IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Detecting Pattern in Crime Analysis and Identifying the Criminals using Big Data Techniques

Reni Hena Helan R[1], Abirami G[2], Sultan Saleem A[3], Divya S[4], Monisha M[5], Kamatchi R[6]
Assistant Professor, Department of Computer Science and Engineering[1,2,3]
Student, Department of Computer Science and Engineering[4,5,6]
Dhanalakshmi College of Engineering, Chennai, India

Abstract: *Criminal behaviour is one of our society's most serious issues. With the resurgence of such activities around the world on a daily basis, crime investigation organisations are finding it increasingly difficult to manage and probe events, either due to a lack of cops or because criminals are outsmarting the investigative process. The traditional police investigative procedure takes a long time to predict criminal profiles, suspect the next prospective crime location, or understand the crime trend. As a result, there is a need to evaluate historical crime trends more efficiently in less time, as well as anticipate future crime location and type. The police department requires a systematic method for quickly evaluating criminal profiles and identifying linked criminals. For criminal activity monitoring, an advanced analytics system is also required to track additional information such as traffic sensors, calls, videos, and police service calls, among other things. We highlighted how Big Data-based data analysis approaches can be used to avoid dealing with such situations in this paper. Furthermore, we have examined various data gathering methodologies, including Volunteered Geographic Information (VGI), Geographic Information System (GIS), and Web 2.0. The prediction based on data gathering and analysis will be the final phase. It will be accomplished through the use of Machine Learning to predict and prevent future crimes.*

Keywords: Crime, Geographic Information System, Big data, Map Reduce

I. INTRODUCTION

Criminal behaviour is a problem that affects people all over the world. Crime incidences have risen dramatically over time, posing a serious threat to our society. With the use of new technologies, criminals have become far wiser than crime investigation authorities (police departments). According to Indian statistics, 95 lakh crimes were registered in 2016. Several factors can contribute to criminal activity. It is harder for agencies to manage and investigate occurrences, either because of a lack of cops or because criminals are more proactive. Every day, they come up with new ways to accomplish it. Traditional police department surveillance and investigative methods take longer to predict criminal profiles, suspect the next prospective crime site, or understand the crime trend. A Geographic Information System (GIS) is a system that belongs to the Geoinformatics domain. This All georeferenced information can be captured, stored, modified, shared, and analysed using this system. It is a larger phrase that refers to a repository of data gathered via online search engines (e.g., Google Earth, Google Maps, Microsoft's Virtual Earth, Wikimedia). Previously, GIS was primarily a business tool, with users using it to access data from an interactive web service. However, it has evolved over time. Users can now update maps, share information, and make changes to it. As a result, they constitute a significant contributor to GIS. Volunteered Geographic Information (VGI) has broadened the scope of georeferenced data available on the internet. It broadens the scope of Geoinformatics even further. Through an online or mobile crime incident reporting tool, or through social media, VGI can assist in obtaining information regarding crimes. VGI, in conjunction with Geographic Information Systems (GIS), can assist in the reporting of small crimes and accidents in local regions, as well as enriching the criminal analysis process. It introduces a novel infrastructure idea that uses georeferenced technologies, user handsets, and a geo database to gather, create, validate, exchange, and analyse data. It can also collect information from users' devices, such as mobile cameras and social media accounts, both locally and at a higher level. At the local level, it can be accomplished by capturing footage with the user's camera and uploading it to a reporting programme. Citizens can record and save information in web portals, or publish posts and tag friends on social media sites, which can aid in more efficiently analysing trends using Big Data, starting at the lowest level. VGI is information supplied by citizens, particularly local

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

residents, which can generate a large amount of data, allowing for more efficient analysis. It is quite simple to post information about difficulties on the web using Web 2.0. In this age of social networking, any information and incidences (even petty crimes) may be simply broadcast on the internet and evaluated using Big Data. It may analyse information by blending data that is officially stored by police departments as well as data collected by users utilising web or mobile crime incident reporting applications and social media networks using Hadoop sophisticated analysis framework of Big Data analytics [5]. With the use of crime patterns, Hadoop can readily forecast future crimes and their locations in a timely manner.

II. PROBLEM STATEMENT

Data mining is sensitive to the quality of input data, which may be incomplete or erroneous (noise, redundant data).Mapping real data to data mining attributes can be difficult in and of itself. Big data is widely used to transform large amounts of unstructured or structured raw data into critical and meaningful information, which aids in the formation of a healthy decision support system for the judiciary and legislature in enforcing law and order and making strategic decisions for the safety and well-being of society.

III. EXISTING SYSTEM

The K-means clustering algorithm has been applied using RDBMS in the existing system, however it has difficulties such as data restrictions, long processing times, and data recovery issues.

IV. PROPOSED SYSTEM

The suggested system uses Hadoop capabilities such as HDFS and map reduce programmes to provide a database with high throughput and minimal maintenance costs. The project will be built in Hadoop utilising joins, partitions, and bucketing techniques, and will be visualised using R Tool. The proposed schema could be expanded with recommendations for actions, appropriate measures, and constructive policies based on the type of offence and the subsequent criminal incidence. This pairing will improve security, monitoring, handling, and the prevention of criminal acts.

V. BIG DATA FRAMEWORK

The Big Data framework is divided into five key components [6]. 1) Resource Manager 2) Cluster computing framework, 3) Data warehousing and computing, 4) Data storing and management, and 5) Data visualisation and analysis Figure 1 shows how these elements fit together.

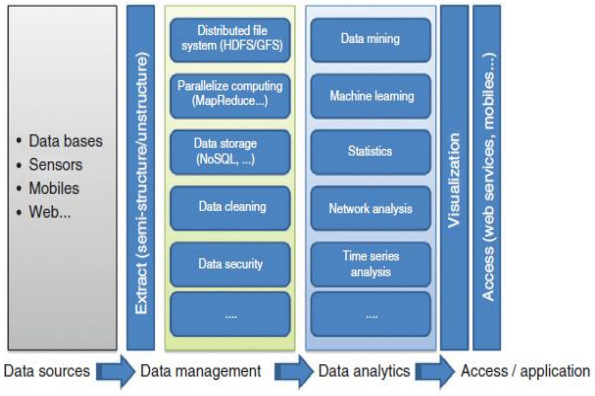

Fig. 1 Big data Framework

5.1 Resource Manager

When it comes to analysing data utilising Big data, resource management is critical. It controls all resource requirements, synchronisation between various resources, increased utilisation, increased performance, decreased management costs, increased Big Data framework compatibility, and reliability. The resource manager synchronises active and inactive

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

nodes while also separating storage and processing components. It includes a Program Master that determines how many processing resources are required to run the complete application, increasing environment orchestration. Mesos and Yarn are two popular resource management and scheduling platforms in Big Data. Mesos and Yarn are designed to maximise cluster resource use by sharing processing frameworks such as Hadoop, MPI, Spark, or several instances of the same framework.

5.2 Cluster Computing Framework

The Data Analytics computing framework is designed for massive clusters of distributed data computation. Batch processing computation and real-time processing computation are two types. MapReduce for offline processing, Hadoop for batch processing, Spark for iterative computing, Hortonworks for batch-oriented processing, Storm for online processing, MPI for high performance, Flume for stream processing, Kafka for real-time event processing, and a data mining and streaming processing framework for S4 are all supported.

5.3 Data Warehousing and Computing

This layer contains data warehouses such as Hive for analysing clusters of data, cluster computing frameworks such as Shark for analysing information using SQL-like queries, and data flow languages such as Python. Pig shortens the time it takes to write queries, Mahout shortens the time it takes to learn new things, and Drill shortens the time it takes to analyse huge databases. The data warehouse is connected to the computing framework via servers, and the computing framework supports scalable No SQL databases like H-base .

5.4 Data Storing and Management

This layer deals with the storage of organised, semi-structured, and unstructured data in databases, as well as challenges including scalability, replication, and increased throughput. It can be classified into two groups. 1) RDBMS (Relational Database Management System) and 2) NoSQL Database Management System.

5.5 Data Visualisation and Analysis

Visualize and view results using interactive dashboards. Several tools, such as D3.js [13], can be used to visualise data. Different software, such as Tableau, RapidMiner, and R Tool, can be used to visualise data. It assists anyone in swiftly analysing, visualising, and sharing data. Tableau may leverage Apache Hive as the defacto standard for SQL access in Hadoop (through ODBC connection). Data visualisation will be used to evaluate patterns in order to make future predictions and forecasts about information.

VI. TECHNOLOGIES/PLATFORMS USED

6.1 Hadoop

Apache Hadoop is a platform for distributed processing of massive data sets across clusters of computers using simple Programming concept, It can scale from a single server to thousands of workstations, each with local computing and storage. HDFS and YARN are the two main components of Apache Hadoop.

6.2 Hive

Hive is a Hadoop-based data warehouse for querying and analysing data. it was Created by Facebook and intended for anyone who knows SQL. Hive employs HiveQL, a SQL-like language, for data searching and analysis. HiveQL's function is to manage and query structured data. It also cuts down on the time it takes to write Hive programmes. This abstracts the complexity of Hadoop. Tables in the Hive database can be partitioned and bucketed. It's simple to add custom mapper/reducer code. Hive uses an inbuilt Apache Derby database to store metadata. The metastore is kept at the hive warehouse. The schemas of the tables are stored in the metastore.

6.3 Sqoop

Sqoop is a command-line utility that allows you to move large volumes of data between HDFS and relational databases. It can used to transport data between relational databases and Hive. Sqoop is a tool for importing and exporting tables

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

and whole databases to HDFS. It creates Java classes that allow you to interact with the data you've imported. We can choose the file format in which data will be loaded to HDFS during importing data.

VII. IMPLEMENTATION

The process will be implemented in stages, as depicted in Figure 2 below,

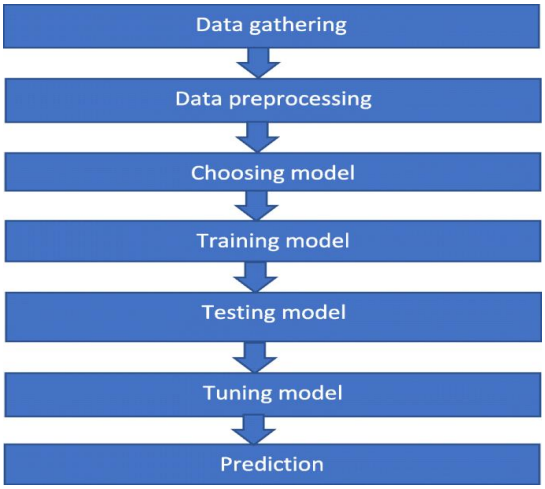

Fig. 2 Implementation Steps

7.1 Data Collection

One of the most crucial jobs in the development of a machine learning model is data collection. The specific dataset is obtained from a newspaper. Unwanted data is also present in the dataset Therefore, you must first preprocess the data to generate the appropriate dataset for your method. We provided information on crime records such as ID, date, time, day, location, x, y, assault, murder, rape, theft and other crimes.

7.2 Data Preprocessing

To analyze it and provide useful results, we collect task-related data based on a set of target factors. Some of the data, on the other hand, may be noisy, with figures that are inaccurate, incomplete, or wrong. As a result, data must be processed before it can be analysed and conclusions drawn. Data pre-processing includes things like data cleansing, data transformation, and data selection.

7.3 Data Analysis

1) MapReduce for analysis: The streaming data obtained from Twitter and Facebook is refined using MapReduce. The data is stored in HDFS in an unstructured manner by Apache Flume. The MapReduce programming model is used to refine data with appropriate logic. There are three classes in a MapReduce programme: Driver, Mapper, and Reducer .HIVE- based analysis: For VGI, Web 2.0, and historical crime datasets, Hive tables will be generated. Data from the VGI application, Twitter and Facebook data, and historical crime data sources will be put into Hive tables from HDFS. When storing data in a Hive table, the Optimized Row Columnar (ORC) file format is employed. It will give quick columnar data access while using minimal storage capacity.

7.2 Data Visualisation

RapidMiner will be used to visualise Hive findings using the Nave Bayes technique. Decision Tree, Nave Bayes, Logistic Regression, Deep Learning, Random Forest, Support Vector Machine, and Fast Large Margin are some of the algorithms available in RapidMiner. In this study, we will utilise the Naive Bayes algorithm to depict crime data in order to see different statistical analyses and predictions of crime in different states. It will be utilised to create an interactive and shareable dashboard that will allow end users to visualise data trends, variations, patterns, and density using graphs and charts.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

VIII. RESULTS

After gathering, processing, importing, and analysing data using various Hadoop framework modules, the data will be visualised and predicted using the Nave Bayes Machine Learning method. This was done using the RapidMiner programme. Data on crime was gathered from the sources listed in the data collecting section.. The data will be saved in a tool that may be used to define and apply models.

Fig. 3 Crime Front End

DATA INFORMATION

	Dates	Category	Descript	DayOfWeek	PdDistric
4	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOC...	Wednesday	PARK
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOC...	Wednesday	NORTHE
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHE
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHE
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHE

Fig. 4 Data Information

DATA PREPROCESSING

	Id	Dates	DayOfWeek	PdDistrict	Address	X	Y
0	0	10-05-15 23:59	Sunday	BAYVIEW	2000 Block of THOMAS AV	-122.3996	37.7351
1	1	10-05-15 23:51	Sunday	BAYVIEW	3RD ST / REVERE AV	-122.3915	37.7324
2	2	10-05-15 23:50	Sunday	NORTHERN	2000 Block of GOUGH ST	-122.4260	37.7922
3	3	10-05-15 23:45	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.4374	37.7214
4	4	10-05-15 23:45	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.4374	37.7214

(884262, 7)

Fig. 5 Data pre-processing

Test DayOfWeek

	DayOfWeek
0	7
1	7
2	7
3	7
4	7
5	7
6	7
7	7
8	7
9	7

Fig. 6 Test Day of Week

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Train DayOfWeek

	DayOfWeek
0	3
1	3
2	3
3	3
4	3
5	3
6	3
7	3
8	3
9	3

Fig. 7 Train Day of Week

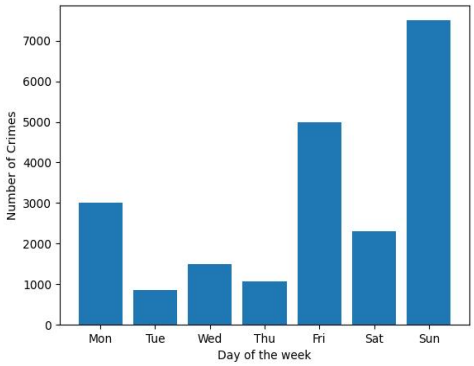

Fig. 8 Day wise Crime Analysis

IX. CONCLUSION

Data from web/mobile crime reporting applications, tags of twitter handles and other social platforms, as well as VGI was utilised to compile the Crime Info repository's crime data set. Taking into account the many characteristics of the crime data, an efficient analysis was carried out, which aided in the study of the incidents as well as the forecast of future crime and its location. All of this was accomplished with the help of the Big Data and Machine Learning framework. According to the analysis, the architecture described in this paper will save the police department money and time by stationing officers in areas that are more sensitive and predictive of crime. It has the ability to analyse at a granular level, which aids in determining the core cause of criminal acts. It will also cut down on the time it takes to investigate a crime because the individual who reported the problem using VGI and Web 2.0 could also be a witness to the crime.

REFERENCES

[1] "Crime Statistics," data.gov.in. [Online]. Available: https://data.gov.in/dataset-group-name/crime-statistics. [Accessed: 07-May-2019].
[2] "Geo BI and Big VGI for Crime Analysis and Report - IEEE Conference Publication." [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7300856. [Accessed: 15-Apr-2019].
[3] J. L. Mohamed Bakillah, "Exploiting Big VGI to Improve Routing and Navigation Services," Big Data, 18-Feb-2014. [Online]. Available: https://www.taylorfrancis.com/. [Accessed: 15-Apr-2019].
[4] R. Broadhurst, P. Grabosky, M. Alazab, B. Bouhours, and S. Chon, "An Analysis of the Nature of Groups Engaged in Cyber Crime," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2461983, Feb. 2014.
[5] J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi, "Big data analysis using Apache Hadoop," in 2013 IEEE 14th International Conference on Information Reuse Integration (IRI), 2013, pp. 700–703.
[6] Ishwarappa and J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," Procedia Computer Sci., vol. 48, pp. 319–324, Jan. 2015.
[7] "Electron: Towards Efficient Resource Management on Heterogeneous Clusters with Apache Mesos - IEEE Conference Publication." [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8030597. [Accessed: 15-Apr-2019].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

[8] "Apache Hadoop 2.7.4 – MapReduce NextGen aka YARN aka MRv2." [Online]. Available: https://hadoop.apache.org/docs/r2.7.4/hadoop-yarn/hadoop-yarn-site/. [Accessed: 15-Apr-2019].
[9] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," J. Big Data, vol. 2, no. 1, p. 8, Oct. 2014.
10. "Big data emerging technologies: A CaseStudy with analyzing twitter data using apache hive - IEEE Conference Publication." [Online].Available:https://ieeexplore.ieee.org/abstract/document/745340 .
[9] T. Phyu, ―Survey of classification techniques in data mining,‖ Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol. IIMECS 2009, March 18 - 20, 2009, Hong Kong.
[10] S.B. Kim, H.C. Rim, D.S. Yook, and H.S. Lim, ―Effective Methods for Improving Naïve Bayes Text Classifiers,‖ In Proceeding of the 7th Pacific Rim International Conference on Artificial Intelligence, Vol.2417, 2002.
[11] S. Sindhiya, and S. Gunasundari, ―A survey on Genetic algorithm based feature selection for disease diagnosis system,‖ IEEE International Conference on Computer Communication and Systems(ICCCS), Feb 20- 21, 2014, Chennai, INDIA.
[12] P. Gera, and R. Vohra, ―Predicting Future Trends in City Crime Using Linear Regression,‖ IJCSMS (International Journal of Computer Science & Management Studies) Vol. 14, Issue 07Publishing Month: July 2014.
[13] L. Ding et al., ―PerpSearch: an integrated crime detection system,‖ 2009 IEEE 161-163 ISI 2009, June 8-11, 2009, Richardson, TX, USA.
[14] K. Bogahawatte, and S. Adikari, ―Intelligent criminal identification system,‖ IEEE 2013 The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka.
[15] A. Babakura, N. Sulaiman, and M. Yusuf, ―Improved method of classification algorithms for crime prediction,‖ International Symposium on Biometrics and Security Technologies (ISBAST) IEEE 2014