Big Data Analysis And Mining of Crime Data Using LSTM

Poonam Dangat[1], Divya More[2], Shalini Roy[3], Rupali Nirgude[4]
[1, 2] Dept of Computer
[1, 2] D Y Patil Institute of Engineering & Technology Ambi, Pune

Abstract- *Big data analytics (BDA) may be a systematic approach for analysing as well as identifying different patterns, relations, as well as trends within a large volume of knowledge. During this paper, we apply BDA to criminal data where exploratory data analysis is conducted for visualization as well as trends prediction. Several the state-of-the art data processing as well as deep learning techniques are used. Following statistical analysis and visualization, some interesting facts and patterns are discovered from criminal data in San Francisco, Chicago, and Philadelphia. The predictive results show that the Prophet model and Keras stateful LSTM perform better than neural network models, where the optimal size of the training data is found to be three years. These promising outcomes will benefit for police departments and enforcement organizations to raise understand crime issues and supply insights which will enable them to trace activities, predict the likelihood of incidents, effectively deploy resources and optimize the choice making process.*

Keywords- Big data analytics (BDA), data mining, data visualization, neural network, time series forecasting.

I. INTRODUCTION

The crime rate is increasing day by day. Since when crime cannot be predicted. It is neither systematic nor random. Also modern technology and hi-tech methods help criminals in acquiring their misdeeds, theft, according to the Crime Records Bureau.[9] There has been a decrease in arson etc. while there have been crimes like murder, sexual exploitation, gang rape etc. Increased. Even though we cannot predict that all can be victims of crime but can predict the place where he is likely to be.

The accuracy of the predicted results cannot be 100 percent assured, but the results show that our application helps reduce crime to some extent by providing security in sensitive areas. So to create such a powerful tool for criminal analysis, we need to report and evaluate crime. Only in the last few decades has technology made a practical solution for a wider audience of law enforcement officers, making local data mining affordable and available. Since the availability of criminal data or records is not limited, we do not use websites, collecting crime data from a variety of sources. This huge data crime record is used as a record to create a database. So the main challenge we face is to develop a good, efficient criminal pattern detection tool to effectively detect crime patterns.

The main challenges we are facing are:

- Increase in crime information that's to be stored and analysed.
- Analysis of knowledge is difficult thanks to data being incomplete and inconsistent.
- Limitation in obtaining crime data records from enforcement department.
- The accuracy of the program depends on the accuracy of the training set.

Discovering patterns and trends in crime is a challenging factor. To identify a pattern, crime analysts take a lot of time, scanning through the data to find out if a particular crime fits into a known pattern. If it does not fit an existing pattern, the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, predict and prevent crime. The steps to analyze crime are:

1. Data Collection.
2. Classification.
3. Pattern Identification.
4. Prediction
5. Visualization

Due to continuous urbanization and increasing population, cities play a crucial central role in our society. However, such developments have also occurred amidst a rise in violent crimes and accidents. To affect such problems, sociologists, analysts, and security institutions have put tons of effort into mining potential patterns and factors. Apply BDA to criminal information where investigative information is examined for perception and pattern expectation. Some terms of art information mining and intensive learning systems are used. After a mediocre investigation and representation, criminal information in San Francisco, Chicago, and Philadelphia reveals some fascinating realities and examples.

II. RELATED WORK

According to Amir Gandomi et. al [1] described inBeyond the hype: Big data concepts, methods, and analytics.Size is the first, and oftentimes, the only dimension that jumps to the mention of big data. It attempts to offer a comprehensive definition of big data that captures its other unique and defined characteristics. Rapid growth and the adoption of big data by the industry have stripped the discourse for popular outlets, forcing the academic press to take hold. Academic journals in many disciplines, which would benefit from the relevant discussion of big data, have not yet covered the subject. This paper presents a consolidated account of big data by integrating the definitions of practitioners and academics. The primary focus of the paper is on the analytical methods used for big data. A distinctive distinguishing feature of this paper is its focus on unstructured data-related analytics, which comprise 95% of big data. This paper highlights the need to develop appropriate and efficient analytical methods to take advantage of large scale heterogeneous data in unstructured text, audio and video formats. This paper also reaffirms the need to develop new tools for predictive analysis for structured big data. Statistical methods in practice were designed to infer from sample data. Large size calls to develop a computationally efficient algorithm of heterogeneity, noise, and structured big data calls that can avoid large data losses, such as spurious correlation.

According to J. Zakiret. al [2] presented inBig data analytics. Today Big Data draws a lot of attention in the IT world. The rapid growth of the Internet and digital economy has led to a rapid increase in demand for data storage and analytics, and the IT department is facing a huge challenge in protecting and analyzing these enhanced versions of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data, rather it is data that includes documents, images, audio, video, and social media content called unstructured data or big Known as data. Big data analytics is a way to extract value from these huge versions of information, and it drives new market opportunities and maximizes customer retention. It primarily focuses on discussing various technologies that work together as Big Data analytics systems to predict future versions, gain insights, take proactive actions, and provide a way to make better strategic decisions. Can help Furthermore this paper analyzes the use, use, and impact of adopting a large value of data, which improves its competitive advantage by using a set of data algorithms for large data sets such as Hadoop and MapReduce.

According to Y.Wang,**et. al [3]** presented inan integrated big data analytics enabled transformation model: Application to health care. A big data analytics-enabled transformation model has been developed based on a practice-based view, revealing the causal relationship between big data analytics capabilities, IT-enabled transformation practices, profit dimensions, and business values. This model was tested in a healthcare setting. By analyzing big data implementation cases, we tried to understand how big data analytics capabilities change organizational practices, thereby generating potential benefits. In addition to conceptually defining four big data analytics capabilities, the model provides a strategic approach to big data analytics. Three important path-to-value chains for health organizations were identified by applying the model, which provides practical insights for managers.

According to U. Thongsatapornwatana,**et. al [4]** presented in a survey of data mining techniques for analyzing crime patterns. In recent years data mining is a technique used to analyze data that is used to analyze previously stored data from various sources to find patterns and trends in crime. Additionally, it can be implemented to increase efficiency in solving crimes faster and can also be applied to automatically notify crimes. However, there are many data mining techniques. To increase crime detection efficiency, it is necessary to select data mining techniques appropriately. This paper reviews the literature on various data mining applications, particularly those applying to solve crimes. The survey also highlights research gaps and the challenges of crime data mining. Additionally, this paper gives information about data patterns in crime so that crime patterns and trends can be used appropriately and can be of help to beginners in the research of crime data mining.

According to W. Raghupathi,et. al [5] presented in Big data analytics in healthcare: Promise and potential. To describe the promise and potential of big data analytics in healthcare. The paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses challenges, and Presents the conclusion. It provides a comprehensive overview of big data analytics for R healthcare researchers and practitioners. : Big data analytics in health services is evolving into a promising field for providing information from very large data sets and improving outcomes while reducing costs. Its efficiency is great; however, challenges remain to be overcome.

According to J. Archenaa,et. al [6] presented in asurvey of big data analytics in healthcare and government. This gives information about how we can uncover additional value from the data generated by the healthcare and government. Large amounts of heterogeneous data are generated by these agencies. But these data became useless without proper data analysis methods. Big data analytics using Hadoop plays an effective role in performing meaningful real-time analysis on vast amounts of data and is able to predict emergency situations before this happens. It deals with big data use cases in healthcare and government.

According to A. Londhe,et. al [7] presented in Platforms for big data analytics: Trend towards hybrid era. The primary objective of this paper is to present a detailed analysis of various platforms suitable for Big Data processing. In this paper, various software frameworks available for Big Data Analytics are surveyed and their strengths and weaknesses are discussed in detail. In addition, widely used data mining algorithms for their optimization for big data analysis are discussed. W.r.t their suitability for handling real-world problems. Future trends of Big Data Processing and Analytics can be estimated by considering the strength of software frameworks and platforms with effective implementation of these well-established and widely used data mining algorithms. Hybrid approaches (integration of two or more platforms) may be more suitable for a specific data mining algorithm and can be highly adaptable to real-time processing as well.

According to W. Grady,et. al [8] presented inAgile big data analytics: AnalyticsOps for data science. Big Data Analytic (BDA) leverages data distribution and parallel processing across a set of system resources. This in particular introduces many new challenges to analytics. The analytics part of the entire lifecycle typically follows a waterfall process - completing one step before the next start. Although attempts have been made to map a variety of analyzes into agile methodologies, the steps are often described as breaking activities in small tasks, while the overall process is still consistent with the step-by-step waterfall. BDA Analytics changes many activities throughout the lifecycle, as well as their orders. The goal of agile analytics is to reach a point of compatibility between the time it takes to generate value from the data and the time it takes to get there. This paper discusses the implications of an agile process for BDA in cleaning, transformation, and analysis.

According toR. Vatrapu,et. al [9] presented inSocial set analysis: A set theoretical approach to Big Data analytics. Current analytical approaches in computational social science can be characterized by four major paradigms: text analysis (information extraction and classification), social network analysis (graph theory), social complexity analysis (complex systems science), and social simulation (cellular automata and agents -) based modeling). However, when it comes to organizational and social units of analysis, one can conceptualize, model, analyze, interpret, and predict social media interactions as associations of ideas, values, identities, and so on. The approach does not exist. To address this limitation based on the sociology of associations and the mathematics of set theory, this paper presents a new approach to big data analytics called social data analysis. Social set analysis consists of a basic framework for the philosophy of computational social science, the theory of social data, conceptual and formal models of social data, and an analytical framework for combining large social data sets with organizational and social data sets. . Three empirical studies of big social data are presented to illustrate and demonstrate social set analysis in terms of fuzzy set-theoretic sentiment analysis, crisp set-theoretic interaction analysis, and event-study-oriented set-theoretic visualization. The implications for big data analytics, the current limitations and future directions of the set-theoretic approach are outlined.

According to Y. Zhang,et. al [10] presented inA big data analytics architecture for cleaner manufacturing andmaintenance processes of complex products.Cleaner production (CP) is considered as one of the most important means for manufacturing enterprises to achieve sustainable production and improve their sustainable competitive advantage. However, implementation of the CP strategy was facing hurdles, such as lack of complete data and valuable knowledge employed to provide better support on product lifecycle management (PLM) and coordination and optimization decisions over the entire CP process. can go. . Fortunately, with the widespread use of smart sensing devices in PLM, real-time and multi-source life cycle big data can now be collected. To make better PLM and CP decisions based on these data, in this paper, a holistic architecture of big data-based analytics for product lifecycle (BDA-PL) was proposed. It integrated big data analytics and service-driven patterns that helped overcome the above obstacles. Under the architecture, availability and access to product-related data and knowledge were achieved. Focusing on the manufacturing and maintenance process of the product lifecycle and developing key technologies to implement big data analytics were developed. The presented architecture was demonstrated by an application scenario, and some observations and conclusions were discussed in details. The results showed that the proposed architecture benefited customers, manufacturers, the environment and even all stages of PLM and effectively promoted the implementation of CP. In addition, managerial implications of the proposed architecture for the four departments were analyzed and discussed. The new CP strategy provided a theoretical and practical basis for the sustainable development of manufacturing enterprises.

III. OBJECTIVES OF SYSTEM

Following are the objectives of the system:

- To store and analyze large volume of crime data.
- To reduce the time of investigation due to complexity issues in investigation by analyzing crime patterns.
- To implement system which capable to handle large dataset.
- To enables, identify users to explore, compare, and analyse evolutionary trends and patterns of crime incidents.

IV. PROPOSED METHODOLOGY

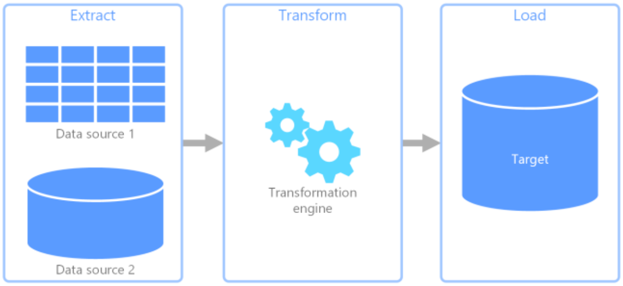

Figure 1 System Architecture

Following are the modules implemented in the proposed system as below:

1. Data Pre-processing

Before implementing any algorithms on our datasets, a series of pre-processing steps are performed for data conditioning as presented below:

- Time is discretized into a couple of columns to allow for time series forecasting for the overall trend within the data.
- For some missing coordinate attributes in Chicago and Philadelphia datasets, we imputed random values sampled from the non-missing values, computed their mean, and then replaced the missing ones.
- The timestamp indicates the date and time of occurrence of each crime, we deduced these attributes into features: Year (2003-2017), Month (1-12), Day (1-31), Hour (0-23), and Minute (0-59).
- We also omit some features that unneeded like incident-Num, coordinate.

2. Data Analysis and Visualization

The three crime datasets we used for analysis are publicly available, which cover 3 cities in US, i.e. San-Francisco, Chicago, and Philadelphia. The San-Francisco crime data contains2,142,685 crime incidents from 2003 to 2017. Data from Chicago has a total number of5,541,398 records, dating back from 2017 to 2003. In the Philadelphia dataset, there are2,371,416 crime incidents which were captured from 2006 to 2017.

A. Featured Attributes

For each entry of crime incidents in the datasets, the following13 featured attributes are included:

1) Incident Num - Case number of each incident;
2) Dates - Date and timestamp of the crime incident;
3) Category - Type of the crime. This is the target/label that we need to predict in the classification stage;
4) Descript - A brief note describing any pertinent details of the crime;
5) Day of Week - Day of the week that crime occurred;
6) PdDistrict - Police Department District ID where the crime is assigned;
7) Resolution – How the crime incident was resolved (with the perpetrator being, say, arrest or booked);
8) Address - The approximate street address of the crime incident;
9) X - Longitude of the location of a crime;
10) Y - Latitude of the location of a crime;
11) Coordinate - Pairs of Longitude and Latitude;
12) Dome - whether crime id domestic or not;
13) Arrest - Arrested or not;

V. SYSTEM ANALYSIS

Algorithm used

Algorithm 1: Long Short-Term Memory (LSTM)

LSTM networks even have long term memory and thus are capable of handling long term dependencies. it's described how LSTM is employed to process indexed data using gate vectors at each position to regulate the passing of data along the sequence. Whenever there's a group of vectors at step t.

Step 1: Read data from current file system using below formula:

$$\sigma(x) = \frac{1}{1 - e^{-x}}$$

Step 2: $\mathrm{ht} = \sigma(\mathrm{W} \times x_t + \mathrm{U} \times h_{t-1} + \mathrm{b})$

Step 3: $\int t_{=} \sigma(\mathrm{W} f \times \mathrm{xt} + \mathrm{U} f \times \mathrm{ht} - 1 + \mathrm{b} f)$

Step 4:it= $\sigma(\mathrm{Wi} \times \mathrm{xt} + \mathrm{Ui} \times \mathrm{ht} - 1 + \mathrm{b}i)$

Step 5:C t= $tanh(\mathrm{WC} \times \mathrm{xt} + \mathrm{UC} \times \mathrm{ht} - 1 + \mathrm{b}C)$

Step 6: Ct= $it \times C\,t + ft + Ct - 1$

Step 7: *ot*= $\sigma(\mathrm{W}o \times \mathrm{xt} + \mathrm{U}o \times \mathrm{ht} - 1 + \mathrm{b}o)$

Step 8: ht=$ot \times \tanh(Ct)$

VI. CONCLUSION

In this project, we used some known approaches to crime analysis and related prediction with Data Mining. Data mining techniques are used to identify criminals also to provide a better future for society and living. In this project what is the progress of best in class best information test; the representation system was used to obtain inaccurate information from three Americans Urban area, which enabled us to distinguish examples and obtain patterns. By investigating Prophet Model, a nervous system model, and computation of intensive learning LSTM, we found that both the Prophet model and the LSTM calculation outperformed the traditional Neural system model.

REFERENCES

[1] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,”Int. J. Inf. Manage., vol. 35, no. 2, pp. 137144, Apr. 2015.

[2] J. Zakir and T. Seymour, “Big data analytics,” Issues Inf. Syst., vol. 16, no. 2, pp. 8190,2015.

[3] Y.Wang, L. Kung,W. Y. C.Wang, and C. G. Cegielski, “An integrated big data analyticsenabledtransformation model: Application to health care,” Inf. Manage., vol. 55, no. 1,pp. 6479, Jan. 2018.

[4] U. Thongsatapornwatana, “A survey of data mining techniques for analyzing crime patterns,”in Proc. 2nd Asian Conf. Defence Technol., Chiang Mai, Thailand, 2016, pp.123128.

[5] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: Promise and potential,” Health Inf. Sci. Syst., vol. 2, no. 1, pp. 110, Feb. 2014.

[6] J. Archenaa and E. A. M. Anita, “A survey of big data analytics in healthcare and government,”Procedia Comput. Sci., vol. 50, pp. 408413, Apr. 2015.

[7] 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. 32353238.

[8] 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. 23312339.

[9] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, “Social set analysis: A settheoretical approach to Big Data analytics,” IEEE Access, vol. 4, pp. 25422571, 2016.

[10] Y. Zhang, S. Ren, Y. Liu, and S. Si, “A big data analytics architecture for cleaner manufacturingand maintenance processes of complex products,” J. Cleaner Prod., vol. 142,no. 2, pp. 626641, Jan. 2017.

[11] E.W. Ngai, A. Gunasekaran, S. F.Wamba, S. Akter, and R. Dubey, “Big data analytics inelectronic markets,” Electron. Markets, vol. 27, no. 3, pp. 243245, Aug. 2017.

[12] Y.-Y. Liu, F.-M. Tseng, and Y.-H. Tseng, “Big Data analytics for forecasting tourismdestination arrivals with the applied Vector Auto regression model,” Technol. ForecastingSocial Change, vol. 130, pp. 123134, May 2018.