CRIME DATA MINING: A GENERAL FRAMEWORK AND APPROACH TO IDENTIFY CRIME PATTERN USING MACHINE LEARNING

BY

MAHMUDA SHARMIN
ID: 151-15-5241
AND

MD. ASLAM KAZI

ID: 151-15-5490

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2018

APPROVAL

This thesis titled **"Crime Data Mining: A General Framework and Approach to identify crime pattern using machine learning"**, submitted by Mahmuda Sharmin and Md.Aslam Kazi to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26 November 2018.

<u>BOARD OF EXAMINERS</u>

_____

Dr. Syed Akhter Hossain   **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

Narayan Ranjan Chakraborty   **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Tarek Habib   **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin   **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Sheikh Abujar, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by: **Co- Supervised by:**

_____   _____

Sheikh Abujar   **Abdus Sattar**
Lecturer   Senior Lecturer
Department of CSE   Department of CSE
Daffodil International University   Daffodil International University

Submitted by:

_____

Mahmuda Sharmin
ID: 151-15-5241
Department of CSE
Daffodil International University

_____

Md.Aslam Kazi
ID: 151-15-5490
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Mr. Sheikh Abujar**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to the Almighty Allah and Head**,** Department of CSE, for his kind help to finish our thesis and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Crime data mining has been a hot topic in the law-enforcement sector as it can solve crime related problems more effectively. In this thesis, we have focused in determining crime patterns using effective data mining algorithms. To classify our raw data we used Naïve Bayes classifier along with NER (Named Entity Recognition) and Co-reference Resolution concept. In order to get higher accuracy, we have trained numerous train data with required keywords. To get the related keywords and train data, both online and offline data are used. For the validation of the process, filed work like talking to "Adabor Thana Police Station" is also performed. Test results are generated based on sample input and results. Our thesis results generates the report which represents crimes stativity of different divisions of Bangladesh. The output data also shows how much crimes are occurring in certain place which can be used for crime pattern analysis for a specific location like an individual district or even for a police station. Upon successful implementation, it our thought that law-enforcement authority will find our findings useful to detect crime sensitive zones of Bangladesh and take precautions so that number of crimes can be reduced.

TABLE OF CONTENTS

CONTENS	**PAGE**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

In general sense, crime means which violate law and rules, regulation or which is harmful or destroy human society or resources or which create problems in human life. Crime, through both its impact on victims and indirect effects on the wider community, has strong links to health and well-being at a population level.

The existence of the crime in a society, it is a challenge to its members due to its harmful affection the order social growth. In fact, its leads to a huge waste of human energy and anenormous economic loss. Therefore, with the advance in the field of criminology and behavioral sciences, efforts are being constantly made to work out a commonly acceptable classification of crimes and criminals for provide a rational basis of punishment for various categories of criminal.

Bangladesh is a small and poor country. It has many problems like, poverty, illiteracy, unemployment, over population etc. So, many cases crimes are related or create to this problem. Such as Theft, Rapine, Terrorism, Juvenile delinquency, kidnapping, Fraud, Money laundering, Drug Trafficking, Extortion, Contract killing, Corruption, Black Marketeering, Political Violence, Abduction, Violation of human rights, Illegal drug manufacturing etc.

In the living world and also Bangladesh are seriously facing cyber-crimes. There are many types of cyber-crime, such as:

Financial crimes, Cyber pornography, sale of illegal articles, online gambling. Intellectual property crimes, Forgery, Cyber blame, theft of information contained in electronic form, cheat data, virus attacks, web hacking, theft of computer system etc.

In this study, we propose a system that will automatically detect the crime. As the detection of crime we will research from previous data and say what kind of crime where can be done in the future.

1.2 Motivation

An ATM booth security guard of Dutch-Bangla Bank was chopped to death in the city's Mohammadpur area early on Sunday 6th October 2013.Thedeceased was identified as Enamul Haq, 25, hailed from Bhauail village, Tarailupazila, Kishoreganj, who was an employee of private security provider Elite Force. He had been residing in a slum in the area. As like there are so many crime in our country that do not have any proper reason or proper research in our country.

We read there are lots of paper in the crime, which this paper tells crime various country or specific nation, but there is no anywhere which write any research or paper in our country.

After reading about the case study something got to mind to make something that can automatically investigate the crime in our country. Since we live in Bangladesh, so for the development of our country and for security, we should work with the crimes of our country. Thus I was interested in analyzing data by classifying them into crime.

1.3 Rationale of the Study

The issue of crime prevention is one of the main human concerns throughout history. Different methods have been applied to reach this goal including social, penal and legal conditional prevention. One of the most complicated issues in our country is the high rate of crime and disorder increase and the incidence of crime and disorder leads to a sense of insecurity and financial difficulties imposed upon society, government, and the country's judiciary.

It is no doubt there are lots of works on crime in others country and these approaches or the processes are being used in many automated system as well as robotics system. But, crime on paper in Bangladesh is very rare. These, actually, take us to work with crime in Bangladesh.

1.4 Research Question

➢ Can we collect raw data for crime?

➢ Can we pre-process the raw data to be used for the Machine Learning approaches?

➢ Can Classifier algorithm be used on the pre-processed data?

➢ Can the Machine Learning process correctly detect or identify the category of the crime?

➢ Where the crime increasing day by day?

➢ What is the category of the crime?

1.5 Expected Output

This research has a large prospective in the present word in Bangladesh. We propose a system that will automatically detect the crime. Where there is more chance of crime, what is the category of crime, which place is more risky, etc. Our system will automatically detect this crime. It help our country people which place is very dangerous. It will make sure the proper condition in the place in our country. We also show which places are very infected crime and also results in percentage for the crime.

1.6 Report Layout

The report will be followed as follows :

Chapter 1 provides the summary of this research. Introductory discussion is the key term of this first chapter. Apart from, what motivated us to do such a research is explained well in this chapter to. The most important part of this chapter is the Rationale of the Study. Then, what are the research questions and what is the expected outcome is discussed in the last section of this chapter.

Chapter 2 covers the discussion on what already done in this domain before. Then the later section of this second chapter shows the scope arisen from their limitation of this field. And very last, the root obstacles or challenges of this research are explained.

Chapter 3 is nothing but the theoretical discussion on this research work. To discuss the theoretical part of the research, this chapter elaborates the statistical methods of this work.

Chapter 4 is related with the outcome of the whole research and the thesis. Some experimental pictures are presents in this chapter to make realize the thesis.

Chapter 5 is based on conclusion topics of the thesis. This chapter is responsible to show the whole thesis report adhering to recommendation. The chapter is closed by showing the limitations of our works that can be the future scope of others who want to work in this field

CHAPTER 2

Background Study

2.1 Introduction

Crime is a global challenge in every nock and corner in the world. Security agencies are facing challenges to investigate the real scenario, criminal behavior, crime pattern, location and environment.

According to crime theory, crime takes place due to some reasons such as state mismanagement and oppression, failure of the court to ensure justice, helplessness of the society towards intolerant power.

In this circumstances, data mining technique can reduce this challenge by crime pattern analysis. From past crime activity we can imagine what kind of crime can happen in the future. It will help in the investigation as it comes to the benefit of the general public.

2.2 Related works

When we decided to research on this section, we looked through online and found some research papers that are related to this topic but they aren't for only our own country Bangladesh, also have so many unsolved challenges and there are many features missing there as those researches are not based on our country.

Peter Phillips et al. showed geospatial co-distribution relations among crime incidents, socio economic, socio demographic, and spatial features and is computationally efficient [1]. They proposed graph based data set representation from heterogeneous areal aggregated dataset. In data pre-processing stage, they used min-max normalization for normalization and generated a weighted directed multigraph to store density values for each region. They also used a global minimum edge weight and kruskal's algorithm.

A descriptive statistical approach based on the quartile (floor-ceil) function which is automatic threshold selection method called ATS for the Minimum Support Threshold (MST) choice value can be the best technique for recognizing crime situation. Omowunmi Isafiade et al. proposed this method. This technique is applicable for pruning

stage in Traditional FL-Growth model (TFPG). By comparing between Revised Frequent Pattern Growth (RFPG) and Traditional FP-Growth (TFPG). We found TFPG, which is the best one [2].

Fuzzy-Rough Oscillatory Region algorithm was used in the subject of crime against women at Tripura. For reducing set of attributes Susmita Roy et al. used fuzzy cognitive map (FCM) and by using direct methods with multiple object, fuzzy membership value lying between (0, 1) made quantitative data from qualitative data [3]. At last, they applied Fuzzy-Rough Oscillatory Region algorithm which helped to draw conclusion.

By implementing Linear Regression, Additive Regression and Decisions Stump algorithms using the same finite set of features on the communities and crime data set Lawrence McClendon et al. tried to prove how effective and accurate the machine learning algorithm used in data mining to predict violent crime patterns [4]. By comparing these algorithm they found Linear Regression is most promising than others and it also could handle randomness in the test samples to a certain extent.

Somayeh Shojaee et al. conducted an experiment by using five different supervised learning algorithms to obtain better supervised learning algorithm to predict crime status by using two different feature selection method [5]. They used Naïve Bayesian, K nearest neighbor (K-NN), neural networks, decision tree and support vector machine (SVM) and compared performance among them. 10 fold cross validation technique was used and finally they found K-NN achieved best accuracy by using Chi-square feature selection technique. Table 1 shows the summary of our findings from different reputed international research papers.

Table 1: Summary of previous research works

Authors	Year	Data	Techniques	Description
Peter Phillips et al. [ma]	2012	108 datasets	min–max normalization, Kruskal's algorithm, Pearson's correlation	Performed co-relation analysis to discover data set. Showed geospatial co-distribution relations

			coefficient or Jaccard's index, density tracing algorithm	among crime incidents, socio economic, socio demographic, and spatial features and is computationally efficient
Andersen And Malleson [23]	2014	Records of over 3 years	Spatial point analysis	Investigate crime displacement by identifying changes in the spatial patterns/ distribution of crime
Grubesic [25]	2006	613 records	Fuzzy clustering	Hot spot detection
Yu at al. [38]	2011	Not mentioned	k-NN, Decision trees (J48), SVM, Neural Network, Nave Bayes, ensemble learning	Prediction of burglary data with different levels of aggregation of historical data (1 month to 10 months)
Dahbur and Muscarello [32]	2003	Not mentioned	Kohonen neural networks and heuristics	Discover patterns of serial crimes; case study on armed robberies
Brown and Hagen [28]	2003	39 records (cases)	Association rule mining	Tool for discovering associations between different crimes
Zubi and Mahmud [39]	2013	350 records	k-means clustering, association rules	Crime analysis of Libyan crime data
Oatley and Ewart [24]	2003	70,000 records	Logistic regression, Neural networks, Bayesian Network	Analysis and prediction of burglary data
Chen et al. [31]	2004	120 records for identity detection;	Co-occurrence,	Deceptive identity

		272 records for network analysis	Hierarchical clustering	detection, criminal network analysis
Xue and Brown [17]	2006	Over 1200 records	Discrete Choice Theory and Clustering	Analyse and predict spatial choice of criminals for residential breaking and entering crimes
Susmita Roy and Sharmistha Bhattacharya [ma]	2015	sample of 500 persons	Feature extraction: fuzzy cognitive map (FCM). Algorithm: , Fuzzy-Rough Oscillatory Region algorithm	Helped to draw conclusion about the pattern of an unknown object
Li et al. [29]	2010	6720 records; 14 crime types	Fuzzy self-organising map, Rule extraction	Data analysis to support decision making; 4 crime trends: typical, gradual increase, sharp increase and wintertime
Wang et al. [37]	2013	4855 records	Series Finder (supervised learning for detecting patterns)	Identification of patterns in housebreaking crimes (from 51 patterns)
Malathi and Baboo [34]	2011	8 years of crime data	DBScan clustering (density based), k-means clustering, Decision trees (C4.5)	Clustering of crime data by crime type and prediction of crime frequency for the following year
Nath [35]	2006	309 records	k-means clustering	Patterns of crime (6 types)

Phillips and Lee [36]	2012	Not mentioned	Graph similarity	Identification and description of crime patterns
Bachner [30]	2013	Not Applicable	Clustering, Social Network Analysis	Overview of predictive modelling
Helbich et al. [22]	2013	200 individual information packages (i.e. emails, transcribed interviews and phone calls)	Text mining and Spatial point analysis	Text mining and spatial analysis to discover new patterns and relationships
Lin and Brown [33]	2006	170 records	Clustering and outlier-based approach	Association of incidents for identification of crimes committed by the same individual; case study on robbery data
Murray and Grubesic [26]	2013	848 records	Non-hierarchical clustering using spatial lag	Hot spot detection

2.3 Research Summary

In general, crime means any wrongdoing done by society. One of the major weaknesses of this definition is to hide the complex problems of society. No society is permanent. In fact, society is constantly changing. Likewise, the trend of crime and its type are constantly changing. Again, the behavior of the crime behavior in the society is noticed. Considering previous crimes, we can say where the crime will be in the future and where there will be no crime.

The dataset shows that militant attacks, terrorism threats, attacks, fighting, bomb attacks, shootings, yaba business, kidnappings, tortures, murders, sabotage, artillery killings,

bombings, clashes, vandalism, beatings, child marriage, suicide, drug trade, Crooked injuries, stab wounds, harassed, theft – are the present alarming crime of Bangladesh.

It is come to know that pattern mining accuracy can be different by using different types of technique. It is essential to know the appropriate technique use of in appropriate time. Then we could get best result from our analysis. Hope this paper could help the investigator, analyst and researcher to get the best result of crime pattern mining.

2.4 Scope of the problem

We have chosen Bangladesh because we are in Bangladesh, it is our motherland. Since research is done in terms of Bangladesh, it is completely new. And that's why we have to face many problems which were cumbrous.

The problems are discussed below:

2.4.1 Time Management

It was a matter of concern that whether we can finish the work in time or not. Because it takes a lot of time to collect and analyze data.

2.4.2 Data collection

We have collected criminal articles from the Prothom-alo paper for collecting data. It took a lot of time because we tried to collect the crime accurately.

2.4.3 Data analysis

We can analyze the data of crime to make predictions and get number of crime. We can also get which crime has the most impact on our country & which crime come most.

2.4.4 Data Sorting

We have received a lot of random information during the collection of data. It is time consuming to find out which of these data is useful and which is less useful.

2.5 Challenges

In every sphere of life there are challenges but we are to overcome the challenges to make something good. The challenges we have faced are given below:

2.5.1 Lack of data

It was one of the biggest problems for us that we spent a huge time to ensure the trusted sites or sources for collecting data.

2.5.2 Problem finding

As some related papers have already been made, we tried our heart and soul to find the bugs and missing features that should have been in the paper.

2.5.3 Real time data collection by doing field work

Since criminal data is with the police, we have not been able to collect them so we have taken help from the newspaper. It was kind of challenging for us.

2.5.4 Tools

Datamining Algorithms such as Naïve Bayes algorithm, NER (Named Entity Recognition) and Co-reference Resolution concept. To implement these techniques, Matlab Desktop software has been used. To visualize the output SPSS, StarUML and Microsoft PowerPoint is used.

CHAPTER 3

Research Methodology

3.1 Introduction

If we think about the major challenge faced by law-enforcement organizations and other intelligence-gathering organizations, what comes first? Well, it is definitely efficiently and accurately analyzing the significant amount of growing crime data. For example, crime placed in different times and places can be really tricky to predict or even determine before it actually by analyzing previous crime records. It is simply not possible because no such active system is there to perform the crime data analysis for the law-enforcement people. There are also a lot of crimes remain hidden or only found in news media which never been solved.

Machine learning is powerful tool which enables data analysis of large databases more efficiently and quickly. This enables the crime investigators determine crime patters, locations, timing and even deciding certain decisions effectively. With help of present day computers, modern data mining techniques can analyze thousands of data within a very short amount of time saving hundreds of hours. In addition, installing and using such programs costs significantly less than hiring trained personnel. If properly optimized, these data mining techniques can be errorless ensuring perfect outcomes than a person, especially who works for a long hours.

In this thesis, we have presented a general and effective framework for crime data mining to identify crime pattern using machine learning. We also measured the effectiveness of this framework using raw data and depicted the result in an orderly fashion.

3.2 Research Subject and Instrumentation

A criminal act can be a wide range of activities starting from a normal civil infraction to something big like mass murder or Bank loot.

Different law-enforcement agencies uses their own as well as the country's well-established standards to identify an act as a crime [6]. While finding a crime the jurisdictional needs and requirements are also followed and maintained. Table 2 lists the crime activities on which we have put our focus to analysis and identify crime pattern throughout the country. These crime-keywords also represent different types of crimes which the local law-enforcement agencies keep track on. We have devised these categories, keywords in different offences which are classified by country's law-enforcement authorities. Among these crimes some are less significant that are dealt with local authorities while others are big enough to draw national and international attentions. The example of such crime can be a 2016's bank robbery in which close to $1 billion USD of Bangladesh bank was taken away [7]. Such crimes happen with less in number but have greater significant value. And these are not out of crime pattern.

Many crimes happen throughout the country while some other happen only in specific regions. For example, theft, robbery, sex-crime, fraud happens throughout the country in both rural and urban areas. Other crimes such as Arson, drug offences, Traffic Violations etc. mostly happens in urban areas. Another important fact that, in many cases the crime pattern remains the same whereas the numbers of some particular crimes are increasing more rapidly than others. Crimes are also spread over more places than it previously was [6]. For these reasons, analyzing and detecting crimes using conventional ways has become more difficult and time consuming.

Table 2: Different Crime Types and their Domain

Crime Types	Crime Domain
চুরি - Stealing	Theft
ডাকাতি - Robbery	Theft
ভাঙচুর - Vandalism	Other Crime

যৌন নির্যাতন - Sexual Abuse	Sex Crime
অগ্নিসংযোগ - Arson	Violent Crime
নাশকতা - Sabotage	Other Crime
ছিনতাই - Snatching	Theft
ইয়াবা - Yaba	Drug Offences
ধর্ষণ - Rape	Sex Crime
ছুরিকাঘাত - Stab	Violent Crime
মাদকের ব্যবসা - Drug Business	Drug Offences
বাল্যবিবাহ - Child Marriage	Social Crime
বোমাবাজি - Bombing	Violent Crime
অপহরণ - Kidnap	Violent Crime
লড়াই - Combat	Violent Crime
সন্ত্রাসবাদী আক্রমণ - Terrorist Attack	Violent Crime
হুমকি - Threat	Other Crime
অর্থপাচার - Money	Fraud

Laundering	
হামলা - Assault	Violent Crime
খুন - Murder	Violent Crime

Whether a crime is less significant than others the amount of socio-economic damage is great. For example, trafficking and transnational in stolen properties can severely impact on business, trade and government revenue [8]. On the other hand, local and reputed gangs, criminal organizations and drug cartels exact a huge financial costs while threatening public health and overall their safety [9].While murder, forcible rape, aggravated assault and robbery are local police matters, terrorism is a global problem, where different country collaborates to minimize the number of terrorists and hence minimizing the number of crimes. So, it is equally important to pay attention in all sorts of crime.

3.3 Data Collection Procedure

In the data collection steps we used both print media and online media to collect as much data as we could. Print media includes different newspapers, magazine and pictorials while online media includes blogs, social media, news sites and RSS feeds. The collected data was initially stored for further analysis. In our data collection procedure, we set certain keywords to find the exact match of the crime that occasionally held in Bangladesh. Table 2 depicted the different crime types that we have used in this thesis.

As there is no predefined database to where we can find our desired we started to look for open source database storage along with the online and offline media. It is good news we have managed to collect most recent data along with data volume from 2014 to 2016. Most of these data were unstructured, and we needed to process it for further purposes. The quality data process involved automated testing, format validation as well manual verification and approval.

Furthermore, we also contacted with the "Adabor Thana Police Station" to know issues regarding crime prediction and data analysis. Their friendly behavior amazed us and they explained how they deal with the common problems, such as finding exact locations, determining and eliminating crimes and anti-social behaviors, court result matching, filling missing information etc. These findings help to further collect and analyze data for better outcome.

Table 3: Description of the Crime Dataset

Filed Name	Data Type	Description of the Field
Crime Name (Type of Crime)	String, Character (Nominal Data)	one of the 20 categories we have previously analyzed
Crime ID	Number (Nominal Data)	ID of that specific crime
Time and Date	Numeric, Date-Time (Interval Data)	Date and Time when the crime occurred
Falls Within	String, Character (Nominal Data)	Data Source Type (Online/Offline)
Location	String, Character (Nominal Data)	Name of the Place where the crime occurred
Crime Domain	String, Character (Nominal Data)	The name of crime domain under which a crime falls

Table 3 shows the attribute and features of the collected data sets. In all our datasets the "Reported By" category is currently print media and online media where print media includes newspapers, pictorials and several types magazine and online media includes blogs, social media, news sites, RSS feeds etc. In the near future the real-time reporting feature will also be included. The "Falls Within" category will also increase as the data volume will rise time to time. The attribute "Location" provides the location of the crime

where it took place. Such data can be the name of the, Road name and any other places with point of interest such as Parking area, super market and shopping area. The "District" provides the name of the district where the crime took place. In future we are interested to put more information like police station name along with the area code. The "Type of Crime" attribute is one of the 20 categories we have previously analyzed with the help of "Adabor Thana Police Station" and depicted in the table 2. The "Frequency of Crime" attribute defines the number of times a particular type of crime has occurred within a specific amount of time. The "Outcome" category tells the final outcome of the crime whether it is under investigation or already found the suspect and completed the investigation. This is most crucial part where we face difficulties to collect the required data as we don't have certain clearances to ask the law-enforce authority. So our data only represents the information what we have found in the online and print media. In where we were unable to determine the outcome the field is left with the word "Unknown".

An instance of the crime data set is one record or data object which can be characterized by the above-mentioned attributes.

The first step that we took after collecting data was to perfectly aggregate the unstructured data into one dataset which was further used for the actual experiment. It contains more than 500 crime records which occurred in different parts of Bangladesh. Most of the missing fields here is the "Outcome" field. The dataset would contain more data if there were no missing fields. However, we are also interested to fill the missing values, but as mentioned earlier, it is not possible to get some sensitive information without proper clearance.

3.4 Analysis of Different Data Mining Algorithms

Traditional data mining techniques which includes classification and prediction, association analysis, outlier analysis and cluster analysis provides best result for a structured data [10]. However, some newer techniques allow identification of crime data in structured as well as in unstructured data. Like other data mining procedures, crime-data mining also raises privacy concerns which results difficulty in data collection

procedures [9]. There are also some automated data mining techniques developed by researchers now are being used in law enforcement as well as in national security applications. Some extraction procedures like Entity Extraction helps to identifies different patterns from video, audio, images and text materials. These techniques has been used for a long time for identifying a specific person, vehicle, address or a significant crime group from law-enforcement authority's reports [11]. With proper machine learning techniques and following data mining algorithms which includes program flow, data structure, organizational comments, and use of proper variables can facilitate secondary investigation by grouping similar activity together and tracing the criminal's behavior. Entity extraction is good for basic crime analysis. However, its results and performance highly depends on the availability of clean input data.

Different clustering techniques can group datasets into classes that has similar characteristics to minimize or maximize intra-class similarity. It is highly useful in crime datamining, for example, it can successfully identify criminals who conduct similar crimes and distinguish different groups that belongs to different crime gangs. These clustering techniques have no predefined classes to be set for assigning items. Sometime researchers utilize the statistics-based space algorithm to successfully and automatically associate objects such as organizations, vehicles and even persons in crime records [6]. To exploit Bank Secrecy Act data in order to support the analysis and detection financial crimes such as money laundering, link analysis technique is used with Financial Crime-Enforcement Network AI System. All these techniques can automate major part of the crime data analysis, but high computational intensity is typically required.

Associate rule mining techniques discover frequently happening items in database and shows the patterns as association rules. It has been used in networking systems to detect intrusion by deriving the association rules from the interaction history if different users. For this reason, this particular technique can be helpful to detect the future potential network attacks [11].

Similar to the association rule mining technique, sequential pattern mining also detects frequently occurring sequential items over different transaction sets happened at different

times. This approach can find intrusion patterns in time-stamped data in order to detect network intrusion. It benefits crime analysis but also requires highly structured and reach data obtain a meaningful result.

Deviation detection technique applies specific measures to study datasets that differs notably from the other datasets. It is also known as outlier detection which can be highly for network intrusion detection, fraud detection, and related other crime analysis.

Techniques, such as string comparator compare textual fields in two database records and then compare the similarity in them. Techniques like this can be used to detect deceptive information including name, social security number and addresses in criminal records [6]. However, the larger the size of the database will be, the more intensive computation is required.

Datamining tools that is used to social network analysis can describe the roles and interactions among different nodes within a conceptual network. Law-enforcement agents and investigators can take benefits from these services to construct a conceptual network to illustrate roles of the criminal, the flow of tangible as well as intangible information and goods, and associations among all these entities. Secondary analysis in such network can reveal the subgroups, critical roles and vulnerabilities inside these networks. Approaches like this helps to visualize the criminal network if observed closely and deeply.

3.5 Crime Data Mining Framework

It has been found that, there are many efforts have been taken using different automated techniques to analyze crime patterns. However, most of them are found without any unifying framework and describing how to properly apply them [6]. More specifically, understanding and maintaining the relation between crime type characteristics and analysis capability can help law-enforcement authorities to effectively use data mining techniques to track and identify crime patterns, trends and even predict the crime. Based on the 20 reputed journals and extensive online as well as offline research, we have developed a general framework which utilizes the advanced machine learning techniques

for crime data mining. The framework is depicted in Figure 3.5.1. The framework shows the data flow between data mining techniques which applied for criminal analysis and the crime showed in Table 2.
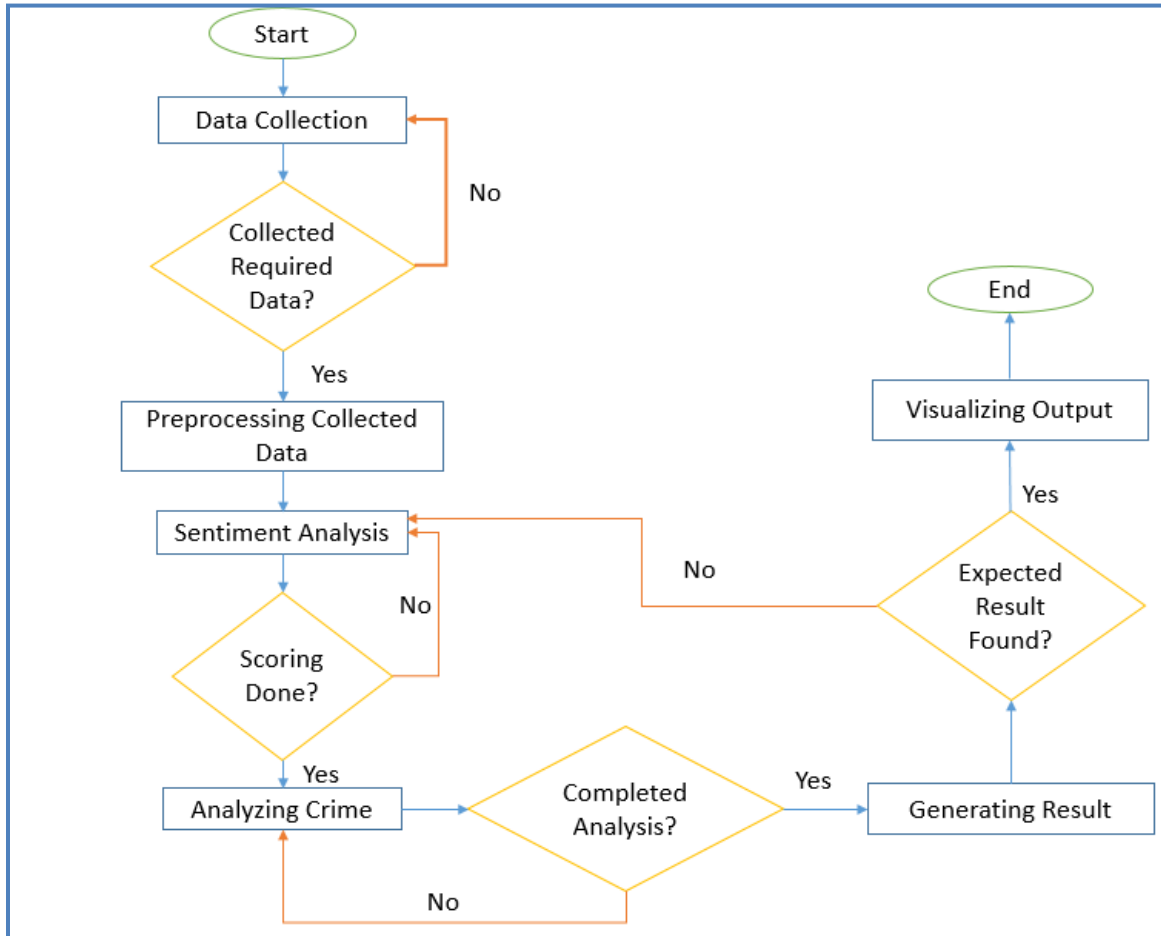

Figure 3.5.1: Crime Data Mining Framework to Identify Crime Patterns

Our crime data mining framework consists of six phases which includes Data Collection, Processing of collected data, Sentiment Analysis, Analyzing Crimes, Generating Results and finally visualizing the Output.

In Data Collection phase, our aim was collect sufficient data to successfully identify the crime patterns. In this phase if it is found that, there was not enough data collected, the phase is repeated to collect more data.

In the next phase we process the collected data and we perform the Sentiment Analysis. If scoring is completed we start the next which is Analyzing the Crime. If scoring is not completed we perform sentiment analysis once again.

After successful analysis of the crime, the system generates the result. If the result matched with the expected result, the output is visualized. If the result does not matched with the expected result and if it seems inappropriate the goes back to Sentiment Analysis again.

This is a complete framework that takes the input generate crime patters more effectively. Upon successful implementation of this framework, we believe that investigators can use this as a tool to analyze crimes, find the frequent crime zones and prevent future crimes by taking necessary precautions.

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

Criminal activities are now happening more often than before. The situation is worse, not just in Bangladesh but throughout the world [12]. Criminals also sometimes form networks to carry out their crimes more successfully. In this regard, it has become a challenge to the law-enforcement authorities to minimize the crimes using traditional ways. The age of Information technology has made our lives easier in many ways. Modern algorithm and different data mining techniques are now being used to solve various problem including crimes as well.

In this thesis we have proposed such data mining techniques to visualize the crime scenario and identify the crime pattern so that law-enforcement authority finds it easier to stop certain crimes in certain places. Our findings also generate crime prone areas which is crucial for crime prediction as well. In this chapter we will discuss the whole crime procedure to identify crime patterns using modern machine learning techniques with useful algorithms. We discuss the procedures following the framework discussed in the previous chapter.

4.2 Experimental Results

We have describes the data collection procedures and statistical analysis part in the previous chapter. We have also mentioned the crime datamining framework. As our entire process is the part of that framework we will start with Data Processing and Sentiment Analysis part here.

4.2.1 Data Pre-processing and Cleaning

Data processing is very important and required to process any data successfully. The term sometimes refer to preprocessing and cleaning raw data. I serves an important purpose and that is makes the data simpler and clean for further processing.

In this thesis, we have collected raw data (crime news) from both online and offline sources. Data that we have collected come out with special characters, commas (,), full

stops (.), breaks (br/) and other characters those are not found in the ASCII table. So at first, we cleaned our data so that we can successfully extract crime information from collected crimes news. Before preprocessing, the accuracy was low and we have found that it has increased significantly.

The following figure (Figure 4.2.1.1) provides a sample input. After the sample output also depicted to visualize the output of the crime analysis.

Sample Input:

Report from: 07.10.2015, Category: Offline Source

```
Man held ◆for killing◆ wife, stepson in Keraniganj
11:42 AM, October 07, 2015. The Daily Star
Police have arrested a man for allegedly killing his wife and a nine-year-old stepson in Keraniganj
at the outskirts of capital Dhaka.
The deceased were Shahera Begum, 30, and her son Faisal, reports Bangla daily Prothom Alo.
The man detained was Ishaq Mia.
Locals say Ishaq recently married Shahera, mother of a daughter and son.
Police suspects that Shahera was stabbed to death while Faisal was strangulated last night.
Ishaq was detained shortly afterwards.
Stepdaughter Jannat was stabbed was sharp weapons. She was taking treatment at
Dhaka Medical College Hospital.
Keraniganj Model Police Station Assistant Sub-Inspector Sohrab Hossain said
the incident might be motivated from family feud.
The dead bodies have been sent for autopsy at Sir Salimullah Medical College Hospital.
```

Figure 4.2.1.1: Sample input before preprocessing

Here the text field we collected as necessary parameters for crime pattern analysis and detection are crime domain, name of the crime from table 2, location and date & time.

The raw text is little messy for these crime news, so before we can perform any analytics on these texts we need to clean things up as there might be some characters which might not be found in ASCII. To clear the text, we have used regular expression. Here is the regex term we have used;

1. **ReplaceNoSpace=re.compile ("(\.)|(\;)|(\:)|(\!)|(\')|(\?)|(\,)|(\")|(\()|(\))|(\[)|(\])")**
2. **ReplaceWithSpace=re.compile ("(<br\s*/><br\s*/>)|(\-)|(\/)")**

Now, after processing the above text, the text is now looks simpler and easier to analysis for further usage. Figure 4.2.1.2 shows the processed text.

```
man held 'for killing' wife stepson in keraniganj
1142 am october 07 2015 the daily star
police have arrested a man for allegedly killing his wife and a nine year old stepson in keraniganj
at the outskirts of capital dhaka
the deceased were shahera begum 30 and her son faisal reports bangla daily prothom alo
the man detained was ishaq mia
locals say ishaq recently married shahera mother of a daughter and son
police suspects that shahera was stabbed to death while faisal was strangulated last night
ishaq was detained shortly afterwards stepdaughter jannat was stabbed was sharp weapons
she was taking treatment at dhaka medical college hospital
keraniganj model police station assistant sub inspector sohrab hossain said
the incident might be motivated from family feud the dead bodies have been
sent for autopsy at sir salimullah medical college hospital
```

Figure 4.2.1.2: Processed text after Pre-processing

Here we have removed any commas, separator, full stops and other special characters those are not helpful and ASCII characters. We have also made the entire texts lowercase so that there will be no mismatch while caparisoning text's ASCII value.

4.2.2 Data Processing, Sentiment Analysis and Crime Detection

For data processing, sentiment analysis and overall, classification we have used Naïve Bayes algorithm which is a statistical as well as supervised learning method. Naïve Bayes is a good probabilistic classifier which based on the input provides a probability distribution of all classes rather than generating a single output [13]. The algorithm successfully classifies a news article into predefined crime types in which the news fits best. What we get here is that,

"What is the probability that a crime news A belongs to a predefined crime type B?"

The reason behind choosing the Naïve Bayes classifier is that, this classifier is relatively simple and it converges faster than logistic regression. Other algorithms like Support Vector Machine (SVM) takes a lot of memory. Because of its performance implementation when the training set increases, the execution speed decreases rapidly [10]. Using the Naïve Bayes algorithm we have created a model after training different crime data like robbery, vandalism, snatching, arson, rape, other sex abuse, burglary,

murder etc. Training here means that we have taught the algorithm using particular inputs in a way that we can test the result using unknown inputs. We used test data to find the accuracy of our model. Unlike some other algorithms, the accuracy increased here as the test set increases. Another significant benefit of Naïve Bayes in our particular scenario is that, it works excellent for small number of training data to find the classification parameters.

Naïve Bayes classifier also helpful in fixing the zero result frequency while calculating a probability like P(F)* P(G/I)*P(H/I)*P(J/I) where P(H/I) = 0.

These type of probability always outputs zero that leads an uncertainty while calculating the results. In order to avoid these circumstances, we have added a (+) 1 while counting every zero values so that we can achieve uniform distribution. Our test results depicts that, Naïve Bayes algorithm works in a more than 80 percent accuracy. To improve the accuracy, we have designed the algorithm to ignore frequently found words like "And", "The", "Of", "OR", and certain phrases. We are still working to increase the accuracy to reach at least 90 percent.

To find offender's name, crime location as well as the victim we have integrate NER (Named Entity Recognition) in the crime analysis. NER is very helpful and frequently used to classify text elements in predefined categories which includes person names, locations, organizations, time and date [14]. So by using this technique in our crime analysis we will be able to get more details crime related information.

Using the NER techniques we can also extract the weapons list offenders used while committing crimes. We have also added the Co-reference Resolution concept to distinguish the reference entities in the text. In machine learning, Co-reference occurs while at least two expressions in a certain text refers to a same person [14].

```
banasree 2 kid murder probe handed to db
1124 am march 07 2016 the daily star
the probe of two children murder in dhaka's banasree who law enforcers said were killed by their mother
has been handed over to detective branch db maruf hossain sorder deputy commissioner media of dhaka
metropolitan police dmp told the daily star this morning of the development
so far five people have been picked up for questioning in this regard the mother
mahfuza malek jasmine is currently in a five day remand for interrogation
nusrat aman aroni 14 and her brother alvi aman 6 were found unconscious at their apartment
on march 01 they were later declared dead in the evening after being taken to dhaka medical college hospital
it was initially believed they died of food poisoning however autopsy report suggested that
the two children were murdered and there were signs of struggle
rapid action battalion rab interrogated the parents and declared that it was the mother who
killed the two children she was said to have been fearful of the children's future father of the children
filed a case against his wife mahfuza in this regard with rampura police station after
about a week the case has been handed over to the detective branch db
```

Figure 4.2.1.1: The pre-processed text input for Crime data analysis

We have extracted all the required and referenced entities from the text. Figure 4.2.1.1 shows the input pre-processed data and Figure 4.2.1.2 shows the output of the crime news analysis.

```
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 14:57:15) [MSC
4)] on win32
Type "help", "copyright", "credits" or "license()" for more
>>>
 RESTART: G:\Study\Mahmuda Sharmin Papia\Final Year Thesis\F
g\Crime Identification\crime_identification2.py
Crime Type:
murder
Location:
dhaka
Month:
march
Year:
2016
>>> |
```

Figure 4.2.1.2: Output of the from input text field

4.2.3 Result Generation and Exporting to Crime_Data Table

For further processing and perfect visualization of our crime pattern analysis we have stored data in the crime_data.csv. The attributes we have collected from successful crime

detection are Crime Type, Location, Month and Year. We have collected these information because we are eager to visualize the crime by different years, months, location and crime types so that it will be easier to law enforcement authority to take appropriate action at right location in right time. The other advantages are, it is very helpful and easy to visualize the crime patterns and interpret the results. The below table (Table 4) shows the data we have generated upon successful crime extraction from crime news.

Table 4: Crime Data generated after Identifying Crimes from Crime News

The dataset we have generated from crime news comes to 300 data set. However, we have collected 500 news among around 200 are turned out that those are not crime news while checking for crime identification. From this crime_data.csv we have generated results to visualize crime prone areas, mostly occurred crimes and year & month wise crime data. The below section provides detail about output visualization.

4.2.4 Output Visualization

From cime-data.csv we have generated earlier is now used to visualize the crime patterns of Bangladesh. For perfect visualization the tools we have used along with Python are;

- Pandas
- Seaborn
- Matplotlib

 All of these are extended libraries to be used for plotting crime data in the desired. Pandas is a library software written for the Python language for data manipulation as well as analysis. In particular, it offers data visualization and operations for manipulating numerical values and time series.

Seaborn is a Python programming data visualization library based on the matplotlib. It provides a high-level interface for drawing very attractive as well informative statistical graphics.

Matplotlib is a plotting library for the Python and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, Qt, GTK+ or wxPython.

With these added libraries we have plot data for;

- Perfect Visualization Of Crime Prone Areas
- Visualization For Crime Prone Months
- Comparison Between Different Years Of Crime
- Perfect Visualization For Different Types Of Crimes
- It Reflects On The Data We Have Analyzed
- Data Those Are Out Of Range Are Discarded
- Depicts Probable Places For Future Crime

So after knowing the crime prone regions we can take more precautions and preventive steps in those areas to prevent any future crime. The law-enforcement authority can set more patrols, put CCTV cameras and take other measures to reduce the crime rate.

The statistical data is showed in figure 4.2.4.1. The X-axis represents the locations of Bangladesh and Y-axis represents the crime rate. It is calculated based on our crime records. From these findings, it is clear that Dhaka is the most crime prone area of Bangladesh. It is to mention that, the graph and the map may change over time as we are still inputting more data.
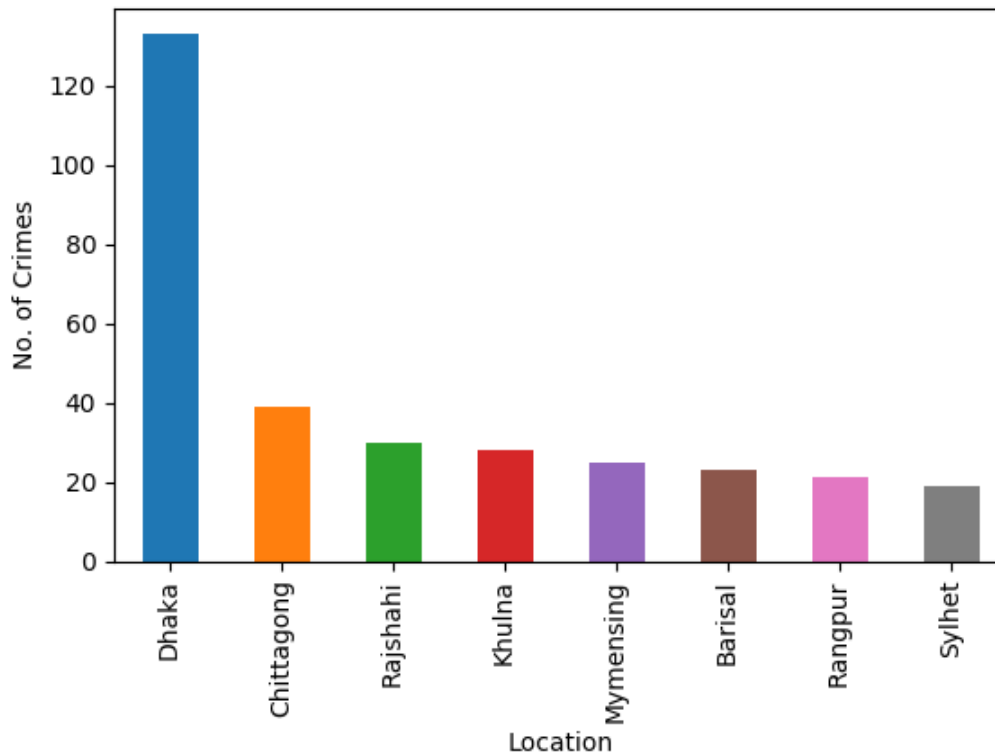
Figure 4.2.4.1: The Statistical Data Regarding Crime Patterns in Different Cities

Figure 4.2.4.2 shows types of crime occurred in different cities of Bangladesh. It is found the crime "Murder" mostly found in news analysis. The reason behind this information might because news regarding murder are published more than other news.
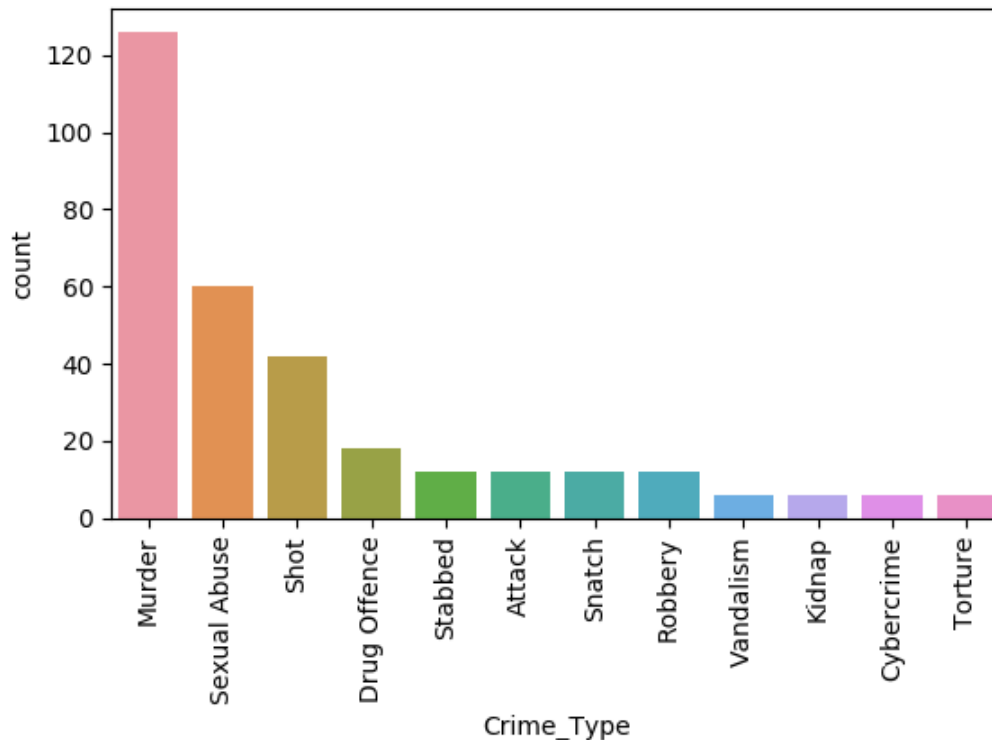
Figure 4.2.4.2: Types of Crime occurred in different cities of Bangladesh

The figure 4.2.4.3 shows information regarding crimes occurred in different years we have collected data from. Finally, the figure 4.2.4.4 shows crimes occurred in different months. It is found that November and December is the most crime prone month of a year, while June, August and September are the lowest crime prone month of a year. Along with location, this information is very useful while preventing crimes. Law enforcement authorities can use this information to be more careful on most crime prone months.

4.3 Descriptive Analysis

Criminals are now more intelligent than ever. It takes huge effort to find criminals and to stop crimes. Modern day data mining Technology can help law-enforcement authority to detect crime patterns so that frequent crimes can be minimized. Our data mining task provides insight of the crime patterns generating frequent crime prone areas with the number of crimes previously occurred. It can help the law enforcement authority to take more steps in those areas which are more affected by frequent crimes. With these data

available it is also now possible to find zones which are more probable place for future crimes and which crimes are more prone to happen.

Our data came from different online and offline media such as newspaper, magazine, social media, online new portal etc. We used Naïve Byes classifier to initially classify and analyze the crime data we got. Then with the help of NER and Co-reference Resolution concept we generated the output from the raw data where text fields are analyzed to get crime names, locations, crime types and time and date. Here, these techniques helps to measure the relation between criminals and the crime patterns. The hierarchical data mining procedure like decision tree helps to generate probability decision by calculating several attributes which are initially used to identify the crime patterns.

As shown in figure 3.5.1, our datamining technique consists of procedures including data processing, classifications, sentiment analysis and analysis of crimes to generate the crime patterns of a certain locations. Our datamining techniques utilizes the 20 keywords depicted in table 2 to identify whether a location is prone to crimes or not.

Along with online and offline data collection we also conducted an effective field study at "Adabor Thana Police Station" to get the insight of the crime analysis. We have also talked with two experts (Sub-Inspectors) there to evaluate the validation of our findings comparing with their knowledge about the crimes of "Dhaka" city. They confirmed that, our findings are accurate and hereby can be helpful and used to determine crime pattern in Bangladesh.

Both of the experts believed that this crime analysis could help identify crime sensitive zones and help law-enforcement authority to take appropriate actions. Significantly, it would provide more knowledge about different crime types, there location as well percentage of each crime types that would require man hours to find manually.

4.4 Summary

An experienced human investigators with proper knowledge can often identify crime patterns by analyzing crimes precisely. However, the number crimes are increasing as

well as the complexity. Human errors are also prone to occur which leads in situation where more time requires to solve case. In the meantime, criminals get more time to escape from the justice by destroying evidence and taking other necessary to prove them innocent. With increased efficiency and errorless computation, crime data mining can facilitate accurate crime patterns enabling investigators to find and stop crimes more effectively.

In order to determine the crime patterns we created or framework with extensive research and field works. We followed the framework step by step to get the desired output.

To follow the framework and to identify crime patterns more effectively we used Naïve Bayes algorithm Concepts along with Pandas, Seaborne, Matplotlib. With these techniques it was possible to identify crime prone areas, crime types, time and date from a generalized crime reports. To depict the result and procedures more visually we have also introduced crime_data.csv which helps to decide whether a location is more crime prone or not. All of these findings help to generate the crime maps showed in figure 4 and statistical data showed in figure 4.2.4.2.

With these findings, it is now possible now possible to determine the most crime prone areas, mostly occurred crimes and their locations. Law-enforcement authority now able to determine the probable place of frequent crimes can take preventive steps so that crimes occurs less frequently than ever.

CHAPTER 5

Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summary

Criminal activities are now happening more often than before. The situation is worse, not just in Bangladesh but throughout the world [12]. Criminals also sometimes form networks to carry out their crimes more successfully. In this regard, it has become a challenge to the law-enforcement authorities to minimize the crimes using traditional ways. The age of Information technology has made our lives easier in many ways. Modern algorithm and different data mining techniques are now being used to solve various problem including crimes as well.

In this thesis we have proposed such data mining techniques to visualize the crime scenario and identify the crime pattern so that law-enforcement authority finds it easier to stop certain crimes in certain places. Our findings also generate crime prone areas which is crucial for crime prediction as well. To follow the framework and to identify crime

patterns more effectively we used Naïve Bayes algorithm Concepts along with Pandas, Seaborne, Matplotlib. With these techniques it was possible to identify crime prone areas, crime types, time and date from a generalized crime reports. To depict the result and procedures more visually we have also introduced crime_data.csv which helps to decide whether a location is more crime prone or not. All of these findings help to generate the crime maps showed in figure 4 and statistical data showed in figure 4.2.4.2.

We are optimistic that now, it is possible to determine the most crime prone areas, mostly occurred crimes and their locations. Law-enforcement authority now able to determine the probable place of frequent crimes can take preventive steps so that crimes occurs in a less frequently.

5.2 Conclusion

Crimes are now more frequent than it was before. Although, there are new and improved techniques and measurement are being used by law-enforcement authorities, criminals are somehow getting their things done. It is now a major concern not just for general people but also law enforcement authority how to deal with these growing number of crimes.

Modern Technology has allowed us with different techniques that edges our life makes it comfortable. Data mining techniques has been used for different problem solving purposes. Crime data mining techniques has been used to aid law-enforcement authorities to deal with crimes more effectively. In this thesis, we have used popular and effective datamining algorithms to classify crime keywords and to generate crime pattern for Bangladesh. We have done the classification using Naïve Bayes classifier which showed around 80 percent accuracy. Using this classifier we have trained a lot of news article along with required keywords. The keywords are generated based on field extensive online and offline results and a field work to our nearest "Adabor Thana Police Station".

Our findings in this research is promising as we are able to generate crime patterns for difference divisions of Bangladesh. If enough data found, we use this technique to generate crime reports for a specific region like a city or a police station. We are

optimistic that, law enforcement authority will find it useful to detect frequent crime zones and take necessary precautions to reduce the crime.

5.3 Recommendation

While performing the thesis we have found that, a large amount of test data and keywords are required if we want greater accuracy. Lack of good amount of raw data is also required to expand the research work for different regions. A high configuration GPU (Graphics Processing Unit) enabled computing device is also required if we want to process huge amount test data in shortest time.

So if we get enough data along with a high performance computer, it will be easier for us to increase accuracy to more than 90 percent. We will also be able to use our technique for a specific regions like city to determine the crime pattern.

5.4 Implication for Future Research

World-wide the number of crimes are increasing as well as the techniques to reduce the crime. Introducing crime data mining in this regard brings new frontier to determine and predict crimes for different places. It is highly useful to determine crime sensitive zones as well. We use our technique to determine crime patterns for different divisions of Bangladesh. The accuracy of our technique is good, but it needs to be increase for better outcome and utilization. So the future scopes of this includes;

- Collecting More Data for Greater Accuracy
- Using More Test Data to Give the Technique More Perfection
- Finding and Using More Crime Keywords
- Implement the Algorithm for Different Cities of Bangladesh
- Working With the Law-Enforcement Authority So That Our Technique Can Be Used For Real Life Purposes

If we will be able to perform all of these future works, we can say that our system will be highly beneficial for finding crime patterns of different regions of Bangladesh as well as predicting the crime so that crimes can reduced easily and effectively.

References:

[1] P. Phillips and I. Lee, "Mining co-distribution patterns for large crime datasets - ScienceDirect." [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417412005945. [Accessed: 30-Jul-2018].

[2] O. Isafiade, A. Bagula, and S. Berman, "A Revised Frequent Pattern Model for Crime Situation Recognition Based on Floor-Ceil Quartile Function," *Procedia Comput. Sci.*, vol. 55, pp. 251–260, Jan. 2015.

[3] S. Roy and S. Bhattacharya, "Application of Fuzzy-Rough Oscillation on the Field of Data Mining (Special Attention to the Crime Against Women at Tripura)," *Procedia Comput. Sci.*, vol. 45, pp. 790–799, Jan. 2015.

[4] L. McClendon and N. Meghanathan, " Using Machine Learning Algorithms to Analyze Crime Data - ReseaarchGate" [Online]. Available: https://www.researchgate.net/publication/275220711_Using_Machine_Learning_Algorithms_to_Analyze_Crime_Data. [Accessed: 30-Jul-2018].

[5] S. Shojaee, A. Mustapha, F. Sidi, and M. A. Jabar, "A Study on Classification Learning Algorithms to Predict Crime Status," p. 10.

[6] Chen, Hsinchun, et al. "Crime data mining: a general framework and some examples." computer 37.4 (2004): 50-56.

[7] Wikipedia. "Crime in Bangladesh" [Online] Available: "https://en.wikipedia.org/wiki/Crime_in_Bangladesh". Last Accessed: November 12, 2018.

[8] W. Chang et al., "An International Perspective on Fighting Cybercrime," Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics, LNCS 2665, Springer-Verlag, 2003, pp. 379-384.

[9] M.Chau, et al, "Extracting Meaningful Entities from Police Narrative Reports, Proc. Nat'l Conf. Digital Government Research, Digital Government Research Center, 2002, pp. 271-275.

[10] R.V. Hauck et al., "Using Coplink to Analyze Criminal-Justice Data," Computer, Mar. 2002, pp. 30-37.

[11] O. de Vel et al., "Mining E-Mail Content for Author Identification Forensics," SIGMOD Record, vol. 30, no. 4, 2001, pp. 55-64.

[12] Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014.

[13] Eibe Frank and Remco R. Bouckaert. "Naive bayes for text classification with unbalanced classes". In Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD'06, pages 503–510, Berlin, Heidelberg, 2006. Springer-Verlag.

[14] Saltos, Ginger, and Mihaela Cocea. "An exploration of crime prediction using data mining on open data." International Journal of Information Technology & Decision Making 16.05 (2017): 1155-1181.