Decoding Cybercrime: Uncovering Trends and Patterns in India Using Machine Learning

Harsh Mishra
Maharaja Agrasen College,
University of Delhi,
Delhi, India
harshm1804@gmail.com

Dr. Latesh Kanoujia
Maharaja Agrasen College,
University of Delhi,
Delhi, India

ABSTRACT

With the dawn of the digital era, India is also grappling with a dangerous threat in the shape of cybercrimes. As the population and institutions more and more rely on the digital platform, the volume and sophistication of offences committed online have sharply increased. Such crimes now reach out to people, institutions, and even essential digital infrastructure.

This study seeks to explore the development and character of cybercrime in India over the past few years. The research will examine data from the National Crime Records Bureau (NCRB) and other available datasets to examine offence trends, geographic differences, and victim characteristics. Using statistical analysis and machine learning approaches, the study will seek to reveal hidden patterns and inferences in the data that may not be apparent through basic observation.

The goal is to determine changes in the nature of cyber crimes that have been reported, determine areas of greatest vulnerability, and determine if there is any correlation between growth in the IT sector and increasing cyber attacks. Having made these connections, the study seeks to provide data that can lead to improved policy-making and enable law enforcement agencies to formulate more effective strategies against cybercrime and to improve India's digital security.

KEYWORDS

Cybercrime, Machine Learning, Trend Analysis, NCRB Data, Digital Security, Clustering, KMeans, Hierarchical

INTRODUCTION

As with the fast-changing digital technologies, India has evolved into a booming digital economy on the global arena. As great as this digital revolution has given much opportunity and benefit, it has also opened India to yet more cybercrimes. While the digital universe keeps expanding, cybercrime has become more of a growing problem, with major implications for citizens, organizations, and critical infrastructure.

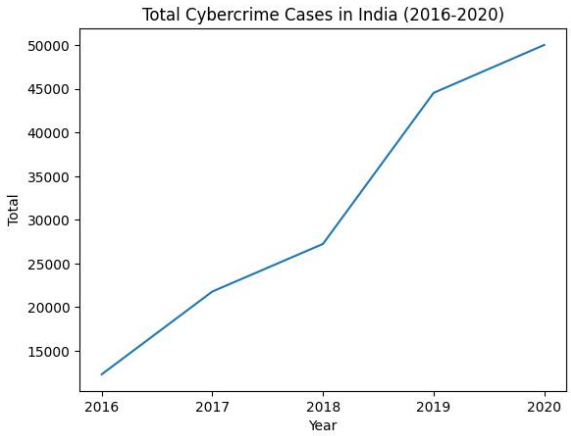

Figure 1: Total Cybercrime Cases reported in India from 2016-2020

Indian cybercrime is distinguished by a mixture of motives spanning from fraud and extortion to sexual exploitation and political manipulation. Though the threat is increasing, little available research and policy consider aggregate crime rates, hence a lack of knowledge regarding the motive-specificity of cybercrime. This paper aims at filling this knowledge gap by discussing various forms of cybercrime motivations which have been reported from the Indian states over the time span from 2016 to 2020.

The primary objectives of this study are as follows:

1. Trend Analysis of Motives of Cybercrime and Growth of the IT Industry: To trend analyze motives of cybercrime in the past across the nation, and study how growth in the IT industry is related to increasing cybercrime. The study will identify how cybercrime is taking place through different motives and geographies and reveal areas with the highest frequency of cybercrime.

2. Indian States Cluster based on Motives of Cybercrime: This entails using KMeans and Hierarchical Clustering to group Indian states according to their patterns of cybercrime motives. The aim is to capture patterns in the spatial distribution of cybercrimes and demarcation of similar cybercrime activity areas for intervention.

3. Policy Recommendations Based on Findings: Policy recommendations from the trend analysis and clustering results will be provided in the study. These will be localized measures for tackling most common cybercrimes in target states, along with enhanced cybersecurity infrastructure and public campaigns in areas of lower reporting.

By achieving these goals, this research will create evidence-based knowledge regarding the change in motives for cybercrime over time in India and how increased digital infrastructure affects cybercrime rates. The results will be used to create tailored policy recommendations that can enhance cybercrime prevention and digital security for Indian states.

BACKGROUND

As the scale and nature of cyber crimes continue to escalate, understanding not just how much cybercrime is happening but why it is happening — the reasons behind it — is essential to creating effective policy interventions.

In order to study this, the research makes use of publicly accessible data from credible government sources. Cybercrime statistics were pulled from data.gov.in, straight from the National Crime Records Bureau (NCRB). This data provides state-wise details of cases of reported cybercrime categorized based on certain motives such as fraud, revenge, sexual exploitation, extortion, etc., from 2016 to 2020. This level of detail allows for more in-depth observations into the kind of cybercrime experienced in different regions of the country.

In order to quantify the relationship between cybercrime and digital advancement, the analysis also includes growth rates on India's IT industry across Indian states. These have been drawn from annual reports and statistics published by the Press Information Bureau (PIB) and concurrent official press releases. These numbers are an indicator to gauge the diffusion of digital infrastructure and activity across each state so that the research can explore potential linkages between cybercrime and diffusion of the digital economy.

By combining motive-specific cybercrime information and IT growth metrics by state, this research seeks to present both quantitative observations about the patterns of cybercrime and qualitative descriptions that can inform region-level policy interventions.

METHODOLOGY

This study adopts a structured, data-driven approach to understand the distribution, evolution, and regional clustering of cybercrime across India. The methodology involves three major components: data collection and preprocessing, exploratory data analysis (EDA), and unsupervised clustering to identify motive-based patterns in cybercrime across Indian states.

1. Data Collection and Preprocessing

1.1 Cybercrime Data

The primary data source utilized in this study was downloaded from data.gov.in, published by the National Crime Records Bureau (NCRB). The data includes state-wise cybercrime figures from 2016-2020, classified based on motives. These range from financial motives such as fraud and extortion, to individual and ideological motives such as revenge, sexual exploitation, and political motives. Some columns, i.e., for terrorism, were published irregularly in each year and hence had to be brought under some broad "Others" column to provide a consistent output. Inconsistency in naming conventions, especially for Union Territories and composite districts like Dadra & Nagar Haveli and Daman & Diu, was rectified by renaming and consistent standardization. Data for each year was cleaned and standardized to a standard format to facilitate multi-year analysis.

1.2 IT Industry Data

State-wise IT industry revenue figures for the years between 2016 and 2020 were gathered from official reports and press releases by the Press Information Bureau (PIB) to check if there can be a

relationship between digital growth and cybercrime. These were used as proxy variables for comparison of all the states' web presence. Although the big states had constant revenue amounts, there were certain Union Territories and small states missing the same and these were excluded from correlation analysis so that the findings remain relevant. The IT dataset was remapped to year – state structure of cybercrime data and afterwards merged on year and state as keys.

1.3 Preparation of Final Dataset

The combined dataset consisted of annual motive-based crime data, total statewide cybercrime cases, and their respective IT revenue figures. Missing values were excluded or grouped into more general motive categories. The motive fields were z-score normalized to eliminate scale bias at clustering so that all motives would contribute equally towards model distance calculations.

2. Exploratory Data Analysis (EDA)

EDA was used in analyzing trends, determining regional trends, and discovering the organizational structure of cybercrime in India. Temporal analysis revealed shifts in the dominance of different motives over time, while geographic analysis depicted the top-ranked states having highest cumulative case ranks and having various motive trends. Correlation was established for the development in the IT sector vs. cumulative number of cybercrimes.

3. Indian States Clustering Based on Cybercrime Motives

3.1 Normalization

Columns based solely on motives were used for clustering, and z-score normalization was applied to all values. This helped prevent the model being too heavily influenced by high-frequency motives such as fraud.

3.2 KMeans Clustering

KMeans Clustering was applied to cluster states based on motive distributions in clusters. Elbow Method was used to identify the ideal number of clusters where the area of decreasing returns on the inertia plot defined k = 6. Each of the clusters was examined separately in order to discover the underlying attributes, i.e., high fraud frequency or rates of revenge and exploitation type cases.

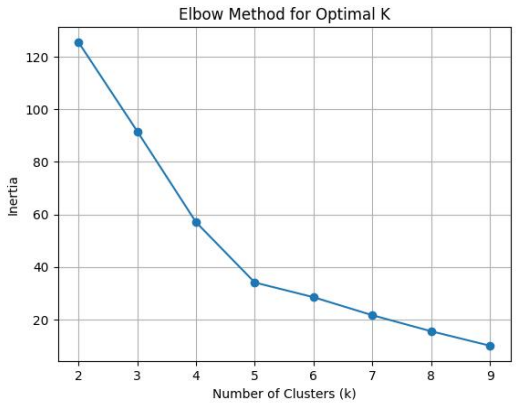

Figure 2: Elbow Method for Optimal K

3.3 Hierarchical Clustering

To check the stability of the KMeans outcome, Agglomerative Hierarchical Clustering using Ward's method was performed. A dendrogram was obtained to show hierarchical relationships between states. Partitioning the tree into six clusters produced outcomes very similar to those of KMeans, establishing the stability of motive-based clustering.

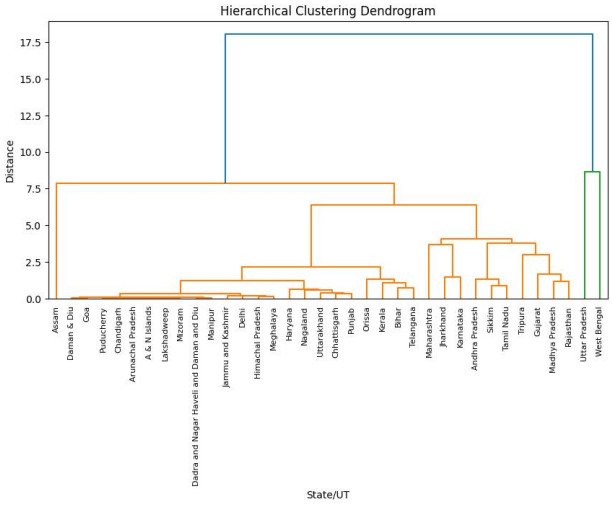

Figure 3: Hierarchical Clustering Dendrogram

3.4 Cluster Profiling and Visualization

Each cluster was characterized by averaging motive-wise values to determine patterns of behavior between states. The clusters were then mapped onto choropleth maps, permitting spatial interpretation of cybercrime trends and cross-comparison of states with similar profiles. Visualization assisted in reinforcing the thesis that geographically close states with similar cybercrime

structures are not necessarily so, due to underlying social or digital rather than locational reasons.

RESULTS

1. Trend Analysis of Cybercrime Motives

Data exploration analysis found changing trends in the motivation of cybercrimes in India from 2016 to 2020. Fraud topped all three categories every year, with the number of the same rising greatly year by year. Other motives like revenge, sexual exploitation, and extortion remained prime contributors but motives like piracy and political manipulation were rare.

Year-by-year motive ratio analysis reveals a year-by-year progression from interpersonal motives (e.g., disrepute or revenge) to monetarily motivated offenses. The trend is probably due to the growing digitalization of services and Internet-based financial transactions in India during this time.

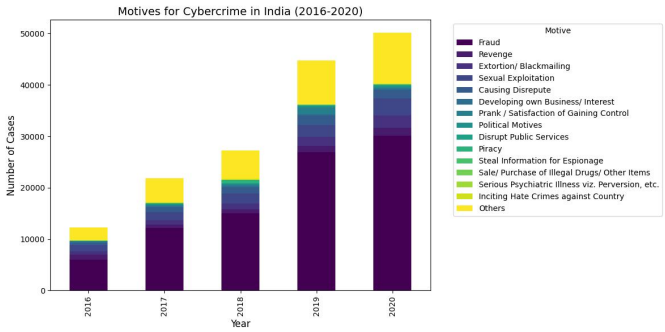

Figure 4: Motives for Cyber Crimes over years

The study also picked West Bengal as the highest-reporting state of cybercrimes over the five-year period. It was repeatedly higher-ranked than other top-reporting states such as Maharashtra, Karnataka, Jharkhand, and Uttar Pradesh. It was noteworthy that motive distribution in West Bengal was plentiful in quantity and wide in scope with high values in most motive categories.

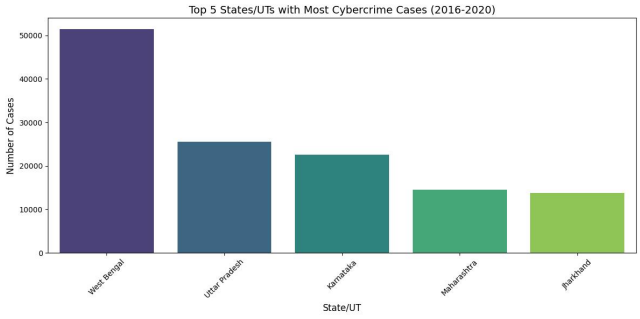

Figure 5: Top 5 States with most Cybercrime Cases

2. Growth of IT Sector and Number of Cybercrime

To examine whether there is any correlation between growth of IT sector and number of cybercrime, the study quantified the correlation between state-wise revenue of IT sector and number of cybercrime cases reported. It was found that the Pearson correlation coefficient was 0.73, which shows strong positive correlation between IT growth and cybercrime.

A linear model was used, which gave an R^2 of 0.528, implying that more than half of the variance in cybercrime cases was statistically explained by IT sector expansion. The IT revenue coefficient was found to be positive and statistically significant, validating the observation that greater digital exposure is related to greater cybercrime.

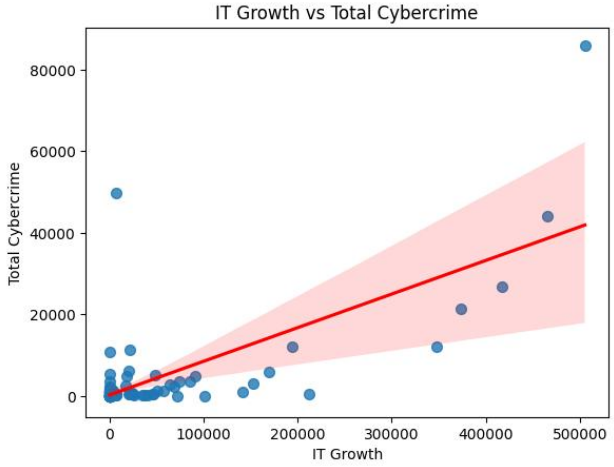

Figure 6: Regression between IT Growth and Total Cybercrime

OLS Regression Results

Dep. Variable:	Total	R-squared:	0.528
Model:	OLS	Adj. R-squared:	0.524
Method:	Least Squares	F-statistic:	132.2
Date:	Mon, 21 Apr 2025	Prob (F-statistic):	5.57e-21
Time:	08:25:45	Log-Likelihood:	-1233.4
No. Observations:	120	AIC:	2471.
Df Residuals:	118	BIC:	2476.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	274.3242	695.303	0.395	0.694	-1102.565	1651.213
IT_Growth	0.0824	0.007	11.496	0.000	0.068	0.097

Omnibus:	150.642	Durbin-Watson:	1.868
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4853.305
Skew:	4.501	Prob(JB):	0.00
Kurtosis:	32.826	Cond. No.	1.04e+05

Figure 7: Regression Results

3. Clustering of Indian States Based on Motive Profiles

Motive-based crime data were employed by the research to apply KMeans and Hierarchical Clustering to group states having analogous cybercrime conduct. Based on the Elbow Method and dendrogram evaluation, the optimal cluster number was six. Both algorithms provided extremely similar to one another clustering, supporting the stability of the identified clusters.

There was a different pattern of behavior in each of the clusters. One cluster, West Bengal alone, had high values for nearly all the motives. A second cluster consisted of states with a strong motive of revenge and sexual exploitation of interpersonal crime. A third cluster had states with low volume of cybercrime for all categories either reflecting low incidence or underreporting.

Assam was a clear cluster unto itself, away from other high-fraud or high-crime states. Its motive pattern was strongly skewed toward extortion, revenge, and sexual exploitation with comparatively lesser focus on fraud relative to other high-volume states. This implies that the cybercrime environment in Assam could be fueled more by social or interpersonal conflicts rather than economic motivations. Its clustering also supports the contention that local level social influences, not national, are involved in cybercrime activity.

Geographically distant states such as Jharkhand and Karnataka were clustered together, indicating common behavioral patterns rather than local proximity.

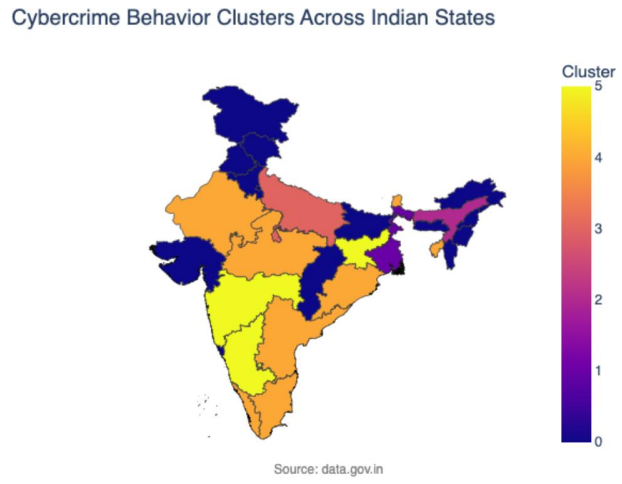

Figure 8: Hierarchical Clustering Results

DISCUSSION

The findings of this research emphasize the complexity and region-specific diversity of cybercrime in India. Though fraud remains the overriding top motive all over India, the study illustrates that cybercrime cannot be tackled as a one-size-fits-all issue. Rather, there are diverse patterns in different states in the incidence and category of cyber crimes registered, highlighting the necessity of finer understanding and more localized intervention.

The trend analysis shows a consistent trend towards socially motivated crimes such as revenge and disrepute to economically driven offenses such as fraud and extortion. This indicates that the integration of India into the digital economy has been on the rise, where online transactions and digital identities have presented other channels of economic exploitation. These high-volume cybercrime states also had more diversified motives, indicating that enhanced digital interaction is accompanied by diversified vulnerabilities.

Among the interesting results of the research is the high positive correlation (0.73) between growth in the IT industry and level of cybercrime, substantiated by a linear regression model which accounts for more than 50% variance. This correlation itself indicates that cybersecurity policy needs to adapt with digital expansion. As the states become increasingly digital, there are increasing opportunities for cybercrime and thus so are digital literacy and enforcement capacity.

The strongest evidence lies in the motive-based Indian state clustering. Six distinct clusters were discovered, which correspond to different patterns of behavior in cybercrime. Both KMeans and Hierarchical Clustering separately yielded very similar conclusions to cluster arrangements. This methodological agreement significantly confirms the validity of the findings. Where two fundamentally different clustering algorithms yield virtually identical groupings, it suggests that clusters are not artifacts of algorithmic choice but point to strong inherent structure in the data. Such consistency gives power to the behavioral interpretations and policy directions they suggest.

Every cluster revealed a different profile:

Cluster 1 included states and Union Territories like A&N Islands, Himachal Pradesh, Punjab, and others with a uniformly low rate of cybercrime for all motives. These have fewer incidents of cyber events or do not have infrastructure and awareness for reporting. The policy priority in these must be enhancing digital literacy,

enhancing awareness of online security, and developing support for reporting and managing cybercrime.

Cluster 2 was dominated by West Bengal itself, the country's leader in cybercrimes. It consistently had extremely high figures in almost all motive categories, indicative of high intensity of online activity and widespread vulnerabilities. This cluster needs a composite approach — consolidation of cyber policing, platform management, tracking of financial frauds, and mass-level public awareness — to address the extent and variety of cases put forward.

Cluster 3 had no other state than Uttar Pradesh, which also had a high number of cybercrime cases. It was distinct from West Bengal in that its motive distribution was more evenly distributed — not predictably too high in any one category, but high everywhere across the board. Such even spread is likely to signal systemic vulnerability rather than single peaks. Uttar Pradesh might benefit from institutional response capacity building, real-time monitoring systems, and statewide cybercrime sensitivities campaigns.

Cluster 4, composed solely of Assam, demonstrated an outstanding concentration in socially inspired cybercrimes — mainly revenge, sexual exploitation, and extortion — with comparatively lower emphasis on fraud. This is a strange pattern that reveals cybercrime in Assam may be socially oriented rather than money-motivated. The initiative here should be accorded top priority to digital gender safety, anti-harassment laws, and sensitization campaigns especially designed to address interpersonal digital abuse.

Cluster 5 related states like Jharkhand, Karnataka, and Maharashtra. These technologically advanced states had robust cybercrime reports with a definite inclination towards financial motives like fraud and extortion. The similarity of motive profiles between geographically disparate states confirms the hypothesis that user behavior and digital exposure, and not geography, drive cybercrime risk. Investment in fraud detection infrastructure, digital forensics departments, and cooperative models among the financial institutions and law enforcement agencies must be made in these states.

Cluster 6 comprised the states of Andhra Pradesh, Tamil Nadu, and Madhya Pradesh. These medium-high volume states saw a mix of economic and social cybercrime motives. Although less extreme than singleton clusters, increasing heterogeneity in motives here is an indicator of increasing levels of cyber threats. Preventive action (like cybersecurity scanning of organizations) needs to be matched in these areas by broader digital hygiene campaigns and victim compensation programs.

The visibility of three singleton clusters — West Bengal, Uttar Pradesh, and Assam — is most prominent. All three states were distinguished from the others by their unique cybercrime profiles: anomalously high volume and variety (West Bengal), uniformly high across categories (Uttar Pradesh), and motive-oriented concentration in interpersonal crime (Assam). Their single-state exceptions highlight state-specific intervention paradigms. Hyper-generalized, national cybercrime policies would be unhelpful in tackling such specially ordered challenges.

This research also verifies that geographic proximity is not a highly effective estimator of cybercrime similarity. Geographically remote states like Jharkhand and Karnataka were grouped together, whereas contiguous states like Assam and West Bengal were distributed across different clusters. This implies that cybercrime activity is influenced more by social, technological, and institutional forces rather than regional or cultural proximity.

Overall, the research shows the practical application of motive-oriented behavior examination of cybercrime. By determining hidden patterns and categorizing states according to similar digital crime patterns, it provides a motive-based framework for policy in regions. The information once again points to the importance of solid data reporting mechanisms and regular surveillance, particularly in states whose motive diversity is changing at a greater pace.

Finally, India's cybercrime response must transition away from a reactive one-for-all approach to a proactive motive-based and regionally focused strategy. The following section responds to these lessons with specific policy proposals intended to respond to the behavior realities uncovered by the data.

POLICY SUGGESTION

The results of this research indicate that Indian cybercrime is not homogenous — it is vastly heterogeneous in motive, size, and regional activity. Drawing on the six behavioral clusters extracted through unsupervised learning, along with IT sector growth cybercrime's high correlation, this section suggests a set of motive-sensitive, focused, and region-sensitive policy interventions.

Cluster 1 – Low Volume / Potentially Underreported (e.g., Punjab, Himachal Pradesh): Low reported cybercrime across these states for all reasons could reflect lower digital activity or underreporting.

Recommendations:

• Conduct awareness drives for reporting and recognizing cybercrime

• Educate local law enforcement officials on the handling of digital evidence

• Establish low-cost cybercrime reporting websites in local languages

Cluster 2 – High Volume and Motive Diversity (WestBengal): West Bengal registered very high values across all motives, indicating underlying systemic weaknesses.

Recommendations:

• Strengthen security of public and private sector digital infrastructure

• Increase cybercrime investigation units with forensic and multilingual capacity

• Collaborate with technology platforms for expedited takedown processes and content moderation

Cluster 3 – Uniformly High Cybercrime (Uttar Pradesh): Uttar Pradesh registered high values across all motives except child exploitation, indicating wide digital risk.

Recommendations:

• Establish district-level integrated cyber response teams

• Establish region-wide public-private partnerships for fraud tracking

• Incorporate digital safety education in state curriculum at school level

Cluster 4 – Socially Driven Motives (Assam): Cybercrime in Assam was focused on motives such as revenge, exploitation, and harassment.

Recommendations:

• Establish helplines and redressal platforms dedicated to gender-based and interpersonal digital abuse

• Work with NGOs and schools for cyberbullying and consent awareness workshops

• Add police training to deal with sensitive child or personal exploitation cases

Cluster 5 – Financial Crime Focus (Jharkhand, Karnataka, Maharashtra):These states exhibited trends characteristic of digitally mature economies with the emphasis on fraud and extortion.

Recommendations:

• Making regular cybersecurity auditing compulsory among startups and IT-enabled services providers

• Encouraging quicker red-flagging of suspicious behavior by banks and payments systems to bring it in line with the law enforcers

• To familiarize the police agencies with monitoring cryptocurrency and online financial frauds

Cluster 6 – Mixed-Motive, Mid-Volume States (states like Tamil Nadu, Andhra Pradesh):This cluster had an equal proportion of financial and social rewards.

Recommendations:

• Establish state cyber training academies for the police

• Launch modular awareness courses for SMEs and educational institutions

• Pilot specialized digital safety initiatives in rural and peri-urban regions

CONCLUSION

As India speeds up digital development, it is more and more important to understand the trend of cybercrime—less in quantity but by motive propelling it. It is a reflective data-driven analysis of motive-based cybercrime trends across Indian states between 2016 and 2020. Through exploratory data analysis, correlation modeling, and unsupervised clustering combined, it sheds light on meaningful behavior patterns of utmost interest to law enforcement and policy-making.

Findings indicate that India's cybercrime is not spread evenly. The reasons are largely dissimilar between places, and there are dissimilar profiles to some states such as West Bengal, Uttar Pradesh, and Assam that warrant interventions of an intended nature. The strong positive relationship between IT sector growth

and cybercrimes underscores the fact that investment needs to be contemporaneous in order to match evolving cybersecurity arrangements in conjunction with developing digital infrastructure.

Clustering by motive profiles, both under KMeans and Hierarchical Clustering, produced six stable and significant behavioral clusters. Consistency between the two clusterings adds strength to the stability of these groupings as a solid framework for region-specific cybercrime response policies. Surprisingly, the study is able to show that being close geographically does not predict similarity in behavior—showing that cybercrime is influenced more heavily by local digital routines, socio-economic composition, and institutional capacity.

In short, the study calls for a move away from boilerplate policy templates to motive-driven, behavior-sensitive, and location-specific interventions. It indicates that cybercrime investigation needs to do more than simply tally cases—it needs to look at the why, not merely the how many

Although the research is constrained by available short time series data and shortages of IT industry statistics in certain regions, it provides opportunities for investigating causality, motive categorization of more depth, and real-time prediction of online threats.

Finally, this paper adds to literature that does not only focus on cybercrime as a technical matter, but rather as an economic, social, and behavioral one—actually one whose very complexity can be addressed meaningfully only when it is conceptualized and dealt with.

LIMITATIONS

This research is limited by the continuity and availability of public data. The cybercrime dataset is five years in length, which limits trend analysis over extended periods and diminishes the efficacy of time series models such as Granger causality. Certain motive categories were reported sporadically or not at all between years and thus had to be grouped or excluded. Data on the IT sector was missing for the majority of Union Territories, further limiting correlation analysis. The research also assumes incidence equals reported cases, potentially not true in low-awareness or underreported areas. Notwithstanding these limitations, the results offer valuable behavioral information on India's cybercrime trend.

ACKNOWLEDGMENTS

I would like to express sincere thanks to the National Crime Records Bureau (NCRB) and the Press Information Bureau (PIB) for their freely available data which made it possible for me to carry out this study. I would also like to thank my guides or faculty members of Maharaja Agrasen College for all their help and encouragement that they provided while doing the research work. Special thanks to Dr. Latesh Kanoujia for their guidance and helpful suggestions.

Finally, I want to thank my colleagues and friends who supported me throughout late research and writing nights.

REFERENCES

[1] U. Kumaran, S. G, A. Narendran, V. P. Meena, S. Kumar Gupta and A. Appaji, "Understanding Cybercrime: A Study of Malware Analysis, Phishing Attacks, and Sentiment Trends," 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2025, pp. 1-9, doi: 10.1109/IITCEE64140.2025.10915529.

[2] Pandey, Hemakshi, et al. "Ensem_SLDR: classification of cybercrime using ensemble learning technique." International Journal of Computer Network and Information Security 14.1 (2022): 81.

[3] P. Datta, S. N. Panda, S. Tanwar and R. K. Kaushal, "A Technical Review Report on Cyber Crimes in India," 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2020, pp. 269-275, doi: 10.1109/ESCI48226.2020.9167567.