IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

Prevention of Harassment of Women by Crime Detection, Analysis and Prediction

Bhavana M S[1*], Bindu B K [2], Bindushree K [3], Chethana D N[4], Kiran Mensinkai [5]

[1,2,3,4,5]Dept. of Computer Science ,East West Institute of technology, Vishveswaraya Technological University, Bangalore, India

Corresponding Author: bhavanams@ewit.edu

DOI: https://doi.org/10.26438/ijcse/v7si15.15 | Available online at: www.ijcseonline.org

Abstract—Sexual harassment in public places is overwhelmingly experienced by women and girls. Sexual harassment is, in fact, the most common form of violence against women and girls and that young women are particularly targeted. Sexual harassment has significant and widespread impacts, both on individuals as well as on society. Sexual harassment in public reduces women and girls' freedom to enjoy public life, and can negatively affect feelings of safety, bodily autonomy and mental health.This project proposes a data-driven method to analyze crime data and behavioral patterns using machine learning algorithms and thus predict emerging crime hotspots for additional police attention.Each community has different crime trends in different areas. These trends are analyzed using machine learning principles which help to predict how crimes against women have significantly increased in various areas of a community. It also helps in rapid visualization and identification of communities which are densely affected with crimes. This approach proves to be quite effective and can also be used for analyzing national crime scenario.

Keywords-K-means Clustering, Random Forest, Google maps GPS, stemming

I. INTRODUCTION

In recent years, acts of assault and violence against women are rising at a menacing rate. With escalation of female employees in industries and other sectors of the commercial market, it is now becoming a necessity for females to travel at late hours and visit distant and isolated locations as a part of their work regime.63 per cent of girls and young women aged 13–21 experience not feeling safe walking home alone, according to the Girls Attitudes Survey 2018. However, the exponential increase in assault, violence and attacks against women in the past few years, is posing a threat to the growth and development of women. Defense isn't the only measure that can suffice against this increasing abuse. A security solution that creates safer environment for women must be devised.

In order to implement this project the data of previous incidents of harassment is collected. The data will then be grouped into clusters and analysed. A data set is created based on this analyzed data. After this in order to determine the crime rate and to predict the crime prone areas various machine learning tools are used.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Tools like clustering, random forests, classification and prediction models are used to establish a working model which learns from the data set. The working modelmakes predictions based on the thorough examination of the data set which may be yes or no, or true or false or a statistical value. Thus the crime rate is predicted and whether a location is dangerous or not is determined. In the last stage the outcome is alerted to the users to take precautions not to venture into the crime scene defenseless.

II. RELATED WORK

Newspaper articles are crawled using a focused crawler and they are classified using a SVM based classifier. Required entities are extracted from classified crime articles and duplicate detection is performed. By using pre-processed data, crime analysis operations are performed in [1] Though the model is effective in its prediction, the training article collection was found to be highly unbalanced because there are large numbers of non-crime articles compare to crime articles when a particular sample from an article population is considered. Madhura Mahajan, KTV Reddy, Manita

Rajput [2] have researched on a location tracking system and a partial wearable that can provide a complete security solution and become a utility that eases the apprehension among women and their family members.Here the GPS module requires at least 4 satellites in its Line Of Sight (LOS) to give proper coordinate readings.

In conclusion, the existing methods provide us solutions to ongoing harassment issues. However, It is a difficult problem to identify the near-duplicate documents in the crime data and data classification is computationally expensive. For a wearable device it requires extra hardware components. In [4] the datasets are from the Internet public datasets. However, one-class SVM for predicting the hotspot crime of location is still slow and computationally expensive.
Our approach integrate two aspects:
(1)Easy harassment reporting through location coordinates.
(2) Predict the crime locations beforehand to prevent harassment

III. METHODOLOGY

The project is designed in a way for easy access for crime reporting. This project requires more Human interaction than completely relying on the electronic devices for the detection. A survey is conducted where the user provides the information about their previous encounters of malicious activities. The user is specified to give complete and accurate details with respect to the encounter which includes the user details, time and location of the incident by providing the accurate GPS location.The algorithm is applied to this data where the relationship is determined by the algorithm to create a model.

Clustering algorithm is used to group the similar data by sorting. Prediction algorithm is used to determine the crime rate by drafting the graphs based on the data provided. The outcome of the regression algorithm is a percentage. The classification algorithm is used to determine specific outcomes of true or false. We use this algorithm to determine if a location is safe or not.

A. Data Collection
Data is collected from the public online through HTML forms. The location of the crime is obtained using the GPS from Google Maps.

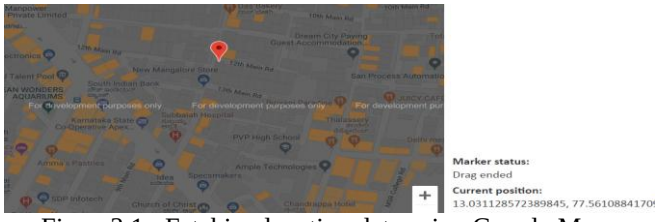

Figure3.1 : Fetching location data using Google Maps

All the data collected from the user is stored in the databse for further processing. The user can drag the marker to select the crime location and the coordinates of that marker is obtained and stored in the database for further processing.

Catcalling	13.087451	77.409454	2019-02-27	16:30:00.000000	A group of boys were cat calling a girl walking on...
Groping	12.999079	77.592537	2019-03-02	14:30:00.000000	A drunk guy behaving very badly in the public.
Groping	12.999079	77.592537	2019-01-08	22:15:00.000000	Near the metro station men were groping a girl pur...
Verbal Abuse	12.950178	77.584618	2019-01-04	19:20:00.000000	A group of men were verbally abusing girls who pas...
Stalking	12.999079	77.592537	2019-03-25	06:36:00.000000	stall
Catcalling	12.950289	77.585045	2019-01-07	21:42:00.000000	Few men were catcalling women who were alone
Verbal Abuse	12.999079	77.592537	2019-03-25	09:35:00.000000	Nothing
Verbal Abuse	13.086518	77.409859	2019-01-31	18:30:00.000000	A boy absuing a girl in the bus stop verbally.
Stalking	12.999079	77.592537	2019-01-17	21:00:00.000000	Boy following a girl in his bike and even blocking...
Catcalling	13.028468	77.539948	2019-01-20	20:30:00.000000	2 boys talking nonsense when girls pass by them.
Groping	13.060879	77.508179	2019-01-26	19:45:00.000000	A drunk guy behaving so badly with a street walker...
Other	12.981524	77.486816	2019-01-22	15:30:00.000000	A guy standing in his bare body

Figure3.2 : User report stored in the database

B. Data Analysis and Feature Selection

1. The data collected is pre-processed to remove irregularities and unwanted data
2. The data is refined by splitting the attributes like date into month and day
3. Data manipulation is done to convert the data to csv form

Algorithm 1 Data Conversion to csv form

START

Input: Data from the database

Output: Data in csv form

4. create a procedure for conversion
5. create a trigger for each row insertion such that the procedure is called when it is triggered
6. set the headers for csv file
7. SET @heading = 'SELECT ''' + right (@heading, len (@heading) -1) + ''' AS CSV' + CHAR(13)
8. Extract each column value and concatenate it in the csv file

9. SET @sqlstrs = 'SELECT CONCAT(' + @cols + ') CSV FROM ' + @tablename

10. seperate each value in the csv file by a coma (,)

END

Once the data has been converted to comma separated values or the csv form the feature selection is made from the data. All machine learning programs run on datasets which are in csv form.

```
50,chethana123@gmail.com,Catcalling,12.945382,77.504021,2019-01-16,21:36:00.000000,"Boys gang catcalling"
51,d.anilkumarkvdoni@gmail.com,"Verbal Abuse",12.981023,77.485687,2019-03-25,21:36:00.000000,"Child Explo
52,chethanadn97@gmail.com,Stalking,12.942789,77.505508,2019-01-21,21:39:00.000000,"Few men were stalking
53,chethanadn12121@gmail.com,Catcalling,12.941924,77.506195,2019-01-24,15:41:00.000000,"Men around usual
54,chethana123@gmail.com,"Verbal Abuse",12.941673,77.505409,2019-01-28,17:36:00.000000,"Men verbal abuse
55,d.anilkumarkvdoni@gmail.com,Stalking,15.130557,76.856224,2019-03-26,09:45:00.000000,"People Abusing"
56,chethanadn97@gmail.com,Other,12.941764,77.505333,2019-01-25,21:48:00.000000,"Some boys were showing su
57,chethanadn97@gmail.com,Catcalling,12.941736,77.505432,2019-01-31,19:05:00.000000,"A group of men catca
58,chethana123@hotmail.com,Catcalling,12.938174,77.509338,2019-02-04,12:09:00.000000,"Catcalling women"
59,chethanadn12121@gmail.com,Catcalling,12.937630,77.500816,2019-02-04,20:28:00.000000,"Catcalling situat
60,thejastm15@gmail.com,Stalking,12.973905,77.607643,2019-02-12,21:30:00.000000,"few mens were stalking a
61,renukasatish15@gmail.com,Groping,13.023291,77.588615,2019-01-01,20:25:00.000000,"People will abuse..."
62,renukasatish15@gmail.com,"Verbal Abuse",12.999079,77.592537,2019-02-08,21:30:00.000000,"Gang of boys w
girl"
63,meghanagowda958@gmail.com,Catcalling,12.999079,77.592537,2019-03-02,16:30:00.000000,"People showing ba
64,madhupriya885@gmail.com,Groping,12.999079,77.592537,2019-01-24,19:30:00.000000,"People are just worst
badly.."
65,meghanas405@gmail.com,"Verbal Abuse",12.999079,77.592537,2019-03-06,18:45:00.000000,"Worst people in s
them"
66,Anonymous@gmail.com,Catcalling,12.999079,77.592537,2019-02-23,20:00:00.000000,"A man was using very ab
local theatre. I have heard so many cases like this happeing near this theatre at night times.  "
67,soundarya8@gmail.com,"Verbal Abuse",13.349539,77.127975,2019-02-14,18:00:00.000000,"Verbal abuse on st
68,abhishekvh9@gmail.com,"Verbal Abuse",12.999079,77.592537,2018-07-26,18:39:00.000000,"People started ab
69,veereshhg@gmail.com,Catcalling,13.337322,77.701355,2019-02-11,17:00:00.000000,"Cat calling at nandi hi
```

Figure3.3 : Data in csv form

Feature Selection

i. Feature selection is done to build the model efficiently
ii. The attributes which are important to train the model are selected as features
iii. The attributes used for feature selection are lattitude, longitude, community area, month and hour

C. Data Clustering

A cluster refers to a collection of data points aggregated together because of certain similarities.The processed data is clustered before training the model. The data is clustered based on location points to analyze the crime rate. On clustering the density of the crime hotspots is measured to trian the model to predict the future hotspotsclusters. We use K-Means Clustering here.Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

We willl define a target number k, which refers to the number of centroids we need in the dataset. A centroid is the imaginary or in this case real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

K means Clustering

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

The data is clustered into location groups to analyze the crime rate of the locations.

With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). K-Means may produce higher clusters than hierarchical clustering. An instance can change cluster (move to another cluster) when the centroids are recomputed.

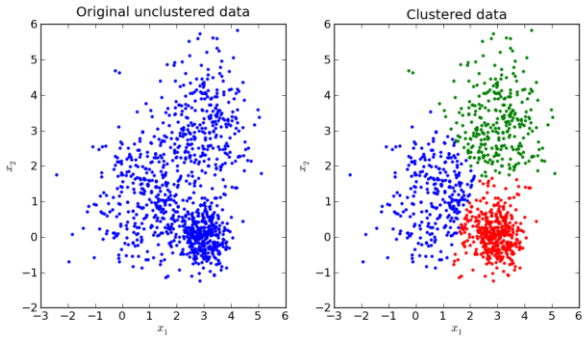

Figure3.4 : Clstering the data based on nearest locations

D. Train prediction model

The clustered data is trained and tested using machine learning algorithm such as the Random Forest Algorithm.

The model is trained to predict the possible future hotspots from the historical data already acquired combined with the new data that will be obtained from the users reporting harassment.The increase/decrease in the crime rate is monitored.

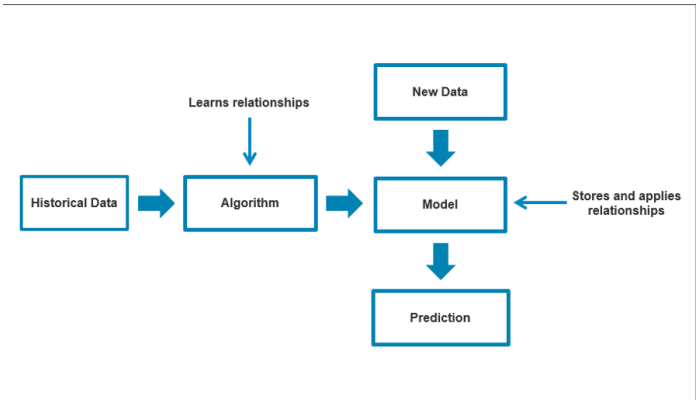

Figure3.5 : Training the prediction model

The prediction model is established by constructing a decision tree. Decision trees have three main parts: a root node, leaf nodes and branches. The root node is the starting point of the tree, and both root and leaf nodes contain questions or criteria to be answered. Branches are arrows connecting nodes, showing the flow from question to answer. Each node typically has two or more nodes extending from it. For example, if the question in the first node requires a "yes" or "no" answer, there will be one leaf node for a "yes" response, and another node for "no."

Once the tree is constructed the gain is calculated to make a decision in the feature.

Gain(T,X) = Entropy(T)—Entropy(T,X)

T = target variable

X = Feature to be split on

Entropy(T,X) = The entropy calculated after the data is split on feature X

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees.

IV. Result and Discussion

In this section, all the steps in each phase of the research methodology are investigated in detail. The project aims to predict crime hotspots and warn the users about it.

The proposed system is tested on the datasets containing information collected by the users through a survey. This research was carried out taking the information of 356 users and their reports on harassment. The obtained information was stored in excel sheets in csv form after processing.

Clustering of location data points is done using K-means clustering. One of the machine learning concept known as Random Forest is used for building the prediction model. It is used in this project for analyzing. For testing, the proposed method 256 user information were used.

From the experimental results we came to know that the user can successfully report the crime by providing the accurate location coordinates. Once the user reports the crime the data is stored in the database and analyzed. The feature selection is made to get accurate results. Once the prediction model predicts the results it is communicated to the users in the webpage. Based on the prediction model the crime rate is determined.

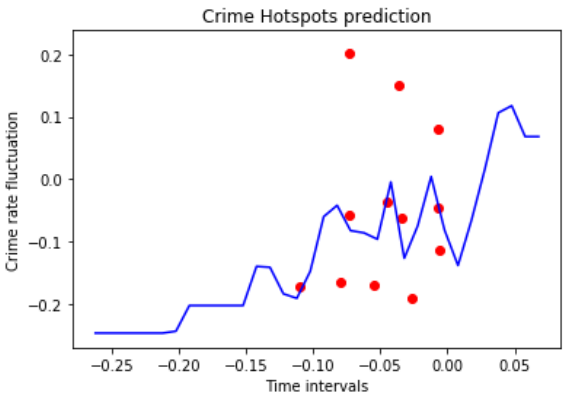

Figure4.1 : Crime rate analysis

The graph shows the crime rate fluctuation against the time intervals based on the predicted model.

V. Conclusion and Future Works

In this paper, The data collection is done through the online survey. The details of the crime are acquired along with the location and time. The location coordinates are obtained with the help of Google Maps GPS. The K-Means and Random Forest Algorithms give more accurate results than the previous systems. The project will help in prevention of harassment by monitring the crime rates.

Acknowledgment

Firstly, we express our sincere thanks to our guide Mr. Kiran M, Assisstant Professor, Department of CSE, EWIT and Dr. Arun Biradar, Head of Department, Computer science and engineering for their moral support. We express our sincere gratitude to our principal Dr. K Chennakeshavalu for his constant support and encouragement, we also thank all the faculties of East West Institute of Technology for their co-operation and support.

References

[1] Mehedee Hassan, Mohammad Zahidur Rahman, "CrimeNews Analysis: Location and Story Detection",2017 20th International Conference of Computerand Information Technology (ICCIT)

[2] R. Arulanandam, B. T. R. Savarimuthu, and M. A.Purvis, "Extracting crime information from online newspaper articles," in Proceedings of the Second Australasian Web Conference - Volume 155,AWC '14, (Darlinghurst, Australia, Australia), pp. 31–38, Australian Computer Society, Inc., 2014.

[3] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, "Crime analytics: Analysis of crimes through newspaper articles," in 2015 Moratuwa Engineering Research Conference (MERCon), pp. 277–282, April 2015.

[4] P. Chamikara, D. Yapa, R. Kodituwakku and J. Gunathilake, "SLSecureNet : intelligent policing using data mining techniques",

International Journal of Soft Computing and Engineering, vol. 2, no. 1, pp. 175-180, 2012.
[5] S. Adhikari and K. Bogahawatte, "Intelligent criminal identification system," in The 8th International Conference on Computer Science & Education, Colombo, Sri Lanka, 2013, pp. 633-638.
[6] M. Choi, "A selective sampling method for imbalanced data learning on support vector machines,", 2010.
[7] K.B.S. Al-Janabi, "A Proposed Framework for Analyzing Crime Data Set using Decision Tree and Simple K-Means Mining Algorithm," in Journal of Kufa for Mathematics and Computer, Vol. 1, No. 3, 2011, pp.8-24.
[8] K. Zhu and J. Zhang, "Predicting the potential locations of the next crime based on data mining," in International Journal of Digital Content Technology and Its Application, Vol. 6, No. 20, 2012, pp. 574-581.
[9] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 427-434.
[10] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns," in 2nd Asian Conference on Defence Technology (ACDT), Jan 2016, pp. 123–128.
[11] G. Yu, S. Shao, and B. Luo, "Mining crime data by using new similarity measure," in Second International Conference on Genetic and Evolutionary Computing, Sept 2008, pp. 389–392.
[12] Punetha, D.; Mehta, V. "Protection of the child/ elderly/ disabled/ pet by smart and intelligent GSM and GPS based automatic tracking and alert system" ,Advances in Computing,Communications and Informatics (ICACCI, 2014 InternationalConference on Year: 2014 pg2349 – 2354
[13] Bingbing Lu, Huaping Zhang, Bin Liu1, Zhonghua Zhao, "Research on UseR Identification Algorithm Based on Massive Multi-site VPN Log", 2017 17th IEEE International Conference on Communication Technology
[14] Ubon Thansatapornwatana, A Survey of Data Mining Techniques For Analyzing Crime Patterns Second Asian Conference onDefense Technology ACDT, IEEE, 2016, ISBN: 978-1-5090-2258-8/16
[15] http://r-statistics.co/Linear-Regression.html.
[16] Machine Learning Tool Kit [Online]. Available: https://github. com/yinlou/mltk
[17] Amazon.com, Inc. (2012) Form 10-K for 2011. Filing date: February 1, 2012. U.S. Securities and Exchange Commission,Washington,DC.
[18] Cattani K, Gilland W, Heese H, Swaminathan J (2006) Boiling frogs: Pricing strategies for a manufacturer adding a direct channel that competes with the traditional channel. Production Oper. Management 15(1):40–56.
[19] Forrester Research, Inc. (2014b) European online retail forecast: 2013 to 2018. Report, May 29. Forrester Research, Inc., Cambridge, MA.
[20] Gans N, van Ryzin G (1999) Dynamic vehicle dispatching: Optimal heavy traffic performance and practical insights. Oper. Res. 47(5):675–692.
[21] Kaplan A (1969) Stock rationing. Management Sci. 15(5):260–267.
[22] Lee H (1987) A multi-echelon inventory model for repairable items with emergency lateral transshipments. Management Sci. 33(10):1302–1316.

Authors Profile

Ms. Bhavana M S is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning.

Ms. Bindu B K is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning.

Ms. Bindushree K is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning.

Ms. Chethana D N is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning.

Mr. Kiran Mensinkai is working as Assisstant Professor in Department of CSE, EWIT. He has 7 years of teaching experience. His area of interest include Machine Learning, Internet of Things, Cloud Computing and Wireless Sensor Network.

