive measures well in advance. Our results have shown that the news feed data can be used for extracting spatiotemporal information about the prediction performance of 68 types of crime. We have achieved a prediction accuracy of 77.49% with our crime prediction models. We have validated our crime prediction model with that of the ARIMA model and found equivalent prediction performance.

KEY CONTRIBUTIONS

The key contributions of this research work include

1. Development of a Spatio-temporal crime analysis visualization tool which can portrait Crime Density In Indian and Bangalore context.
2. Classified various crime data for effective investigations and verified the crime data with official sources (Appendix-1).
3. The framework and the KDE algorithm was implemented and tested against a Casestudy which contains multiple scenarios like Indian Bangalore context.
4. To predict and evaluate the crimes using forecasting techniques.

PUBLICATION FROM THESIS CONTRIBUTION

Research work showcases the implementation of Spatio-temporal crime analysis on social media data. This research aims to develop a crime visualization tool that can able to perform the analysis of different crimes, provide the Crime density concerning Geo-Spatial, and forecasting time series data for different crimes in Indian and Bangalore perceptive.

Geo-Spatial Crime analysis Using Newsfeed data in Indian Context

The output of this research shown as crime hotspots. Other outcomes include crime maps and cartographic work for different geographic areas of Bangalore and India. The shift in criminal activity for one year period (2017 Jan to 2018 Jan) are compared and analyzed. The clustering of the crimes based on location is also done. A dropdown menu can be used to choose the crime type and view the corresponding visualization. The results help police in planning patrol in the identified locations. The experimental results show that models have a high level of accuracy, as verified with government records. Figure 6.1 shows the procedural steps followed in the work and detailed results discussed in the Results and discussion. Figure 6.2 Shows the Geo-Spatial Crime identifications in the Indian context with hotspot identification.

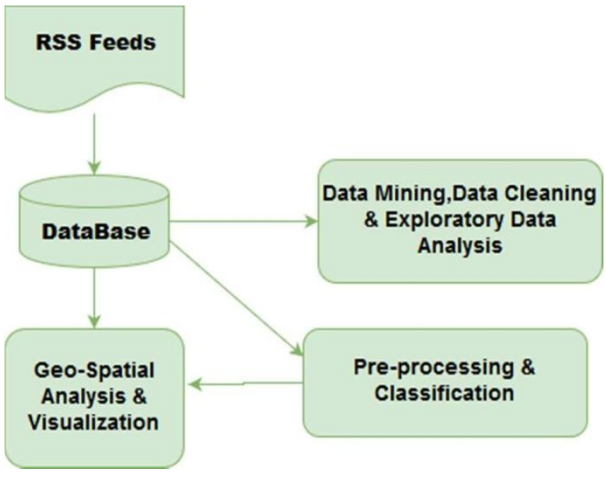


Figure 6.1 The flow of Geo-Spatial analysis and visualization

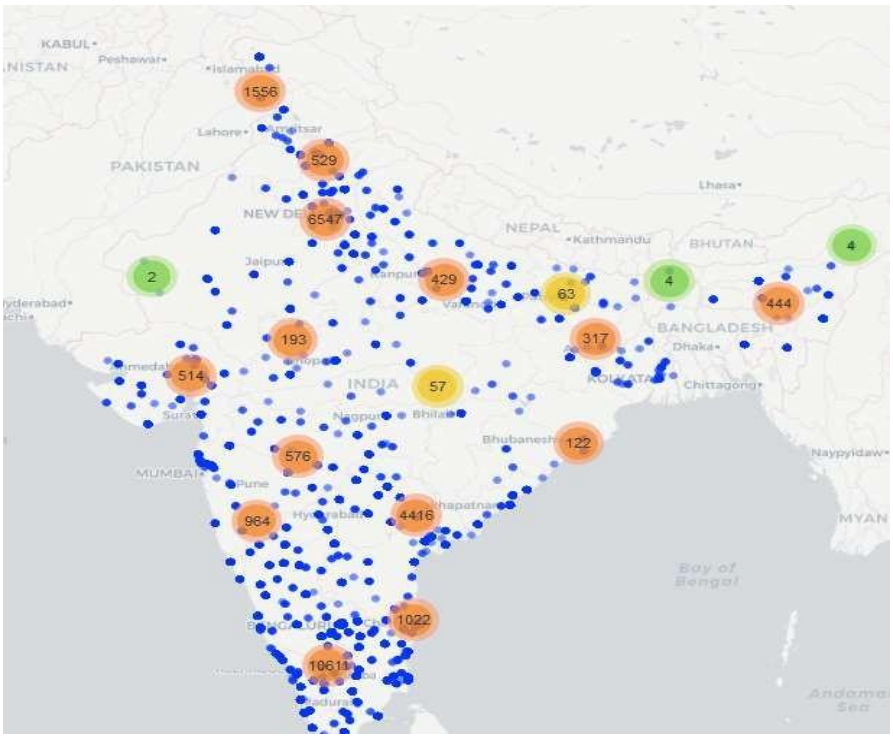


Figure 6.2 Geo-Spatial crime hot spots in India

Geospatial crime analysis to determine crime density using KDE for the Indian context

The research has produced in this section as an outcome in the form of the Density of different crimes in the Indian and Bangalore context. The shift in criminal activity for one year period (2017 Jan to 2018 Jan) are compared and analyzed. It also shows the clustering of various crime types with respect to commercial areas. Figure 6.3 shows the procedural steps for the density of crime identification in the Indian context. Our results have shown that the spatiotemporal analysis of news feed data can increase the prediction performance of 16 types of crime.

This indicates the potential to improve the planning for scarce resources in law enforcement agencies. Figure 6.4 shows the Visualization of Density identification in Indian context detailed results discussed in the Results and discussion.

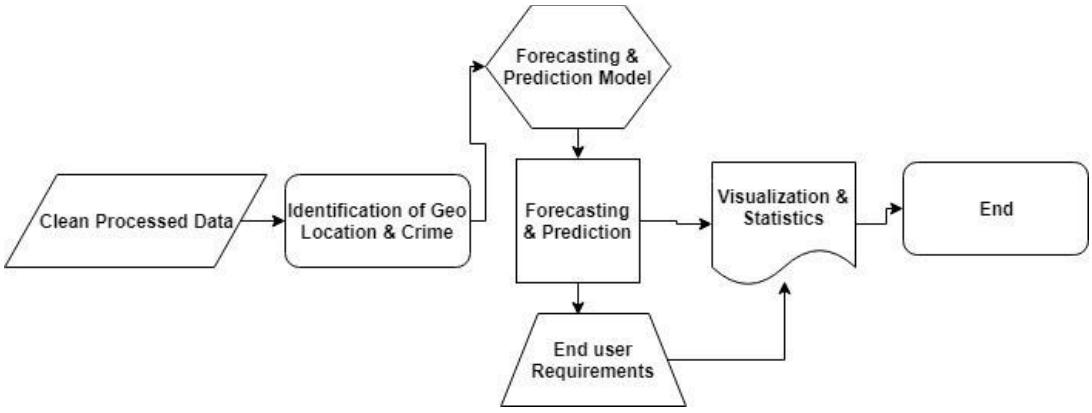


Figure 6.3 Data post-processing and visualization architecture

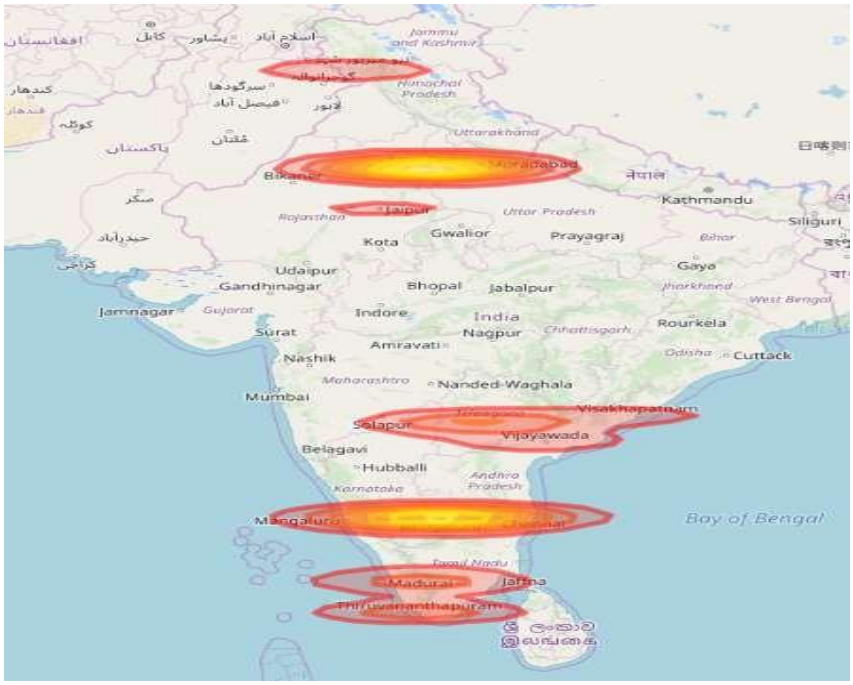


Figure 6.4. Crime Density identification using KDE

Spatio-Temporal Crime Analysis Using KDE and ARIMA Models in Indian Context

This section examines the news feed data collected from various sources regarding crime in India and Bangalore city. The crimes are then classified on the geographic density and the crime patterns such as time of day to identify and visualize the distribution of national and regional crime such as theft, murder, alcoholism, assault,

etc. In total, 68 types of crime-related dictionary keywords are classified into 6 classes based on the news feed data collected for 1 year. Kernel Density Estimation method is used to identify the hotspots of crime. With the help of the ARIMA model, time series prediction performed on the data. The diversity of crime patterns visualized in a customizable way with the help of a data-mining platform.

6.3 CONCLUSIONS ON RESEARCH WORK

In this research, 1-year crime data has used under the context of Indian and Bangalore crimes. A total of 68 types of crime keywords are identified, and they classified into six groups. The quality of the input newsfeed data has been compared and validated with that of RTI data.KDE is used for density analysis and compared with the ARIMA model. The proposed model predicts the possible crimes for the time span of six hours, one day, three Days, one Week, and 1 Month to help the crime authority to take preventive measures well in advance. Our results have shown that the news feed data can use for extracting spatiotemporal information about the prediction performance of 68 types of crime research had achieved a prediction accuracy of 77.49% with our crime prediction models. The outcome of research had validated our crime prediction model with that of the ARIMA model and found equivalent prediction performance. In the future, our work can be extended to topic modelling in text analysis to reduce the false acceptance ratio. Additional features such as socio-economic characteristics of the population can include in the news feed analysis. Alternative technologies such as NoSQL can use to scale the system replacing PostgreSQL.With the help of our system, police authorities in Bangalore can deploy their resources effectively. This application will result in a reduction of effort and improvement in crime response rates.

Time series analysis enables the prediction of crime rates in the same location in the future. Along with the present scope of our project, which is a prediction of the crime-

prone areas, this research can also predict the estimated time for the crime to take place as a future scope. Along with this, one can try to predict the location of the crime. This research will test the accuracy of frequent-item sets and predictions based on different test sets. So the system will automatically learn the changing patterns in crime by examining the crime patterns. Also, crime factors change over time. By sifting through the crime data, researchers have to identify new factors that lead to crime. Since research is considering only some limited factors, full accuracy cannot be achieved. For getting better results in prediction, research has to find more crime attributes.

REFERENCES

[1]. Agarwal Jyothi, Renuka Nagpal and Rajni Sehgal., "Crime Analysis using KMeans Clustering". International Journal of Computer Applications vol.83,no.4, pp.1-4, December 2013.

[2]. Ahishakiye, E., Taremwa, D., Omulo, E. O., Nairobi-Kenya, G. P. O., & Niyonzima, I., "Crime Prediction Using Decision Tree (J48) Classification Algorithm Analysis". International Journal of Computer and Information Technology, vol. 6, no.03, pp.188-195,2017.

[3]. Ahsan Morshed, Pei-Wei Tsai, Prem Prakash Jayaraman, Timos Sellis, Dimitrios Georgakopoulos, Sam Burke, Shane Joachim, Ming-Sheng Quah, Stefan Tsvetkov, Jason Liew, and Corey Jenkins., " VisCrime: A Crime Visualisation System for Crime Trajectory from Multi-Dimensional Sources ", WSDM '19 Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining,pp.802-805,2019.

[4]. Anna L. Buczak and Christopher M. Gifford, "Fuzzy association rule mining for community crime pattern discovery", In ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD '10). Association for Computing Machinery, New York, NY, USA, Article 2, pp.1–10,2010.

[5]. Andrey Bogomolov and Bruno Lepri "Predicting Crime Hotspots Using Aggregated and Anonymized Data on People Dynamics" Proceedings in conference on the scientific analysis of mobile phone datasets, MIT Media Lab, pp.7-10, April 2015.

[6]. Angers, J., Biswas, A., & Maiti, R, "Bayesian Forecasting for Time Series of Categorical Data," Journal Of Forecasting, vol.36,no.3,pp.217-229, 2016.

[7]. Anthony J. Corso, Gondy Leroy and Abdulkareem Alsusdais "Toward Predictive Crime Analysis via Social Media, Big Data, and GIS" Proceedings of In iConference 2015.

[8]. Ardis Hanson, “Integrating Geographic Information Systems into Library Services: A Guide for Academic Libraries” pp.175-201,2008.

[9]. Azeez and D. J. Aravindhar, "Hybrid approach to crime prediction using deep learning," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, pp. 1701-1710, 2015.

[10]. Aziz Nasridinov, Sun-Young Ihm, Young-Ho Park, “A Decision Tree-Based Classification Model for Crime Prediction”, Information Technology Convergence. Lecture Notes in Electrical Engineering, vol 253, pp. 531-538, 2013.

[11]. Bao, P., Shen, H. W., Chen, W., & Cheng, X. Q, “Cumulative effect in information diffusion: Empirical study on a microblogging network”. PLoS One, vol.8,no.10,2013.

[12]. Behrens, K., & Robert-Nicoud, F, “Survival of the Fittest in Cities: Urbanisation and Inequality,” The Economic Journal, vol.124,no.581, pp.1371-1400, 2014.

[13]. Berestycki, H., Wei, J., & Winter, “Existence of Symmetric and Asymmetric Spikes for a Crime Hotspot Model”, SIAM Journal On Mathematical Analysis, vol.46,no.1, pp.691-719, 2014.

[14]. Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., & Pentland, A, “Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics”, Big data,vol.3,no.3, pp.148-158,2015.

[15]. Bowers, Kate, Martin Newton, and Richard Nutter. “A GIS-linked database for monitoring repeat domestic burglary”, Mapping and Analysing Crime DataLessons from Research and Practice, pp.120-137, 2001.

[16]. Catlett, C., Cesario, E., Talia, D., & Vinci, A,. “A Data-driven Approach for Spatio-Temporal Crime Predictions in Smart Cities”, 2018 IEEE International Conference on Smart Computing (SMARTCOMP) pp. 17-24,2018.

[17]. Chainey, S., & Radcliffe, J, “GIS and crime mapping”, IEEE Journal, vol.45,no.3, pp.115-118, 2018.

[18]. Chaolun Xia , Raz Schwartz ,Ke Xie ,Adam Krebs, Andrew Langdon , Jeremy Ting and Mor Naaman, “City Beat: real-time social media visualization of hyperlocal city data”, International World Wide Web Conferences Steering Committee Republic and Canton of Geneva, Switzerland, pp. 167-170, April 2014.

[19]. Charpentier and E. Gallic, "Kernel Density Estimation with Ripley's Circumferential Correction", SSRN Electronic Journal, 2014.

[20]. Chen, P., & Yuan, H, “Forecasting crime using the arima model”, Fifth International Conference On Fuzzy Systems And Knowledge Discovery, vol.5,no.10, pp.627-630, 2008. 132

[21]. Clancey, G., Kent, J., Lyons, A., & Westcott, H, “Crime and crime prevention in an Australian growth centre”, Crime Prevention And Community Safety, vol.19,no.1, pp.17-30, 2017.

[22]. C. S. Marzan, M. C. Baculo and R. de Dios Bulos, "Time Series Analysis and Crime Pattern Forecasting of City Crime Data", ICACS,pp.113-118,2017.

[23]. D. Kolsrud, "A Time-Simultaneous Prediction Box for a Multivariate Time Series", Journal of Forecasting, vol.34, no.8, pp.675-693,2015.

[24]. D. Shahaf and C. Guestrin, “Connecting the dots between news articles”, 16th ACM/SIGKDD international conference on Knowledge discovery and data mining, pp. 623-632,2010.

[25]. Eck, J., & Weisburd, D, “Crime and place, crime prevention studies”, Criminal Justice Press, 1995.

[26]. Ellen G. Cohn; “Weather And Crime”, The British Journal of Criminology, vol.30, no.1, pp.51–64, 1990.

[27]. F. Ahmad, M. Uddin and L. Goparaju, "Role of Geospatial technology in Crime Mapping: A case study of Jharkhand state of India", American Journal of Geographical Research and Reviews (AJGRR), vol. 1, no. 15, 2018.

[28]. F. Ahmad, S. Syal and M. Tinna, "Criminal Policing Using Rossmo's Equation by Applying Local Crime Sentiment", Advances in Intelligent Systems and Computing, vol. 542, 2019.

[29]. F. Jiang, X. Yang and S. Li, "Comparison of Forecasting India's Energy Demand Using an MGM, ARIMA Model, MGM-ARIMA Model, and BP Neural Network Model", Sustainability,vol.10,no.7,pp.1-17,2018.

[30]. F. Kamalov, "Kernel density estimation based sampling for imbalanced class distribution", Information Sciences, vol.34, no.5,pp.1-15,2019.

[31]. F. Lo, "Time-Space Analysis of Facebook News Feeds", Inspiring Critical Thought, vol.10, no.2, pp. 1-15, 2018.

[32]. Fayyad, U, " Knowledge Discovery and Data Mining: Towards a unifying framework", 2Nd Int. Conf. On Knowledge Discovery And Data Mining, vol.45,no.3,pp.112-115,2012.