International Journal of Data Science and Analysis
2021; 7(3): 51-59
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210703.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Crime Data Analysis, Visualization and Prediction Using LSTM

Mufeeda Manengadan[1, 2], **Silpa Nandanan**[1, 2], **Neethu Subash**[1, 2]

[1]Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India

[2]Department of Computer Science and Engineering, Mar Athanasius College of Engineering Kothamangalam, Kerala, India

Email address:
mufeedamf98@gmail.com (M. Manengadan), silpanandanan96@gmail.com (S. Nandanan), neethusubash@mace.ac.in (N. Subash)

To cite this article:
Mufeeda Manengadan, Silpa Nandanan, Neethu Subash. Crime Data Analysis, Visualization and Prediction using LSTM. *International Journal of Data Science and Analysis*. Vol. 7, No. 3, 2021, pp. 51-59. doi: 10.11648/j.ijdsa.20210703.11

Received: January 25, 2021; **Accepted**: April 19, 2021; **Published**: May 8, 2021

Abstract: Crimes are common social problems that can even affect the quality of life, even the economic growth of a country. Big Data Analytics (BDA) is used for analyzing and identifying different crime patterns, their relations, and the trends within a large amount of crime data. Here, BDA is applied to criminal data in which, data analysis is conducted for the purpose of visualization. Big data analytics and visualization techniques were utilized to analyze crime big data within the different parts of India. Here, we have taken all the states of Indian for analysis, visualization and prediction. The series of operations performed are data collection, data pre-processing, visualization and trends prediction, in which LSTM model is used. The data includes different cases of crimes with in different years and the crimes such as crime against women and children in which, kidnap, murder, rape. The predictive results show that the LSTM perform better than neural network models. Hence, the generated outcomes will benefit for police and law enforcement organizations to clearly understand crime issues and that will help them to track activities, predict the similar incidents, and optimize the decision making process.

Keywords: Big Data Analytics, Crime Data, Crime Data Analysis, Visualization, Prediction, LSTM

1. Introduction

There are several crimes that are happening in our country. But many of the people might not be aware of such crimes that are occurring in the different parts of the world. The crime related activities can severely affect socio-economic activities of a society too. Thus, definitely there is a need for a system that can provide all the necessary information's to the people. The primary aim of crime data analysis is to assist the operations of a police departments as well as enforcement departments. This may include criminal investigation, crime prevention, reduction strategies and problem solving. The different operations that are performed for the purpose of crime data analysis are data collection, data pre-processing, visualization and trends prediction.

After data collection and pre-processing, including data filtering and normalization, Google maps based geo-mapping of the features are implemented for visualization of the statistical results and time series modeling are utilized for future trends analysis. For the entire process we took the crime data that has been occurred in the different parts of our country. This data also includes the year wise information about the different crimes. The different crimes that are happening around us can be alerted always as well as the paper can also represent different crimes such as crime against women and children, murder, kidnap. Thus through this we can easily identify the crime prone areas.

Big Data analytics is that the method of collecting, organizing and analyzing massive sets of data to discover patterns and other useful information [1]. Big Data analytics can better help the organizations to understand the information in the crime data. There are several ways to analysis such a huge amount of data. The detailed visualization of the crime data has also pictorially represented with in this paper.

The issues regarding the crime pattern deals with predicting the hidden crimes with in the country. The crime rate is increasing day by day and the crime patterns are always changing. Thus, the behaviors in crime are difficult to be predicted. As a result, crime prediction with in an area was not an easy task. But now a days it has become more popular to use different methodologies for such a purpose [21, 22].

The predictive model which is based on a neural network Long Short-Term Memory (LSTM), where a small group of attributes are trained, which further enables the prediction of the class label in the validation stage [26, 27]. This shows a high percentage of prediction accuracy also. The LSTM model is being widely used and it is preferred more than the other neural network models since it is easier to handle.

2. Related Work

2.1. Crime Data Analysis

During crime analysis, the input data is important to be used in training and testing process. The training of the process is used to conduct the crime model and the testing of the process is used to validate the algorithm [19, 21, 22]. Input data can be obtained from the criminal records with the help of the government, agencies, etc. As a consequence, the collected data is large volumes of data and it is also in the unstructured data formats. The collected data is stored into different databases [9, 10, 17].

The many researches are concerning to solve the problem of handling such a huge data. Hence to overcome such a difficulty, of accessing the knowledge from large volumes of data, several other methods may be useful to integrate data [20].

2.2. Crime Pattern Discovery and Prediction

The concept of combination of FP growth and Pearson correlation of two stages is being used efficiently. A Crime tree will be constructed according to an FP tree with the input value as Crime Binary Data (CBD) where it forms crime paths with respect to the state and year accordingly [2]. To generate the maximum frequent crime sets FP MAX algorithm is followed where crime tree (CT) is given as an input which in turn returns Maximum Frequent Crime Type (MFCT). Through FP max maximum frequent crime sets are generated for total country and also for each state individually. The maximum frequent crime sets developed for each state is undergone knowledge discovery process. This gives a proper analysis in type of crimes in each state and it is easy to take preventive steps according to that. Correlation plays a key role in data analysis [3-5].

In dataset Correlation for the crime sets of total Indian states is done to get the weightage of a certain type of crime, in determining its crime Intensity point. The frequent crime sets of our country which are generated from above FP Max is taken as input for the correlation analysis. It undergoes two stages in determining weightage of a crime type [6]. The use of Pearson correlation coefficient analysis to find the linear correlation between frequent crime types [7, 8].

3. Proposed Work

The data refers to State wise persons arrested under crime against children and women by crime head which cover all states in India. The crime data contains crime incidents from 2001 to 2015. The various crimes including murder, infanticides, against women and children occurring in different states of India are visualized with in the maps, which even includes year of crimes. Data Analysis, visualization and prediction operations are used to show them analytical relationships among different attributes in the huge amount of dataset. LSTM Recurrent neural network algorithm is employed to forecast trends with the largest accuracy. Detailed analysis of the dataset is performed as follows.

3.1. Data Collection

First we collect the crime data's and verify the features of attributes. For each arrival of crime incidents in the datasets, the upcoming featured attributes are included:

1) Category- Type of the crime.
2) Descript - A short note describing details of the crime;
3) Dates - Date and time of the crime incident;
4) X - Longitude of the location of a crime;
5) Y - Latitude of the location of a crime;
6) States - Crime happened States in India;
7) Year - which year the case happened;
8) Coordinate - Pairs of Longitude and Latitude.

3.2. Data Preprocessing

It is the cleaning process of tangled data sets for doing operations and analysis. Before initiating any algorithms and operations on our datasets, a series of preprocessing steps are performed for data conditioning as presented below:

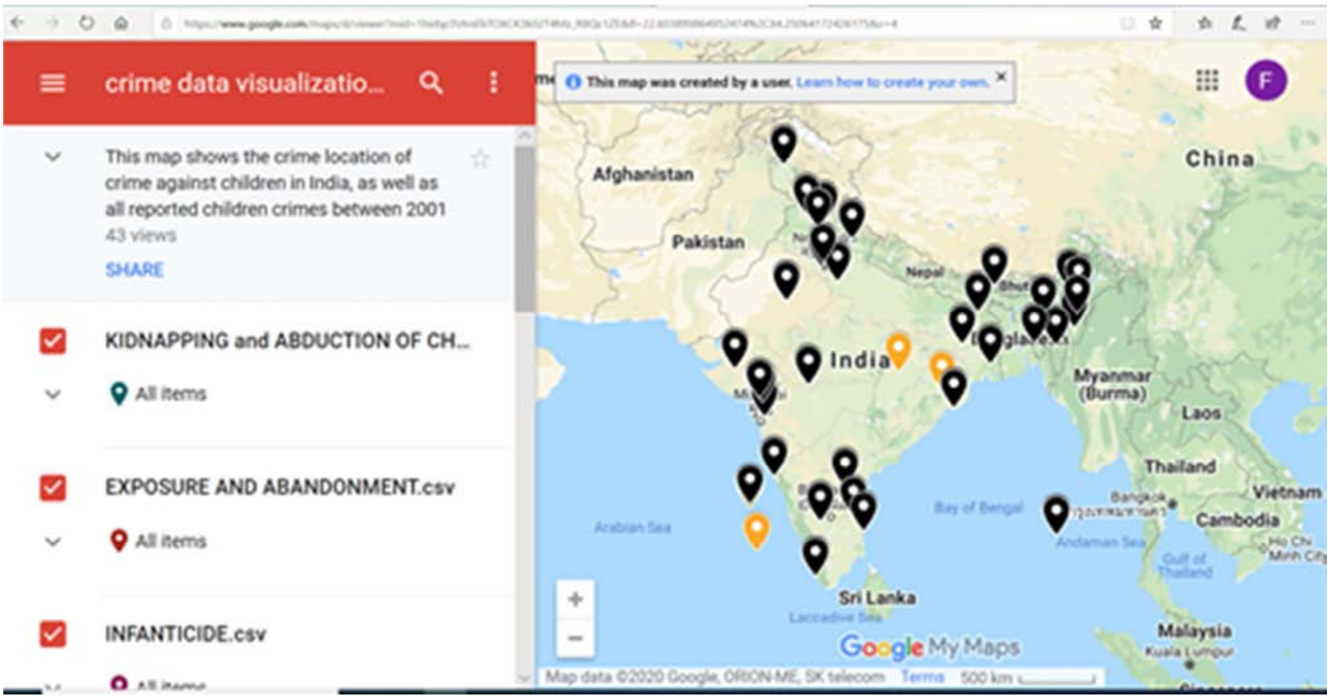
1. For some missing coordinate attributes in Indian datasets, assigned uncertain values representative from the non-missing values, computed their mean, and then replaced the missing ones.
2. We also avoid few features that needless.

3.3. Narrative Visualization

After the Data collection and Featured Attributes we perform data preprocessing and negotiating. It is the process of cleaning tangled data sets for doing operations and analysis. Take brief summary of the data frame; It helps obtain a quick overview of the data set. The summary includes a list of all columns with their data types and the number of non-null values in each column. And since our values are not null we don't have to fill the missing values.

Considering the geographic nature of the crime incidents, an interactive map based on Google map was used for data visualization, where crime incidents are grouped according to their latitude/longitude information. For geographic nature visualization, we create a custom map, then we add the data sets and crime details to the map (Figure 1).

From figure 2 it each point show the different crimes happened in India. Each points show the complete details of the particular crime.



crime data visualizatio...

This map shows the crime location of crime against children in India, as well as all reported children crimes between 2001

43 views

SHARE

☑ KIDNAPPING and ABDUCTION OF CH...

All items

☑ EXPOSURE AND ABANDONMENT.csv

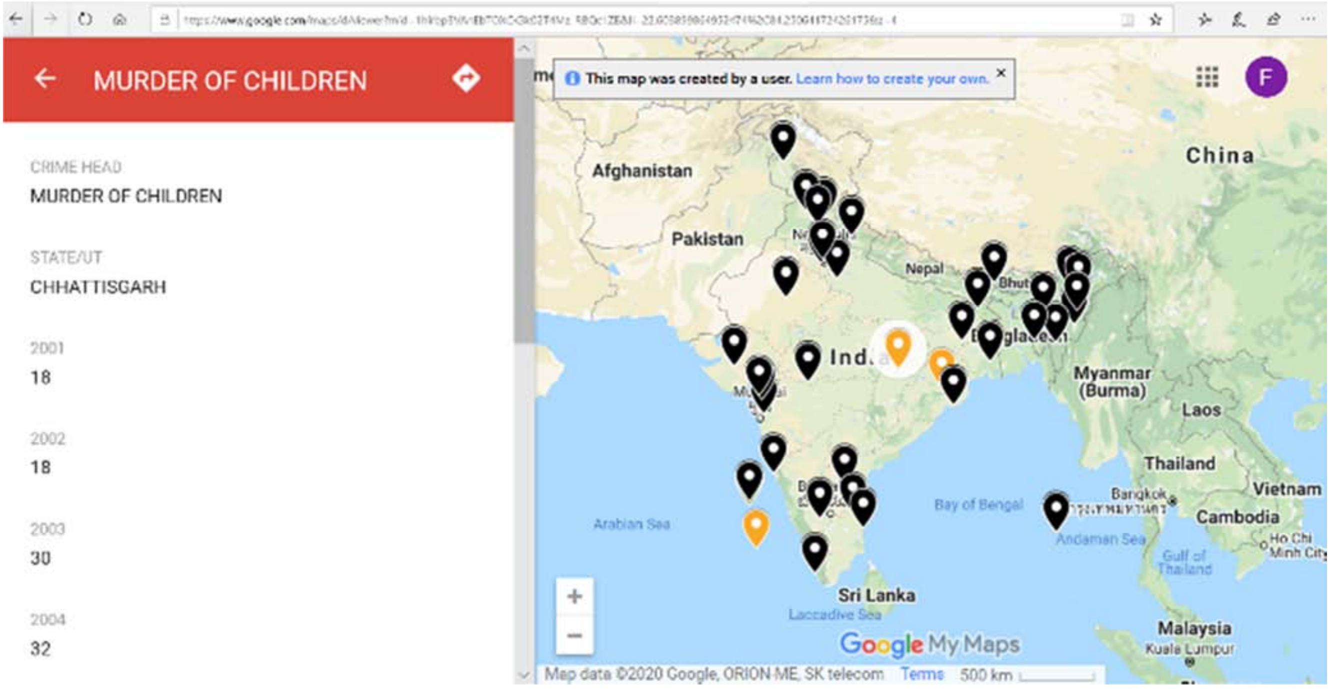
All items

☑ INFANTICIDE.csv

All items

This map was created by a user. Learn how to create your own.

Figure 1. *Crimes in Google Map.*



MURDER OF CHILDREN

CRIME HEAD
MURDER OF CHILDREN

STATE/UT
CHHATTISGARH

2001
18

2002
18

2003
30

2004
32

This map was created by a user. Learn how to create your own.

Figure 2. *Crimes Details in Google Map.*

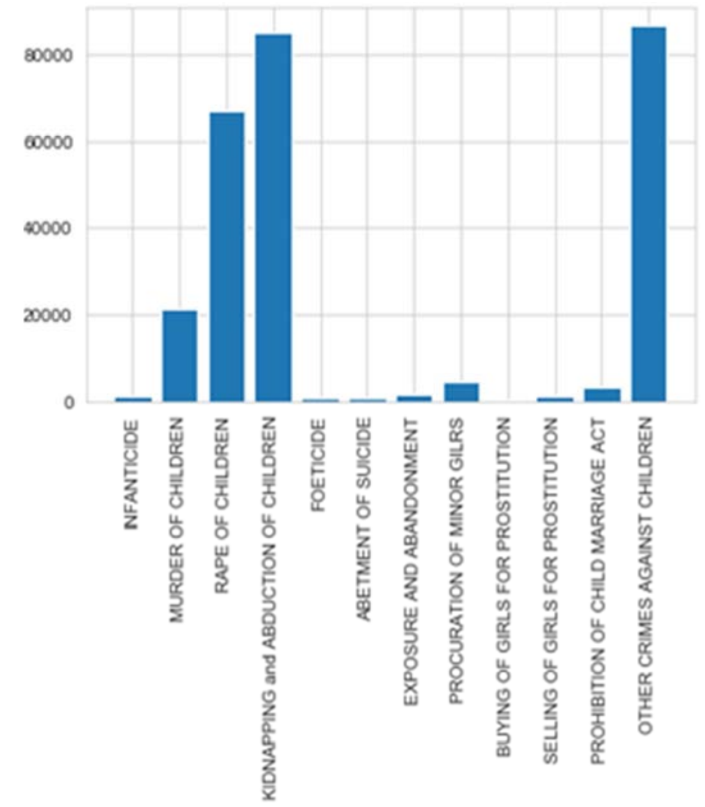


Figure 3. *Relevant crimes in India.*

Visualization of data helps to understanding the trends, outliers, and patterns in the data summarized crime incidents in each year for the Indian states and know which crime was most prevalent in India.

From Figure 3, we can see that the ‘other crimes’ category has the highest number followed by “Kidnapping” and “Rape” of children. So, we’ll focus on these types as they are more ubiquitous. We’ll analyze specific crimes one by one to get better insights. First on the list is kidnapping.

By seeing the graph Figure 4 we can point out that Uttar Pradesh has the highest number of kidnapping with a total of 30,625 cases which is 38 percentage of total states. This is a horrible number for a state, there has been problems. Because of such high rates, people are spreading fake rumors of child abduction and kidnapping. And due to this situation, the state has announced: “Charges under the National Security Act (NSA) will be pressed against the accused if such cases are reported in the future. “Now, we’ll visualize the second-highest crime i.e. Rape of children.

The graph in figure 5 shows that Madhya Pradesh records for most rape cases. To fight this, the state has adopted law to award death sentences to those found guilty of raping minors, a landmark decision in a state that recorded the highest number of child rapes. Our outcome for the objective that is Uttar Pradesh with kidnapping being the most occurring crime [11, 12]. A picture visualization (Figure 6) of states and the total number of crimes. It shows that Uttar Pradesh has the highest number of crimes.

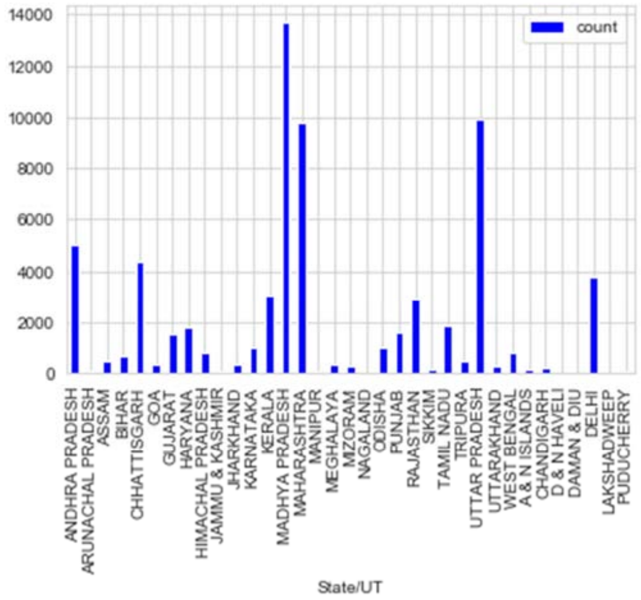


Figure 5. *Specified Crime Rape.*

Among the categories of reported crime incidents available in the datasets, the distribution of these categories is heavily skewed. As such, we focus mainly on the frequently occurred crimes and plot their distributions as percentages in figure 7. It gives an frequently occurred crimes in India.

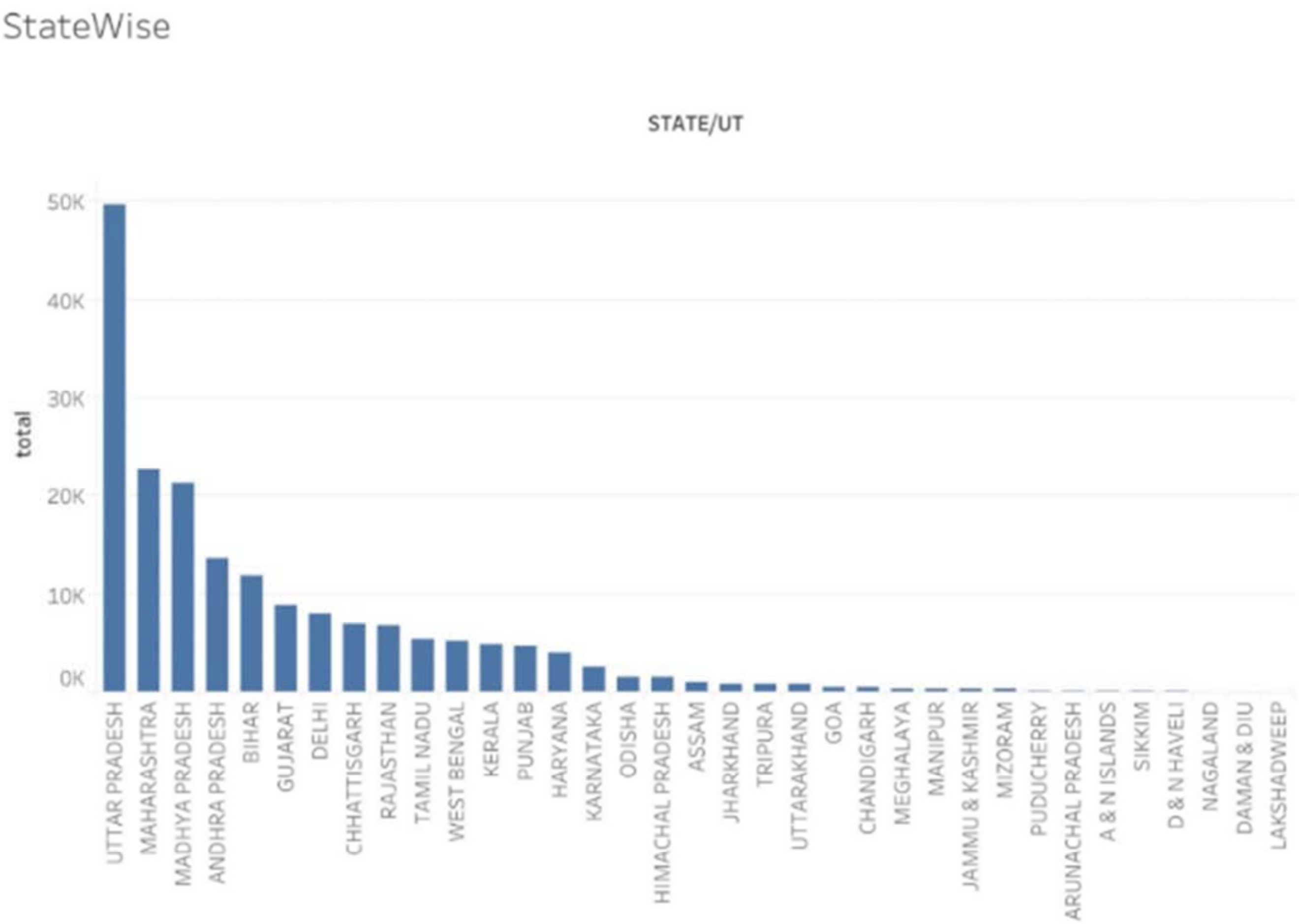


Figure 6. *States and Total number of crimes.*

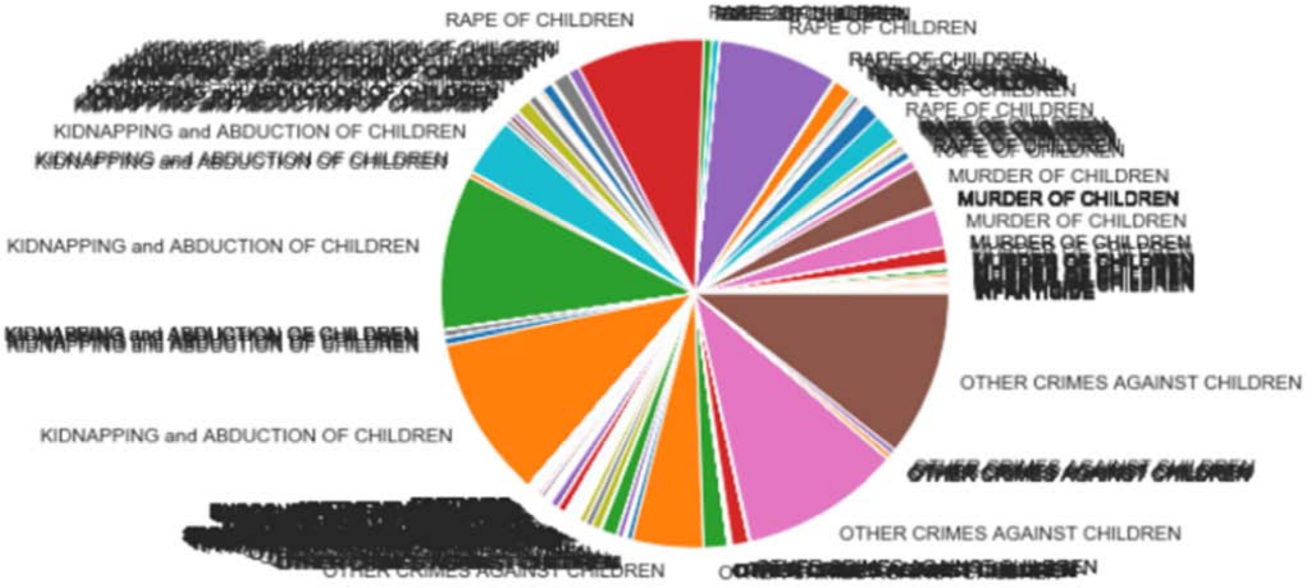


Figure 7. *Frequently Occurred Crimes.*

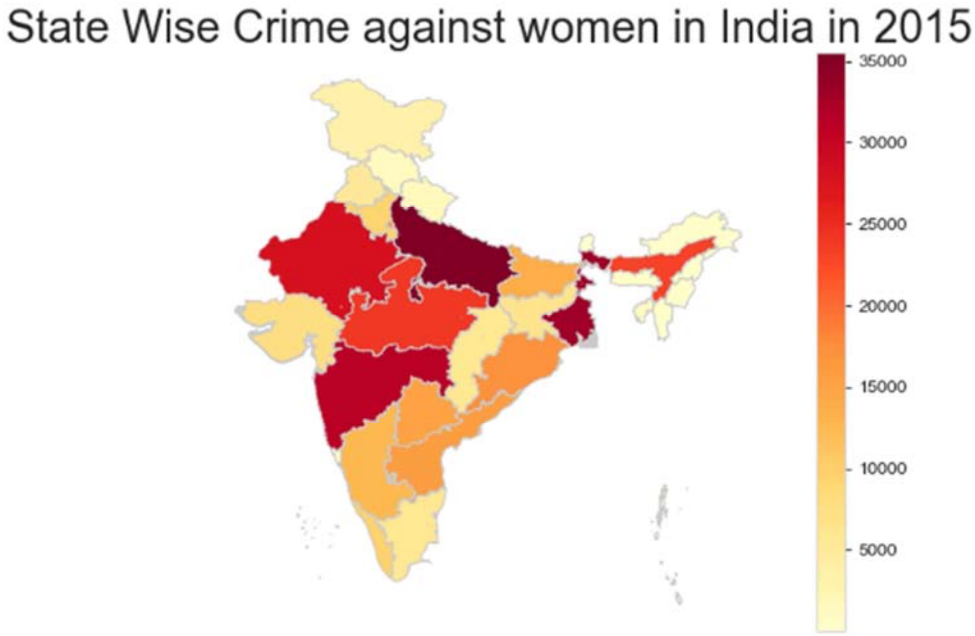


Figure 8. *State wise crime against women in India.*

Figure 8 Plot the data for crimes against women on a choropleth map of India with respect to each state. Read the Indian map shape file with district boundaries in a Geo data frame and open the datasets. Join both data frames by state names. Plot the chloropeth map.

4. Methodology

LSTM - Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) which is used in the field of deep learning. It is different from other feed forward neural networks, LSTM has feedback connections [26, 27]. The working of LSTM can be explained based on single data points as well as entire sequences of data and its processing. The different units within LSTM model includes a cell, an output gate, an input gate and a forget gate. The cell memorize values during each time intervals and the information can flow in and out based on the gates. The model can generate the future values of a time series and it can be trained using the datasets. As usual, the data gets split into training data and test data so we can later assess how well the final model performs. We take 80% of the dataset as a training data and 20 % as a text data.

For time series involves auto correlation, i.e. the presence of correlation between the time series and lagged versions of itself, LSTMs are particular utilize in prediction due to their capability of maintaining the state whilst recognizing patterns over the time series. The recurrent architecture enables the states to be persisted, or communicate between updated weights as each epoch progresses. Moreover, the LSTM cell architecture can enhance the RNN by enabling long term persistence in addition to short term.

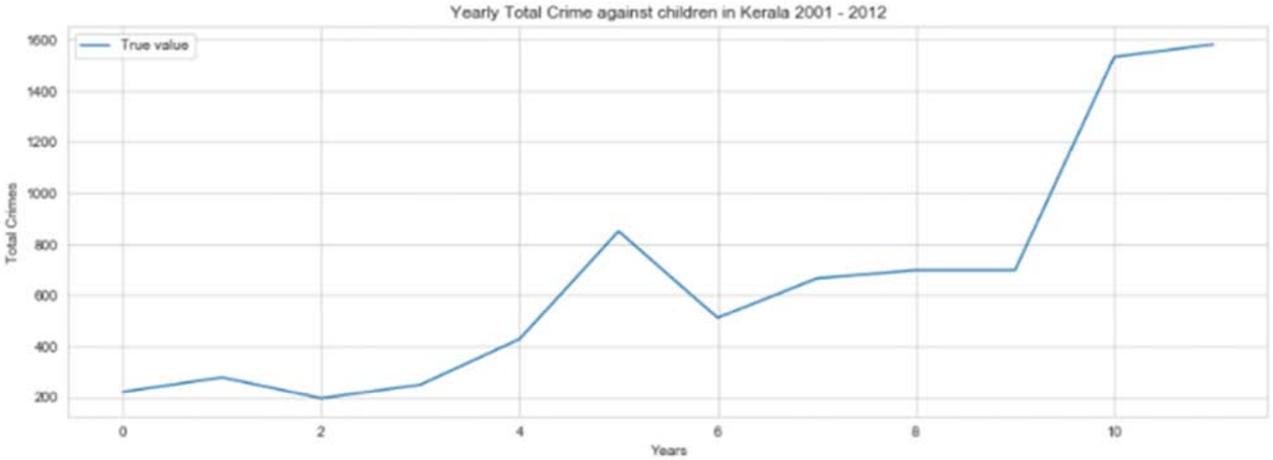
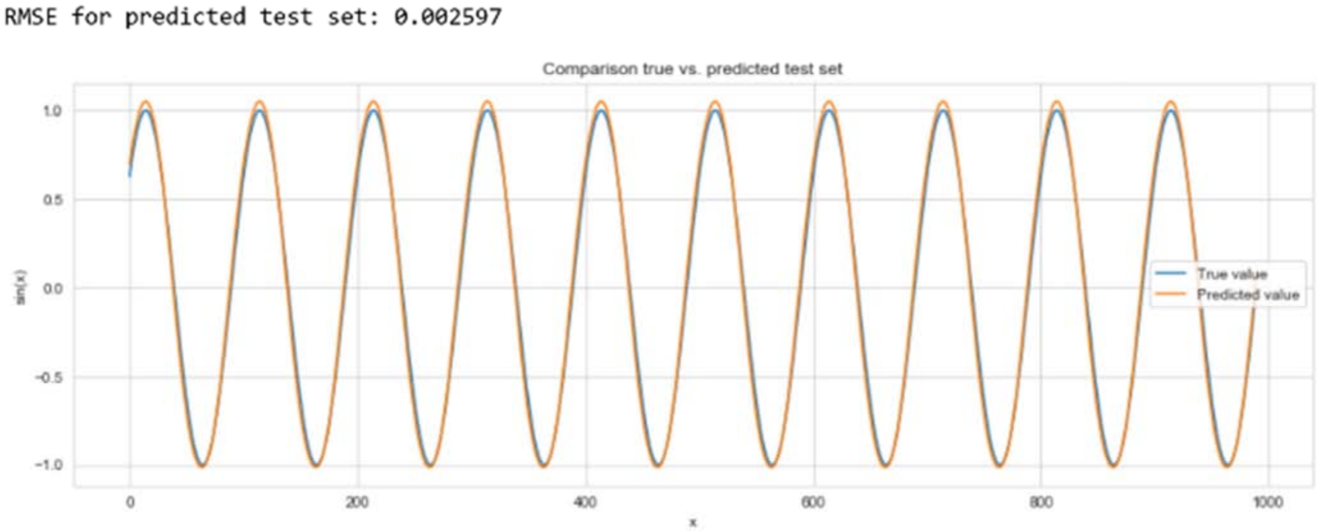


Figure 9. *Yearly total crime against children Kerala.*



```
RMSE for predicted test set: 0.002597
```

Figure 10. Comparison true vs predicted test set.

From figure 9 Y axis takes Total crimes and X axis takes years. The trends of crimes in Kerala shown in the figure. In LSTM the actual data and the predicted data always keep similarity. There is a chance of little bit error. By exploring LSTM model, found that LSTM perform better than CNN. We also found the optimal time period for the training data sample [23-25].

LSTM can be used for classifying, processing and making predictions based on time series data. The dependencies between the elements in the input sequence is being tracked by the cells. The functionality of input gate is to control the extent to which a new value flows into the cell [28, 29]. The forget gates can control the extent to which a value remains in the cell. Similarly, output gate controls the extent to which the value in the cell is used to compute the outcome activation of the LSTM unit. The activation function is represented using the logistic sigmoid function [13, 18].

5. Performance Evaluation

The performance measures for proposed system are evaluated relevant to the performance parameter Root Mean Square Error (RMSE). In the existing Neural Network Model containing higher errors than LSTM. Time series forecast these models to predict crime trends. For performance evaluation, the Root Mean Square Error (RMSE) used in terms of parameter in different sizes of training samples [14-16].

From Figure 10 we can identify that the true value and predicted value is almost same and here we got the RMSE value for predicted test set is 0.002597. It is a very low error value. The optimal time period for crime trends forecasting is 10 years where the RMSE is the minimum. The results also showed that LSTM model performed better than traditional neural network models. The neural network seems has lower RMSE but the correlation between predicted values and the real ones is low. The visualization of the trends in Figure 9 and Figure 10 also conforms this conclusion.

6. Conclusion

This paper investigated that the LSTM neural networks based approach is for predicting the future class labels of a crime incidents. To evaluate the performance of our method, we use a data set that collects all the necessary information of crime indices in the different states of India. These data are generated, which will be beneficial for the police and the enforcement department. Before applying the LSTM model, we used a pre-processing and analysis of the entire crime data. Finally we represented the different states with different crimes such as crime against women, children, murder, and kidnap.

Big Data Analytics and visualization techniques were utilized to analyze crime big data from different Indian states, which allowed us to identify patterns and obtain trends. By exploring the deep learning algorithm LSTM, we found that LSTM performs better than other conventional neural network model. Also found the optimal time period for the training will take many more years, in order to achieve the best prediction of trends in terms of Real Mean Square Error, RMSE correlation. Optimal parameters for the LSTM models are also determined. Other results explained above will provide new insights to crime trends and will assist both police departments and law enforcement agencies in their decision making.

Google map visualization helps a person; who entering in the new place to get a whole idea of the state's crime details. So he can take precautions. Finally, the performance analysis of the proposed system by considering the RMSE is done.

Acknowledgements

We appreciate everyone who helped us in our graduate studies on both professional and personal levels. We passionately acknowledge ours supervisor Prof. Neethu Subash, for providing guidance throughout the study. Finally, a special thanks to our Families for understanding and support

in our studies.

References

[1] Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi, and Qiao yuan Liu, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of CrimeData", IEEE Transactions on Creative CommonsAttribution 2019.

[2] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.

[3] Z. Jia, C. Shen, Y. Chen, T. Yu, X. Guan, and X. Yi, "Big-data analysis of multi-source logs for anomaly detection on network-based system," in Proc. 13th IEEE Conf. Autom. Sci. Eng. (CASE), Xi'an, China, Aug. 2017, pp. 1136–1141.

[4] M. Huda, A. Maseleno, M. Siregar, R. Ahmad, K. A. Jasmi, N. H. N. Muhamad, and P. Atmotiyoso, "Big data emerging technology: Insights into innovative environment for online learning resources," Int. J. Emerg. Technol. Learn., vol. 13, no. 1, pp. 23–36, Jan. 2018.

[5] J. Zakir and T. Seymour, "Big data analytics," Issues Inf. Syst., vol. 16, no. 2, pp. 81–90, 2015.

[6] Y. Wang, L. Kung, W. Y. C. Wang, and C. G. Cegielski, "An integrated big data analytics-enabled transformation model: Application to health care," Inf. Manage., vol. 55, no. 1, pp. 64–79, Jan. 2018.

[7] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1–10, Feb. 2014.

[8] J. Archenaa and E. A. M. Anita, "A survey of big data analytics in healthcare and government," Procedia Comput. Sci., vol. 50, pp. 408–413, Apr. 2015.

[9] A. Londhe and P. Rao, "Platforms for big data analytics: Trend towards hybrid era," in Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS), Chennai, 2017, pp. 3235–3238.

[10] W. Grady, H. Parker, and A. Payne, "Agile big data analytics: AnalyticsOps for data science," in Proc. IEEE Int. Conf. Big Data, Boston, MA, USA, Dec. 2017, pp. 2331–2339.

[11] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to Big Data analytics," IEEE Access, vol. 4, pp. 2542–2571, 2016.

[12] Y. Zhang, S. Ren, Y. Liu, and S. Si, "A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products," J. Cleaner Prod., vol. 142, no. 2, pp. 626–641, Jan. 2017.

[13] E. W. Ngai, A. Gunasekaran, S. F. Wamba, S. Akter, and R. Dubey, "Big data analytics in electronic markets," Electron. Markets, vol. 27, no. 3, pp. 243–245, Aug. 2017.

[14] Y.-Y. Liu, F.-M. Tseng, and Y.-H. Tseng, "Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model," Technol. Forecasting Social Change, vol. 130, pp. 123–134, May 2018.

[15] D. Fisher, M. Czerwinski, S. Drucker, and R. DeLine, "Interactions with big data analytics," Interactions, vol. 19, no. 3, pp. 50–59, Jun. 2012.

[16] S. Musa, "Smart cities — A road map for development," IEEE Potentials, vol. 37, no. 2, pp. 19–23, Mar./Apr. 2018.

[17] S. Yadav, A. Yadav, R. Vishwakarma, N. Yadav and M. Timbadia, Crime pattern detection, analysis prediction," in Proc. IEEE Int. Conf. Electron., Commun. Aerosp. Technol., Coimbatore, India, Apr. 2017, pp. 225–230.

[18] N. Baloian, C. E. Bassaletti, M. Fernandez, O. Figueroa, P. Fuentes, R. Manasevich, ´ M. Orchard, S. Penafiel, J. A. Pino, and M. Vergara, "Crime prediction using patterns ˜and context," in Proc. 21st IEEE Int. Conf. Comput. Supported Cooperat. Work Design, Wellington, New Zealand, Apr. 2017, pp. 2–9.

[19] X. Zhao and J. Tang, "Exploring transfer learning for crime prediction,"in Proc. IEEE Int. Conf. Data Mining Workshops, New Orleans, LA, USA, Nov. 2017, pp. 1158–1159.

[20] S. Wu, J. Male, and E. Dragut, "Spatial-temporal campus crime pattern mining from, historical alert messages," in Proc. Int. Conf. Comput., Netw. Commun., Santa Clara, CA, USA, 2017, pp. 778–782.

[21] K. R. S. Vineeth, T. Pradhan, and A. Pandey, "A novel approach for intelligent crime pattern discovery and prediction," in Proc. Int. Conf. Adv. Commun. Control Comput. Technol., Ramanathapuram, India, 2016, pp. 531–538.

[22] C. R. Rodr´ıguez, D. M. Gomez, and M. A. M. Rey, "Forecasting time series from clustering by a memetic differential fuzzy approach: An application to crime prediction," in Proc. IEEE Symp. Ser. Comput. Intell., Honolulu, HI, USA, Nov./Dec. 2017, pp. 1–8.

[23] A. Joshi, A. S. Sabitha, and T. Choudhury, "Crime analysis using K-means clustering," in Proc. 3rd Int. Conf. Comput. Intell. Netw., Odisha, India, 2017, pp. 33–39.

[24] N. M. M. Noor, W. M. F. W. Nawawi, and A. F. Ghazali, "Supporting decision making in situational crime prevention using fuzzy association rule," in Proc. Int. Conf. Comput., Control, Informat. Appl. (IC3INA), Jakarta, Indonesia, 2013, pp. 225–229.

[25] M. Injadat, F. Salo, and A. B. Nassif, "Data mining techniques in social media: A survey," Neurocomputing, vol. 214, pp. 654–670, Nov. 2016.

[26] Z. Zhao, S. Tu, J. Shi, and R. Rao, "Time-weighted LSTM model with redefined labeling for stock trend prediction," in Proc. IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI, Boston, MA, USA, Nov. 2017, pp. 1210–1217.

[27] J. Dai, G. Sheng, X. Jiang, and H. Song, "LSTM networks for the trend prediction of gases dissolved in power transformer insulation oil," in Proc. 12th Int. Conf. Properties Appl. Dielectr. Mater., Xi'an, China, 2018, pp. 666–669.

[28] H. Kashef, M. Abdel-Nasser, and K. Mahmoud, "Power loss estimation in smart grids using a neural network model," in Proc. Int. Conf. Innov. Trends Comput. Eng. (ITCE), Aswan, Egypt, 2018, pp. 258–263.

[29] J. Peral, A. Ferrandez, D. Gil, E. Kauffmann, and H. Mora, "A review of the analytics ´ techniques for an efficient management of online forums: An architecture proposal," IEEE Access, vol. 7, pp. 12220–12240, 2019.

Biography

Mufeeda M received Bachelor of Technology degree and specialized in Computer Science and Engineering from University College of Engineering, Thodupuzha, Kerala affiliated to Mahatma Gandhi University in 2019 and currently pursuing Master of Technology, specializing in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Machine Learning, Image processing and Data Science.

Silpa Nandanan received Bachelor of Technology degree and specialized in Computer Science and Engineering from University College of Engineering, Thodupuzha, Kerala affiliated to Mahatma Gandhi University in 2019 and currently pursuing Master of Technology, specializing in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Machine Learning, Data Mining and BigData.

Neethu Subash is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in 2011 and M-Tech in 2014 in Computer Science and Engineering from Mahathma Gandhi university. She is interested in the areas of Cyber Security solutions using Block Chain and Machine Learning.