AN EXPLORATORY STUDY OF HARASSMENT USING PUBLICLY AVAILABLE DATA

*S Thangapriya

Assistant Professor, Department of Computer Science with Data Science, SRM Arts and Science College, Affiliated to the University of Madras, Kattankulathur, Chengalpattu-603203, priyasaravananrex@gmail.com

R Barath Ramesh

Assistant Professor, Department of Computer Science with Artificial Intelligence, SRM Arts and Science College, Affiliated to the University of Madras, Kattankulathur, Chengalpattu-603203, barathr110@gmail.com

Abstract

Harassment is one of the biggest issues in the real world; it remains a persistent problem, and many researchers are still conducting studies on it. Due to a lack of awareness, social pressure, and various other reasons, victims often struggle to find solutions. Many people silently suffer from this issue, even within their own families. Under time pressure or stress, some individuals may even commit such acts. This paper highlights some publicly available datasets that can be used by researchers to gain deeper insights into the issue. As the data contains sensitive information, care must be taken when handling it. With recent advancements in technology, such work can now be implemented more easily. However, identity protection is a major concern, as most people are reluctant to share their experiences publicly. This work primarily utilizes Kaggle repository provide large volumes of standardized, trustworthy data, enabling researchers to begin their work without spending time on data collection. Exploring such data is essential for fostering new innovations. This study marks an initial step toward addressing the issue more broadly, opening a pathway for researchers to design frameworks that combine data-driven AI models with secure blockchain-enabled infrastructures. Such integration not only advances the technological response to harassment but also establishes a foundation for future interdisciplinary research aimed at building safer, trust-based environments.

Keywords: sexual harassment, exploratory data analysis, crime against women

Introduction

Sexual harassment can be predicted using crime analysis datasets. In order to ensure that women and men can live peaceful, sustainable lives in society, such issues must be addressed and corrected. By identifying these problems, legal authorities can monitor and protect vulnerable areas more effectively. Nowadays, Exploratory Data Analysis (EDA) techniques in data science offer a better understanding of such complex issues. EDA can provide more insightful results compared to traditional methods from other domains. A deeper understanding of the data can be achieved through EDA. Most reported harassment-related

crimes include rape, kidnapping, domestic violence by family members, human trafficking, molestation, sexual harassment, and more. This is a major societal issue that involves handling fragile and highly sensitive data.

Methodology

For this analysis, the dataset contains 396 rows and 12 columns. The dataset name is "crimes-against-women-from-2011-2021-in-india". In total, the dataset includes data from 36 states. Table 1 shows that Step-by-Step Procedure for Exploratory Data Analysis (EDA) Using Crime Against Women Dataset (2011–2021)

Table 1: Procedure for EDA using Crime Data

Step	Description
Step 1	Import the libraries such as numpy, pandas, os, scipy.stats, seaborn, and matplotlib.pyplot.
Step 2	Import the dataset: **Crime Against Women from 2011–21 in India**.
Step 3	Preprocess the data to check for any **missing** or **null** values.
Step 4	Provide the **preprocessed dataset** as input for EDA (Exploratory Data Analysis).
Step 5	Use **Pseudocode 1** to view **state-wise sexual crimes** in a particular state.
Step 6	**Repeat** the process to visualize the crime data for **each state**.
Step 7	Generate **plots** using data science techniques such as **ggplot**, **scatter plots**, and **charts**.
Step 8	**End** the process.

Pseudocode 1 describes the step-by-step procedure to visualize crime trends against women across Indian states using EDA techniques. The pseudo code describes a process for analyzing crime data state by state. It begins by initializing a list of states to analyze. For each state in the list, the dataset is filtered to include only the data for that particular state. Any unnecessary columns, such as the state name, are removed to keep the data clean. The 'YEAR' column is then used as the x-axis, while selected crime types are used as the y-axis for plotting. A correlation is calculated between different crime types to understand how they are related, and this is displayed as a heatmap. The crime data for each state is then plotted with appropriate titles, legends, and labels. Finally, the plots are displayed to visually interpret the crime trends and patterns in that state.

Pseudo Code 1:State-wise Crime Analysis

Start

Initialize list of states

For each state in the list:

    Filter dataset for the current state

    Remove unwanted columns (e.g., STATES)

    Set 'YEAR' as x-axis

    Select relevant crime data as y-axis

    Calculate correlation between crime types

    Plot the correlation as a heatmap

    Plot the data for that state

    Add title, legend, and state label

    Show the plots

End

Result Analysis

The analysis of crimes against women revealed several important findings. The top 10 states with the highest number of assault cases were identified, with a pie chart illustrating that these states contribute significantly to the overall reported cases and shown in Figure 1.

Table 2:State-wise Distribution of K&A and DD Cases Against Women

State	K&A	DD	Total	Remarks
Uttar Pradesh	101,701	21,357	**123,058**	**Highest**
Bihar	65,137	13,568	78,705	
Assam	55,094	1,742	56,836	
Madhya Pradesh	50,457	5,834	56,291	
UTTAR PRADESH	32,155	19,258	51,413	
West Bengal	37,848	4,006	41,854	
Maharashtra	38,061	2,175	40,236	
Rajasthan	36,059	3,695	39,754	
Delhi UT	31,251	1,106	32,357	

Odisha	23,928	3,956	**27,884**	**Lowest**

The result shows which states have areas particularly affected the most by all types of crimes faced by women. These insights offer a clearer understanding of crime patterns against women across India and help in recognizing areas that may require stronger preventive measures and targeted interventions. Uttar Pradesh reports the highest total number of crimes against women with 123,058 cases. Odisha records the lowest with 27,884 cases.
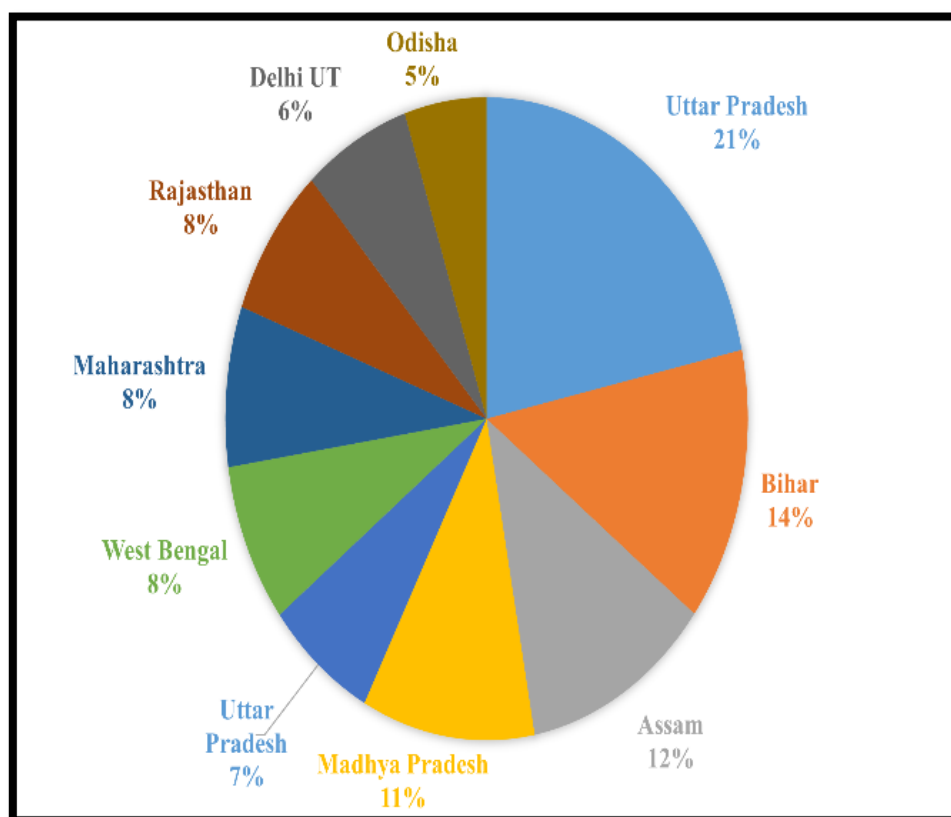
Figure 1 : State wise Sexual Harassment against Women

Conclusion

A sharp increase is observed after 2019, with total crimes against women peaking above 20,000 cases by 2021. Among the categories, cruelty by husband or his relatives consistently shows the highest number of cases, with a steep rise after 2019, reaching nearly 10,000 cases in 2021. Similarly, cases of sexual harassment and assault on modesty show a noticeable increase post-2019. In contrast, crimes like rape and kidnapping remain moderate and steady over the years, while dowry deaths, human trafficking, and immoral trafficking stay relatively low but stable. The overall trend highlights a growing concern for domestic violence and public safety in Tamil Nadu, possibly influenced by increased awareness, better reporting mechanisms, or changes in social behaviour and legal reforms during that period. In future, the work can be extended in many ways. We can use deep learning methods to identify

whether the data is correct or deeply faked. Next, the dataset can be updated using current problems—this work still uses old data. The data can be collected in real time, and immediate actions or switches can be integrated with CCTV systems with crime-prone zones. The work can be implemented in workplace environments to protect women more effectively. In this work, a single state is shown; in future, the data can be extended more using advanced technologies, and privacy should be considered, so blockchain can be used. Cryptographic algorithms can be used to protect the data.

References

[1] Sharma, K., and Gupta, A. "Analysis of Crimes Against Women in India Using Machine Learning Algorithms." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, no. 2, 2023, pp. 115–121.

[2] Das, R., and Saha, P. "A Deep Learning Approach for Crime Prediction Based on Real-Time Surveillance Data." Procedia Computer Science, vol. 215, 2023, pp. 1090–1097.

[3] Ramesh, N., and Bhandari, S. "Blockchain Technology for Ensuring Data Privacy in Surveillance Systems." Journal of Cybersecurity and Information Management, vol. 12, no. 1, 2024, pp. 34–42.

[4] Kumari, P., and Verma, M. "Crime Data Analysis of Women Safety Using Data Visualization Techniques." Journal of Emerging Technologies and Innovative Research, vol. 11, no. 3, 2024, pp. 23–30.

[5] Singh, R., and Yadav, V. "Predicting Gender-Based Violence Using Artificial Intelligence and Big Data." IEEE Access, vol. 11, 2023, pp. 66725–66734.

[6] Meena, S., and Rathi, A. "Smart Surveillance System for Women's Safety Using IoT and ML." International Journal of Innovative Research in Technology, vol. 10, no. 1, Jan. 2024, pp. 114–120.

[7] Ghosh, A., and Tripathi, S. "Securing Smart City Surveillance Data Using Cryptographic Blockchain Techniques." International Journal of Security and Networks, vol. 19, no. 2, 2024, pp.98–106.