International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Behavior Analysis and Crime Prediction using Big Data and Machine Learning

Pranay Jha, Raman Jha, Ashok Sharma

***Abstract*: Criminal activity is one of the major problems in our society. With the revival of such activities globally every day, it is quite difficult to manage and investigate the incidents by crime investigation agencies either because of less head counts of cops or criminals are smarter than investigation process. Traditional process of investigation for police department takes quite longer to predict about the criminal profiles, to suspect the next future crime location, or to know the pattern of crime.**

Therefore, there is need to analyze the historical crime patterns more effectively in minimum time, and predicting the future location and type of crime. Police department needs a systematic way for analyzing criminal profile easily and find the associated criminals who can be associated to that crime. Advanced analytics system is also required to track other information such as traffic sensors, calls, videos, police service calls etc. for monitoring the criminal activities. In this paper, we have discussed how Big Data based data analytical approaches can be used to prevent the deal with such cases. In addition to it, we have also discussed different data collection approaches have been discussed which includes Volunteered Geographic Information (VGI) along with Geographic Information System (GIS) and Web 2.0. Last phase will be the prediction based on data collection and analysis. It will be done with using Machine Learning to predict and avoid the future crime.

***Index Terms*: Big Data, VGI, Web 2.0, Data Visualization, Hadoop, Crime Analysis, Machine Learning, RapidMiner**

I. INTRODUCTION

Criminal activity is a worldwide common problem. Over the time, Crime incidents are increasing drastically and it is a significant threat for our society. Criminal are pretty smarter than crime investigation agencies (police department) with the help of new technologies. As per the statistics in India, A total of 95 Lakh crimes were reported in 2016 [1]. Criminal activities can be caused from several reasons. It is quite difficult to manage and investigate the incidents by agencies either due to lack of head counts of cops or criminals are more proactive. They are finding new way of doing it every day. Traditional method of monitoring and investigation for police department takes longer time to predict about the criminal profiles, to suspect the next future crime location, or to know the pattern of crime.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Pranay Jha*, Research Scholar, School of Computer Application, Lovely Professional University, Punjab, India.
Raman Jha, Senior System Engineer, Infosys, Pune, India.
Dr Ashok Sharma, Associate Professor, School of Computer Science Engineering, Lovely Professional University, Punjab, India.

Geographic Information System (GIS) is a system which is related to the broader domain of Geoinformatics [2]. This system is designed to capture, store, modify, share and analyze all georeferenced information. It is a broader term which is base of information retrieved with the help of online engines (e.g. Google Earth, Google Map, Microsoft's Virtual Earth, Wikimedia). Earlier GIS was massive tool for enterprise only and users were using it to retrieve data of interest from the interactive online service. But it has changed over the time. Now users can edit maps, share and modify information. Hence, they are also major contributor of GIS. Volunteered Geographic Information (VGI) has increased the diversity of georeferenced information available online [3]. It increases the broader domain of Geoinformatics vaster. VGI can help in gathering information about crimes through web or mobile crime incident reporting application or through social network. VGI along with Geographic Information System (GIS) can help in reporting of minor crimes accident also from the local areas and enrich the crime analysis process. It pro- vides a new concept of infrastructure to collect, create, validate, share and analyze information through georeferenced technology, user's handsets and Geo database. It can also collect information captured by user's devices such as mobile camera and social network accounts at the local level as well as on higher level. At local level it can be achieved through footage captured by user's camera and post it to reporting application. Citizens can capture and store information in web portals or create post and tag friends on social network which can help in analyzing patterns using Big Data more efficiently starting from the lower level itself. VGI is the information contributed by citizen especially local people which can generate a high amount of information which can make analysis more efficient. With the help of Web 2.0, it is very easy to post information regarding issues on the web [4]. In an era of social networking, any information and incidents (even minor crimes) can be posted directly on the web and analyzed very easily with the help of Big Data. Using Hadoop advanced analysis framework of Big Data analytics, it can analyze the information by mixing up the data which is stored officially by police department as well as collected through users using web or mobile crime incident reporting application and social network [5]. Hadoop can predict future crimes and their location very easily with the help of crime pat- terns in a time effective manner. It deal with the analysis of the reasons of the accidents occurred in the past, the crime prone areas, burglary occurred, crime pattern, a crime specific to some area, age group involved in the crime,

reason behind the crime and then can predict the areas where a specific crime is associated with, crime which will occur in future, can find some ways to stop a specific age group crime, monitor the crime, can patrol cops at location where there are more chances of crime, tracking of all information etc. Predictive analysis is done through the Machine Learning using RapidMiner with processing the historical crime patterns with the help of Hadoop framework which will include MapReduce processing, Hive processing, Sqoop importing and exporting of data etc. With the help of this analysis, Police department can conserve cost, efforts and time.

II. BIG DATA FRAMEWORK

Big Data framework can be categorized into 5 major components [6].

1) Resource manager, 2) Cluster computing framework, 3) Data warehousing and computing, 4) Data storing and management, 5) Data visualization and analysis. These components can be understood by Figure 1.

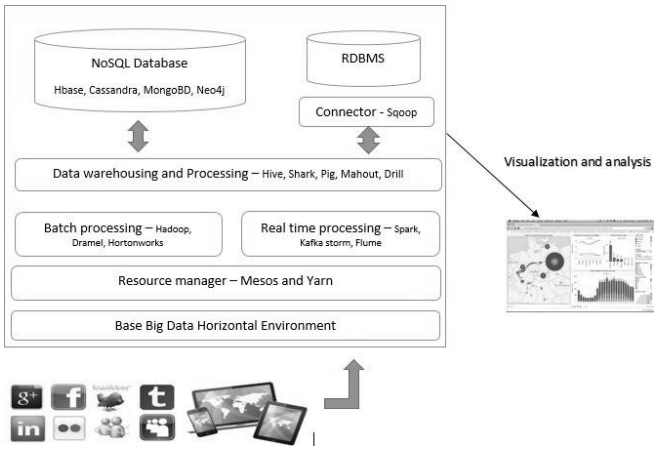

Fig. 1: Big Data Framework

A. Resource Manager

Resource management is a very important aspects in analyzing the data using Big data. It manages all the resource's requirement, synchronization between these resources, increase utilization, increase performance, decrease management cost, increase interoperability and reliability of the Big Data framework. Resource manager synchronizes active and passive nodes, separate storing from processing components. It includes an Application Master which figures out how many processing resources is required for executing the entire application hence increase orchestration in environment. Common resource management and scheduling platform in Big Data are – Mesos and Yarn. Mesos and Yarn aims to increase utilization of resource of cluster by sharing among multiple processing frameworks such as Hadoop, MPI, Spark or multiple instance of same framework [7] [8].

B. Cluster Computing Frameworks

Data Analytics computing framework is for computation of large cluster of distributed data. It can be categorized as Batch processing computation and Real time processing computation [9]. It supports offline processing MapReduce, Hadoop for batch processing, Spark for iterative computing, Horton works for batch-oriented processing, Storm form online processing, MPI for high performance, Flume for stream processing, Kafka for real time event processing and data mining and streaming processing framework for S4.

C. Data Warehousing and Computing

This layer contains data warehouse i.e. Hive for analyzing of cluster of data, cluster computing framework such as Shark which uses SQL like query for analyzing information, data flow language e.g. Pig which decreases the development time for writing queries, Mahout for machine learning and Drill for appropriate analysis of large-scale database. The data warehouse is connected via servers whereas computing framework supports scalable No SQL database such as Hbase [10].

D. Data Storing and Management

This layer refers to the storing of structured, semi-structured and unstructured data in database and its management issues such as scalability, replication, higher throughput etc. It can be divided into two categories. 1) Relational Database Management System (RDBMS), and 2) NoSQL Database Management System [11].

RDBMS is for structured data only means the data which is in form of tables and it is fast. RDBMS can interact with data warehouse layer through connector known as Sqoop. It is designed for efficiently relocating bulk data between structured data and Hadoop. Sqoop can import the data from RDBMS to HDFS and can also export it from HDFS to RDBMS. It can directly import the data into Hive table from RDBMS. Whereas on the other hand NoSQL provides the best fit for unstructured as well-structured data. NoSQL Database Management System enables real time, column oriented, distributed, scalable, robust, sparse and sorted collection of data [12].

E. Data Visualization and Analysis

The result can be visualized and reported with the help of interactive dashboards. Data visualization can be done by using several tools such as D3.js [13]. Data can also be visualized by using different software e.g. Tableau, RapidMiner or R Tool. It helps anyone to quickly analyze, visualize and share information [10]. Tableau can use Apache Hive (via ODBC connection) as the defacto standard for SQL access in Hadoop [4]. Patterns are to be analyzed through data visualization for future prediction and forecasting about information.

III. ARCHITECTURE

The architecture defines all the process comes under analyzing VGI and Web 2.0 captured data for effective crime analysis [14][15]. The main layers of architecture are described below which can be understood by Figure 2:

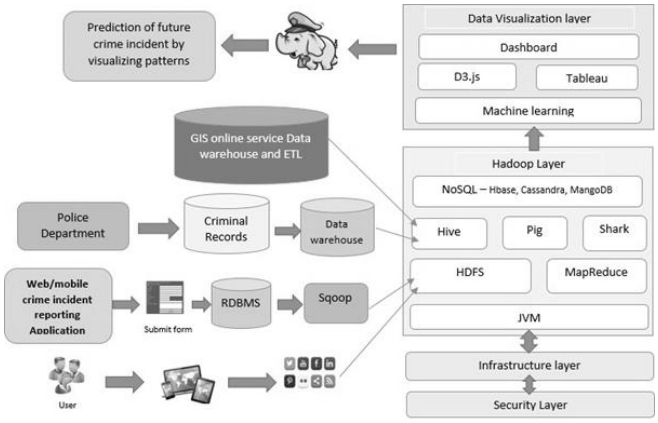

Fig. 2: Architecture of Big Data Analysis

A. Data Collection and Integration Layer

Data can be collected by different sources. It can be collected from the social networking sites where people can post crime incidents, tag friends, share information and chat with friend. The analysis process will be easier if they will do all these activities with a specific keyword all the time. The second way can be by using a web or mobile crime incident report application in which they can fill their information along with the crime incident information and submit web form which can automatically save into police department warehouse. Integration of the data can be done by integrating these data along with police department crime database and GIS online services ware- house [16] [17].

B. Security Layer

This is one of the most focused layers in crime analysis architecture. Firewall should be there in the system for proving security. Encryption of the citizen's personal information should be done using a secure algorithm. Authentication should be done based two factor authentications by using bio metric identity along with password and authorization is only given to those who have the specific access [18].

C. Infrastructure Layer

This layer contains infrastructure requirement to support Big Data horizontal and scalable environment. This layer makes Big data environment highly fault tolerant, high throughput, suitable to analyze large dataset and reliable [19].

D. Hadoop Layer

This layer allows for processing of bulk data sets across cluster using programming components HDFS and MapReduce. HDFS is a distributed file system to store data whereas MapReduce is a programming model which is used in processing of large datasets. Scalable, Sorted and distributed database can be created using NoSQL Database Management System such as Hbase, Cassandra and MongoDB [20].

E. Analysis Layer

Hadoop uses Hive, Pig, Shark and MLlib which is chosen based on type of data collected. For analysis of data collected from VGI and web or mobile crime incident reporting application, Since the data is stored in the RDBMS it is to be first imported in HDFS of Hadoop framework through connector, Sqoop and then it is analyzed using Hive. The data which is collected from tags or posts on social networking sites is first streamed through Flume. Flume is a reliable and distributed service for collecting, moving, and aggregating large amounts of log data very efficiently. After streaming of data, the data is stored in Hadoop Distributed File System and then it is processed by MapReduce simple programming model which can be written in Java as well as in any other supportive language such as Python. The data collected from Police department crime database is mixed (JOIN operation) with database gathered from VGI and web or mobile crime reporting application to find more efficient crime patterns.

IV. MODULES AND TECHNOLOGIES USED

A. Hadoop

Hadoop framework which helps in processing of bulk datasets across clusters of commodity hardware using programming model known as MapReduce [21]. There are two core components of Hadoop:

1) Hadoop Distributed File System (HDFS)

HDFS works on a master-slave architecture. There are two components of HDFS:

- Namenode – This works as a master node on which helping running the job tracker. It is master of the system and contains metadata about each datanode. It also maintains and manages the blocks exists on the datanodes. It does not store any data, and only captures the metadata information in the form of namespaces and edit logs.
- Datanode - This is slave node which is deployed on each machine of cluster. It helps in storing the actual data. It is responsible for sending and receiving read and write request for the clients. Task trackers are associated to it. It continuously sends heartbeat to namenode and in this way namenode has the information for each datanode that it is alive or not. In datanode, replication of data is there for fault-tolerance. Rack awareness concept is used to ensure lesser latency and to increase fault tolerance.

2) MapReduce

- MapReduce works in following stages:
- Input splits — Firstly, large cluster of datasets split into many input splits based on various configuration in Hadoop.
- Map — Map processes the input splits data and generate key- value pair.
- Sorter — In this stage data is sort based on the key.
- Combiner — It helps in lessen up the traffic between mapper and reducer.
- Shuffler — This is the only phase in which data shuffles from mapper to reducer.
- Partitioner — It helps in deciding which key will go to which reducer.
- Reducer — Reducer processes the data and gives us the final output as per the logic written [22].

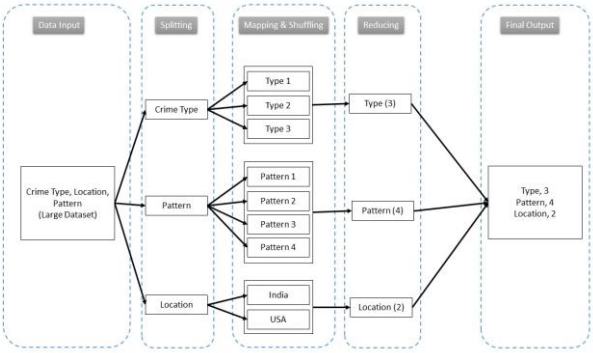

Fig. 3: MapReduce Framework

B. Hive

Hive is a data warehouse and built on Hadoop for querying and analysis of data. It is originally developed by Facebook, targeted towards users comfortable with SQL. Hive uses SQL-like language known as HiveQL, for querying and analysis of data. Role of HiveQL is to manage and querying structured data. It also reduces the efforts in developing the scripts using Hive. It abstracts the complexity of Hadoop. Hive tables can be partitioned and bucketed. It is easy to plug in custom mapper/reducer code. Hive stores metadata in an embedded Apache Derby database. Metastore are stored in hive warehouse. Metastore is a place where we store schemas of the tables [23].

C. Sqoop:

Sqoop is a command line program which allows effectively transfer of large data between HDFS and relational data. It can also be used for transferring data between Hive and relational databases. Sqoop helps in import/export tables or entire databases to HDFS. It generates java classes which allow you to interact with imported data. While importing data, we can also decide the file format in which data will be imported to HDFS [24].

D. Twitter 4j/Facebook API:

Twitter 4j is java library for accessing information using Twitter API. With the help of Twitter 4j, we can easily integrate application with twitter sources. We can access information from twitter with the help of Consumer key, Access key and Access tokens. In the same way, Facebook API can also be used to access information from Facebook. These two API plays an important role in accessing information using Web 2.0.

E. RapidMiner:

RapidMiner is a data analysis, prediction, and visualization tool used for creating an interactive and shareable dashboard with the help of which we can visualize the trend, variations, pattern and density of data with the help of graph and charts. The output from the Hive can be imported to RapidMiner for more data visualization [25].

V. IMPLEMENTATION

Implementation of the research will be in different phases showing in figure 4 below.

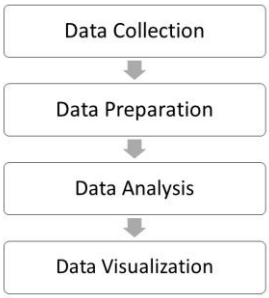

Fig. 4: Implementation Phases

A. Data Collection

An introductory series of questions has been classified for gathering the data. Data collection has mainly based on Crime data. Other ways of collecting data from different sources of VGI and Web 2.0. The data collection process including clarifications for several years from below sources.

1. *VGI - Web or mobile crime incident reporting application or through social network*
2. *Web 2.0 - Twitter, Facebook*
3. *CrimeInfo - Existing Crime Dataset*

We have used the crime related dataset available on the internet listed below.

- https://data.gov.in/data
- http://ncrb.gov.in
- https://data.gov.in/catalog/cognizable-crimes-under-indi a n-penal-code-ipc-crimes-different-crime-heads

B. Data Preparation:

Data preparation will be done from the following data sources:

1) Volunteered Geographic Information (VGI)

In this process, data will be collected with the help of a crime incident reporting application where volunteers or local people can submit various type of crime incidents by submitting location and short description of crime. Data from the application will be stored in relational databases. With the help of Sqoop command, Data can be easily imported from relational databases to Hadoop distributed file system. Sqoop is a simple command line interface used for effective transfer of bulk data between Hadoop distributed file system and relational data. It can also be used for directly transferring data between Hive and relational databases.

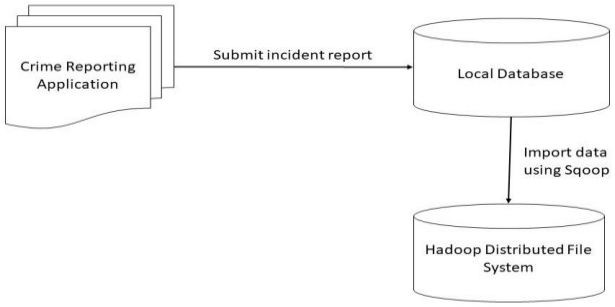

Fig. 5: Import VGI data to HDFS using Sqoop

2) Web 2.0

In this process, with the help of Apache Flume, streaming data can be ingested from Twitter and Facebook API sources and will be stored in HDFS with the help of memory channels. Web 2.0 fetched data will be in the form of unstructured data. This data can be refined into useful information with the help of MapReduce programming model.

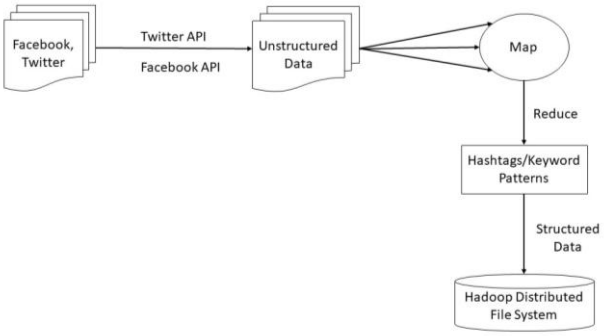

Fig. 6: Import Web 2.0 data in HDFS using MapReduce

3) CrimeInfo

Historical crime datasets will be fetched with the help of Crime- Info sources. It will export the data in Comma separated value (CSV) format. With the help of basic hadoop commands, the CSV format file can be stored in HDFS.

Using the data above from different sources, we will create three HIVE tables for VGI, Web 2.0, and CrimeInfo dataset. HDFS data will be stored in those hive table

C. Data Analysis

1) Analysis using MapReduce

MapReduce is being used for refining the streaming data fetched from Twitter and Facebook. Apache flume stores the data in HDFS in un- structured format. To refine data with required logic, MapReduce programming model is being used. A MapReduce program consists of three classes: Driver, Mapper and Reducer class.

2) Analysis using HIVE

Hive tables will be created for VGI, Web 2.0 and historical crime datasets. Data will be loaded in Hive tables from the data stored in HDFS from VGI application, Twitter and Facebook data and historical crime data sources. Optimized Row Columnar (ORC) file format is being used while storing data in Hive table. It will provide fast columnar access of data and consume less space for storage.

Table join, partitioning and bucketing concept based on logics will be applied on Hive tables for refinement of data such that it can be analyzed and queried in efficient, easy and less time-consuming way using Hive Query language (HiveQL). With the help of Hive fast and advance analytics, crime pattern can be recognized as showing in figure 7.

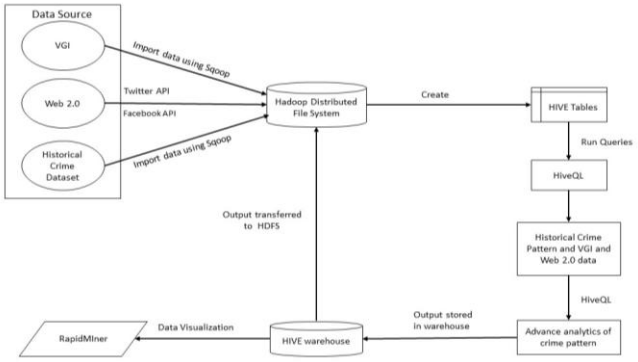

Fig. 7: Data Analysis and Visualization

D. Data Visualization using RapidMiner:

For easy visualization of Hive results, Naïve Bayes algorithm will be used through RapidMiner. RapidMiner provides different algorithms such as Decision Tree, Naïve Bayes, Logistic Regression, Deep Learning, Random Forest, Support Vector Machine, and Fast Large Margin. In this paper, we will use Naive Bayes algorithm as showing in figure 7 to visualize the crime data to see different statistical analysis and prediction of crime in different states. It will be used for creating an interactive and shareable dashboard with the help of which end users can visualize the trend, variations, pattern and density of data with the help of graph and charts.

VI. CHALLENGES

Though the architecture defines an effective and robust way of analyzing crime incidents by integrating VGI, data stored from web or mobile crime incident reporting application with database of police department criminal records by which police can analyses crime from the lowest level to the highest level, but it also possesses some challenges in achieving such analysis which are listed below.

A. Security

This is the most focused part for any system. In this architecture for crime analysis, security is most critical part as citizen who is sharing information can be given death threats by criminals. Therefore, Personally Identified Information (PII) should be kept secure by using two factor authentications, firewall on the database servers and encryption of every communication followed. Logs should be checked regularly for any hacking activities.

B. Infrastructure Cost

The benefit of analyzing information with Big Data is substantially very high but the infrastructure cost for storing large cluster of data sets, achieving replication in Big Data, cost for data visualization is very much.

Big Data technology should be chosen with Cost-Benefits ratio in mind.

C. Spam or Refuted Data

Sometimes this can also possible that unauthorized users are sharing wrong information to misguide police department or any spam data is generating automatically. These conditions should be handled with two factor authentications of identity.

VII. RESULT

After collecting, preparing, importing data, and Analysis using the data using different modules of Hadoop framework, visualization and prediction will be through Naïve Bayes Machine learning algorithm. RapidMiner tool has been used to perform this. Crime data has been collected through sources mentioned in data collection section. Import the data to RapidMiner as showing in below figure. It will be store the data in tool which can be used to define and apply models.

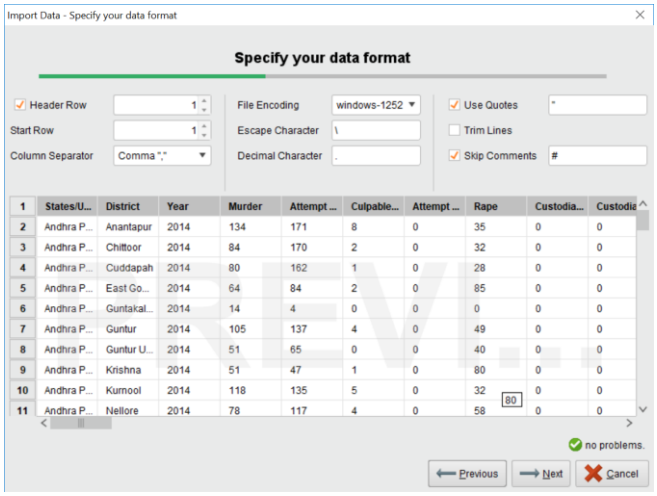

Fig. 8: Import Data

Once data import gets complete, Create a new blank process in Design view page. Select different operators such as Retrive_Crime2014, Set Role, Naïve Bayes, Apply Model, and Performance as showing in figure 9. Connect the operators and run the model. Retrieve operator will fetch the imported data which can be used for further analysis.

Fig. 9: Create Process and Run the Model

Filter out any NULL value if exists in data. We verified it through Statistics view as showing in Figure 10. Here, you will able to see the data if classifier is missing.

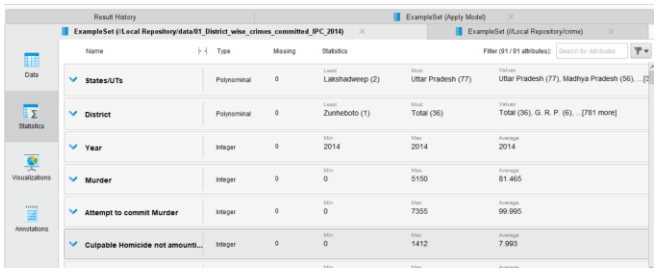

Fig. 10: Filter Null Data

After reviewing the data, we used Auto Model to see the results. We have statistics available according to different attributes. Clicked on Auto Model option to see the results accordingly as showing in Figure 11(a), 12(b), 13(c), and 14(d).

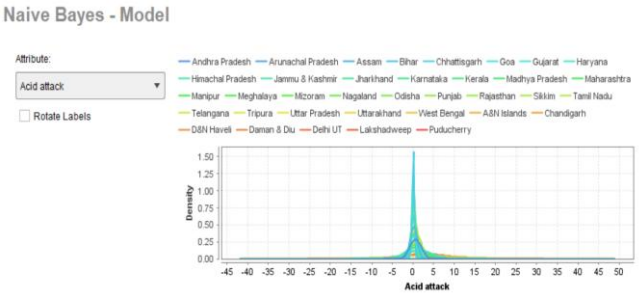

Fig. 11: Crime based result (a)

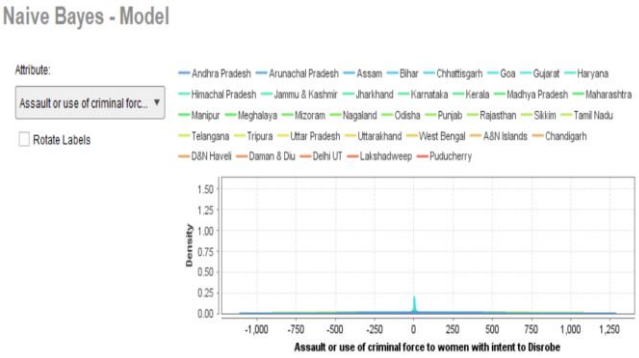

Fig. 12: Crime based result (b)

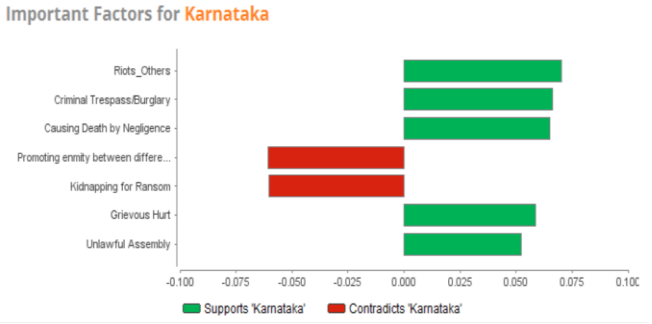

Fig. 13: State wise result (c)

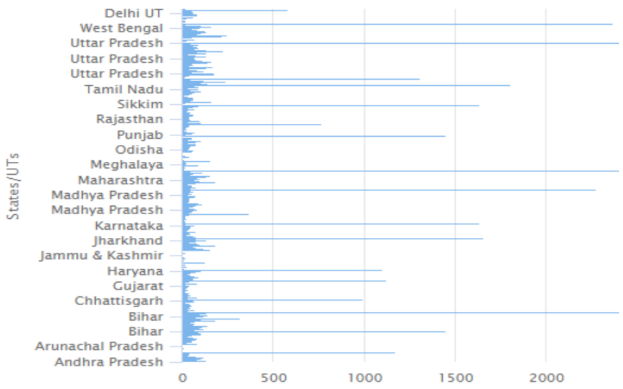

Fig. 14: Location wise Result(d)

We will also check accuracy of model which can be checked in %Performance as showing in below figure 15. It will ensure and validate the performance of Model and prediction.

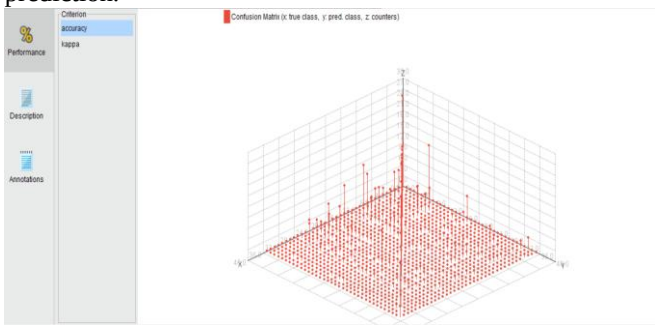

Fig. 15: Accuracy of Model

On the basis of statistics available, Naïve Bayes will predict the crime as showing in below figure 16 and Table 1.

Naive Bayes - Predictions

Row No.	States/UTs	prediction(S...	confidence(Andhra Pradesh)	confidence(Arunach...	confidence(...	confidence(...	confidence(...	con
1	Andhra Prade...	Chhattisgarh	0.000	0	0	0	1.000	0
2	Andhra Prade...	Andhra Prade...	1.000	0	0	0	0.000	0
3	Andhra Prade...	Delhi UT	0.035	0	0	0	0	0
4	Andhra Prade...	Andhra Prade...	1.000	0	0	0	0	0
5	Andhra Prade...	Haryana	0.000	0	0	0	0	0
6	Andhra Prade...	Haryana	0.000	0	0	0	0	0
7	Andhra Prade...	Tamil Nadu	0	0	0	0	0	0
8	Andhra Prade...	Gujarat	0.002	0	0	0	0.047	0

Fig. 16: Prediction

Table 1: Prediction based on crime type

INSULT TO MODESTY OF WOMEN	PREPARATION AND ASSEMBLY FOR DACOITY	MURDER	prediction (MURDER)
138.0	0.0	96.0	51.2
84.0	0.0	72.0	51.2
2.0	0.0	2.0	51.2
90.0	0.0	120.0	51.2
276.0	0.0	101.0	51.2
297.0	0.0	103.0	51.2
274.0	1.0	56.0	51.2
98.0	0.0	191.0	51.2
73.0	0.0	128.0	51.2
313.0	0.0	98.0	51.2

VIII. CONCLUSION

With the help of VGI, data collected through web/mobile crime reporting applications, from tags of twitter handles and other social platforms, and also with the crime data set available on CrimeInfo repository. Considering the different attributes of the crime data, efficient analysis has been done which has helped in analyzing the incidents and prediction about the future crime and its location. These all have been done using the Big Data and Machine learning framework.

The architecture proposed in this paper will reduce cost and efforts of the police department by patrolling the cops in area which are more sensitive and predictive for crime as per the analysis.

It can analyze at the granular level which will help to know root up the cause of crime incidents. It will also reduce the time of investigation for crime as the person who has reported the issue using VGI and Web 2.0, can also be witness of the crime.

Big Data is appropriate framework for analyzing crime data as it provides high throughput, fault tolerance, analyze very large data sets, process on commodity hardware and generate reliable results, whereas the Machine learning's Naïve Bayes algorithm is to do a better prediction using the available dataset.

REFERENCES

1. "Crime Statistics," *data.gov.in*. [Online]. Available: https://data.gov.in/dataset-group-name/crime-statistics. [Accessed: 07-May-2019].
2. "GeoBI 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 Comput. 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].
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/7453400. [Accessed: 07-May-2019].
11. R. Agrawal, A. Borgida, and H. V. Jagadish, "Efficient Management of Transitive Relationships in Large Data and Knowledge Bases," in *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 1989, pp. 253–262.

12. V. N. Gudivada, D. Rao, and V. V. Raghavan, "NoSQL Systems for Big Data Management," in *2014 IEEE World Congress on Services*, 2014, pp. 190–197.
13. D. Keim, H. Qu, and K. Ma, "Big-Data Visualization," *IEEE Comput. Graph. Appl.*, vol. 33, no. 4, pp. 20–21, Jul. 2013.
14. N. Sawant and H. Shah, "Big Data Application Architecture," in *Big Data Application Architecture Q & A: A Problem-Solution Approach*, N. Sawant and H. Shah, Eds. Berkeley, CA: Apress, 2013, pp. 9–28.
15. D. Ghosh, S. A. Chun, B. Shafiq, and N. R. Adam, "Big Data-based Smart City Platform: Real-Time Crime Analysis," in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, New York, NY, USA, 2016, pp. 58–66.
16. S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop*. Packt Publishing Ltd, 2013.
17. "Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies - Sandelowski - 2000 - Research in Nursing & Health - Wiley Online Library." [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/1098-240X(200006)23:3%3C246::AID-NUR9%3E3.0.CO;2-H. [Accessed: 07-May-2019].
18. "Big Data and Hadoop-a Study in Security Perspective - ScienceDirect." [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091500592X. [Accessed: 07-May-2019].
19. P. Jha, D. A. Sharma, R. Jha, and S. Kaur, "How can your IT Infrastructure Withstand the Pressure of Digitalization?," vol. 7, no. 1, p. 6, 2019.
20. A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–781, Oct. 2013.
21. A. Jain and V. Bhatnagar, "Crime Data Analysis Using Pig with Hadoop," *Procedia Comput. Sci.*, vol. 78, pp. 571–578, Jan. 2016.
22. J. Dittrich and J.-A. Quiané-Ruiz, "Efficient Big Data Processing in Hadoop MapReduce," *Proc VLDB Endow*, vol. 5, no. 12, pp. 2014–2015, Aug. 2012.
23. A. Menon, "Big Data @ Facebook," in *Proceedings of the 2012 Workshop on Management of Big Data Systems*, New York, NY, USA, 2012, pp. 31–32.
24. R. Casado and M. Younas, "Emerging trends and technologies in big data processing," *Concurr. Comput. Pract. Exp.*, vol. 27, no. 8, pp. 2078–2091, 2015.
25. S. Lu *et al.*, "A Framework for Cloud-Based Large-Scale Data Analytics and Visualization: Case Study on Multiscale Climate Data," in *2011 IEEE Third International Conference on Cloud Computing Technology and Science*, 2011, pp. 618–622.

AUTHORS PROFILE

Pranay Jha is a Research Scholar in Lovely Professional University. He is working in IBM as an Infrastructure Architect for Virtualization and Cloud platform. He has been in IT profession from last 12 years and carrying experience on numerous products along with core experience on Virtualization and Cloud technology. He is a member of the Cloud Computing Innovation Council of India (CCICI) and VMware User Group (VMUG. He is an Independent founder and blogger at http://vmwareinsight.com. He has been awarded as VMware vExpert x 4 (16/17/18/19) and holding several industrial certificates including VCIX-DCV, VCAP5/6-DCD, VCAP5-DCA, VCP7-CMA VCP5/6-DCV, VCA-DCV, VCA-Cloud, VSP, VCE-CIA, MCITP, MCSE, MCSA(Messaging), IBM Certified Architect, Open Group Certified Architect.

Raman Jha is working in Infosys, currently designated as a Senior System Engineer. He has been working on technologies such as Big Data and RPA with knowledge in Spark transformations and actions, handling data frames, different file formats, compression codecs, Flume, Kafka, Spark streaming, Oozie, Hive, Sqoop, Scala, and AssistEdge RPA.

Dr Ashok Sharma is an Associate Professor in Lovely Professional University, with having 15 years of experience in academic. He has completed his Ph.D. from MANIT Bhopal.