OPEN ACCESS Freely available online

International Journal of Advancements in Technology

Research Article

Multilayered Map Reduce Framework to Link Cybernetic Vulnerabilities and Cybernetic Laws from E-News Articles

Sudhandradevi P*, Bhuvaneswari V

Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu, India

ABSTRACT

The growth of technology evaluation and the influence of smart gazettes, which have a very complex structure, the amount of data in an organization, E-commerce, and ERP explode. When data is processed as described, it becomes the engine of every individual. According to projections from 2025, social media, I_oT, streaming data, and geodata will generate 80% of unstructured data, and there will be 4.8 billion tech enthusiasts. The most popular social media trend allows users to access publicly available data. Hackers are highly qualified in both the web space and the dark web, and the rise of complexity and digitization of this public access will cause loopholes in legislation. The major goal of this study is to gather information about the cyber vulnerability of electronic news. Data collection, text standardization, and feature extraction were all part of the initial step. In the second step, Map reduce was used to obtain demographic insights using a multi-layered categorization strategy. Cybercrime is classified using a classifier technique, and the model has a 53 percent accuracy rate. Phishing is a result of cyber weaknesses, and it has been discovered in a higher number of metropolitan cities. Men, rather than women, make up the majority of crime victims. Individuals should be made aware of secure access to websites and media, according to the findings of the study. People should be aware of cyber vulnerabilities, as well as cyber laws enacted under the IPC, the IT Act 2000, and CERT-in.

Keywords: Cyber vulnerability; CERT-In; DT classifier; Indian penal code; Map reduce; Social media; Text mormalization

INTRODUCTION

The Internet has reduced the entire planet to a single little shell. The use of digital surfaces has increased as a result of technological advancements, digital repositories, and social media. The world has entered the big data era; data is rapidly accumulating, and it is estimated that by 2025, the total volume of big data will have increased from 4.4 zettabytes to 44 zettabytes, or 44 trillion gigabytes. The volume of data has grown at a faster rate from many sources, rendering traditional processing systems ineffective. Intrusion detection, according to helps users secures their systems in cyberspace by alerting them [1]. For security analysis, big data will analyze heterogeneous data and a large volume of data, as well as block network traffic. According to the author, big data uses frameworks like Hadoop and Spark to handle the problem of storing and manipulating heterogeneous data [1,2].

Social media is a type of computer-based technology that allows people to share information and ideas in a virtual setting. It is intended to let users to instantly exchange content via smart devices, such as personal data, images, videos, documents, blogging, social games, and social networks. According to projections, the number of social media users in the United States will reach 257 million by 2023. Burnap et al. [3], used machine learning categories to describe cybernetic hate speech on Twitter. Ethnicity, religion, and more generic responses are used to categorize textual data. For anticipating the anticipated spread of cyber hate speech on Twitter, the data is trained and tested using classification algorithms to categories the optimal utilizing a combination of probabilistic, rule-based, geographical based classifiers [3].

According to a study on global tolls, cybercrime would cost $6 trillion in damages by 2021. The potential for cyberattacks should be used to organize the creation of anti-cybercrime policies. The processing of data in a secure manner is a vital aspect of digitalization. Data breaches have grown in importance as a type of cybercrime that can cause harm to people in either a trustworthy

Correspondence to: Sudhandradevi P, Department of Mechanical Engineering, Indian Institute of Technology Patna, Bihar, India, Tel: 9050999578; E-mail: psudhandradevi@gmail.com

Received: 18-Mar-2022, Manuscript No. IJOAT-22-16296; **Editor assigned:** 23-Mar-2022, Pre QC No. IJOAT-22-16296 (PQ); **Reviewed:** 06-Apr- 2022, QC No. IJOAT-22-16296; **Revised:** 11-Apr-2022, Manuscript No. IJOAT-22-16296 (R); **Published:** 18-Apr-2022. DOI:10.35248/0976-4860-22.13.176.

Citation: Sudhandradevi P, Bhuvaneswari V (2022) Multilayered Map Reduce Framework to Link Cybernetic Vulnerabilities and Cybernetic Laws from E-News Articles. Int J Adv Technol. 13:176.

or untrustworthy setting. Unauthorized access, harmful code, denial of service, and theft all contribute to computer crime. Heinous content, Internet fraud, and misconduct are examples of traditional crimes. Access to control technologies, cryptography, surveillance, assurance tools, India's legislative approach, and the IPC's core purpose of increasing security awareness in India are all technological solutions [4]. Briefs by traditional crime, online fraud, email spoofing, and virus/worm assault, Trojan virus, web hacking, hacking, and computer abuse are all examples of cybercrime [5]. Phishing, email spoofing, and phone phishing are the experimental analyses of the numerous cyberattacks provided in the research. A phone phishing investigation targets 50 employees and their bank account information. In this percentage, 32% of employees used electronic banking to share their credentials. Employees revealed their passwords in 16 percent of cases, while 52 percent refused to do so. According to the report, emails are hacked due to two concerns: they are misleading and consumers overestimate the security and confidentiality of emails [5].

Provided a security analysis solution for cyber security in the context of big data [6], Because of the continuous expansion of data, the type and frequencies of cyber-attacks are increasing at an exponential rate. The worldwide network is connected to a secure system in every country. In 2000, 45 million people were affected by cybercrime, according to a poll. Cybercrime includes spamming, botnets, Denial of Service (DoS), phishing, malware, and website threats. As a result of big data, cyber attackers' hacking skills have improved. They split the data into passive and active users and implemented security analytics using BDA technology [6].

Users are completely ignorant of the occurrence of cybercrime. Preventing cybercrime and attacks is the responsibility of law enforcement agencies such as the Computer Emergency Readiness Team (CERT-In) and the Government of India. The Indian Penal Code (IPC) sections of the IT Act 2000 provide legal cyber punishment/penalty in India for certain cybercrime attacks. Cybercrime, issues, trends, and difficulties, as well as solutions, were the focus of Kandpal et al. [7]. According to polls, cybercrime increased by 57.1 percent in 2012.61.0 percent of IT Act offenders and 45.2 percent of IPC Sections offenders were between the ages of 18 and 30. According to CERT-in 78, the 308371 government and other websites were hacked between 2011 and 2013. Updating computers, choosing a secure password, protecting computers, being social media savvy, and securing a wireless network are all important skills to reduce cybercrime threats [8].

METHODS

A multi-layered technique was used to implement the architecture. Data are gathered in the first layer from various news sources and IPC Laws. Data were extracted and processed in a structured format in the second layer. The Crime corpus is created for storage in the third tier. Cybercrime is classified into demographic data such as gender, geographic attributes, cyber offences, and criminal legislation in the fourth layer. The study was released as an analytical report charts and statistics reports at the bottom layer. The critical reflections on big data security by aim to identify the risks and problems of data security as well as the social and economic benefits of big data [9]. They definitively define the dangers of big data as a loss of confidentiality, integrity, and availability. The problem will be remedied by improving human capital through know-how and inventions, relational capital through improved client relationships, and structural capital through management process modifications. Big data was primarily concerned with technological, sociological, and ethical ramifications, as well as moral judgments [9]. Depicts the multi-layer architecture of cybercrime vulnerability (Figure 1).

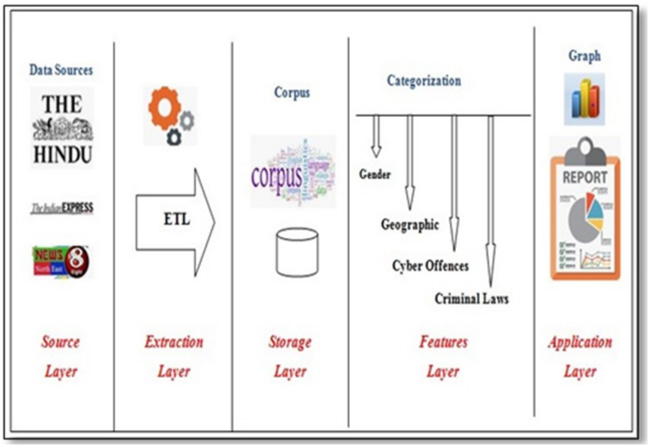


Figure 1: Multi-layered architecture.

Using the Map reduce technique, a multi-layered architecture was used to associate the labels with the document label. Data are divided into two categories: socioeconomic cybercrime and domain-based cybercrime, using a tiered method. Table 1 shows multi-layered data extraction [9].

Table 1: Data extraction: Multi-layered approach.

Data sources	Socioeconomic cybercrime	Domain-based cybercrime	
Article report	Cyber offence	Cyber article label	IPC crime label
	Geographic location	Geographic parameters	State
	Gender	Cyber victim	Cyber criminals
Crime report (IPC)	Crime laws	Cert-in, IPC/IT act section	Punishment

The objective of this research is to create a framework for Cyber Crime Attacks (CCA). Cybercrime attacks are classified using the cyber vulnerabilities found in articles and cyber legislation found in the IPC/IT sections (CVCL). Figure 2 describes the methodology of Cyber Crime Attacks (CCA) [10].

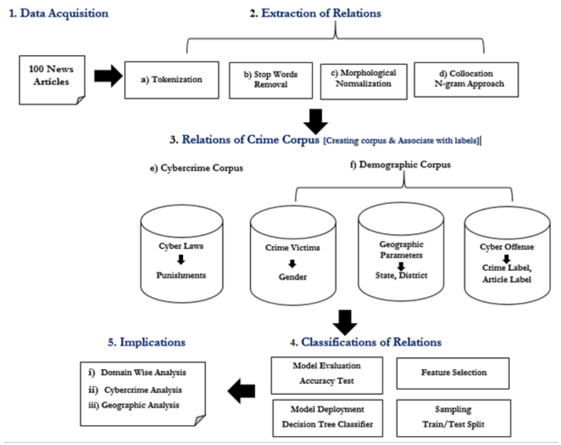

1. Data Acquisition
2. Extraction of Relations
100 News Articles
a) Tokenization
b) Stop Words Removal
c) Morphological Normalization
d) Collocation N-gram Approach
3. Relations of Crime Corpus [Creating corpus & Associate with labels]|
e) Cybercrime Corpus
f) Demographic Corpus
Cyber Laws
Punishments
Crime Victims
Gender
Geographic Parameters
State, District
Cyber Offense
Crime Label, Article Label
5. Implications
4. Classifications of Relations
i) Domain Wise Analysis
ii) Cybercrime Analysis
iii) Geographic Analysis
Model Evaluation Accuracy Test
Feature Selection
Model Deployment Decision Tree Classifier
Sampling Train/Test Split

Figure 2: Framework: Cyber Crime Attacks (CCA).

i) Techniques for building a database of cyber-vulnerabilities and laws.

ii) The multi-layered cyber vulnerability approach considers demographic aspects as well as crimes.

iii) Classify cybercrime by category and test the model using the classifier technique.

Data acquisition

Cybercrime news is gathered from e-news stories and presented systematically. Most of the news pieces are unstructured and clear.

E-news dataset

A manual data collection of 100 instances of newspaper diaries is performed. Article Labels, Headlines, Content, States, Year, Gender, and URL are the attributes of the dataset.

Cyber law dataset

The Cyber law dataset is a collection of cyber law sections and punishments obtained from India's cybercrime initiatives. The Cyber law dataset contains 78 cases, with Crime labels as the first attribute, followed by sections and sanctions. As per India rules, the crime section and punishments are CERTIn and the Indian Penalty Code (IPC) IT Act 2000. Table 2 shows the statistical reports for both the datasets.

Table 2: Data extraction: Multi-layered approach.

Reports	Feature extraction	E-news dataset	Cyber law dataset
Data acquisition report	Number of sources	2	2
	Sources	The Indian express, The Hindu	CERT-in, IPC/ IT Act 2000
Domain-level reports	Crime labels	Email spoofing, clone finger prints, facebook, debit card, social media, twitter, hacking, whatsapp, online crime, ransomware, harassment, internet fraud, data-wiping virus	Phishing, profile hacking, cyber bullying, web jacking, online scams, software piracy, cyber-stalking
	Number of instances	100	78
Dataset instance reports	Number of attributes	7	3
	Attributes	Article label, headlines, content, district, states, year, url	Crime label, IPC sections, punishment

Extraction of relations

In the context of the newspaper, punctuation marks, special characters, and images are determined. To provide a more user-friendly and distinct representation for higher-level modelling, which is also legible by humans. It's tokenized, with stop words eliminated, morphological normalization, and collocation. The text pre-processing phase is illustrated in Figure 3.

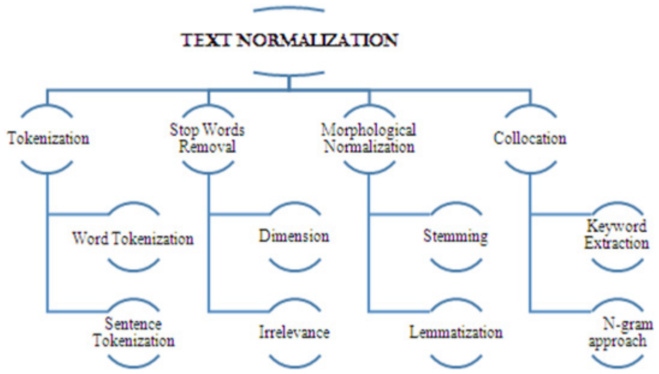
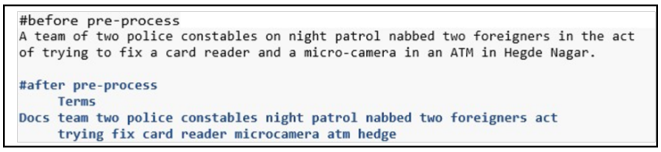


Figure 3: Extractions of relations.

The text is standardized in this part using NLP algorithms. To achieve a consistent structure, the first step is to remove the punctuation expressions and convert them to a lower case. They're listed below.

Tokenization

Tokenization divides the text into bits based on how it is used in the context. The "Sentence tokenization" or "Word Tokenization" techniques can be used to tokenize the context. Word tokenization is done with this task (Figure 4).



```
#before pre-process
A team of two police constables on night patrol nabbed two foreigners in the act
of trying to fix a card reader and a micro-camera in an ATM in Hegde Nagar.

#after pre-process
     Terms
Docs team two police constables night patrol nabbed two foreigners act
     trying fix card reader microcamera atm hedge
```

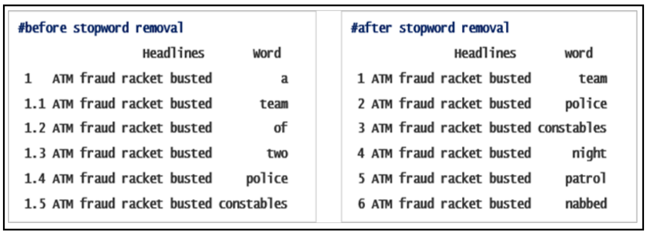
Figure 4: Pre-processing technique.

The document is cleaned using the pre-processing technique in step #1. Because the punctuation marks in the document take up unnecessary memory, they are eliminated. The documents have been changed to lower case for unique identification. Converting to a lower expression can sometimes modify the context's semantic meaning. Part-of-speech refines this issue by revealing the world's syntactic behavior.

Stop words removal

Stop word removal is used to locate data that isn't relevant. The information containing words has no sense and is also noisy. Dimension is utilized to greatly minimize the tokenization document.

In #2, unique space characters are used to tokenize words. Following tokenization, a few common terms arise, which add noise to the study and have little meaning. These words are known as stopwords or empty words. Removing stopwords also allows for a large reduction in the number of tokens in documents, as well as a reduction in the feature dimension is represented (Figure 5).



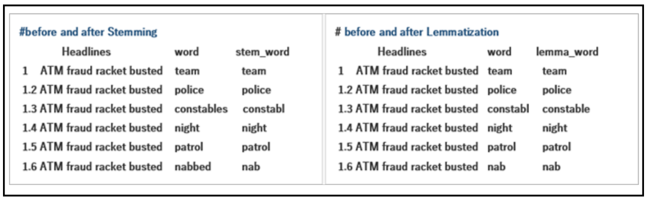
```
#before stopword removal
                  Headlines       Word
1   ATM fraud racket busted          a
1.1 ATM fraud racket busted       team
1.2 ATM fraud racket busted         of
1.3 ATM fraud racket busted        two
1.4 ATM fraud racket busted     police
1.5 ATM fraud racket busted constables
```

```
#after stopword removal
                Headlines       word
1 ATM fraud racket busted       team
2 ATM fraud racket busted     police
3 ATM fraud racket busted constables
4 ATM fraud racket busted      night
5 ATM fraud racket busted     patrol
6 ATM fraud racket busted     nabbed
```

Figure 5: Stop words or empty words.

Morphological normalization

The method is known as morphomes, and it seeks to find stem words. Stemming is a normalization approach that removes common suffixes from a term's output and returns the term's underlying word. Lemmatization is the correct application of the word's morphological analys is vocabulary (Figure 6).



```
#before and after Stemming
           Headlines        word        stem_word
1   ATM fraud racket busted team        team
1.2 ATM fraud racket busted police      police
1.3 ATM fraud racket busted constables  constabl
1.4 ATM fraud racket busted night       night
1.5 ATM fraud racket busted patrol      patrol
1.6 ATM fraud racket busted nabbed      nab
```

```
# before and after Lemmatization
           Headlines        word      lemma_word
1   ATM fraud racket busted team      team
1.2 ATM fraud racket busted police    police
1.3 ATM fraud racket busted constabl  constable
1.4 ATM fraud racket busted night     night
1.5 ATM fraud racket busted patrol    patrol
1.6 ATM fraud racket busted nab       nab
```

Figure 6: Stemming and Lemmatization of the word morphological analysis.

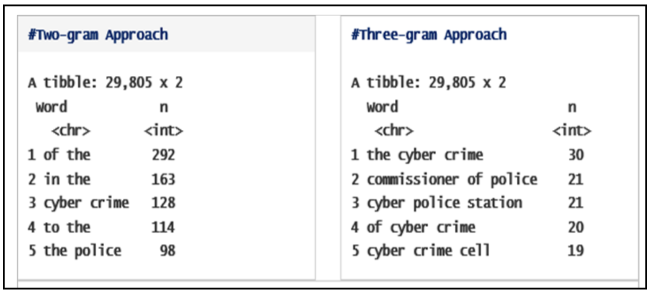
Building up from smaller meaning-bearing pieces is what normalization is all about. The first document in #3 contains two stem words, such as "constables" and "nabbed." 'constabl' and 'nab' are the results of eliminating the suffix. The word 'constabl' is incorrect. It will attempt to return the dictionary format as 'constable' after removing the inflectional endings.

Crime corpus of relations: Label extraction

The keyword is retrieved from the n-gram approach in collocation to discover the most relevant terms and better insight into the examined material.

Collocation: Bigram (two-gram) approach and trigram (three-gram) approach

After pre-processing the newspaper articles, a corpus was constructed for labeling cybercrime categories with article label domains. The frequency of document terms is calculated for all documents, and important terms are identified using the n-gram method. IDF (t, D)=logN/|d€D:t€D|, where N is the number of documents in the corpus (N=|D|), T is the number of occurrences, and |d € D:t€D| is the number of document terms (Figure 7).



```
#Two-gram Approach

A tibble: 29,805 x 2
 Word            n
   <chr>      <int>
1 of the        292
2 in the        163
3 cyber crime   128
4 to the        114
5 the police     98
```

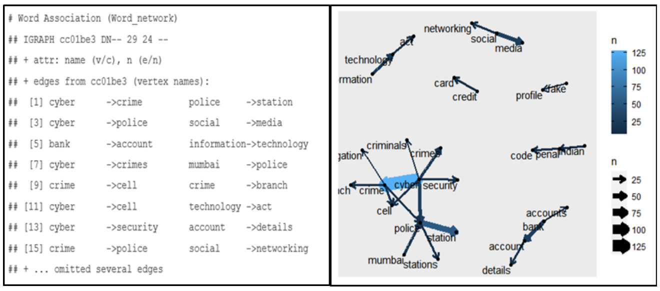
```
#Three-gram Approach

A tibble: 29,805 x 2
 Word                       n
   <chr>                  <int>
1 the cyber crime          30
2 commissioner of police   21
3 cyber police station     21
4 of cyber crime           20
5 cyber crime cell         19
```

Figure 7: Bigram (two-gram) approach and trigram (three-gram) approach.

Narrow down the news content based on a certain theme, such as crime labels or article label, to have a better understanding of the behavioral insights. Collocation aids in the retrieval of two or more words that are extremely likely to occur together. News the documents contain content that is closely related to the term 'cyber,' such as 'cybercrime,' 'cyberspace,' 'cyber criminals,' 'cyber security,' and so on. The bi-gram (two-gram) or tri-gram (three-gram) approaches might be used depending on the findings. The sentence in 1, 2, 4 documents does not give any significant phrase in the #4 bi-gram technique. To infer more insights, always prefer the tri-gram approach based on behavior analysis.

Association of cyber news-visualizing bigrams approach

The relationships between the news terms are visualized using a 'network' or 'graph' to find them all at once. In #5, phrases connected to cybercrime, such as 'cybercrime,' 'police,' 'bank,' and 'crime,' are nodes that are frequently followed by others. #5 shows how to see a graph (Figure 8).



```
# Word Association (Word_network)
## IGRAPH cc01be3 DN-- 29 24 --
## + attr: name (v/c), n (e/n)
## + edges from cc01be3 (vertex names):
##  [1] cyber      ->crime         police     ->station
##  [3] cyber      ->police        social     ->media
##  [5] bank       ->account       information->technology
##  [7] cyber      ->crimes        mumbai     ->police
##  [9] crime      ->cell          crime      ->branch
## [11] cyber      ->cell          technology ->act
## [13] cyber      ->security      account    ->details
## [15] crime      ->police        social     ->networking
## + ... omitted several edges
```

Figure 8: Association of cyber news-visualizing bigrams approach.

Cybercrime binary classification

The current traffic between People-centric and Techno-centric has clearly demonstrated the complexity of cybercrime and its impact. These two words are linked to the cyber-factor vs. the human element of criminal crimes. Figure 9 shows the categories of cybercrime based on cyber enabled and cyber dependent [11].

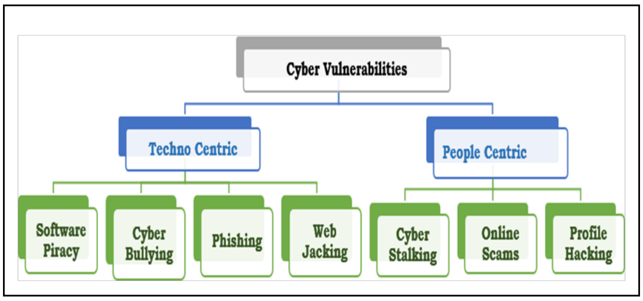


Figure 9: Cybercrime label classification.

Semantic mapping model-corpus creation

To execute semantic modelling, associate the label with the map reduce technique. The Cyber vulnerable label corpa is mapped with IPC Section [Law Punishment] using a map reduces technique. Shows the key (Crime labels) and value (Accusation) pair (Figure 10).

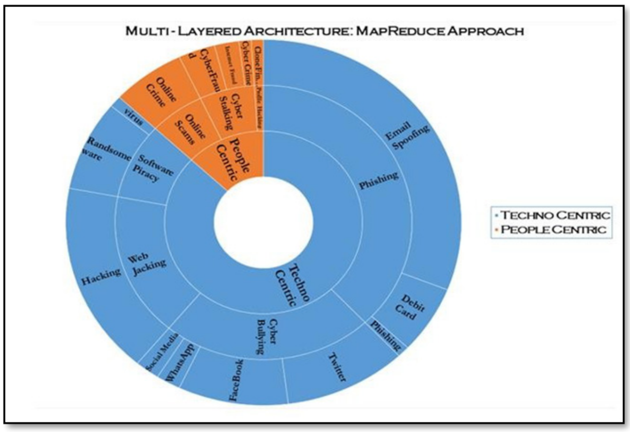
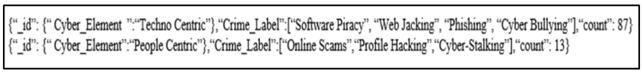


Figure 10: Multi-Layered Architecture: mapreduce Approach.

A multi-layered architecture is used to model the dataset. To determine the Crime vulnerabilities, the qualities are modelled from a domain perspective. The various outcomes of the semantic match are discussed in the following sections.

#6 mapping of cyber elements and crime label

In case #6, 87 percent of cyber vulnerabilities are classified as "techno centric", and they occur as a result of gadgets such as mobile phones, laptops, and smart devices. "Software piracy, web jacking, phishing, and cyber bullying" are the key contributions of technology aspects, while "online scams", Profile hacking, and cyber-stalking" are the main contributions of people-based factors (Figure 11).

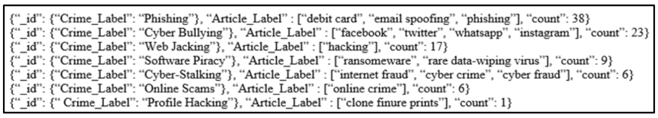


```
{"_id": {" Cyber_Element ":"Techno Centric"},"Crime_Label":["Software Piracy", "Web Jacking", "Phishing", "Cyber Bullying"],"count": 87}
{"_id": {" Cyber_Element":"People Centric"},"Crime_Label":["Online Scams","Profile Hacking","Cyber-Stalking"],"count": 13}
```

Figure 11: Mapping of cyber elements and crime label.

#7 Mapping of Cyber Labels, Article Labels

#7 implies that the majority of susceptible crime is reported as "Phishing". According to the data, Phishing and kinds accounted for 38% of all offences. Email spoofing and debit card access are both common vulnerabilities. Furthermore, because of the prevalence of social media apps such as Facebook, WhatsApp, Twitter, and Instagram, 23 percent of offences were reported as "Cyber Bullying" (Figure 12).

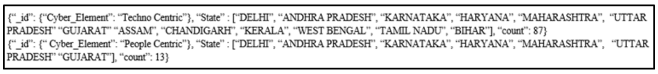


```
{"_id": {"Crime_Label": "Phishing"}, "Article_Label" : ["debit card", "email spoofing", "phishing"], "count": 38}
{"_id": {"Crime_Label": "Cyber Bullying"}, "Article_Label" : ["facebook", "twitter", "whatsapp", "instagram"], "count": 23}
{"_id": {"Crime_Label": "Web Jacking"}, "Article_Label" : ["hacking"], "count": 17}
{"_id": {"Crime_Label": "Software Piracy"}, "Article_Label" : ["ransomeware", "rare data-wiping virus"], "count": 9}
{"_id": {"Crime_Label": "Cyber-Stalking"}, "Article_Label" : ["internet fraud", "cyber crime", "cyber fraud"], "count": 6}
{"_id": {"Crime_Label": "Online Scams"}, "Article_Label" : ["online crime"], "count": 6}
{"_id": {" Crime_Label": "Profile Hacking"}, "Article_Label" : ["clone finure prints"], "count": 1}
```

Figure 12: Mapping of cyber labels, article labels

#8 state-wise mapping of cyber elements

#8 It reports on the geographical study of cyber susceptibility victims in India. It alludes to the fact that the majority of crimes in India occur as a result of technological influences. In comparison, the western zone of India was hit by 48 percent of crimes, while the south was hit by 23 percent. India's most affected states are Maharashtra and Delhi. Both are affected by technology and are centered on people (Figure 13).

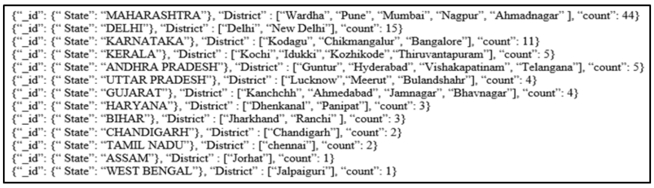


```
{"_id": {"Cyber_Element": "Techno Centric"}, "State" : ["DELHI", "ANDHRA PRADESH", "KARNATAKA", "HARYANA", "MAHARASHTRA",  "UTTAR
PRADESH" "GUJARAT" "ASSAM", "CHANDIGARH", "KERALA", "WEST BENGAL", "TAMIL NADU", "BIHAR"], "count": 87}
{"_id": {" Cyber_Element": "People Centric"}, "State" : ["DELHI", "ANDHRA PRADESH", "KARNATAKA", "HARYANA", "MAHARASHTRA",   "UTTAR
PRADESH" "GUJARAT"], "count": 13}
```

Figure 13: State-wise mapping of cyber elements.

#9 mapping of states and district

#9 refers to Maharashtra, which is heavily impacted by cybercrime in India's Western zone. The central zone of Delhi is the second most afflicted, and Karnataka is the most affected state in India's southern zone. In the northern section of India, the number of crimes is the lowest (Figure 14).



```
{"_id": {" State": "MAHARASHTRA"}, "District" : ["Wardha", "Pune", "Mumbai", "Nagpur", "Ahmadnagar" ], "count": 44}
{"_id": {" State": "DELHI"}, "District" : ["Delhi", "New Delhi"], "count": 15}
{"_id": {" State": "KARNATAKA"}, "District" : ["Kodagu", "Chikmangalur", "Bangalore"], "count": 11}
{"_id": {" State": "KERALA"}, "District" : ["Kochi","Idukki","Kozhikode","Thiruvantapuram"], "count": 5}
{"_id": {" State": "ANDHRA PRADESH"}, "District" : ["Guntur", "Hyderabad", "Vishakapatinam", "Telangana"], "count": 5}
{"_id": {" State": "UTTAR PRADESH"}, "District" : ["Lucknow","Meerut", "Bulandshahr"], "count": 4}
{"_id": {" State": "GUJARAT"}, "District" : ["Kanchchh", "Ahmedabad", "Jamnagar", "Bhavnagar"], "count": 4}
{"_id": {" State": "HARYANA"}, "District" : ["Dhenkanal", "Panipat"], "count": 3}
{"_id": {" State": "BIHAR"}, "District" : ["Jharkhand", "Ranchi" ], "count": 3}
{"_id": {" State": "CHANDIGARH"}, "District" : ["Chandigarh"], "count": 2}
{"_id": {" State": "TAMIL NADU"}, "District" : ["chennai"], "count": 2}
{"_id": {" State": "ASSAM"}, "District" : ["Jorhat"], "count": 1}
{"_id": {" State": "WEST BENGAL"}, "District" : ["Jalpaiguri"], "count": 1}
```

Figure 14: Mapping of states and district.

RESULTS

Classification of relations: decision tree classifier

To classify cybercrime, the methodology is tested using a decision tree classifier. Various metrics, such as Gini and Entropy, are used to validate the decision tree. Table 3 shows the decision tree, which is divided into two phases.

Table 3: Phases of decision tree.

Building phase	Operational phase
Import the dataset	Make predictions
Pre-process the dataset	Measure performance
Split the data based on the train/ test ratio	-
Build the model for classifier	-

Decision tree algorithm-pseudo code

i) Assign all the training data to the root node

ii) Partition input data recursively based on selected attributes

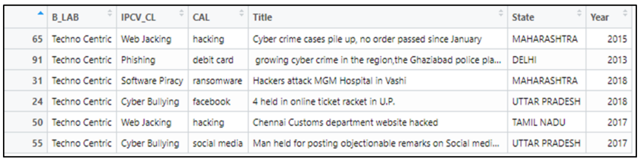
iii) Test attributes at each node are selected based on a heuristic or statistical measure

iv) Conditions to stop partition

v) If all samples belong to the same class, then no further partitioning

Decision tree: Outcome

In this dataset, the crime label (IPCV_CL) is a subset based on domain expertise (B_LAB). Here the dependent variable is crime label (IPCV_CL) and the independent variable is Article_label (CAL). Article_label is fetched from articles and linked to Crime vulnerability activity (IPCV_CL). The data is pre-processed by removing null values and missing values. The structure of the dataset is given in below Figure 15.



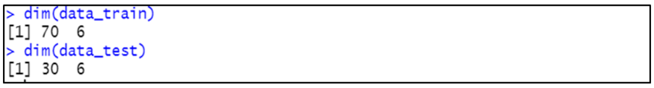
	B_LAB	IPCV_CL	CAL	Title	State	Year
65	Techno Centric	Web Jacking	hacking	Cyber crime cases pile up, no order passed since January	MAHARASHTRA	2015
91	Techno Centric	Phishing	debit card	growing cyber crime in the region,the Ghaziabad police pla...	DELHI	2013
31	Techno Centric	Software Piracy	ransomware	Hackers attack MGM Hospital in Vashi	MAHARASHTRA	2018
24	Techno Centric	Cyber Bullying	facebook	4 held in online ticket racket in U.P.	UTTAR PRADESH	2018
50	Techno Centric	Web Jacking	hacking	Chennai Customs department website hacked	TAMIL NADU	2017
55	Techno Centric	Cyber Bullying	social media	Man held for posting objectionable remarks on Social medi...	UTTAR PRADESH	2017

Figure 15: Dataset.

Note: Crime label (IPCV_CL); Domain expertise (B_LAB).

Split of train/test

Random sampling was used to divide the dataset into the train and test sets. A random id is generated so that the value can be split in the ratio at random. 70% of the data is used to train the model, while 30% of the data is used to improve the forecast (Figure 16).

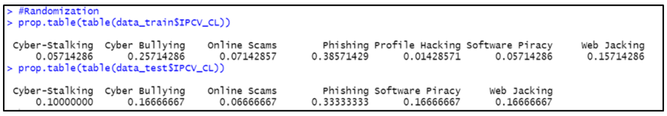


```
> dim(data_train)
[1] 70  6
> dim(data_test)
[1] 30  6
```

Figure 16: Split of Train/Test

Randomization of train/test split

Create two data frames depending on the train and test split, using random sampling. To begin, construct the model using the training data. 70 observations are in the train set (data train) and 30 observations are in the test set (data test) based on the ratio. The probability of the randomization method is given in Figure 17 [12].



```
> #Randomization
> prop.table(table(data_train$IPCV_CL))

 Cyber-Stalking  Cyber Bullying    Online Scams        Phishing Profile Hacking Software Piracy     Web Jacking 
     0.05714286      0.25714286      0.07142857      0.38571429      0.01428571      0.05714286      0.15714286 
> prop.table(table(data_test$IPCV_CL))

 Cyber-Stalking  Cyber Bullying    Online Scams        Phishing Software Piracy     Web Jacking 
     0.10000000      0.16666667      0.06666667      0.33333333      0.16666667      0.16666667 
```

Figure 17: Randomization of train/test split.

Classification of cyber vulnerabilities

Create a criminal element-based model. It depicts the percentage of the tool that is a major victim of cyberattack. This node checks for "Phishing," "Web Jacking," "Software Piracy," "Cyber-Stalking," "Cyber Bullying," "Online Scams," and "Profile Hacking" in the element. If true, then shift to the left child node of the root node. It will delve into the features and determine which ones have an impact on the risk of criminal factors. Figure 18 shows how the article label is categorized once the decision tree model is built based on the criminal label.

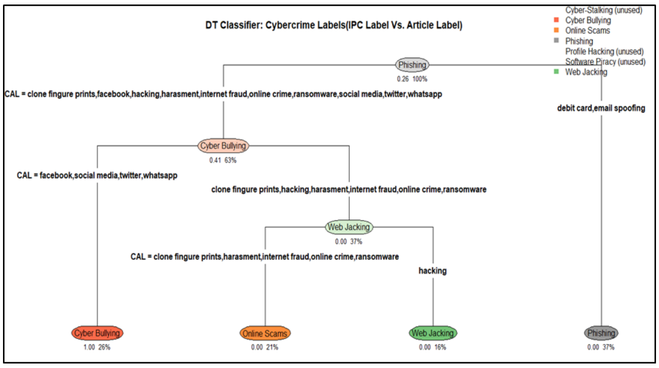
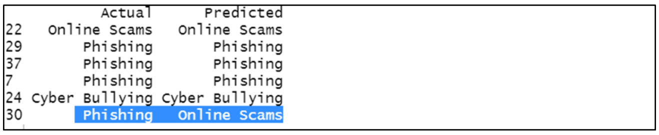

DT Classifier: Cybercrime Labels(IPC Label Vs. Article Label)
Cyber-Stalking (unused)
Cyber Bullying
Online Scams
Phishing
Profile Hacking (unused)
Software Piracy (unused)
Web Jacking
Phishing
0.26 100%
CAL = clone fingure prints,facebook,hacking,harasment,internet fraud,online crime,ransomware,social media,twitter,whatsapp
debit card,email spoofing
Cyber Bullying
0.41 63%
CAL = facebook,social media,twitter,whatsapp
clone fingure prints,hacking,harasment,internet fraud,online crime,ransomware
Web Jacking
0.00 37%
CAL = clone fingure prints,harasment,internet fraud,online crime,ransomware
hacking
Cyber Bullying
1.00 26%
Online Scams
0.00 21%
Web Jacking
0.00 16%
Phishing
0.00 37%

Figure 18: DT classifier: Cybercrime labels [IPC label vs. Article label].

Accuracy test

The confusion matrix is used to evaluate the classification accuracy. The true negative is accounted for. Each column represents a predicted target, whereas each row represents an actual target. It calculated the projected value for test data using the train data as input. In #14 there is a comparison between the expected and actual value. The model's correctness is assessed using a fine-tuning technique, and this model achieves a 53 percent accuracy level (Figure 19).



```
             Actual        Predicted
22     Online Scams     Online Scams
29         Phishing         Phishing
37         Phishing         Phishing
7          Phishing         Phishing
24  Cyber Bullying  Cyber Bullying
30         Phishing     Online Scams
```

Figure 19: Accuracy test.

DISCUSSION

The purpose of this paper is to examine the crime that was reported in the news article. The unstructured data retrieved from the news story was converted to structured data using multi-level crime categorization label mapping of the IPC legal code and a demographic of location and gender. Based on the analysis of the dataset (CVCL), the following conclusion can be drawn. The most common criminal article is "Phishing," which is followed by Bullying and Cyber Stalking. "Jacking, webspace, and the last few are allegations of profile hacking and data theft," according to the typical criminal record.

Qualitative insights-cyber news

The basic qualitative insights from the news given in the article are depicted using a word cloud. The cloud is visualized based on the word occurrence. Dark typefaces are used to highlight words with a high frequency. Small typefaces are used for the terms with the fewest occurrences. Figure 20 shows one example of this. The most frequently used terms in a news item or 'crime,' 'bank,' 'account,' and 'online.'

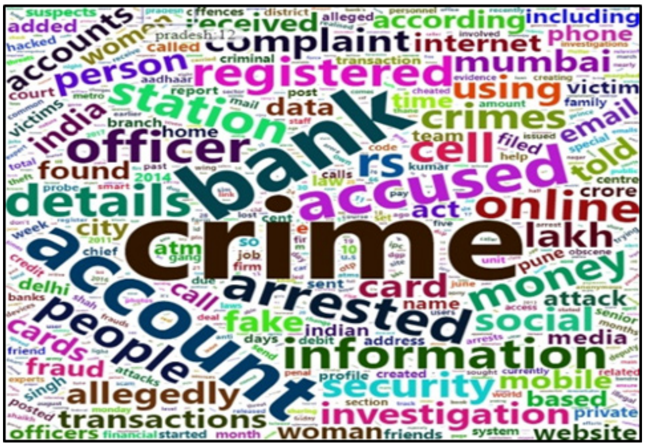


Figure 20: Qualitative insights: Cyber news.

Domain wise inference

Email spoofing is determined to be the most common type of crime, and social networking sites are also used to fund cyberattacks. Financial services are the target of cyberattacks. This project aims to map cyber vulnerabilities in accordance with the IPC, IT Act 2000, as well as to comprehend the penalties and sections of the "IPC-Section 465".

Location-wise analysis

Maharashtra is one of the most vulnerable states in India to cybercrime. The central zone of Delhi is the second most afflicted, whereas Karnataka is the most hit in the southern zone. Figure 21 depicts a crime analysis by location.

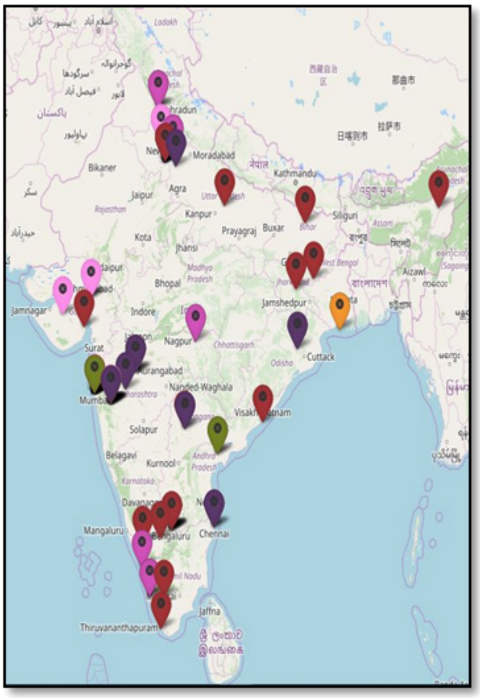


Figure 21: Cyber vulnerabilities-India

Gender wise analysis

In comparison to women, most men are reported to be quite vulnerable. It has received a lot of attention in Mumbai and New Delhi. Andhra Pradesh, Tamil Nadu, and Kerala are the southern states with the lowest reported incidences.

Year-wise analysis

Because of the tendencies of social networking and digitalized technologies in India, cyber susceptible crimes have increased in 2018.

CONCLUSION

Cybercrime vulnerability to the IPC Section is recommended as a specific goal in this article. The dataset was built using news stories extracted from top news publications that dealt with cybercrime. For the years 2012-2018, 100 articles have been downloaded. Cybercrime label mapping, IPC Sections, and demographics are all classified using a framework using a methodology. The Decision Tree Classifier was used to examine the results of the crime classification approach, and it was discovered that 100% of the crime labels are classed as Article labels.

LIMITATIONS

• The dataset has a minimum size that needs to be expanded, which is one of the study work's drawbacks.

• The corpus developed for the research effort can be used to create an automation web crawler.

• This study does not take into account the mapping of security features in other cloud systems.

FUTURE SCOPE

On future research, an automatic web crawler will be used.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work.

CONFLICT OF INTEREST

I P SUDHANDRADEVI, declare that no funds, grants, or other support were received during the preparation of this manuscript.

REFERENCES

1. Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. Journal of Big Data. 2015;2(3):1-41.
2. Chen CLP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences. 2014;275:314-47.
3. Burnap P, Williams ML. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet. 2015;7(2):223-42.
4. Pathak PB, Nanded YM. A dangerous trend of cybercrime: ransomware growing challenge. Int j adv Res Comput Sci Eng Inf technol. 2016;5(2):371-3.
5. Gunjan VK, Kumar A, Avdhanam S. A survey of cyber crime in India. Int Conf Adv Comput. 2013;pp.1-6.
6. Mahmood T, Afzal U. Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. Int Conf Adv Comput.2013;pp.129-134.
7. Kandpal V, Singh RK. Latest face of cybercrime and its prevention in india. Int j basic appl sci. 2013;2(4):150-6.
8. Mittal S, Singh A. A Study of Cyber Crime and Perpetration of Cyber Crime in India. InCyber Law, Privacy, and Security: Concepts, Methodologies, Tools, and Applications 2019;pp.1080-1096.
9. La TM, Dumay J, Rea MA. Breaching intellectual capital: critical reflections on Big Data security. Meditari Accountancy Research. 20189;26(3):463-482.
10. Porcedda MG, Wall DS. Data science, data crime and the law. InResearch Handbook in Data Science and Law 2018 Dec 28. Edward Elgar Publishing.
11. Ibrahim S. Social and contextual taxonomy of cybercrime: Socioeconomic theory of Nigerian cybercriminals. Int J Law Crime Justice. 2016;47:44-57.
12. Chandrala S. Interactive Clinical Dashboards Using RStudio®. PharmaSUG Proceedings. Ephicacy Consulting Group Inc. 2021.