SHAHJALAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

UNDERGRADUATE THESIS

Sachetan: A Crowd-source Based Personal Safety Application

Author:

SHADMAN Habib

OZAYER Islam

Supervisor:

Sheikh NABIL Mohammad

A thesis submitted in partial fulfillment of the requirements for the degree of BSc(Eng)[0.3cm] *in the*

Department or CSE

September 11, 2018

Declaration of Authorship

We, Shadman Habib and Ozayer Islam, declare that this thesis titled, "Sachetan: A Crowd-source Based Personal Safety Application " and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

Crimes and criminals are one of the ever-growing problems in Bangladesh. Despite the upmost efforts from our govt., this difficult situation is still haunting our country. Due to the fast growth of digitalization and computerization, nowadays almost all problems are being handled in scientific ways. So, we are also trying to build a proper system which can help all of us to find any pattern available in those crimes, analyze them and also possibly predict any future crime and thus possibly prevent them from even occurring. Our system is based on an android device named "Sachetan" which is a crowd-source based application that collects, analyzes and visualizes the crime spots around the city. We used data mining technique for predicting crime and finding pattern. We focused on reported crimes both by newspapers and the users of this app. ...

Acknowledgements

We would like to thank our supervisor Sheikh Nabil Mohammad for his instruction and guidanceduring this work. He's has been really patient in teaching us and it wouldn't have been possible without his help.

We are also very much grateful to our co-supervisor Moqsadur Rahman for his guidance in our thesis procedure

We would also like to thank our senior Sakhawat Hossain Saimon, Mustafizur Rahman Nebir, Mithun Das, Nishikanto Sarkar Shimul and Tanvir Islam Preom, for selflessly aiding us with their knowledge.

9th September, 2018

Shadman Habib

Ozayer Islam

...

Contents

List of Figures

List of Tables

Dedicated to those young souls that passed away from the horrible road accident in Dhaka a few days back.…

Chapter 1

Introduction

1.1 Our Motivation

Nowadays crime is a very common threat to every individual person and the common economy of our country. Not a single day goes by that doesn't include any news of crime. Despite all these crimes happening around us, we can not say that those crimes follow a general pattern neither can we say that they are random. But one thing we know for sure is that they possess a great threat to all of us each and every moment that passes by. Crimes like murder, robbery, theft, mugging, rape etc. are so frequent that now we are hoping to find patterns between them. Besides these crimes, road accidents are also posing a great threat to our lives as a lot of people die daily due to these unfortunate events. A report [1] by daily star shows that over 2400 people died from road accidents in 2018 only in our country! That shows how devastating this problem is becoming. Also, the crime report [2]came that university student from SUST was stabbed to death by muggers over dinner money! So, we can understand how serious the problem is in our neighborhood, city and country. Our goal is to reduce these crimes somehow with the help of technology and with the involvement of law enforcement authority. As there is not much work done for reducing crime in our country, we studied a healthy amount of papers related to crime and learn various characteristics, methodology, types, symptoms of different crimes and various ways of crime pattern generation procedure and prediction of future crime. We can help as long as we get enough information about the crimes. Of course, we cannot predict results with 100 percent accuracy, but with more and more dataset, we can do enough to predict the crimes. Collecting those data is never

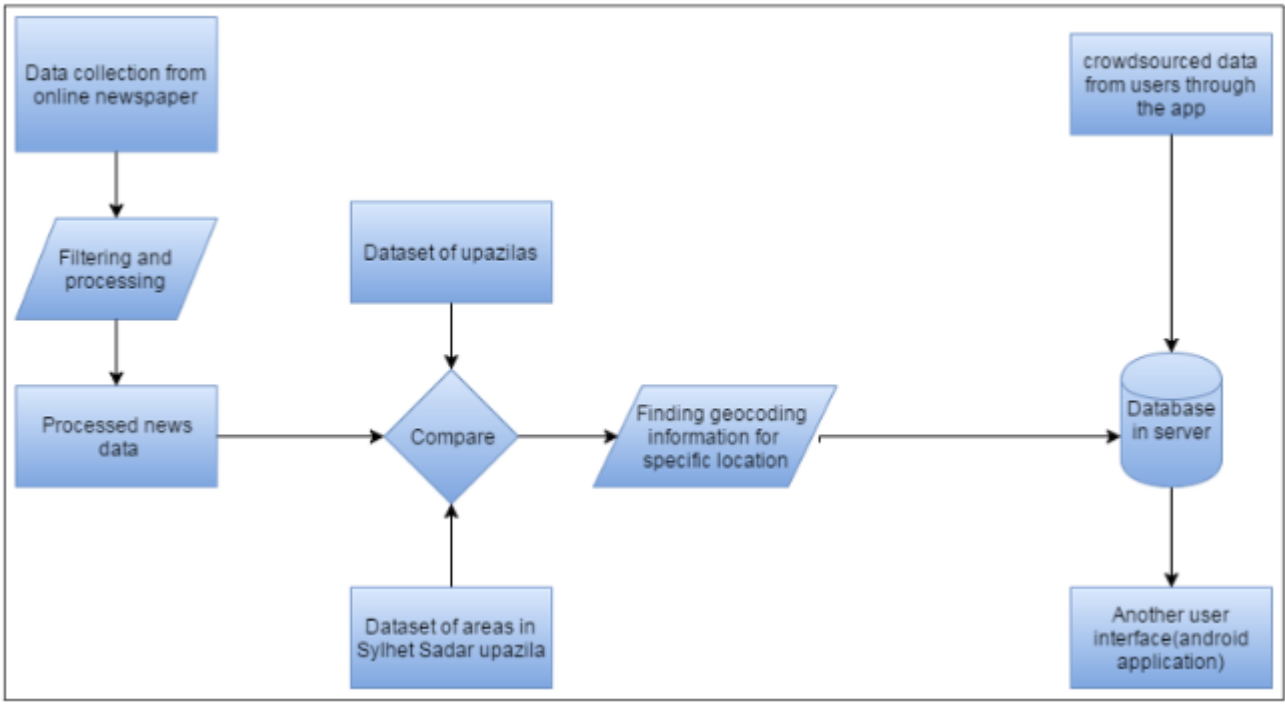

FIGURE 1.1: Basic System design of the app

an easy job, we had to face a lot of difficulties while collecting the data. Among those difficulties, few of them are:

- Lack of help from law enforcement authority.
- Lack of spontaneous volunteers to report any crime.
- Reluctant people who denies to report the crimes they faced

After collecting data, the next step for us was to process them properly in order to make the best use of them. As the data were inconsistent and random, we had to pre-process them to run any classification. Only then we could generate any pattern or predict any future crime based on them.

1.2 Basic System

Our system first collects data from various resources, then pre-processes, visualize the crime spots and finally run classifications. Those can be visualized as below:

This platform is in continuous upgrade stage as we are developing the model gradually. We hope to continue our service and give the best possible prediction by generating best possible patterns of the crimes.

Chapter 2

Related Works

Research in crime section is an important topic all over the world. But we can see very few results in front of us that have been made public. In England, Cambridge Police department has developed a system called "Series Finder" for finding the patterns in burglary. The data included means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to another break-ins. Using nine known crime series of burglaries Series Finder recovered most of the crimes within these patterns and also identified nine additional crimes. Their system showed accuracy up to 80 percent which is quite impressive. In 2014, two researchers from India built a system based on data mining where such predictions were made, but their main work was visualization by heat map which is also available in our app. Besides these, there are also some works available which will be discussed properly in literature review section.

Chapter 3

Previous Works on the App

Previously we focused on the 'Sachetan' app. This app was developed a few years ago by our seniors. The development progress was not documented properly. So, we had to work quite hard to get the proper understanding of this project. The app had some bug issues which we fixed. Our work on the app:

- Cleaning the app
- Sign up problem fixed
- Crashing fixed
- Alternate path added
- Login page problem fixed
- Server related problem fixed
- Password recovery option added

Chapter 4

Literature Review

4.1 Crowd-Sourcing

Our project Sachetan relies on data. Data collection is the first and foremost part of our research. Huge data is needed to process data mining and generate new information. We introduced the sources of data. News data can be extracted from different newspapers by using crawler. But these data are not enough as the quantity is less and also all news of the negative incidents occurring in different locations can't be found in only newspaper. Here comes the need of crowdsourcing [3], by which we can get user data. It is beneficial because user living in a specific place can give a rare information that the newspapers might not have. So, crowdsourcing plays a vital role in our research.

What is Crowdsourcing

Crowdsourcing involves obtaining work, information or opinions from a large group of people who submit their data via the Internet, social media and smartphone apps. People engaged in crowdsourcing sometimes work as paid freelancers, while others perform small tasks on a voluntary basis. For example, traffic apps encourage drivers to report accidents and other roadway incidents to provide realtime updated information to app users.

Category of Crowdsourcing

According to an article the literature on crowdsourcing can be categorized into four parts. They are - Application, Algorithm, Performance and Dataset.

Application:

Generally, there are two groups of users in a crowdsourcing site requesters and user.

A list of available tasks is exhibited with associating time and period. Maintaining this condition, workers compete to give the best solution. Again, crowdsourcing application can be categorized into four types. They are voting system, game, information sharing system and creative system.

Algorithm:

A crowdsourcing system design can be easily formalized by an algorithm. Solution of some theoretical challenging problems are provided by internet users. They can be game theory or any algorithm base problem.

Dataset:

If a large dataset is needed for further research then crowdsourcing is a good solution to collect huge amount of data in short time. For example, in case of making a corpus huge amount of word is needed. We can offer users to provide word. In some case we can give them special offer.

Performance:

It provides a chance to make an existing technique better.

There are four types of crowdsourcing:

- Crowdsourcing design
- Crowdfunding
- Microtask
- Open innovation

Advantages:

The advantages of crowdsourcing include cost savings, speed and the ability to work with people who have skills that an in-house team may not have. If a task typically takes one employee a week to perform, a business can cut the turnaround time to a matter of hours by breaking the job up into many smaller parts and giving those segments to a crowd of workers. Companies that need some jobs done only on occasion, such as coding or graphic design, can crowdsource those tasks and avoid the expense of a full-time in-house employee.

Disadvantages:

It is unrealistic to expect that a group of people will give data at a same time. People

want quick output of their input. But it needs some time to process crowdsourcing data. Sometimes to attract people into providing data, they must be offered with some advantages in return. There is a big problem of false information also. People provide false information often and that creates a huge problem of getting proper result.

4.2 Application Based Work

This is one of the main parts of our research as we had to read through a lot of papers regarding our works. We know that this is completely a new section of research in our country and thus not a lot of existing works were found. Still tried to cover all those works and saw the advantages and limitations of them. Out of many papers, we choose those which looked good and promising enough in our perspective. Different mobile apps are also developed to help these victims by sending signal or help message to security monitor center. Among these apps, SOS Response [4], Not your Baby App [5], Circleof6 [6], Safetipin [7], Abhaya , OnWatch[8], SheSecure etc. are mention worthy. In some of these apps, GPS location is also provided. This approach in some cases mitigates the harassment, but not necessarily prevents it.

Therefore, scholars were interested in analyzing crime hot spots and burning times which are becoming a major component of the work of criminologists, crime analysts and crime prevention practitioners from the past decade. As social networks are the common platform of people to freely provide private information, for instance on their current situation, opinions etc. others try to use this opportunity to improve the prediction of crimes . With the help of crowdsourcing some attempts were made so that people can both provide crime-related information and be aware about their locations. Hollaback!, a social movement organization mainly works for raising awareness about street harassments against women. Protibadi is also developed for sharing street harassment experiences for Bangladeshi women. It allows a user to report the location along with a description. HarassMap collects and summarizes different types of harassments in a map mainly in Egypt. Another Bangladeshi movement, SafeStreet is mentionable which empowers women in public places against sexual harassment. They also enable a woman to find a safe path

that has less harassment hazard, at any point of time via their app.

4.2.1 Details

There are quite a number of works done before this app. Among those "Protibadi" and "Safestreet" are mention worthy. We read all the papers related to these works thoroughly and learnt some key features of these kinds of apps. Both these apps are focused on helping women against sexual harassments in public places. Those who faced harassments can share their experiences anonymously through the apps. Below are the special features that these apps had:

1. In the "Safestreet" app, background running is also enabled in order to capture real time harassments really quick. This way he/she can add report on that incident later.

2. In the "Protibadi" app, some additional features like web interfacing and SMS based reporting is also available which is great for crowdsourcing. These features brought some positive reviews from the female users according to the publisher.

4.2.2 Limitations

As a pioneering work, both of the apps did a marvelous job. But they also have some limitations which encouraged us to develop our very own app regarding crime. Among the limitations, the biggest one is that the apps are not available anymore! This is really bad to see that these apps are not available in Google Play Store. Even we started on this project 6 months earlier, the apps were not available even back then! We could only read the papers and learnt from them only! Beside this, the app is focused on sexual harassments only to safe the women from facing the heinous crime. They do not deal with other daily crimes like mugging, hijacking etc. In our country these crimes occur parallelly. So, a better solution was just a need of time.

4.3 Data Analysis

All of these works described in the upper section was mainly for data collecting and visualizing crime spots. None of them were for analyzing the dataset or going further with them. But there are also some papers related to data mining which were of great help to us. The summary of those papers and their model is given below:

1. Last year a group of researchers from KUET in Bangladesh published a journal regarding data mining in Crime analyzation. They trained a dataset and then predicted next possible crime happening around the country. Their dataset was collected from the official website of Dhaka Metropolitan Police Website and the data mining technique that were used was Clustering, Evaluation, Association, Prediction and Trend Analysis.

   Generally, crime is categorized into four categories according to Federal Bureau of Investigation (FBI): Murder, Forcible Rape, Robbery and aggravated assault. But in the paper the crimes were categorized as Dacoit, Robbery, Murder, Women and Child Repression, Kidnapping, Burglary and Theft. They used Linear Regression Model to generate they predictions.

   Result: They had achieved quite a good result hitting the accuracy markers up to 90 percent. But their success was mainly because of their dataset which was obtained from the Dhaka Metropolitan Police website. The data was well structured and also basic with only a few features to be considered. Below there is a table showing their accuracy for murders only. Like this table, they have few more tables showing other crimes like Kidnapping, Women oppressions etc.

   Problem with this model: The main problem with their approach is that in real life it is quite hard to find the data of the crimes in details. Their system could only predict the crimes in numbers which was not really necessary. What we need are the locations, time of incidents, criminal types etc. Only then people can be aware of those crimes and be safe. Like the Table above, all other results from their research was mainly in numbers for different region but not any

Region	Actual no. of Murder	Predicted no. of Murder
DMP	262	255
CMP	120	107
KMP	22	39
RMP	22	37
BMP	15	28
SMP	44	46
Dhaka Range	1395	1243
Chittagong Range	792	757
Sylhet Range	277	366
Khulna Range	520	486
Barisal Range	209	256
Rajshahi Range	463	406
Rangpur Range	349	416

FIGURE 4.1: Results of the prediction

specific location or any specific time. So, in our app, we are working to predict the locations and specific times.

2. Two Indian Researchers published a paper on crime detection and prediction back in 2014. Their motto was same as us, they wanted to classify the crimes into different class and then upon getting another crime report, the system finds out whether it can be classified into the existing classes or not. They collected crime data from various news reports, blogs, websites etc. After the news is collected, the keywords are extracted and they form some classes that later works when classifying the data. For example, they took VIP Presence, weather attributes, area sensitivity, notable event, presence of criminal groups etc. as crime factors. After collecting the data and Classifying them into those groups, they ran apriori algorithm to generate any pattern or connection available in those crimes. It mines the frequent crime patterns for a place. So, if there is a pattern for a crime then occurring the same pattern may result in a crime occurrence-that's what they proposed. For predicting any future crime occurrence, they also used decision tree. Decision trees are mainly for categorizing the data, so when a new news is crawled, where to classify the data is selected by this tree. The tree determines:

   - Which variable to split at a node.

- Decision to stop or split
- Assign terminal nodes

After that, finally, they visualized all the crimes around the city so that people can see actually where the crimes are occurring frequently, which zone is safer and which is not. They used heat maps to visualize the crime zones. Now, As the classifier, they used Naïve-Bayes which is one if the most common and powerful classifier. Naïve-Bayes is a supervised machine learning method that works as a statistical method for classification too. It is a probabilistic classifier which when given an input gives a probability distribution of all classes rather than giving a single output. It is mainly used to classify the crawled news to the best fitted class. From this classification, what we get is, "What is the probability that a crime type document D belongs to a given class C?". They used Naïve-Bayes mainly because of its convergence is quicker than logistic regression. Also, it is easier to implement Naïve-Bayes than SVM. The Pseudocode for Naïve-Bayes is Given Below:

Now, for the crawling of words from a news, they used NER (Named Entity Recognition) technique effectively. What it does is extract some keywords like name, location, weapon etc. in JSON format from a news and save them in the database. It is shown in the example below:

News (Input): "The bike borne chain snatchers targeted two women pedestrians in Sanpada and Panvel on May 6, 2014, Tuesday and robbed their gold ornaments. While, 60-year-old woman's gold chain worth Rs 20,000 was snatched by the bike's pillion rider around 3.45 pm, while she was walking on the street near HDFC bank in sector-14, Sanpada, yet another woman from Khalapur was targeted by the pillion rider while she was walking along the road near old Thane naka in Panvel. The thief snatched away her gold necklace worth Rs 67,500. In both the incidents, robbery case under Section 392 and 34 has been registered at Turbhe and Panvel police stations respectively."

News (Input):

"A pillion bike rider snatched away a gold mangalsutra worth Rs 85,000 of a 60-year-old woman pedestrian in sector 19, Kharghar on Friday. The victim, Shakuntala

Algorithm 1 Pseudocode

1. Given training data set D which consists of documents belonging to different class say class A and B.
2. Calculate the prior probability of class A=number of objects of class A / total number of objects
   Calculate the prior probability of class B=number of objects of class B / total number of objects
3. Find ni, the total number of word frequency of each class.
   na= the total number of word frequency of class A.
   nb= the total number of word frequency of class B.
4. Find conditional probability of keyword occurrence given a class.
   P(word1 / class A) = wordcount / ni(A)
   P(word1 / class B) =wordcount / ni(B)
   P(word2 / class A) = wordcount / ni(A)
   P(word2 / class B) =wordcount / ni(B)
   ……
   ……
   P(wordn / class B) =wordcount / ni(B)
5. Avoid zero frequency problems by applying uniform distribution.
6. Classify a new document C based on the probability P(C /W).
   a) Find P(A / W) = P(A) * P(word1/ class A) * P(word2/ class A) …..* P(wordn / class A).
   b) Find P(B / W) = P(B) * P(word1 / class B) * P(word2/ class B) …..* P(wordn / class B).
7. Assign document to class that has higher probability.

FIGURE 4.2: Pseudocode of Naive-Bayes

```
{
  "nerList": [
    {
      "location": "Vashi"
    },
    {
      "location": "MUMBAI"
    },
    {
      "location": "Sanpada"
    },
    {
      "location": "Sanpada"
    },
    {
      "location": "Panvel"
    },
    {
      "date": "May 6,2014"
    },
    {
      "date": "Tuesday"
    }
  ]
}
```


FIGURE 4.3: NER Output

Mande, was walking towards a vegetable outlet around 9.40am, when a bike came close to her and the pillion rider snatched her mangalsutra. A robbery case has been registered at Kharghar police station."

Output: The NER generates the following output:

Output from Coreference Resolution:

Then they reach for pattern identification where they had to identify trends and patterns in crime. They used Apriori algorithm for finding crimes and determined association rules to get general trends in the database. They extracted crime pattern for a particular place. Corresponding to each location they took the attributes of that place like VIP presence, weather attributes, area sensitivity, notable event, presence of criminal groups etc. After taking a sample list of 100 news for a place, they applied Apriori algorithm which mined the frequent crime patterns for a place. So if there is a pattern in which crime occurred then if that pattern occur again for a place then

```
["The victim" -> "Shakuntala Mande"
 "her mangalsutra" -> "a gold mangalsutra"
 "Kharghar" -> "Kharghar on Friday"
 "her" -> "Shakuntala Mande"
 "the pillion rider" -> "A pillion bike rider"]
```

FIGURE 4.4: Output of coreference resolution

there is probability for crime occurrence in that place. These papers were helpful for better knowledge gain of crime related works. Again, there is a mention of heat map for better data visualization because of the following reason:

- Numeric and category-based color images.
- Gradient color range.
- Analyze only the data we want.
- Out of range data is automatically discarded.

We can detect the most criminally active places by analyzing crime occurrence frequency and by the analysis, we can say that places like airport, temples, bus station, railway stations, bank, casino, jewelry shops, bar, ATM, airport, bus station, highways etc has high rate of crime. So, people around that places should be aware about crimes more.

Chapter 5

Methodology

Our work is based on crime data analysis where we can get crime patterns and predict time accurately. Data collection, data preprocessing, feature enriching, applying different machine learning methods are the steps of works we have done. We process our dataset after analyzing important thesis papers. We add column of frequency of each crime type based on a specific hour of day which helps people more for better understand the crimes and take necessary measures. Our working procedure is as follows:

5.1 Data Collection

From previous works on this topic, data was collected mostly from individuals living in Sylhet district which was random and less in number to get a pattern of crime occurrence. Again, like the developed countries, where each police department has crime data digitalization process, our country still doesn't have the system on that detailing. There is some data available which covers only number of different crimes occurred which is not efficient for pattern generation and predicting crime. After studying a good number of thesis papers on worldwide crime datasets, we understood the characteristic of crime occurrence more deeply and got the vision of which attributes are more likely to cause crime. We travelled on different online dataset portals where we could get our desired dataset and then we collected the dataset.

5.1.1 Features

Each entry in our training data set is about a specific crime, and contains the following information:

- • occurrenceyear
- • occurrencemonth
- • occurrenceday
- • occurrencedayofyear
- • occurrencedayofweek
- • occurrencehour
- • MCI
- • Division
- • Neighbourhood
- • Premisetype

5.1.2 Preprocessing the data

We preprocessed our data before applying machine learning algorithm on it. The steps were:

- • Feature dropping: We dropped those features which were unimportant and had limited significance in a real-world scenario of which type of crime was occurred. The features we dropped are: Latitude, Longitude, Hood-ID, Offence Description. We dropped Reported date, Reported Time, Reported Month as these were only necessary for police department and not for crime prediction analysis.
- • We selected our desired feature list and factorize each columns of the dataset which encode the object as an enumerated type or categorical variable.

- • There are several categorical features in the dataset and we replaced that with OneHotEncoding which outputs sparse matrix where each column corresponds to one possible value of one feature.
- • We kept only the crimes occurred in recent times for better prediction accuracy of present and future crimes.

5.1.3 Feature Enrichment

Suppose, a user wants to know, on a specific time-range of a specific day of the week, the occurrence frequency of a specific crime. For this to be happen, we add an extra feature named "occurrencefrequency" whose values were calculated from the two columns from existing csv file ("occurrencedayofweek", "occurrencehour") which gives the frequency of occurrence on a specific hour of a specific day. We think this improves our dataset for the classification purpose and by having it in the 'Target' value users should be benefited more.

- At first, we selected the two columns, 'occurrencedayofweek' and 'occurrencehour'.
- • Then we checked for each unique record of the combination of two columns.
- • We counted the total number of each unique record by initializing arrays for each days of the week and increase the value of arrays for the specific hour of a day.
- • We add a column to the dataset called 'occurrencefreq' where for every specific hour of a specific day, we add the value of the specific array.
- • By this, we get the occurrence frequency of that specific time.

5.1.4 Pseudo-code

Initialize seven arrays for seven days of the week $for i := 0 to (total_rows - 1)$

begin

$if dataframe['occurrencedayofweek'] is one of the seven days of the week$

$do for j := 0 to 24$

$begin$

$if dataframe[`occurrencehour']isanyof 0-24(j)$

$doincrementvalueofarray[j]by 1$

end

end

$fori := 0to(total_rows - 1)$

$begin$

$if dataframe[`occurrencedayofweek']isoneofthesevendaysoftheweek$

$doinsertdataframe[`occurrencefrequency'] = array[dataframe[`occurrencehour']]end$

5.2 Apply Methods

After preprocessing the dataset, we applied various machine learning approaches on them to measure accuracy, confusion matrix and classification report. We make the crime name column which is "MCI" in the csv file, the target variable and the other columns as the feature variable. Below are the algorithms we used with the scores mentioned above:

5.2.1 Random Forest Classifier

Random Forests is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data. It uses randomness when making split decision to avoid overfitting on the training data. Accuracy: 75.87 percent

Confusion Matrix:

TABLE 5.1: Confusion Matrix for Random Forest Classifier

[	392	449	7	3	59
	211	1995	68	3	130
	26	282	159	2	45
	19	67	3	4	13
	91	257	22	4	2987]

Classification Report:

After Encoding:

TABLE 5.2: Classification Report of Random Forest Classifier

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.53	0.43	0.48	910
Assault	0.65	0.83	0.73	2407
Robbery	0.61	0.31	0.41	514
Theft Over	0.25	0.04	0.07	106
Auto Theft	0.92	0.89	0.91	3361
avg / total	0.75	0.76	0.75	7298

Accuracy: 78.02 percent

Confusion Matrix:

TABLE 5.3: After Encoding, Confusion Matrix

[	434	424	7	1	44
	173	2047	55	1	131
	25	275	159	1	54
	20	68	1	3	14
	55	239	14	2	3051]

Classification Report:

TABLE 5.4: After Encoding, Classifier Report

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.61	0.48	0.54	910
Assault	0.67	0.85	0.75	2407
Robbery	0.67	0.31	0.42	514
Theft Over	0.38	0.03	0.05	106
Auto Theft	0.93	0.91	0.92	3361
avg / total	0.78	0.78	0.77	7298

5.2.2 Gradient Boosting Classifier

Gradient boosting technique has a noticeable attention for its prediction speed and accuracy, especially with large and complex data. It performs bad when encoding

the columns, so it is not shown here. Accuracy: 68.65 percent

Confusion Matrix:

TABLE 5.5: Confusion Matrix for Gradient Boosting Classifier

[	121	691	0	0	98
	82	2008	0	0	317
	3	375	0	0	136
	7	87	0	0	12
	106	374	0	0	2881]

Classification Report:

TABLE 5.6: Classifier report of Gradient Boosting Classifier

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.38	0.13	0.2	910
Assault	0.57	0.83	0.68	2407
Robbery	0	0	0	514
Theft Over	0	0	0	106
Auto Theft	0.84	0.86	0.85	3361
avg / total	0.62	0.69	0.64	7298

5.2.3 Decision tree

A decision tree is a decision support tool that uses a tree-like graph of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Dividing efficiently based on maximum information gain is key to decision tree classifier.

Accuracy: 70.35 percent

Confusion Matrix:

TABLE 5.7: Confusion Matrix of Decision Tree

[	398	318	59	31	104
	342	1612	208	49	196
	45	204	181	14	70
	24	44	10	11	17
	93	231	77	28	2932]

Classification Report:

TABLE 5.8: Classifier Report of Decision Tree

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.44	0.44	0.44	910
Assault	0.67	0.67	0.67	2407
Robbery	0.34	0.35	0.35	514
Theft Over	0.08	0.1	0.09	106
Auto Theft	0.88	0.87	0.88	3361
avg / total	0.71	0.7	0.71	7298

After Encoding:

Accuracy: 72.46 Percent

Confusion Matrix:

TABLE 5.9: After Encoding, Confusion Matrix

[	341	348	32	19	80
	321	1691	152	46	197
	40	210	180	13	71
	23	57	4	8	14
	87	215	67	14	2978]

Classification Report:

TABLE 5.10: After Encoding, Classifier Report

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.48	0.47	0.48	910
Assault	0.67	0.7	0.69	2407
Robbery	0.41	0.35	0.38	514
Theft Over	0.08	0.08	0.08	106
Auto Theft	0.89	0.89	0.89	3361
avg / total	0.72	0.72	0.72	7298

5.2.4 Linear SVC Classifier

With a linear Support Vector Machine, it is possible to take a peek at the top features. A linear SVM creates a hyperplane that uses support vectors to maximize the distance between classes.

Accuracy: 50.27

Confusion Matrix:

TABLE 5.11: Confusion Matrix of SVM

[	179	709	1	0	21
	356	1966	5	0	80
	54	438	3	0	19
	12	90	1	0	3
	344	1491	5	0	1521]

Classification Report:

TABLE 5.12: Classifier Report of SVM

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.19	0.2	0.19	910
Assault	0.42	0.82	0.55	2407
Robbery	0.2	0.01	0.01	514
Theft Over	0	0	0	106
Auto Theft	0.93	0.45	0.61	3361
avg / total	0.71	0.5	0.49	7298

After Encoding:

Accuracy: 74.83

Confusion Matrix:

TABLE 5.13: After Encoding, Confusion Matrix

[	366	506	4	0	34
	193	2089	28	0	97
	33	396	34	0	51
	21	72	1	0	12
	89	291	9	0	2972]

Classification Report:

TABLE 5.14: After Encoding, Classifier Report

Type of Crimes	precision	recall	f1-score	support
Break and Enter	0.52	0.4	0.45	910
Assault	0.62	0.87	0.73	2407
Robbery	0.45	0.07	0.12	514
Theft Over	0	0	0	106
Auto Theft	0.94	0.88	0.91	3361
avg / total	0.73	0.75	0.72	7298

So, from the above methods, we can see clearly that Random Forest Classifier best fits the dataset and give us maximum accuracy score on all the terms.

5.2.5 Occurrence Frequency Prediction

On the above work, we get patterns for crime types. Now, we will get patterns for crime occurrence frequency. On the Target variable we keep the column 'occurrencefreq' and all the other columns are listed as feature columns. We use different algorithms for this as mentioned above. When pattern is generated, one can input the day of the week and hour of the day and he can have the knowledge about which crime occurs most on that time.

Chapter 6

Future Works

- • Criminal Profiling: Criminal profiling is the concept which helps the crime investigators to record the characteristics of criminals. To provide crime investigators with a social and psychological assessment of the offender, criminal profiling is necessary. For gaining maximum detailing of the criminals, we have to analyze the criminal backgrounds and criminal records. For the ease of work, we can use some visualization techniques to represent the criminal details in a human understandable form. For representing criminal data, we can use graph database like Neo4j. We can conclude certain attributes of criminals like name, hair-colour, eye-colour, nationality, blood group, age, marital status, whether member of any criminal groups etc.

- • After analyzing a lot, we understand that, crime prediction and pattern generation will be more accurate if the dataset has features like: Criminal Organization, Name of victim/offender, categories and sub-categories of crime occurrence, type of vehicle offender used, type of weapons offender used.

- • We had no real dataset to work with. But once a proper dataset is generated, then we should see a positive result of our research. So, for the future, the main job is collecting continuous crime data around the city of our country. For this, like the developed countries, there should be a digital system for data collection and maintenance.

- • The other thing that can help our data collection is to promote our app and encourage people to send crime data that they have in their knowledge.

- • Once the dataset is complete, then we can enlist more detail information into the app. Users can get knowledge about more detail information on different types of crime occurrence based on different types of methods. That will make the app our countries first android application for general safety awareness with its own Artificial Intelligence. We are eagerly looking forward to this day.

Chapter 7

Conclusion

Crimes are irresistible in the prospect of our country. But we can try to reduce it as much as we can for ourselves, for our country. For this, we have to fight with everything we have, we have to contribute from every possible aspect. And for a generation like this, using data science is almost the best possible way to do so. Data mining is being applied for the betterment of many other sectors like weather prediction, sports, agriculture etc. Now its high time we used it in fighting the crimes. We are approaching towards it one step at a time. At this time, one platform is ready to be used for it as our app is nothing but all about it. With time, when enough data will be collected, then all these analyses is expected to come in handy for predicting next crime spots and thus preventing the crime. So, we can say that success of our "Sachetan" app depends largely on the feedback of the users which we collect as crime reports. After analyzing the data properly, with the help of law enforcement authority, we hope to stop a lot of crimes around us.

Chapter 8

Reference

1. https://www.thedailystar.net/country/bangladesh-road-accidents-in-2018-over-2400-deaths-on-roads-report-1598827
2. https://www.dhakatribune.com/bangladesh/nation/2018/03/26/sust-student-stabbed-death-sylhet/
3. https://dailycrowdsource.com/training/crowdsourcing/what-is-crowdsourcing/
4. https://globalnews.ca/news/540110/app-aims-to-increase-safety-for-b-c-women-at-risk-from-domestic-violence/
5. http://www.gender-focus.com/2013/05/18/new-apps-tackle-dating-violence-street-harassment/
6. http://www.circleof6app.com/
7. http://safetipin.com/
8. www.onwatchoncampus.com/