A Probe on Crime Data in Various Domains

Sukumar P, Robert L

Abstract**:** ***The word 'crime' is not a word; there is a lot of pain and loss behind it. The word 'crime' is the most horrible word that scares human society. Even as this article is being read, somewhere in a corner of the world, a man is currently suffering as a result of a crime. Crime has changed its shape over time according to the dimension of men. The system of detection of crimes and punishment of criminals has reached considerable development. Efforts have been made to prevent crimes in the past, but no amicable solution has been reached. To date, there have been lots of strategies to prevent this. The intention of this research article is to analyze the research articles that have been used in various criminal cases/activities that have taken place at different locations, and to identify the criteria, concepts, tools, conclusions, credentials and shortcomings found in those articles.***

Keywords: Crime, Data Mining, Crime Mining, National Crime Record Bureau (NCRB), Online News Papers, Police Narrative reports, Social Media.

I. INTRODUCTION

"Crime" the term was derived from Latin language [19]. The eminent Greek philosopher Aristotle said "Poverty is the parent of crime". Poverty is one of the main attributes of crime issues; nowadays crime is a very serious social issue. According to the National Crime Record Bureau (NCRB) Crime Report 2017, there were 5007044 crimes committed in India [20]. In today's environment, the work of the Police Department is very difficult; crimes and its data are growing rapidly. It is very challenging for the police to detect and investigate crimes and convict the accused. Moreover, police are working hard to prevent crimes. It is better to prevent the crime before it happens than to take strict/ punishable measures after the occurrence of a crime. This is beautifully portrayed in the classical Tamil poem Thirukkural as

> "வருமுன்னர்க் காவாதான் வாழ்க்கை எரிமுன்னர்
> வைத்தூறு போலக் கெடும்."

These lines imply that the prosperity of a person who does not timely guard against faults, will perish like straw before fire. [21]. It is very important to systematize the crimes and store them and investigate whether the crimes are repeated to a lager extent. This study reviews the research findings of researchers who analyzed such data. In this research article, four different crime data source-based articles have been considered namely National Crime Record Bureau (NCRB), Online News Papers, Police Narrative reports, and Social Media.

The content of this article is as follows: Section II discusses about the National Crime Record Bureau (NCRB), Section III revolves around the Online News Papers, Section IV focusses on the Police Narrative reports based articles criteria, and Section V provides an insight on the Social Media based articles concepts, tools, conclusions, credentials and shortcomings and Sections VI elaborates the calculation of this article.

II. NATIONAL CRIME RECORD BUREAU (NCRB)

All the countries in the world systematically collect and store information on criminal activity in their country, within their legal framework. The research report of the researchers analyzed with such information is reviewed and categorized as follows: the elements of identification, concepts, tools, conclusions, credentials and shortcomings.

A. Crime pattern Detection, Analysis & prediction

Sunil Yadav et al. [15] developed a model for crime prediction to improve the accuracy. In this model data set are collected from police department that contain, the number of individuals arrested and number of individual crimes committed with various other attributes. WEKA pre-processing techniques were applied to clean the unnecessary data. K-means algorithm was used to cluster the data such as high number of individuals committed in crime and, low number of individuals committed in crime. Apriori algorithm (association mining) was used to discover the frequent item set. In this model Apriori is applied in clustered data and that data is divided into high and low values. The outcome of Apriori demonstrated a relationship among the persons captured during year and person released in the same year. In this model Naïve Bayes algorithm was used to classify data like, number of crimes committed by a specific age group. In this model linier/linear regression is applied to predict the number of persons released from the crime against the number of hearings finalized throughout the year.

B. Crime detection and criminal identification in India using data mining techniques

Devendra Kumar Tayal et. al. [7] proposed the intergraded system of crime detection and criminal identification by applying data mining techniques for Indian cites.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

Sukumar P*, Research Scholar, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamilnadu, India Email: sugumarskp@gmail.com
Dr. L. Robert, Associate Professor, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamilnadu, India Email: robertatgac@gmail.com

In this system data are collected from NCRB (National Crime Record Bureau), CPJ (Committee to protect journalists), and other Web source, from the period of 2002-2012. Data pre-processing techniques are applied to clean and convert the data into a structured format. In this proposed model seven Indian cities were selected namely, Bengaluru, Delhi, Hyderabad, Kolkata, Mumbai, Jaipur and Pune and K-means clustering algorithm was applied to cluster the cities based on crime rate.

KNN (K-Nearest Neighbor) algorithm was then applied to classify the clustered data based on similar ones. KNN examined the previous crimes and discovery similarity by matching with the current crimes based on the number of nearest neighbors. The overall system measured an accuracy of crime clusters as 93.62% and 93.9%.

C. Crime Analysis Based on Association Rules Using Apriori Algorithm

Mehmet Sevri et al. [11]. displayed an itemized Crime Analysis that depended on data mining Association Rules. The creator applied the Apriori Algorithm to make an association rule to remove the connection among present and past Crime violations. In this framework data collection was done from NIBRS (National Incident-Based Reporting System) crime repository and criminal records from USA. Dataset contained 48 types of crime taken from nearly 5 million crime stories. The crime dataset contained various attributes namely: State, Population Group, Incident Date and Hour, Location type, weapon / force, Type property loss, property description, sex of victim, Relationship of victim to offender, offender age and sex. The proposed framework arranged the dataset into five phases namely Data pre-processing, data encoding, creating transactions in the dataset, creating frequent item sets which provide minimum support value and creating association rules which provide minimum confidence value out of the created item sets. In data pre-processing phase, the dataset attributes are examined using frequency analysis method to reduce irrelevant attributes and nominal values are converted into numerical values for program optimization. Data pre-processing outcome results are then coded in data encoding phase. In transaction creation phase, Apriori algorithm was applied to create the transactions dataset, followed by identifying frequent itemset that provides the minimum support value, and in the final phase association rules were framed to provide the minimum confidence value from the frequent item sets. This overall framework covers a detailed analysis of crimes.

D. Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data

Zakaria Suliman Zubi et al. [18]. Proposed a framework for crime and criminal analysis. In this system, dataset was collected from the Libyan national crime data. Dataset contained more than 350 crimes and criminal records. WEKA mining software Numeric-To-Nominal function was applied to pre-process the collected data. In this framework K Means clustering algorithm was applied to group the crimes and criminals. It additionally gave the general statistical factual information about Criminal age versus Crime types. In this framework Apriori Algorithm was used for Association rule mining to identify frequent crime rate. Finally, both the algorithms displayed a promising outcome.

Table – I: Classification report of National Crime Record Bureau

Author	Input Source	Approach/ Methods	Techniques /Tools	Outcome of Research	Merits	Demerits
Sunil Yadav et al. [15]	Crime Records in Police Department	Association Mining, Clustering Classification, Correlation and Regression	Apriori Algorithm, K-mean, WEKA Tool, R Tool, Naive Bayes algorithm, Liner regression model	Crime accuracy prediction	This model predicated; Each 10 rape case trials were finished and only 2.5 to 3 people are sentenced for the rape charge.	Accuracy metrics are not deeply discussed.
Devendra Kumar Tayal et al. [7]	NCRB, CPJ (Committee to protect journalists)	Data Extraction (DE), Data pre-processing (DP), Clustering, Google map representation, classification, Crime detection	NetBeans (JAVA Tool), WEKA, K-means, KNN classification	Crime detection and criminal identification	This model predicted; Delhi has the highest number of rape rate between other Indian cities. These helped the cities in providing tight security for women.	This model still needs enhancement on data collection, the accuracy of crime classification, and other security measures.
Mehmet Sevri et al. [11]	NIBRS Crime Database Criminal Records in USA (Year of 2013)	Association Rule	Apriori Algorithm Python	Extraction of the relationship and characteristics of Criminal Records	The proposed framework identified 300 frequent item sets and 368 solid relationships. This system applied Association rules to find the Minimum Support Value whose value was 0.05 and the minimum confidence value was 0.06	In this system, 5 million crime reports were analysed but it only identified 300 frequent itemset. Other missing outlier's data are not discussed.

Zakaria Suliman Zubi et al. [18]	Libyan National Crime Data	Clustering, Association rules	Simple K Means Algorithm, Apriori Algorithm, WEKA mining software, Confusion matrix, Google Map API	Statistical report of crime and criminal analyses	This proposed system identified 32,33,34,36 aged group criminals committed the high rates crime and 15, 16, 43,44,51,52 aged group criminals committed the low rates.	In this proposed system data pre-processing phase results are not discussed.

III. ONLINE NEWS PAPERS

In the current digital world, web media reaches people instantly. It is an excellent source of information for people. The following are the findings of researchers who have primarily investigated crime data using such data. The table also classifies the concepts, tools, results, credentials, and shortcomings of the concepts used.

A. Crime Profiler: Crime Information Extraction and Visualization from News Media

Tirthankar Dasgupta et al. [17]. proposed NLP technique to pre-process the digital data collected from different sources namely, WikiCrimes, WordNet, and a subset of the news resources and developed crime ontology for Crime Profiling and also proposed a semi-supervised learning technique by applying SVM algorithm into categories namely, "Nature of crime", "Criminal", "Victim", "Enforcement" news. They crawled around 3000 crime news documents, and presently have achieved an F-Measure of around 64% in detecting accused name, 72% in detecting crime type, 88% in detecting crime location and 87% in detecting date and time of occurrence of the crime event. At the end, the commonly observed errors and the crimes in different geographical locations were visualized.

B. Crime Analytics: Analysis of Crimes through Newspaper Articles

Jayaweeran Isuru et al. [8]. proposed an online newspapers-based crime analysis system. In this system Crawler4j was used to collect source from Srilankan English newspapers namely, Daily Mirror, The Island, and Ceylon Today. SVM (Support Vector Machine) based Classifiers was used to classify the news articles into crime and non-crime. The proposed system performs foremost mechanisms like, Entity extractor, duplication detector, Crime analyzer, and GUI based Visualizer. Entity extractor extracts the entities from each article like crime date, location, police, court, victims count etc. GATE (General Architecture for Text Engineering), techniques were used to pre-process the text. Duplicate Detector used 64-bit sim hash values to remove the duplicate newspaper articles from the database. Crime analyzer proposed operations such as, like hotspot detection, Crime Comparison, and Crime pattern Visualization. The proposed system achieves some valuable accuracies in article classification (95.7163 %), and Entity Extraction (79.33%). Duplication detection detected that average precision value was 95% and recall value was 96 %. They overall proposed system achieved some decent performance in Crime analysis system.

C. Web News Mining in an Evolving Framework

José Antonio Iglesias et al. [9]. proposed an approach to classify different kind of news articles and clustered various topics of news. In this framework hundreds of News articles data are collected from New York Time online newspapers. NYT articles are clustered in seven different phases like, Travel, technology, Sports, Health, Science, Business and Art. In this system 500 articles per each cluster head from the overall 3500 web news articles were collected. In this proposed model different cluster head like, Health vs. Science, Science vs. technology, Health vs. Sports. Etc were taken into account. This proposed framework had two major phases namely, Term Extraction and Evolving classification. Term Extraction model is classified in two phases, Term Generation and Term Filtering. Term Generation Phase used RapidMiner to generate and pre-process the data. Term filtering phase finds the outlier of the irrelevant data and removes it from the source and also updates the new categories into the news articles in the corpus. Evolving Classification phase is divided into modules namely, Creation of the Evolving Fuzzy Rules and Web New Classification. Evolving Fuzzy rules phase, creates the fuzzy rules for the news articles, calculates the distance between two news articles, updates all the prototypes, inserts the newly found categories and, finally removes the unnecessary prototypes. Web news classification phase classifies the new news articles, although every news articles are signified by one or more prototypes. In this classifier previously analyzed categories are compared with the entire prototype using cosine distance, smallest distance and closest similarity.

D. Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers

Srinivasa K et al. [14]. proposed a framework to construct an information base of Crime Entities from online newspapers without duplicity and information loss. In this system data are collected from three Indian newspapers namely, Indian Express, Times of India, and Deccan Chronicles. In this framework manually crime related terms of corpus are created. First phase of framework verified all the headlines of the collected news articles. If it did not match the crime corpus focus is moved onto the next phase. In this phase latent Dirichlet allocation (LDA) algorithm was applied to find the probability distribution of the headlines. Probability Distribution of the Headlines (PDH) values are matched with the threshold values kept on the headlines and updates the Crime Corpus otherwise the News articles are rejected. The proposed NLP rule-based techniques are used to extract the entities from the selected Newspapers. This upgraded NLP framework reduced the inaccurately tagged text and untagged text. In the Tokenization and POS tagging phase, articles were tokenized into sentence and words. NLTK POS tag was used to tag the tokenized words of each sentence.

In the Named entity Recognition and Relationship Extraction phase, the relationship of previously tagged entities based on SVO (Subject-Verb-Object) rule is identified. In this proposed system contextual similarity measure and semantic similarity measures are applied to identify the similarity between the text events.

E. Extracting Crime Information from Online Newspaper Articles

Arulanandam Rexy et.al. [4]. presented a framework for mining the crime location from online newspapers. In this model data are collected from 70 online newspapers articles from Otago Daily Times (New Zealand), Sydney Morning Herald (Australia) and The Hindu (India). First phase of this framework built the crime corpus based on theft related articles.

In the Sentence tokenization phase, NLTK tools were used to split articles into individual sentences. In the Location identification phase Named Entity Recognition algorithms were used to identify the location from individual sentence. Feature identification phase was used to label identified location as CLS (crime location sentence) and as NO-CLS (not a crime location sentence), Label assignment phase was used for feature identification. In this phase labels were manually assigned to the unidentified sentence. In the Training CRF (Conditional

Random Field) phase, CRF algorithm was used to learn the weights between Feature identification and label assignment phase and a new model was created. When new dataset was used, this model automatically labelled the sentence from the article. In the Sentence classification phase, automatically assigned labels were compared with manually assigned labels with the help of precision, recall and f-score and identifies accuracy metrics.

Table – II: Classification report of online news papers

Author	Input Source	Approach/ Methods	Techniques / Tools / Algorithms	Outcome of Research	Merits	Demerits
Tirthankar Dasgupta et. al. [17]	News Article	NLP technique. Classification	NLP, Semi-supervised learning technique (SVM) Visualizing (Chart, Wheal)	Crime Information Extraction and Visualization	This literature helps to learn different NLP techniques to extract Entities from data	Still need accuracy for detecting accused names because accuracy is below 65%.
Jayaweeran Isuru et. al. [8]	Srilankan English News Paper (Daily Mirror, Island, and Ceylon today)	Entity extractor, duplication detector, Crime analyser, and GUI based Visualiser	SVM classifier, Crawler4j, Jsoup, Hybrid sampling techniques, Grid Search, SMOTE Up Sampling, Ensemble Boosting, Ensemble Bagging, DEC (Different Error Cost), Hybrid Sampling techniques, Machine Lanning, Context Pattern.	Hotspot detection, Crime Comparison, and Crime pattern Visualization	The Crime analysis proposed system handled techniques such as, Crime analysis system SVM classifier, Crawler4j, Jsoup, Hybrid sampling techniques, Grid Search, SMOTE Up Sampling, Ensemble Boosting, Ensemble Bagging, DEC (Different Error Cost), Hybrid Sampling techniques, Machine Lanning, Context Pattern, GATE, OpenNLP and Google Map API. These techniques are useful for further continuation of the research.	Entity extraction extracts the accuracy values as Crime Location 82%, Crime Type 82%, and Crime Date 74%. This model still needs some improvements in accuracy.
José Antonio Iglesias et al. [9]	Online News	Term Generation Team extraction Information Retrieval Evolving Classification	RapidMiner Tokenization Stopword removal Stemming or lemmatization Snowball Stemmer TF-IDF Evolving Fuzzy System eClass0 classifier C4.5 classifier Naïve Bayes Classifier ANN, SVM KNN	Various Classified categorises of News Articles	eClass0 is the well performed classifier and this classifier did not store all streamed data into the memory. It was simple, efficient and rule based to evolve. This framework provides good results namely, Health and Science news articles category 4 rules was created, Health, Science and Sports news articles category 16 rules was Created. Arts, Business, Health, Science, Sports, and Travel news articles category 32 rules was Created	Artificial Neural Network groundwork procedure is very time consuming

Srinivasa K et al. [14]	Online Newspapers	Tokenization and POS Tagging Named Entity Extraction Relationship Extraction	NLTK, LDA Algorithm Python 3 Library OWL 2.0 ontologies DBpedia SPARKQL Word2Vec	Crime Knowledge base	This proposed framework applied present techniques namely, NLP and Word2Vec. This system presented knowledge graphs for events and also showed knowledge graphs after semantic merging.	Primarily 100 crime related keywords used were to create the corpus. Updated crime terms result are not discussed.
Arulanandam Rexy et.al. [4]	Otago Daily Times Sydney Morning Herald The Hindu	Corpus building, Sentence tokenization, Location identification, Feature identification, Label assignment, Training CRF, Sentence classification	Mozenda Web Screen Scrapper Tool, Punkt Tokenizer, NLTK toolkit, Named Entity Recognition, Stanford NER, Conditional Random Field, CRF classifier, LBJ Tagger, python regular expressions,	Crime Location Identification	This proposed framework provided a promising accuracy of the results revealed. New Zealand articles differs from 84% to 90%. Other countries of India and Australia articles differs from73% to 75%	Sentence classification phase comparison of automatically assigned labelled with manually assigned results are not discussed.

IV. POLICE NARRATIVE REPORTS

One of the most important duties of police is to collect information on the crime scene and record the first report of the victims and to report where the crime has taken place. The following is the review report of the researchers who analysed the data as primary data. The table also classifies the concepts, tools, results, credentials, and shortcomings of the concepts used.

A. Crime Information Extraction from Police and witness narrative reports

Chih Hao Ku et al. [5]. developed a web-oriented reporting system that was associated with NLP (Natural Language Processing) techniques. This proposed system collected data from cognitive interview reports that were used to extract the valued entities. The proposed system created a lexicon (huge crime explicit lexicon) that was developed from UCR (Uniform Crime Report), Wikipedia, MSN Encarta, Frame Net Serious Wheels. Every category of lexicon included enormous sub-lexicons. The lexicon was also used Collins Cobulid, to hold or eliminate a term, whereas the term was overlapped. The proposed system applied the GATE (General Architecture for Text Engineering) modules for data pre-processing. They also developed own JAPR rules and Gazetteer lists to extract the entities. The overall system results for precision, police narrative was 94%, witness narrative was 96 %, results for recall percentage was police narrative 85% and witness narrative 90%.

B. An intelligent Analysis of a City Crime Data Using Data Mining

Malathi A et al. [10] analyzed a city crime data to predict crime trends from 2000 to 2009 by collecting Crime data from Tamilnadu police department. The filtering phase removed the unwanted elements from the dataset. WEKA tool was used for clustering the crime data to the size of the city population and the next phase predicted the crime future trend by forecasting crime rate between one year to the following year and applied data mining strategies to extend changes into what's to come. The proposed model applied attributes namely murder, rape, theft, cheating and other IPC Crime to cluster crime cities that had a similar trend, and then next year cluster information to classify records, and these were joined with the state poverty data to create a classifier that predicted future crime trends.

C. A Distance measure for determining similarity between criminal investigations

Tim K. Cocx et al. [16]. presented a distance measure for determining similarity between criminal investigations. This framework contained four phases, commercial text miner, Table transformation unit, distance calculator and visualization. In this system data were collected in police narrative reports. Text miner techniques were applied to extract the entities and labelled extracted entities and finally stored it into high dimensional table. High dimensional table was used to compare crime cases and find the common similarity in crime. In this system, distance measure techniques were used to calculate the distance between similarities of criminal cases. This framework produced some good results: 28 police investigation, and extracted 152,280 unique entities by proposed distance measure, calculated the distance and visualized it in a clustered image.

D. Big Data-Based Smart City Platform: Real – Time Crime Analysis

Debopriya Ghosh et. al. [6]. presented a knowledge-based framework for smart city platform in New York. In this model data are collected from NJ Crime History and FIR. Data acquisition phase collected 4000 different types of incidents in various time period. The raw data that was kept in the repository contained a lot of unrelated and counterfeit data. Data pre-processing techniques were applied to remove punctuations; stop words and words were stemmed. In this model machine learning algorithms were applied to classify the crime incident automatically.

Document indexing phase applied Term Frequency – Inverse Document Frequency (TF-IDF) weighting function to train and validate the text documents. Cosine – Normalization was used to normalize TF-IDF Results. In this model Latent Semantic Analysis algorithms were used to extract specific crime type like "theft". Latent Dirichlet Allocation algorithms were applied to model the separate topics. Text Categorization phase applied multiple learning algorithms such as Random Forest, Support Vector Machine (SVM), Scaled Linear Discriminant Analysis (SLDA), MAXENT (Maximum Entropy Classifier), and Artificial Neural Network (ANN). This model provided some promising results: Named Entity Recognition phase extracted the entities from text and stored into database.

This overall framework framed a relationship between public, location, incidents, and vehicles from the previous crimes and provided new knowledge to study the further new investigations.

V. SOCIAL MEDIA

Social websites are a major source of crime information. Social networking sites like YouTube, twitter, Facebook and Instagram share a variety of data daily. In just a minute, in online 481,000 tweets are sent to twitter, 973,000 users are logged on to Facebook, 3.7 million searches are done on Google Chrome, and 4.3 million videos are viewed on YouTube. These numbers reveal the rapid growth of digital data exposed [22]. The articles reviewed by analysts who used such crime data are as follows. The table also classifies the concepts, tools, results, credentials, and shortcomings of the concepts used.

A. Language Usage on Twitter Predicts Crime Rates

Abdulaziz Almehmadi et al. [3]. proposed to predict the crime rates from public twitter data in Houston and New York City. Twitter Streaming API was used to collect the tweets from Houston (2.5 Million) and New York City (3 Million), for each tweet extract tweet's text, Creation date and languages. WEKA 3.7.10 Machine Learning Toolkit was used to pre-process the data and construct a 2-dimension matrix (document, attribute). Automatic Feature selection was used to search 400 features. SVM classifier was used to classify the language as an offensive or non-offensive one. Classified data was divided into two chunks, training and testing. In the training phase, 1000 data are manually labelled based on vulgar words. In the evaluation phase cross- validation partitioned the data into k equal size subsamples and the model was built on k-1 of the folds. This model identified a number of statistically significant correlations between crime rate and the mode of language used.

Table – III: Classification Report of Police Narrative Reports

Author	Input Source	Approach/ Methods	Techniques /Tools	Outcome of Research	Merits	Demerits
Chih Hao Ku et al. [5]	Police and witness narrative reports, Lexicon: UCR Wikipedia, MSN Encarta, FrameNet Serious Wheels	Information Extraction Cognitive Review Lexicon Developed	Collins Cobulid Tokenizer Sentence Spiller POS tagger Noun Chunks Ortho-matcher JAPR rule, Gazetteer lists	Online reporting system that combines NLP	The proposed system was used for various techniques in the phase of information Extraction. They succeeded 100% precision results for extracting information in the entities fields of Personal property, Physical Condition, clothes and vehicles.	The proposed system encountered the bottommost precision and recall results in entities Age was (58% and 54%).
Malathi A et al. [10]	City Crime data from Tamilnadu police department	Data Collection, Data pre-processing, Clustering, Classification, Crime trend prediction	Weka Tool DBScan EM (Expectation Maximization) K-Means Clustering	Predicted the Crime trends.	The crime analysed system predicted violent crimes against women namely, Rape, Sexual Harassment and Dowry Death. These were down in 2000, and unfortunately increased in 2004 to 2008.	In this model there is no discussion on why crimes rates against women increased in 2004 to 2008 and What are the social attributes involved?
Tim K. Cocx et al. [16]	Police narrative reports	Text Mining Distance Measure Associative array clustering technique	INFO-NS Program SPSS Lexi quest text mining tool Neural Network	Criminal similarity distance measured	Proposed a new distance measure	Poor quality of data accuracy in data collection phase, still in want of an improvement.
Debopriya Ghosh et. al. [6]	NJ Crime History, FIR	Data pre-processing, Document Indexing, Topic Modelling, Named Entity Recognition, Text Categorization.	TF-IDF, Cosine – Normalization Latent Semantic Analysis, Latent Dirichlet Allocation, R- Text Tool, Random Forest, SVM, SLDA, MAXENT, ANN, R- Shiny Package, MySQL Database.	Smarter Safer City	This framework compared multiple machines learning algorithms. These algorithms provided some promising results.	Limited number of vocabularies were used in Entity Extraction Phase

B. Mining Social Media Content for Crime Prediction

Aghababaei Somayyeh et al. [1]. proposed a model for crime trend prediction. This model retrieved crime rates from four urban areas of the United States namely Chicago, Philadelphia, San Francisco, and Houston. Topic-based sampling strategies utilized explicit keywords or hashtags to gather tweets from Twitter API based on related geographic zone. The training data and input data of this model were derived from crime indexes. Binary classification model was used to reduce the trends' prediction. The linear SVM Classifier was used to predict given data, regardless of whether the crime index will go up or down in the future.

This model evaluates the predicted user-generated content in social media, no keywords and exact terms correlated to the crime. A comprehensive experiment was conducted in different cities of the United States to predict crime trends.

The Complete outcomes showed that there was an association between before posted tweets and crime rates in the standpoint time frame. This proposed model used only twitter content, the correlated relationship content and crime trends that had been exposed. The F-measure values of prediction model are labelled grounded on “trend” and “mean”, Chicago (mean 0.71, trend 0.62), Philadelphia (mean 0.63, trend 0.7), San Francisco (mean 0.52, trend 0.58), and Houston (mean 0.53, trend 0.53). This prediction model compares the correlation between content and crime trends in the different crime incidents.

C. Crime Analysis and Prediction Using Data Mining

Sathyadevan Shiju et al. [12]. proposed a system that was used to predict the high possibility for crime incidence and visualize crime zones. In this system data are collected from various sources namely, News Sites, Blogs, Media, RSS Feeds, and the collected data was stored into Database (Mango DB). In this system, Naïve Bayes Algorithm was used to build the trained crime data model and it showed over 90 % accuracy. NER (Named Entity Recognition) techniques are used to extract the entity in text namely, person names, organizations, location, date and time. This model applied the Apriori association rule mining algorithm to find the frequent crimes patterns in a particular region. The crime pattern was used to construct a model for the decision tree. Techniques were then used to build a model by tanning the frequent patterns. Finally, these models graphically represented the crime-prone areas.

D. Mining Twitter data for crime trend prediction

Aghababaei Somayyeh et al. [2]. proposed a model for crime trend prediction. This proposed model collected data from Chicago data portal and Twitter and the model definitions

were its own training data. This proposed system applied the binary classification algorithm that was used to classify the crime index from input data. This system proposed a temporal topic model by applying a Latent Dirichlet Allocation (LDA) algorithm to renew entire vocabulary in various time frames. In this model, outcome of crime types recommended a solid relationship between the content of twitter and the trend of crime rates. This proposed model visualized the frequent terms distributions for the top 20 topics, the gathered topics showed the various characteristics in terms of document-topic distribution and the F-measure values for different crime types.

E. An Approach to Build a Database for Crimes in India Using Twitter

Sinha Ranu et al. [13]. developed a real-time Crime repository. In this proposed system nearly 21k tweets(data) were collected from twitter news channels account and newspapers twitter account namely, @ndtv, @TimesNow, @the_hindu, @IndianExpress, and @DDNewsLive. This model handled the R language and applied rtweet library to access the Twitter API, extracted the tweets and created the model for crime database. Crime keywords were used to categorise the tweets. Data pre-processing techniques were used to clean and remove the duplicate tweets. The proposed model used the openNLP package to extract the name and location in the tweets and finally stored into the database. The stored data visualized the High crime rate zones. The overall proposed system developed a model for the Name-Entity Recognition.

Table- IV: Classification report of Social Media

Author	Input Source	Approach/ Methods	Technique s /Tools	Outcome of Research	Merits	Demerits
Almehmadi Abdulaziz et al. [3]	Twitter Tweets	Streaming	API, SVM, MongoDB, NoSQL	Crime rates model	They suggest better predictable crimes rates, used an opinion finder and Google-Profile of Mood States (GPOMS), for each tweet, overall mood in terms of being positive, negative.	The label assignment in text mining was not fully automated process: it was done by manually and still need improved techniques. In this analysis only few available fields of twitter data are used, if they used more fields, information possibly could have caught some interesting patterns.

Aghababaei Somayyeh et al. [1]	Twitter tweets	Data Collection Predication Model	linear SVM Classifier	Crime prediction model	The outcomes prediction model concluded, the correlation between content and crime trends in the different crime incidents.	In this research limited number of crime index are handled, further analysis is expected to look at the joining of other financial lists and geographical data, which connect with offender behaviours. These displayed the correlation with different incidents.
Sathyadevan Shiju et al. [12]	Web Sites, New Sites, blogs, Social media, RSS Feeds	Data Collection, Classification, Pattern identification, Prediction, Visualization, Crime Profiling	Mango DB, Naïve Bayes, SVM, NER, Decision tree, Neo4j	Crime Prone reigns	The proposed system predicts crime prone areas in India.	The proposed system predicts crime regions only, and does not consider appropriate time.
Aghababaei Somayyeh et al. [2]	Chicago data portal and Twitter	Labelling Topic modelling Term-Topic Distribution Document-Topic Distribution F-measure	Binary classification Latent Dirichlet Allocation (LDA)	crime trend prediction model	This proposed model applied Binary classification algorithm to label the crime index and Latent Dirichlet Allocation (LDA) algorithms to model the topics.	This proposed model handled a lot of trained data
Sinha Ranu et al. [13]	news channel twitter posted tweets	Keyword based search of crime Cleaning of data Extract name and location	openNLP R language rtweet library	Crime data repository	The proposed system developed a model for the Name-Entity Recognition to extract the Name and location from tweets.	The proposed system, accuracies are not much in the case of input and extracted data. The data pre-processing techniques not much used in an elaborate manner.

VI. CONCLUSION

This review article evaluates the statements made by researchers with four major crime data sources, the National Crime Record Bureau (NCRB), Online News Papers, Police Narrative Reports, and Social Media. Data collection, data pre-processing, Entity Extraction, and Classification are the more complex and challenging in Crime evaluation, and there is a need for significant improvement on Data collection and Entity Extraction. Researchers have applied current popular techniques, methods and tools. This study helps to find out that many crime analyses have been done under a lot of research in English newspapers. The researcher did not use other criminal newspapers to study crime. Much of this research explores the possibilities and problems of exploring other language newspapers. This study opens the door for new research on crime data mining.

REFERENCES

1. Aghababaei Somayyeh, and Masoud Makrehchi. "Mining social media content for crime prediction." In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 526-531. IEEE, 2016.
2. Aghababaei Somayyeh, and Masoud Makrehchi. "Mining Twitter data for crime trend prediction." *Intelligent Data Analysis*, *Vol.* 22, pp. 117-141, 2018
3. Almehmadi Abdulaziz, Zeinab Joudaki, and RoozbehJalali. "Language usage on Twitter predicts crime rates." In *Proceedings of the 10th International Conference on Security of Information and Networks*, pp. 307-310. ACM, 2017.
4. Arulanandam Rexy, Bastin Tony Roy Savarimuthu, and Maryam A. Purvis. "Extracting crime information from online newspaper articles." In *Proceedings of the second australasian web conference-volume 155*, pp. 31-38. Australian Computer Society, Inc., 2014**.**
5. Chih Hao ku, Alicia Iriberri, and Gondy Leroy. "Crime information extraction from police and witness narrative reports." In *2008 IEEE Conference on Technologies for Homeland Security*, pp. 193-198. IEEE, 2008**.**
6. Debopriya Ghosh, Soon Chun, Basit Shafiq, and Nabil R. Adam. "Big data-based smart city platform: Real-time crime analysis." In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, pp. 58-66. ACM, 2016**.**
7. Devendra Kumar Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, and Nikhil Tyagi. "Crime detection and criminal identification in India using data mining techniques." *AI & society*, vol. 30, no. 1, 2015
8. Jayaweera Isuru, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera, and Adeesha Wijayasiri. "Crime analytics: Analysis of crimes through newspaper articles." In *2015 Moratuwa Engineering Research Conference (MERCon)*, pp. 277-282. IEEE, 2015**.**
9. José Antonio Iglesias, Alexandra Tiemblo, AgapitoLedezma, and Araceli Sanchis. "Web news mining in an evolving framework." *Information Fusion*, vol. 28, pp. 90-98, 2016.
10. Malathi A., S. S. Babboo, and A. Anbarasi. "An intelligent analysis of a city crime data using data mining." In *International conference information electronic engineering*, vol. 6, pp. 130-134. 2011.
11. Mehmet Sevri, HacerKaracan, and M. Ali Akcayol, "Crime analysis based on association rules using apriori algorithm." *International Journal of Information and Electronics Engineering*, Vol. 7, no. 3 pp. 99,2017.
12. Sathyadevan Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." In *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, pp. 406-412. IEEE, 2014**.**
13. Sinha Ranu, Mohit Kumar, and Saptarsi Goswami. "An Approach to Build a Database for Crimes in India Using Twitter." In *International Conference on Computational Intelligence, Communications, and Business Analytics*, pp. 150-160. Springer, Singapore, 2017**.**
14. Srinivasa K., and P. SanthiThilagam. "Crime base: Towards building a knowledge base for crime entities and their relationships from online newspapers." *Information Processing & Management*, Vol. 56, no. 6, Elsevier**,**2019
15. Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, and Nikhilesh Yadav. "Crime pattern detection, analysis & prediction." In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, pp. 225-230. IEEE, 2017**.**
16. Tim K. Cocx, and Walter A. Kosters. "A distance measure for determining similarity between criminal investigations." In *Industrial Conference on Data Mining*, pp. 511-525. Springer, Berlin, Heidelberg, 2006.
17. Tirthankar Dasgupta, AbirNaskar, RupsaSaha, and Lipika Dey. "Crime Profiler: crime information extraction and visualization from news media." In *Proceedings of the International Conference on Web Intelligence*, pp. 541-549. ACM, 2017.

18. Zakaria Suliman Zubi, and Ayman AltaherMahmmud. "Using data mining techniques to analyze crime patterns in the Libyan national crime data." *Recent advances in image, audio and signal processing, Vol.* 8, pp. 79-85**.** 2014
19. https://en.wikipedia.org/wiki/Crime
20. http://ncrb.gov.in/StatPublications/CII/CII2017/cii2017.html
21. https://www.ytamizh.com/thirukural/kural-435/
22. https://www.visualcapitalist.com/internet-minute-2018/

AUTHORS PROFILE

Mr. P. Sukumar pursued Bachelor of Computer Science from Government Arts College of (Autonomous), Coimbatore, India, in the year 2011. Master of Computer Applications from Government Arts College of (Autonomous), Coimbatore, India, in the year 2014. He has received M.Phil. in the field Data Mining and Currently pursuing Ph.D. full time research scholar in Government Arts College of (Autonomous), Coimbatore, India.

Dr. L. Robert is an Associate Professor in the Department of Computer Science, Government Arts College, Coimbatore, India. He was a faculty member of the Computer science and Information System Department, King Saudi University, Riyadh, KSA. He Received his tertiary education from St. Joseph College, Trichy, Obtaining B.Sc. Computer Science in 1993 and M.Sc. Computer Science in 1995, both receiving College Management Medals for outstanding academic performance. Another feather in the cap is clearing SET and NET. He Completed his Ph.D. in Computer Science from PSG college of Technology, India, in 2008.

His area of research is Data Compression and Data Management. He has science successfully guide four Ph.D. Scholars. In the area of publication, he has authored more then 20 papers in national and international peer reviewed journals, and more than 30 papers in national and international conference.