# Area-Wise Crime Rate Detection using Text and Web Intelligence

**Tanya Joon, Tanvi Aggarwal, Tarishi, Vishesh Tandon, Students, B.Tech., Computer Science and Engineering, The NorthCap University, Gurugram**

***Abstract*** *Crime is defined as an act harmful not only to the individual involved, but also to the community. It is also a forbidden act that is punishable by law. Crimes are social nuisances that place heavy financial burdens on society. Crimes rates are increasing day by day as we see in our daily lives. Something must be done in order to control or detect the crimes and help government and public to take the required steps and precaution in order to lower down the rate of crime. In this paper, we propose an approach for the design and implementation of crime rate detection for Indian cities and hence, we are able draw a clear view about the crimes in the different cities. Our approach is divided into 5 modules – Scraping, Data Extraction, Data Preprocessing, Lemmatizations, Heap Maps. First module Scraping in this we scrap the articles which are required using BeautifulSoup package. Second module Data extraction extracted the city names from the articles where crimes happened in India. Third module Data Preprocessing preprocessed the URLs required by using NewsPlease and Tokenization each sentence is partitioned into a list of words and we remove the stop words. Fourth module Lemmatization we made the dictionary of the cities and then also checked manually. Fifth module Heap Maps by plotting the Heap Maps, will help to analyze which city is prone to high rate of crimes.*

**Keywords** Lemmatizations . Tokenization . Heat Maps . NewsPlease . BeautifulSoup .

## I. INTRODUCTION

Crime is an offense against the society that is often prosecuted and punishable by the law. It has been observed that criminals commit crimes at any place and in any form. Crimes remain as headlines of the news for the very long time; sentiments of crores of people of India are attached to these crimes in sympathy, and many campaigns are raised in the protest them. This reveals that crimes terrifically affect not only the victims but also the people of the country.

So, the check on crimes and target to the criminals are inevitable that need to be performed by the law enforcement agencies to secure the country. These agencies along with additional computer data analysts are responsible for unambiguous and competent crime investigation from the voluminous crime data. Therefore, an approach for crime detection and using Text and Web Intelligence for Indian cities.

Our approach is divided into 5 modules – Scraping, Data Extraction, Data Preprocessing, Lemmatizations, Heap Maps. First module Scraping in this we scrap the articles which are required using BeautifulSoup package. Second module Data extraction extracted the city names from the articles where crimes happened in India. Third module Data Preprocessing preprocessed the URLs required cleans, integrates and reduces the extracted crime data into structured crime instances by using NewsPlease and Tokenization each sentence is partitioned into a list of words and we remove the stop words. Fourth module Lemmatization we made the dictionary of the cities and then also checked manually. Fifth module Heap Maps

by plotting the Heap Maps, will help to analyze which city is prone to high rate of crimes.

Our approach contributes in the betterment of the society by helping the investigating agencies in crime detection and criminals' identification, and thus reducing the crime rates. [5]

## II. LITERATURE SURVEY

Crimes in India are stoked up at an alarming rate, and criminals are opting for queer activities to commit them. Newspapers, Web blogs, etc. are day to day filled with various crime incidents. Some of the mystified crimes that occurred in Indian last couple of years.

A professor was beaten to death by his own students in Ujjain, Madhya Pradesh. A gang of nine taxi drivers from Gurgaon, Haryana, robbed and killed at least 35 people after offering them lift. Unruly mob stripped and molested a girl in full public view at the Gateway of India, Mumbai, on the New Year eve. Days after horrible Nithari killing, 4 decomposed bodies of children were recovered from abandoned go-down in Punjab. Sexually assaulted teenage girls in the Kashmir valley are still struggling to cope up with trauma. These incidents reveal how crimes are becoming a growing blight in India and have become a dominant fact of an Indian life as well.

Some responsible factors that prevail in India for sheer increase in crimes are poverty, migration, unemployment, frustration, starvation, illiteracy, corruption, nepotism, inflation, etc. Impact of such crimes is that today people living in India now focus their eyes toward crime investigation agencies and security agencies to check and control crimes. Currently, physical investigation by agencies has the probability to ignore and neglect the supportive crime features.

Most of these agencies are searching manually database of criminals, which is a tedious process and takes much more time. Few of them work with the help of computer data analysts and are responsible for crime detection, criminal identification and prediction, and crime verification to ensure safety to the citizens of India. To contribute in this aspect, we propose our Area – wise Crime Rate Detection approach using Text and Web Intelligence for Indian cities by consideration of selected crime features. Our methodology can help these agencies to filter crime database to find out the most probable criminals. This will save a lot of time for the agencies.

## III. RELATED WORK

### *A. Crime Mining*

Some results on crime mining have been made through using data mining techniques. Chen et al.[1] applied data mining techniques to study crime cases, which mainly concerned entity extraction, pattern clustering, classification and social network analysis. Abraham et al. [2] proposed a method to employ log files as history data to search relationship by using the frequency occurrence of incidents.

### *B. Event Oriented*

Construction Event extraction is the process to extract attributes and relationship in web pages. Some researchers have proposed ideas of event-oriented construction for processing events. Lin [3] presented a method for information retrieval based on event ontology for event elements such as location, time etc. Zarri [4] proposed a method to append events for the concept of ontology to be closer to the goal of semantic web.

## IV. METHODOLOGY

### Algorithms and Packages Used

BeautifulSoup

Web scraping is a computer software technique of extracting information from websites. This

technique mostly focuses on the transformation of unstructured data (HTML format) on the web into structured data (database or spreadsheet). Beautiful Soup is a Python library for pulling data out of HTML and XML files. Beautiful Soup 3 only works on Python 2.x, but Beautiful Soup 4 also works on Python 3.x. Beautiful Soup 4 is faster, has more features, and works with third-party parsers like lxml and html5lib. Beautiful Soup parses anything you give it and does the tree traversal stuff for you. You can use it to find all the links of a website. Find all the links whose URLs match "foo.com". Find the table heading that's got bold text, then give me that text. Find every "a" element that has an href attribute etc. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

It can be an excellent option depending on what you need to get done. Some key features: provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree.

NewsPlease

NewsPlease is an open source, easy-to-use news crawler that extracts structured information from almost any news website. It can follow recursively internal hyperlinks and read RSS feeds to fetch both most recent and old, archived articles. You only need to provide the root URL of the news website. Furthermore, its API allows developers to access the extraction functionality within their software. news-please also implements a workflow optimized for the news archive provided by commoncrawl.org, allowing users to efficiently crawl and extract news articles including various filter options.

Tokenization

Each sentence is partitioned into a list of words and we remove the stop words. Stop words are frequently occurring, insignificant words that appear in a database record, article or a web page, etc.

Stop Words

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. Natural Language Processing with Python Natural language processing is a research field that presents many challenges such as natural language understanding. Text may contain stop words like 'the', 'is', 'are'.

In computing, stop words are words which are filtered out before or after processing of natural language data. Other search engines remove some of the most common words—including lexical words, such as "want"—from a query in order to improve performance. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is also a part of queries and Internet search engine.

Lemmatization

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning.

Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research.

**Dataset**

Day by day, crime records are expanding in size which gives rise to some key problems such as data storage and data reliability. Data storage aids in efficient storing of the crime database. But data storage needs always not be compatible and feasible due to computers' processing limitations.

Data need not be reliable when an agency has the incorrect crime data.

We have made our own dataset by using different techniques above mentioned.

| | Crimelist |
|---|---|
| 1 | |
| 2 | 'fraud' |
| 3 | 'mug' |
| 4 | 'arson' |
| 5 | 'assassination' |
| 6 | 'assault' |
| 7 | 'blackmail' |
| 8 | 'burglary' |
| 9 | 'dacoity' |
| 10 | 'dowry' |
| 11 | 'forgery' |
| 12 | 'harassment' |
| 13 | 'hijack' |
| 14 | 'homicide' |
| 15 | 'kidnap' |
| 16 | 'manslaughter' |
| 17 | 'molest' |
| 18 | murder' |
| 19 | 'pickpocket' |
| 20 | 'rape' |
| 21 | 'riot' |
| 22 | 'robbery' |
| 23 | 'shoplift' |
| 24 | 'smuggle' |
| 25 | 'stalk' |
| 26 | 'theft' |

Table 1 : Crime List

| | |
|---|---|
| 'Delhi' | 'Gurgaon' |
| 'Agra' | 'Ghaziabad' |
| 'Aurangabad' | 'Ahmedabad' |
| 'Cuttack' | 'Agartala' |
| 'Madurai' | 'Amritsar' |
| 'Mangaluru' | 'Navi-mumbai' |
| 'Goa' | 'Bengaluru' |
| 'Mysuru' | 'Vijayawada' |
| 'Lucknow' | 'Vadodara' |
| 'Mumbai' | 'Nagpur' |
| 'Kochi' | 'Kolkata' |
| 'Rajkot' | 'Faridabad' |
| 'Thiruvananthapuram' | 'Surat' |
| 'Pune' | 'Noida' |
| 'Jaipur' | 'Patna' |
| 'Hyderabad' | 'Thane' |
| 'Nashik' | 'Allahabad' |
| 'Meerut' | 'Bareilly' |
| 'Salem' | 'Bhubaneswar' |
| 'Chandigarh' | 'Bhopal' |
| 'Ludhiana' | 'Chennai' |
| 'Coimbatore' | 'Indore' |
| 'Visakhapatnam' | 'Dehradun' |
| 'Kozhikode' | 'Mumbai |

Table 2 : City List

**Modules involved:**

*A. Data Scraping*

Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer. It's one of the most efficient ways to get data from the web, and in some cases to channel that data to another website.

We have done scraping of the articles related to crimes happened in India using BeautifulSoup package in python.

*B. Data Extraction*

Data extraction is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage.

We have extracted the city names from the articles where crimes have happened in India in an area.

*C. Data Preprocessing*

Data preprocessing is the conversion of data into usable and desired form. This conversion or “processing” is carried out using a predefined sequence of operations either manually or automatically. Most of the data processing is done by using computers and thus done automatically.

The URLs scraped are cleaned, integrated and reduced into structured crime instances and tokenized each sentence is partitioned into a list of words and we remove the stop words.

*D. Lemmatization*

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word.

In this we made a dictionary of the cities and then also checked manually.

*E. Plotting of Heat Map*

A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors.

The seaborn python package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator's requirement. It is a bit like looking a data table from above. It is useful to display a general view of numerical data, not to extract specific data point.

## V. RESULTS

```
Done  14
Delhi: Rapes, abductions remain top challenges
['rape', 'kidnap', 'theft', 'murder', 'robbery', 'dacoity', 'burglary', 'molest
Delhi
(28.6273928, 77.1716954)


rape        2
kidnap      2
theft       2
murder      3
robbery     3
dacoity     1
burglary    2
molest      2
Name: Delhi, dtype: int64


Done  15
Bangalore: Woman, her daughter and parents found dead inside house
[]
Bengaluru
(12.9791198, 77.5912997)


Series([], Name: Bengaluru, dtype: int64)


Done  16
CCTV camera captures daring theft in Bengaluru's Yeshwanthpur
['theft', 'arson']
Bengaluru
(12.9791198, 77.5912997)


theft    1
arson    1
Name: Bengaluru, dtype: int64
```

Fig 1

The news article extracted, and the sentences are preprocessed and stop words being removed. The analyzation of crimes in each city, how many times the crime has occurred in a city and longitude and latitude has been also plotted.

| | murder | kidnap | burglary | mug | molest | pickpocket | rape |
|---|---|---|---|---|---|---|---|
| Delhi | 6 | 3 | 2 | 0 | 3 | 0 | 3 |
| Goa | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Thiruvananthapuram | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Salem | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gurgaon | 5 | 3 | 0 | 0 | 0 | 0 | 4 |

Fig 2

```
      murder  kidnap  burglary  mug  molest  pickpocket  rape  arson  \
Delhi      6       3         2    0       3           0     3      0

      hijack  fraud   ...    blackmail  forgery  dacoity  assault  \
Delhi      0      0   ...            1        0        1        1

      assassination  dowry  harassment  riot  stalk  homicide
Delhi             0      0           0     0      0         0

[1 rows x 25 columns]]

In [9]: [df.loc['Gurgaon':'Gurgaon', :]]
Out[9]:
        murder  kidnap  burglary  mug  molest  pickpocket  rape  arson  \
Gurgaon      5       3         0    0       0           0     4      0

        hijack  fraud   ...    blackmail  forgery  dacoity  assault  \
Gurgaon      0      1   ...            1        0        0        3

        assassination  dowry  harassment  riot  stalk  homicide
Gurgaon             0      1           2     0      0         0
```

Fig 3

The city wise table made according to how many times a crime has occurred and we should take preventive measures to avoid them.

The Heat Map is plotted using the longitudes and latitudes. It describes which area is more prone to crimes and this will help the people to be more secure will entering to this zone.

The redder the area is more prone to crimes rather than area marked with green.

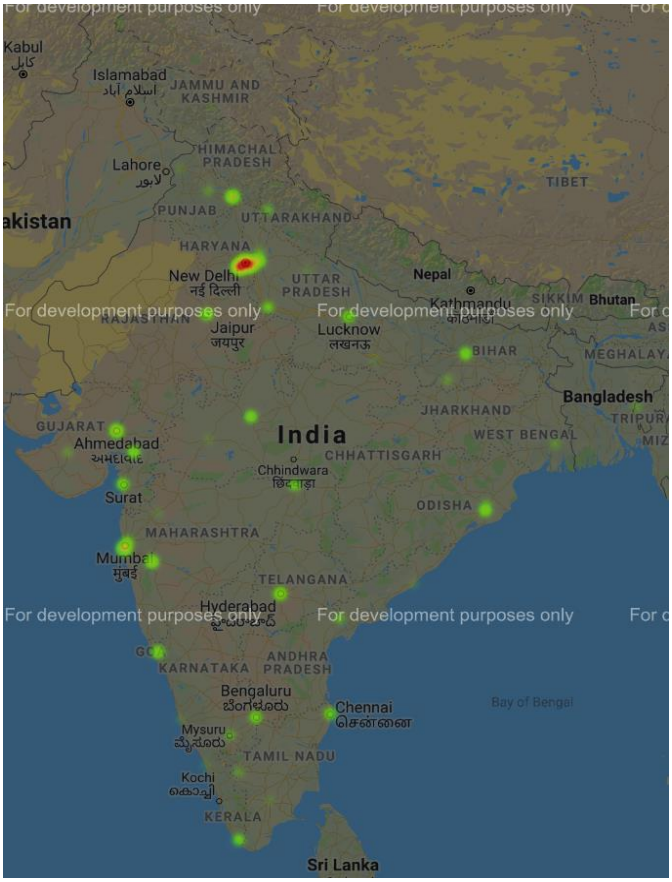


Fig 4

we have made table and plotted Heat Map which will result in analyzation of crime rate in different cities and areas in India which will help citizens to take precautionary measures before entering into the crime prone areas.

```
Done  19
Odisha police under scanner for under-reporting encounter deaths
['robbery', 'dacoity', 'molest', 'rape', 'murder']
Bhubaneswar
(20.2665668, 85.8437586)


robbery    1
dacoity    1
molest     1
rape       1
murder     1
Name: Bhubaneswar, dtype: int64


Done  20
UP sees marginal decline in cases of heinous crime
['murder', 'rape', 'dacoity', 'kidnap', 'burglary', 'riot', 'dowry']
Agra
(27.1752554, 78.0098161)


murder     1
rape       1
dacoity    1
kidnap     1
burglary   1
riot       1
dowry      1
Name: Agra, dtype: int64


Done  21
Man arrested for harassing former office colleague
['stalk']
Hyderabad
(17.3616079, 78.4746286)
```

Fig 5

## VI. CONCLUSION

Crimes in India are rising at an alarming rate because of the factors such as increase in poverty, migration, unemployment, frustration, illiteracy and corruption. Investigating agencies can utilize our proposed data mining tool to ease their crime investigation process. Area – wise crime rate detection can speed up the crime solving process by processing and filtering the voluminous crime data within a short span of time. Thus, aid the law enforcement agencies to enforce the security of citizens of India.

So, to contribute toward combating crimes and to identify criminals, we propose an integrated technology using Text and Web Intelligence for Indian cities.

In this, using different packages and algorithms

## VII. LIMITATIONS

In this, we have referred only one website, but we refer to multiple links then there can be chance of data redundancy like on various links same news will be there. This can cause data errors and our result and analyzation will be tampered.

## VIII. FUTURE WORK

In future, we can enhance data privacy, reliability, accuracy and other security measures of our crime-based data mining system. Password protected user interface is designed to access the Text and Web Intelligence tool. We shall also collaborate with security agencies in India.

## IX. REFRENCES

[1] Hsinchun Chen, Wingyan Chung, Yi Qin, et al. Crime Data Mining: An Overview and Case Studies. Proceeding of the 2003 annual national conference on Digital government research, Boston, M.A, 2003, pp 1-5.

[2] T. Abraham and O. de Vel. Investigating profiling with computer forensic log data and association rules. Proc. Of the IEEE International Conference on Data Mining (ICDM'06), 2006, pp 11-18.

[3] H. F. Lin and J. M. Liang Event based ontology design for retrieving digital archives on human religious self-help consulting. Proc. Of 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, 2005 pp. 453-475.

[4] G. P. Zarri. Semantic web and Knowledge Representation, Proc. Of the 13th International Workshop on Database and Expert System Applications (DEXA'02), 2002, pp. 1529-4188.

[5] Tayal, D.K., Jain, A., Arora, S. et al. AI & Soc (2015) 30: 117. © Springer-Verlag London 2014