Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 27 (2024), No. 8, pp. 2255–2271
DOI : 10.47974/JDMSC-1822

A deep learning model to predict and classify crime rate using tweets

Pranav Dass [†]
Department of Computer Science
Shyam Lal College
University of Delhi
Delhi 110032
India

Vedika Gupta [*]
Jindal Global Business School
O. P. Jindal Global University
Sonipat 131001
Haryana
India

Shreya Dhingra [§]
Rohan Arora [‡]
Piyush Katariya [@]
Adarsh Kumar [#]
Department of Computer Science Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi 110063
India

Abstract

Crime is a detrimental socioeconomic issue that impacts individuals globally. Crime significantly affects a nation's standard of living, financial well-being, and standing. Over the past few years, there has been a significant increase in the crime rate. Law enforcement must implement proactive measures to mitigate crime rates. Enhanced technologies and innovative methods are required to bolster crime analytics and safeguard communities.

[†] *E-mail:* **pdasscs@shyamlal.du.ac.in**
[*] *E-mail:* **vgupta2@jgu.edu.in (Corresponding Author)**
[§] *E-mail:* **shreyadhingra54@gmail.com**
[‡] *E-mail:* **rohanaroraaashi@gmail.com**
[@] *E-mail:* **piyushkatariya95@gmail.com**
[#] *E-mail:* **adarsh12k@gmail.com**

Precise and up-to-date crime predictions can reduce crime rates, but they provide a complex problem for scientists due to the multiple factors that drive crime occurrences. This study employs a range of visualisation techniques and machine learning algorithms to predict the spread of crime across a vast geographical area. The datasets were evaluated and presented in the initial phase, based on their relevance. Subsequently, machine learning algorithms were employed to extract insights from extensive datasets and uncover concealed correlations within the data. This information was subsequently utilised to identify and analyse crime patterns, providing crime analysts with valuable tools for crime prediction. Consequently, this approach proves advantageous in the realm of crime prevention. Every day, a substantial amount of criminal activities are committed. The information in this instance includes both the date and the crime rate for the relevant years. The crime rate in this project is determined by employing various distinct criminal classifications. Utilising historical data, we employ the RNN, LSTM, and GRU algorithms to forecast the future percentage of the crime rate. The programme receives the date as an input and produces the crime rate percentage for that specific year.

Subject Classification: *(2010) 68Txx (Artificial intelligence), 68Uxx (Computing methodologies and applications), 68Vxx (Computer science support for mathematical research and practice).*

Keywords: *Crime, Machine learning, Deep learning, NCRB, Metropolitan cities, Twitter.*

1. Introduction

An intentional act that might result in physical or psychological suffering, as well as financial harm or loss, is considered a crime. Depending on the gravity of the offence, the government or other authorities may decide to punish the perpetrator of the crime. The incidence of illegal activities as well as the variety of such activities are growing, which is why law enforcement organisations are being forced to devise effective strategies to discourage them. Traditional approaches to solving crimes, which are both slow and ineffective, are unable to provide results in the current environment, which is characterised by a dramatically rising rate of criminal activity. Crime control techniques must address the following key questions: a) the identity of the perpetrator, b) the identity of the target, c) the type of crime being committed, d) the specific location of the crime, and e) the time at which the crime will occur. The purpose of this study is to make a prediction about the occurrence of a specific category of criminal activity by analysing historical data. The research will concentrate on the geographical and temporal characteristics of the crime rather than the individuals who are engaged as victims or offenders. The primary objective in criminology is to determine the key spots within the city that require the highest allocation of law enforcement resources by the agency. A significant obstacle encountered by law enforcement agencies is

the acquisition of precise crime predictions to effectively allocate patrols and resources, hence improving crime prevention and police response times [19].

As per routine activity theory, the majority of crimes take place when three conditions are met: the presence of a determined offender, the availability of a vulnerable target, and a deficiency in victim safeguarding. In accordance with the rational choice theory, a potential criminal will weigh the potential benefits of successfully carrying out the criminal conduct against the possibility of getting captured. Based on this analysis, the prospective criminal will then arrive at a rational decision regarding whether or not to proceed with the illegal act. Both perspectives concur that a crime transpires when an individual who possesses the intention to commit it is afforded the opportunity to do so. Typically, law enforcement officers utilise regional maps to document events and mark the location of each one with a pin. By analysing these maps, researchers can identify these patterns and effectively predict the locations with a higher probability of crime, known as hotspots.

To harness these extensive phenomena as early indicators of crime, more intricate methodologies, such as machine learning, must be employed instead of basic mapping techniques. Various machine learning techniques, such as Random Forests, Naive Bayes, and Support Vector Machines (SVMs), have been employed to predict the occurrence of crimes in a certain area and to identify crime hotspots [20]. The proficiency of the analyst in organising data and creating customised characteristics that precisely depict the challenge at hand is vital for the triumph of a machine learning investigation [18]. Deep learning employs an algorithm that pulls features from raw data, thereby surpassing the limitations of traditional machine learning approaches [17]. Undoubtedly, this increase in benefit is accompanied by a substantial trade-off in terms of computational intricacy and consumption of raw data. Considering this, we analyse deep learning architectures for forecasting crime hotspots and offer design suggestions.

The remaining work presented in the paper is organised in the following manner. In the forthcoming Section 2, the focus will be on examining the previous research and projects conducted in the field of crime rate prediction and classification, which share similarities with the current study. Section 3 provides a detailed discussion of the dataset, methods, and technology employed in constructing the crime rate detection model. The experimental data, along with a comparative table, are presented in Section 4. Section 5 presents the conclusion of this research work.

2. Literature Survey

In this paper, Aravindan, S., E. Anusuya, and M. Ashok Kumar developed a GUI model for Chennai and (SMLT) supervised machine learning technique model [1] which finds the best model among SML algorithms by getting more than 90% accuracy for determining crime rate by PSA and year. The dataset is taken from the online Indian police department, split into training and testing in 7:3, and applied to Random Forest, SVC, KNN, Decision tree, and logistic regression algorithms. The research gap is not comparing with deep learning models and not incorporating this model into a real-time system for crime rate prediction.

Wang and Bao created a deep learning ST-Resnet model that surpasses the fully-ternary ST-Resnet, ARIMA, KNN, and HA models in terms of performance. The authors utilised the abundant and accurate data to calibrate the spatial-temporal residual network for predicting crime distribution in Los Angeles at an hourly resolution, specifically in blocks of neighbourhood size. The experiments, in addition to comparisons with other prediction approaches, demonstrate that the proposed model is superior in terms of accuracy [2]. Ultimately, the authors propose a three-fold solution to address the problem of resource consumption in real-world implementation.

Tabedzki and Christian conducted a study where they built a machine learning method to forecast crime-related data in Philadelphia, United States [3]. The problem was broken up into three independent parts: identifying the occurrence of a crime, evaluating the chance that a crime would occur, and identifying the type of crime that is most likely to occur. A number of different techniques, including logistic regression, KNN, ordinal regression, and tree approaches, were utilised in order to train the datasets in order to generate crime predictions that were accurate and suitable for the situation. In addition to this, they offered a visual depiction in the form of a map, which represented diverse crime categories throughout a variety of Philadelphia neighbourhoods during a particular period of time. Each category of criminal activity was represented by a distinct colour. In order to accurately depict the full pattern of criminal activity in Philadelphia over a certain period of time, a wide variety of criminal activities, ranging from physical assaults to internet deceit, were added into the study. An astounding 69% accuracy was displayed by their algorithm when it came to predicting the probability of a crime, and it also demonstrated a 47% accuracy when it came to calculating the number of offences that occurred within a range of one to 32%.

Bandekar's work mostly centred on the investigation and creation of machine learning algorithms with the intention of lowering the rate of criminal activity in India. Techniques from the field of machine learning were utilised in order to discover the pattern correlations that existed amid a massive amount of data. By examining the frequency of previous criminal episodes in particular places, the major purpose of the study was to make predictions about future criminal behaviour [4]. Bayesian neural networks, the Levenberg-Marquardt algorithm, and a scaled algorithm were some of the approaches that were utilised in the process of analysing and interpreting the dataset. For example, when compared to the other two methods, the scaled algorithm displayed significantly better performance than the other two. Based on the findings of a statistical analysis that included correlation, analysis of variance, and graphs, it has been determined that the application of the scaled technique has the ability to lower the crime rate by 78%, with a precision of 0.78. The research conducted by Hossain and Sohrab, on the other hand, proposes a method for predicting criminal activity by analysing a dataset that contains information on previous crimes and the patterns to which they have consistently occurred. Two different machine learning methods, namely decision trees and KNNs, are utilised by the system that has been proposed [5]. The usage of more sophisticated methods, such as the random forest algorithm and adaptive boosting, contributed to an increase in the accuracy of the prediction model. For the purpose of improving the performance of the model, the crimes were separated into two categories: both common and unusual. The crimes that fallen into the frequent category were the ones that occurred the most frequently, whilst the crimes that fell into the rare category were the ones that occurred the least frequently. Information concerning criminal acts that took place in San Francisco during a period of twelve years was made available to the system that was being designed. In order to achieve a level of accuracy of 99.16%, a combination of undersampling and oversampling techniques, in conjunction with the random forest algorithm, were utilised.

Kim and Suhong conducted a study analysing crime statistics from the past 15 years in Vancouver, Canada, with the aim of determining its predictability. This criminal investigation, which relied on machine learning techniques, included essential components such as the collection of data, the classification of data, the recognition of patterns, the interpretation of predictions, and the visualisation of data. The K-nearest neighbour (KNN) and boosted decision tree approaches were utilised in order to do additional analysis on the crime dataset. The researchers found

that by utilising machine learning algorithms, they were able to forecast crime with a precision that ranged from 39 percent to 44 percent. This was determined by studying 560,000 crime statistics that spanned from 2003 to 2018 [6]. Furthermore, the scientists extrapolated that the precision might be improved by meticulously adjusting both the algorithms and the crime data for specific use situations. Although the precision was inadequate as a prognostic model, the authors concluded that it could be improved.

Bharati and her co-authors conducted an analysis on a comprehensive dataset including numerous criminal incidents. They utilised various factors to make predictions about the specific type of crime that is likely to transpire in the immediate future. Crime prediction was conducted on a Chicago crime dataset using machine learning and data science methodologies. The crime dataset include information such as the description of the crime scene, the kind of crime, the date and time of occurrence, and precise geographical coordinates. Various machine learning algorithms, including KNN classification, logistic regression, decision trees, random forest, support vector machine (SVM), and Bayesian approaches, were tested. The model with the highest accuracy was chosen for training. The KNN classification demonstrated a superior accuracy of approximately 0.787 [7]. The primary objective of this study is to demonstrate the application of machine learning in law enforcement to enhance crime prediction, identification, and resolution, ultimately leading to a decrease in criminal activity.

Kang conducted a comprehensive study of crime using a novel and multifaceted technique. This research investigates a feature-level data fusion technique that utilises a deep neural network (DNN) to accurately predict criminal incidents. The technique combines data from several domains and incorporates environmental context information to enhance the reliability of the forecasts. A collection of photographs, demographic and meteorological data, and data obtained from an internet database of Chicago crime statistics are all included in this collection. Regression analysis, kernel density estimation (KDE), and support vector machines (SVM) are all methods that are utilised in the process of crime prediction [8]. There were three distinct stages involved in the technique: the first stage was the collection of data, the second stage was statistical analysis to determine the link between crime incidents and the data obtained, and the third stage was the accurate prediction of future crime incidents. The DNN model takes into account environmental context, temporal characteristics, and spatial information to create its model. The suggested deep neural network (DNN) model displays superior performance with

an astounding accuracy of 84.25%. This accuracy surpasses that of the SVM and KDE models, which attained accuracies of 67.01% and 66.33% respectively.

Stalidis devised a framework to comprehend and elucidate these uncertainties, drawing upon John Dewey's notions of intension and extension, and applying mathematical set theory to address this issue [9]. The participants engaged in a discussion regarding the influence of their model in aiding law enforcement authorities in analysing bias offences for several objectives, such as prevention, statistical documentation, and criminal litigation. A comprehensive analysis of crime classification and prediction investigation is offered, utilising a deep learning architecture developed by Nolan and James. The researchers confirmed the efficacy of deep learning algorithms in this domain and utilised publicly available police report data [10] to propose recommendations for the development and training of deep learning systems for crime area prediction.

The STAC algorithm identifies the most concentrated points on the map and subsequently fits each point to the standard deviation ellipse. Through the examination of the dimensions and arrangement of the ellipse, analysts are able to make inferences regarding the characteristics of prospective clusters of criminal activity [11]. Thematic cartography involves dividing the map into distinct regions and representing illicit activity as individual spots on the map. Subsequently, these locations can be incorporated into the geographic unit area and assigned a colour based on the quantity of infractions that fall within the specified range of [12].

Kianhmer and Alhajj [13] employed Support Vector Machines (SVM) as a machine learning technique to forecast the positioning of access points. Yu et al. conducted a comparative analysis of the effectiveness of Support Vector Machines (SVM) [14] in relation to various machine learning techniques, such as Naive Bayes and Random Forest. They noted that in instances of a residential burglary, events that occurred in a particular location are prone to reoccur. Table 1 enlists analysis of existing research papers in tabular format.

Table 1

Analysis of Various research papers

S. No	Paper	Approach	Methodology	Results
1	[1]	Machine Learning	Used split of training and testing in 7:3, and applied to Random Forest, SVC, KNN, Decision tree, and logistic regression algorithms	This GUI approach showed an accuracy of 90%.
2	[2]	Deep Learning	The authors developed deep learning ST-Resnet model which outperforms the fully-ternary ST-Resnet, ARIMA, KNN, and HA model.	The results showed an accuracy of 75% respectively
3	[3]	Machine Learning	Used ML models like k-Nearest Neighbors(k-NN), logistic regression for predicting crime-related statistics in Philadelphia, United States	They predicted likelihood of a crime with 69% accuracy, number of crimes ranging from one to 32% with a 47% accuracy.
4	[4]	Machine Learning	Used ML models, analyzed and interpreted using approaches such as Bayesian neural networks, the Levenberg Marquardt algorithm, and a scaled algorithm, with the scaled algorithm outperforming the other two techniques	Graphs revealed that the crime rate decreased by 78% using the scaled method with an accuracy of 0.78.
5	[5]	Machine Learning	Used ML models such as decision trees, random forest algorithm, adaptive boosting and KNNs.	The accuracy was raised to 99.16% by combining under sampling and oversampling approaches with the random forest algorithm.
6	[6]	Machine Learning	K-nearest neighbor (KNN) and boosted decision tree methods, they looked at 560,000 crime statistics from 2003 to 2018, and found that using machine learning algorithms.	They predicted crime with an accuracy of 39% to 44%.

Contd...

7	[7]	Machine Learning	Used ML models such as KNN classification, logistic regression, decision trees, random forest, a support vector machine (SVM), and Bayesian techniques.	With an accuracy of about 0.787, the KNN classification proved to be the most accurate.
8	[8]	Deep Learning Machine Learning	Used Regression analysis, kernel density estimation (KDE), support vector machines (SVM), and DNN model.	The DNN model outperforms with an accuracy of 84.25%, compared to 67.01% and 66.33% for the SVM and KDE models, respectively.
9	[9]	Machine Learning	Developed a framework to understand and clarify these ambiguities based on John Dewey's concepts of intension and extension and their own application of mathematical set theory to this problem	Obtained accuracy 82% on their test set
10	[10]	Deep Learning	Used Convolutional Neural Network (CNN) for detection.	Got an accuracy of 92.7%
11	[11]	Deep Learning and Machine Learning models	Used densest points to detect on the map, and then each point is fitted to the standard deviation ellipse.	Got an accuracy of 90.7%
12	[12]	Deep Learning	The map is divided into border areas, and illegal activities are placed on the map in the form of points	Got an accuracy of 96%
13	[13]	Machine Learning models	Used SVM in machine learning methods to predict the location of access points.	SVM showed accuracy of 90% followed by Adaboost and Random Forest.
14	[14]	Machine Learning	Used Naive Bayes and Random Forest for detection.	Got an accuracy of 93%

Contd...

15	[15]	Deep Learning models	Used an "inline" algorithm to modify its previous predictions when future predictions are wrong.	Estimates best weights for the predictions of members of the combined set with accuracy of 91%.
16	[16]	Deep Learning models	Used greedy search and pruning algorithms to combine DSTP to form an integrated spatio-temporal pattern (ESTP)	They tested this method to predict residential burglaries and achieved 80% accuracy.

3. Data Gathering and Preprocessing

The initial goal was to determine which types of crimes are more commonly mentioned and discussed among the general public. We utilize Twitter for this. Using hashtags like #crime, #CrimeAgainstHumanity, and others, a data collection of tweets was constructed and analyzed to see which crimes are debated the most among the general public.

The Twitter tweets are collected using the Python library and the Twitter API. The data collection includes tweets retrieved using hashtags such as #crime and #CrimeAgainstHumanity from January 1 to September 2, 2020. Figure 1 depicts a sample of Tweets fetched using relevant hashtags.

	Text	Time
0	Shri. Adesh Gupta ji President, BJP Delhi has ...	01-01-2020
1	Murders, rapes, kidnapping: How Covid affected...	01-01-2020
2	The father killed his 10-year-old son along wi...	01-01-2020
3	@bainsindian: Madam Shiv Senni ho gayi ho😕\nME...	01-01-2020
4	» Average 77 #rape cases daily reported in #In...	01-01-2020
...	...	...
24510	Respected @narendramodi ji we consider u as ou...	30-09-2020
24511	@matrixxmedia: An Indian woman allegedly assau...	30-09-2020
24512	Headless Body of Man Found in Plastic Bag in N...	30-09-2020
24513	@DigitalShakti @NCWIndia @Facebook @AutobotInf...	30-09-2020
24514	@apradhan1968: Dear @dtptraffic please take no...	30-09-2020

24515 rows × 2 columns

Figure 1

Tweets fetched using #crime, #CrimeAgainstHumanity, etc.

The top five most talked crime categories were then chosen for additional study and crime rate forecast based on the findings of the Twitter analysis. Predictions of crime rates are made for several of India's main metropolitan cities, including Ahmedabad, Delhi, Mumbai, Chennai, and Kolkata. For the years 2016 to 2020, precise data was gathered from the Indian government agency National Crime Record Bureau (NCRB). Figure 2 depicts a Snapshot of Dataset created using data provided by NCRB for 5 metropolitan cities of India.

Year - 2016

Cities	Murder			Crime Against Women			Kidnapping and Abduction			Theft			Accident		
	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of crime
Ahmedaba	103	103	1.6	131	132	4.4	376	392	5.9	2624	2624	41.3	376	392	5.9
Delhi	479	490	2.9	3746	3762	49.4	5925	6281	36.3	126467	126474	775.2	5925	6281	36.3
Mumbai	147	152	0.8	2183	2199	25.6	1949	2027	10.6	9839	9853	53.4	1949	2027	10.6
Chennai	133	139	1.5	63	63	1.5	34	35	0.4	3070	3070	35.3	34	35	0.4
Kolkata	78	105	0.6	348	348	5.1	290	305	2	4186	4186	29.7	290	305	2

Year - 2017

Cities	Murder			Crime Against Women			Kidnapping and Abduction			Theft			Accident		
	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of crime
Ahmedaba	90	90	1.4	224	224	3.5	263	263	4.1	3246	3246	51.1	346	346	5.4
Delhi	400	428	2.5	2548	2712	15.6	5203	5703	31.9	161818	165987	991.8	1317	1451	8.1
Mumbai	127	129	0.7	1713	1750	9.3	2159	2282	11.7	9718	9972	52.8	492	492	2.7
Chennai	154	162	1.8	71	71	0.8	54	54	0.6	4158	4158	47.8	234	239	1.7
Kolkata	64	64	0.5	309	314	2.2	374	395	2.7	3493	3567	24.8	1312	1404	15.1

Year - 2018

Cities	Murder			Crime Against Women			Kidnapping and Abduction			Theft			Accident		
	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of crime
Ahmedaba	98	103	1.5	208	208	6.9	277	279	4.4	2837	2845	44.7	373	383	5.9
Delhi	416	460	2.5	2353	2366	31.1	5124	5323	31.4	178953	182610	1096.9	1379	1481	8.5
Mumbai	164	166	0.9	2038	2038	23.9	2202	2228	12	9468	9545	51.4	366	401	2
Chennai	172	174	2	83	83	1.9	56	57	0.6	3891	3891	44.7	1267	1284	14.6
Kolkata	55	59	0.4	372	372	5.5	378	398	2.7	2608	2608	18.5	274	286	1.9

Year - 2019

Cities	Murder			Crime Against Women			Kidnapping and Abduction			Theft			Accident		
	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of crime
Ahmedaba	81	82	1.3	205	207	3.2	248	256	3.9	2907	2907	45.8	421	442	6.6
Delhi	505	520	3.1	2311	2326	14.2	5746	5948	35.2	242642	246877	1487.2	1379	1561	8.5
Mumbai	168	170	0.9	2069	2098	11.2	2102	2185	11.4	2608	2608	18.5	374	452	2
Chennai	172	177	2	89	89	1	49	51	0.6	3618	3618	41.6	1229	1262	14.1
Kolkata	55	59	0.4	372	372	2.6	378	398	2.7	8582	8585	46.6	274	286	1.9

Year - 2020

Cities	Murder			Crime Against Women			Kidnapping and Abduction			Theft			Accident		
	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of cri	Incident	Victim	Rate of crime
Ahmedaba	70	70	1.1	167	168	2.6	181	187	2.8	2389	2393	37.6	328	336	5.2
Delhi	461	476	2.8	1799	1805	11	4011	4134	24.6	175442	178228	1075.3	1130	1156	6.9
Mumbai	148	149	0.8	1507	1509	8.2	1173	1179	6.4	6234	6258	33.9	303	336	1.6
Chennai	150	160	1.7	61	61	0.7	37	39	0.4	4788	4788	55.1	855	872	9.8
Kolkata	53	55	0.4	304	304	2.2	308	308	2.2	1325	1325	9.4	204	218	1.4

Figure 2

Snapshot of Dataset created using data provided by NCRB for 5 metropolitan cities of India.

4. Methodology

To begin, a list of ten main kinds of crimes covered by the Indian Penal Code (IPC) was provided, including antisocial behavior, theft, child abuse, crime against women, cybercrime, fraud, hate crime slavery, accident, murder, and terrorism. After then, a data set of tweets was constructed based on certain specific hashtags connected to each of the above-mentioned crime types in order to train a machine learning model that can categorize the crime based on tweets. Logistic regression, naive

	Text	Time	clean_text
0	Shri. Adesh Gupta ji President, BJP Delhi has ...	01-01-2020	shri adesh gupta ji president bjp delhi presen...
1	Murders, rapes, kidnapping: How Covid affected...	01-01-2020	murders rape kidnapping how covid affect crime...
2	The father killed his 10-year-old son along wi...	01-01-2020	the father kill 10-year-old son along wife gir...
3	@bainsindian: Madam Shiv Senni ho gayi ho😕\nME...	01-01-2020	madam shiv senni ho gayi ho😕 mea already reach...
4	» Average 77 #rape cases daily reported in #In...	01-01-2020	» average 77 rape case daily report india 2020...
...	...	...	...
24510	Respected @narendramodi ji we consider u as ou...	30-09-2020	respected ji consider u head head family i ' a...
24511	@matrixxmedia: An Indian woman allegedly assau...	30-09-2020	an indian woman allegedly assault rap mumbai f...
24512	Headless Body of Man Found in Plastic Bag in N...	30-09-2020	headless body man found plastic bag navi mumba...
24513	@DigitalShakti @NCWIndia @Facebook @AutobotInf...	30-09-2020	NaN
24514	@apradhan1968: Dear @dtptraffic please take no...	30-09-2020	dear please take note fall traffic signal gate...

24515 rows × 3 columns

Figure 3

Dataset after processing

Bayes, decision trees, random forest, and support vector machine were the machine learning models employed for categorization. Figure 3 shows the sample dataset after preprocessing

Following the training, the machine learning model chosen was SVM, which had the highest accuracy of 94.552 percent, and after applying the model to the second data set, it was discovered that crimes such as accident, murder, kidnapping, crime against women, and theft are frequently discussed among the population on Twitter.

A data set is now constructed from the NCRB data set for Metropolitan cities for only those detected crime kinds, depending on the most talked crime types observed. Now for each metropolitan city, a crime rate vector is built in such a way that, based on previous year's rates of crime, the crime rate for the following year may be projected. This vector is then input into multiple Deep Learning models, such as Simple RNN, LSTM, and GRU, to determine the trend behind each metropolitan city's crime rates and forecast them for the coming year.

5. Results

With increasing number of criminal offences all across India, it is important to monitor and keep a track of these cases so that they can be prevented in future. The project revolves around getting authentic crime records data all across five major metropolitan cities of India- Ahmedabad, Delhi, Mumbai, Chennai and Kolkata. Further we analyze crime rates specific to five IPC listed crime categories which includes murder, kidnapping and abduction, crime against women, theft and accident.

These crime categories and cities have been chosen on the basis of Twitter analytics for finding out the most hyped crimes and the areas in which they generally occur. Further crime rate has been predicted for the year 2021 using various algorithms.

Crime records were taken from NCRB official records and have been visualized for a time span of 2016 to 2020. Figure 4 depicts crime records of murders across these five metropolitan cities from the year 2016 to 2020. It can be seen clearly that maximum number of murder cases have been reported in Delhi in the year 2019 whereas Ahmedabad has seen a decline through the years. Figure 5 and Figure 7 depicts Delhi as the crime capital for Kidnapping as well as Theft cases. Figure 6 clearly depicts Delhi and Mumbai have maximum contribution towards crime against women but the crime rate has gone down in 2020 due to COVID-19 imposed lockdowns. Accident cases have gone down over the years due to properly framed traffic rules except for Chennai as shown in Figure 8.

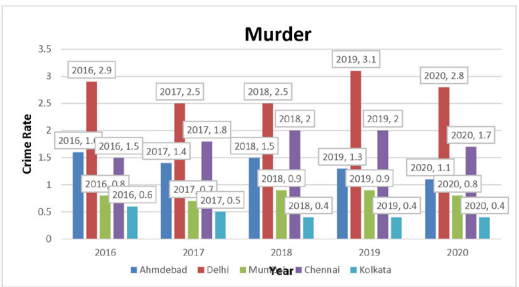

Figure 4

Murder cases across five metropolitan cities in through the years 2016 to 2020

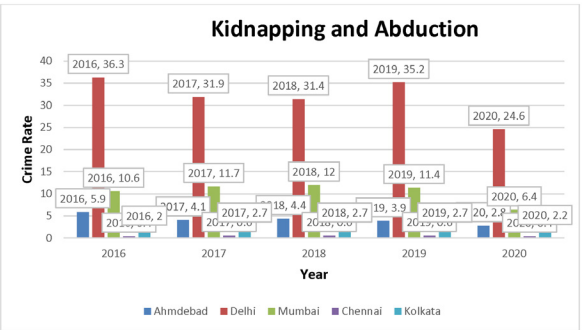

Figure 5

Kidnapping cases across five metropolitan cities in through the years 2016 to 2020

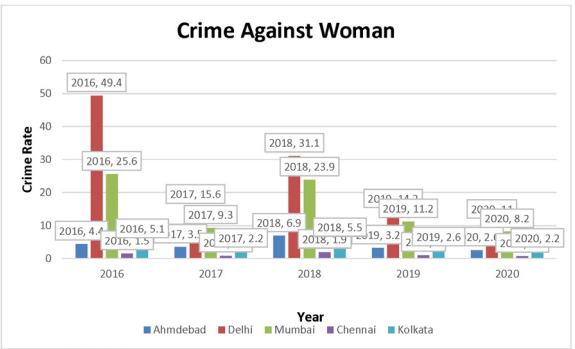

Figure 6

Crime against women across five metropolitan cities in through the years 2016 to 2020

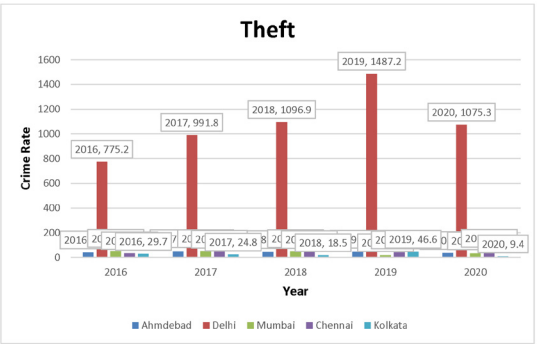

Figure 7

Theft cases across five metropolitan cities in through the years 2016 to 2020

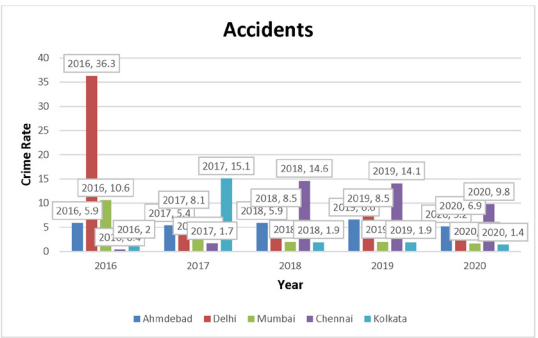

Figure 8

Accident cases across five metropolitan cities in through the years 2016 to 2020

Table 2
Classification accuracy of Machine learning algorithms on Twitter data

Machine learning algorithms	Classification accuracy
Naive Bayes	0.95
Logistic Regression	0.97
Decision Tree	0.97
Random Forest	0.93
SVM	0.97

Results from Twitter analytics

Data obtained over the years for all Indian states under all IPC crime heads was very vast to be processed. Hence, Twitter analytics was used to find out the most talked about crimes and the areas in which they occur. The accuracy of all the algorithms used of classification has been depicted in Table 2.

Conclusion

Based on the requirements of the situation, we performed an analysis on the datasets and displayed them in the first stage of this study. After that, machine learning algorithms were utilised to extract knowledge from these massive datasets and uncover hidden relationships among the data. This information was then utilised to report and uncover crime patterns. This information is helpful for crime analysts to analyse these crime networks by utilising a variety of interactive visualisations for the purpose of crime prediction, which ultimately contributes to the prevention of crime. There are a significant number of criminal acts that are committed on a daily basis. In this particular instance, the dataset includes not only the date but also the crime rate for the years that are under consideration. When calculating the crime rate for this project, numerous distinct types of criminal activity are taken into consideration. For the purpose of estimating the percentage of the crime rate in the future, we use the RNN, LSTM, and GRU algorithms to the data that we have already collected. The date is provided as an input to the algorithm, and the conclusion that is reached is the percentage of the crime rate for that particular year.

References

[1] S. Aravindan, E. Anusuya, and M. Ashok Kumar, "GUI based prediction of crime rate using machine learning approach," (2020).

[2] B. Wang, P. Yin, A. L. Bertozzi, P. J. Brantingham, S. J. Osher, and J. Xin, "Deep learning for real-time crime forecasting and its ternarization," *Chinese Annals of Mathematics, Series B,* vol. 40, no. 6, pp. 949-966 (2019).

[3] C. Tabedzki, A. Thirumalaiswamy, P. van Vliet, S. Agarwal, and S. Sun, "Yo home to Bel-Air: Predicting crime on the streets of Philadelphia," University of Pennsylvania, CIS, 520 (2018).

[4] S. R. Bandekar and C. Vijayalakshmi, "Design and analysis of machine learning algorithms for the reduction of crime rates in India," *Procedia Computer Science,* vol. 172, pp. 122-127 (2020).

[5] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, Mar. 26–27, 2020, Revised Selected Papers 1, pp. 277-289, Springer Singapore (2020).

[6] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime analysis through machine learning," in Proc. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420, IEEE, Nov. (2018).

[7] A. Bharati and R. A. K. Sarvanaguru, "Crime prediction and analysis using machine learning," *Int. Res. J. Eng. Technol.*, vol. 5, no. 9, pp. 1037-1042 (2018).

[8] H.-W. Kang and H.-B. Kang, "Prediction of crime occurrence from multi-modal data using deep learning," PLoS One, vol. 12, no. 4, p. e0176244 (2017).

[9] P. Stalidis, T. Semertzidis, and P. Daras, "Examining deep learning architectures for crime classification and prediction," arXiv preprint arXiv:1812.00602 (2018).

[10] J. J. Nolan III, J. McDevitt, S. Cronin, and A. Farrell, "Learning to see hate crimes: A framework for understanding and clarifying ambiguities in bias crime classification," *Criminal Justice Studies*, vol. 17, no. 1, pp. 91-105 (2004).

[11] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security Journal*, vol. 21, no. 1, pp. 4-28 (2008).

[12] D. Williamson, S. McLafferty, P. McGuire, T. Ross, J. Mollenkopf, V. Goldsmith, and S. Quinn, Tools in the spatial analysis of crime. Mapping and analysing crime data, A. Hirschfield and K. Bowers, Eds. London and New York: Taylor & Francis, vol. 1, p. 187 (2001).

[13] K. Kianmehr and R. Alhajj, "Effectiveness of support vector machine for crime hot-spots prediction," *Appl. Artif. Intell.,* vol. 22, no. 5, pp. 433-458 (2008).

[14] C. H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in Proc. 2011 IEEE 11th Int. Conf. Data Mining Workshops, pp. 779-786, IEEE, Dec. (2011).

[15] J. Xu, P. N. Tan, J. Zhou, and L. Luo, "Online multi-task learning framework for ensemble forecasting," *IEEE Trans. Knowl. Data Eng*., vol. 29, no. 6, pp. 1268-1280 (2017).

[16] C. H. Yu, W. Ding, M. Morabito, and P. Chen, "Hierarchical spatiotemporal pattern discovery and predictive modeling," *IEEE Trans. Knowl. Data Eng*., vol. 28, no. 4, pp. 979-993 (2015).

[17] A. R. Pathak, M. Pandey, and S. Rautaray, "Empirical evaluation of deep learning models for sentiment analysis," *J. Stat. Manag. Syst.,* vol. 22, no. 4, pp. 741-752 (2019).

[18] A. Jain and A. Ghosh, "Novel insights into data mining to improve the specificity of pharmacovigilance and prevent adverse drug reactions in psychiatric patients," *Asia Pacific J. Health Manag.,* vol. 16, no. 3, pp. 130-136 (2021).

[19] N. Jain, V. Gupta, U. Tariq, and D. J. Hemanth, "Fast violence recognition in video surveillance by integrating object detection and Conv-LSTM," *Int. J. Artif. Intell. Tools* (2022), doi: 10.1142/S0218213023400183.

[20] V. Gupta, N. Jain, H. Garg, S. Jhunthra, S. Mohan, A. H. Omar, and A. Ahmadian, "Predicting attributes-based movie success through ensemble machine learning," *Multimedia Tools Appl.*, pp. 1-30 (2022).

Received March, 2023