Crime Density using News Article Analysis

Shailendra Kr. Gupta - 2016CSB1059
Shreyanshu Shekhar - 2016CSB1060

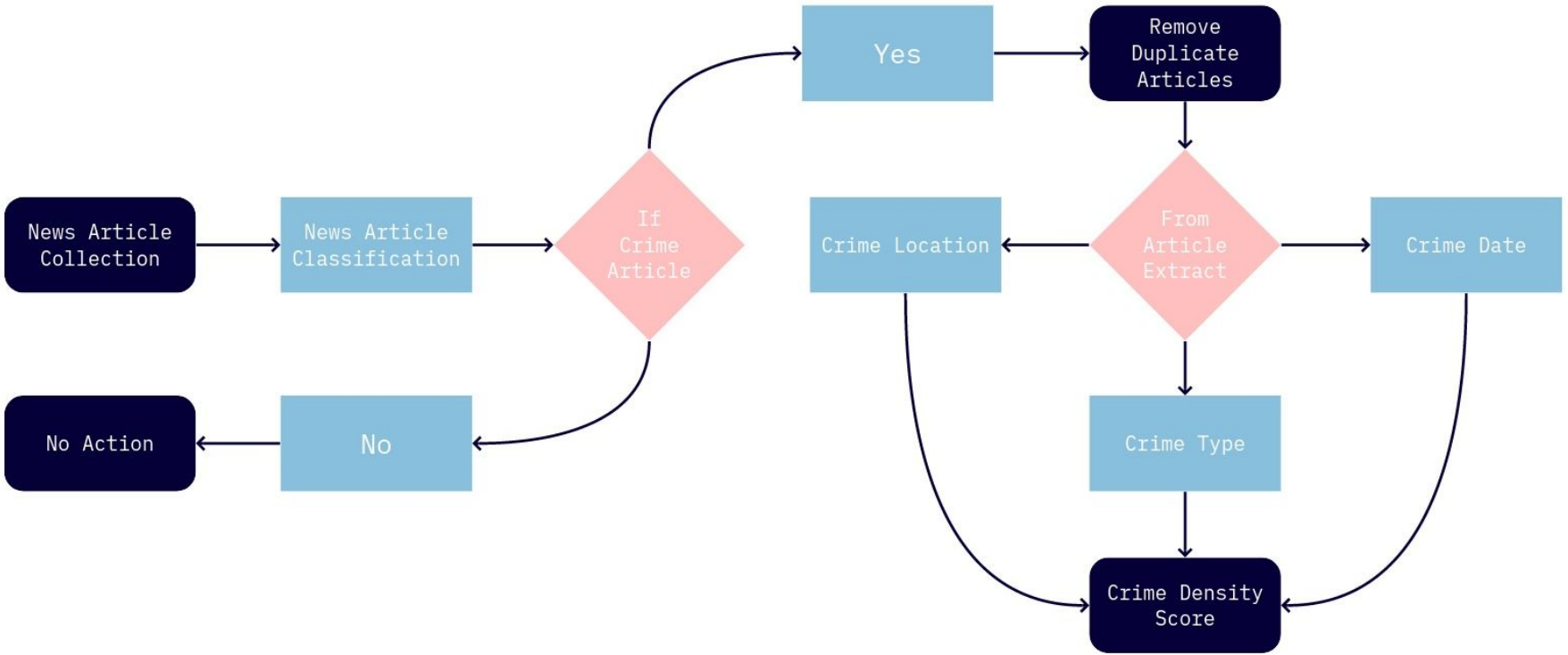
Supervisor:-

Neeraj Goel
Mukesh Saini

Problem Statement

- Main objective:
  - Generate a heat map
  - Based on crime rate
- Use Case:
  - Finding safest route/place
  - Dynamically organising police force
  - Predicting the future occurrence of crime

Proposed Plan



Challenges We Faced

- Data and ground truth (unlabeled data)
- Rating severity of crime (very subjective)
- Finding location of crime (similar location and person name)
- Repeated articles of same crime
- Different coverage of a crime at different locations

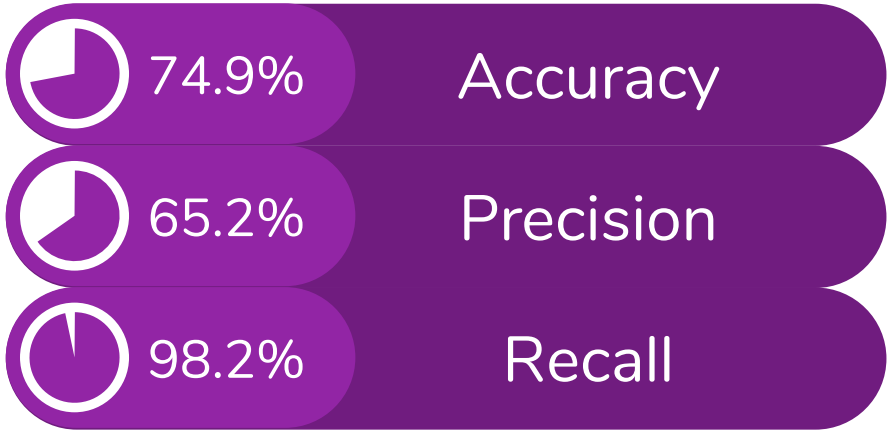
Progress with data

- Crawler
  - Newspaper3k
  - 50k and still collecting
  - News sites:
    - TOI
    - Hindu
    - NDTV
    - News18
    - India Today
    - Hindustan Times
- Interface
  - http://172.26.5.254/login.php
  - Php based web interface
  - Total 1118 tagged
    - 572 Crime
    - 546 Non-Crime

News Classification(method 1)

- Words selection
  - Major Crime Words
  - Synonyms addition
  - Assigning score to each word manually
- Finding similar meaning words in the article by averaging
  - WUP(Wu Palmer) similarity - based on taxonomy depth
- Final **CrimeClassificationScore** calculation using the assigned score to synonyms
- Threshold to segregate crime and non-crime articles(empirically)

Result: News Classification(method 1)



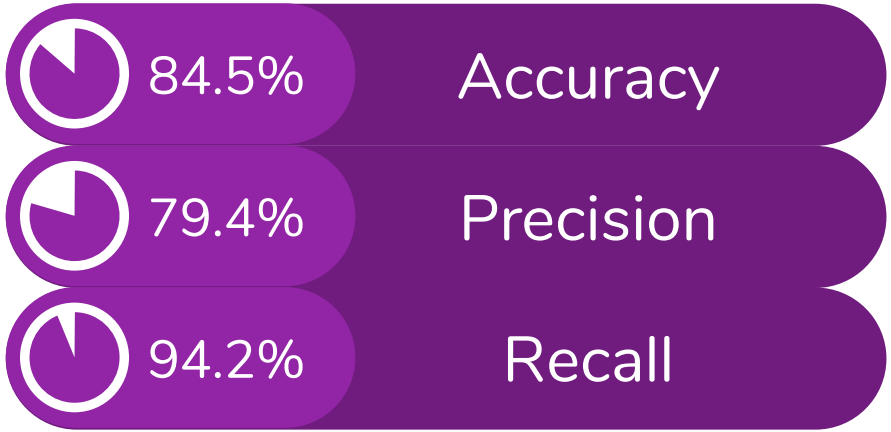
News Classification(method 2)

- Words selection
  - Major Crime Words and their Synonyms
  - Assigning **Ambiguity Score** to each word
    - In how many different context a word can be used in
    - **‘Hit** the road’ or **‘Hit** a man’
  - **Crime Score** of a word = 1 / **Ambiguity Score**
- Analyzing Article’s Title and Body separately
- Even if the text has atleast one of the crime word in our list
  - It will be classified as crime

News Classification(method 2)

- If we are only analyzing the Article's Title
  - **Accuracy** is 81% but **Recall** is 75%
- If we are only analyzing the Article's Body
  - **Accuracy** is 67% but **Recall** is 99%
- We want both, high **Accuracy** and high **Recall**
- So we combined both the method
  - If both model gives different results
  - For example one says Crime and other says Non-Crime
  - Then we apply threshold on Crime Score of that article

Result: News Classification(method 2)



Duplicate News Detection

- Based on following factors:
  - Text Similarity
  - Entities Similarity
    - Persons, Locations
  - Crime Type
- Yet to experiment on weights of each factors on similarity score
- Ground data, need to prepare

Location Extraction

- Entity taggers are widely used tool for information extraction
  - Eg. LBJ Tagger, Stanford tagger, NLTK NER chunker,etc
- We have used following two taggers:
  - NLTK NER chunker
    - Pattern based NER chunker
  - Stanford Tagger
    - Trained CRF model especially for PERSON, LOCATION, ORGANIZATION tags.

Location Extraction

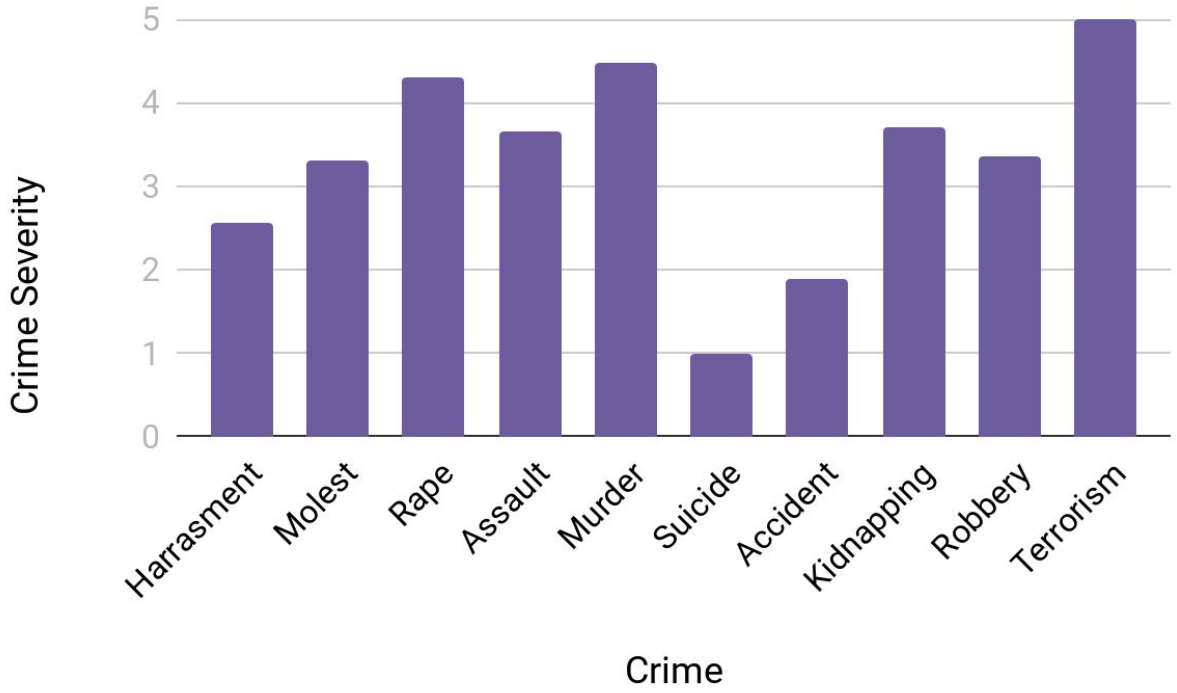
- Created some Lists
  - Created List of some commonly used tags in location names(Tag list)
    - Eg. Patel **nagar**, Pahar**ganj**, Anand **vihar**, Chandni **chowk**, Elliot **beach** etc.
  - Created List of words used before/after location entities(common words list) by reading news articles
    - Eg. in, near, from, at, etc.
- Process
  - Extract out all entities
  - If entity contains any word from tag list -> consider location
  - Else for other entities check the presence of common words
- Performed different combination of taggers

	METHODS	N-1 pred(%)	N-2 pred(%)	Total Location Prediction
1	NLTK NER - direct entity	32	74	403/1022
2	NLTK NER - with processing	58	89	696/1022
3	Stanford - direct entity	53	89	654/1022
4	Stanford - with processing	56	89	668/1022
5	BOTH - direct entity	55	87	660/1022
5	BOTH - with processing	59	90	703/1022

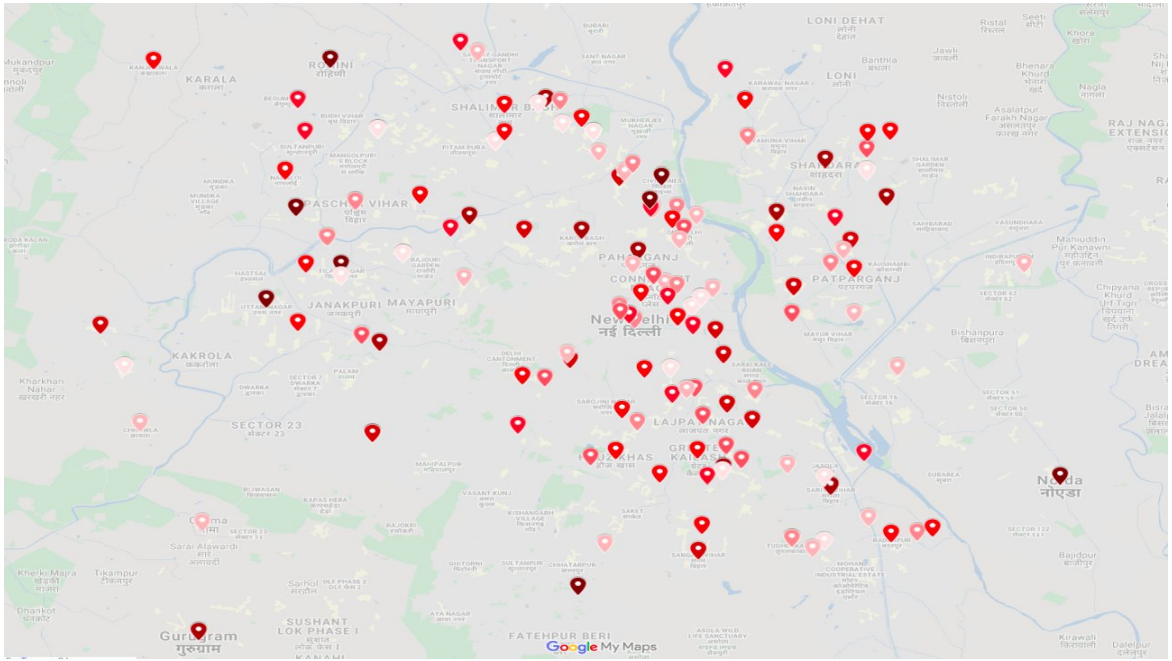
Crime Word Extraction

- Major classes of crime in which we are classifying our Article into are:
  - Harassment • Molest • Rape • Physical Assault • Kidnapping
  - Suicide • Accident • Murder • Theft and Robbery • Terrorism
- Created list of crime words specific to each class of crime
  - E.g. if Homicide appears in the article then it will belong to Murder Class
- Every word in the article votes for their respective Crime Class
- The Class having maximum votes is assigned to that article
- **Accuracy** on 10 Classes is **76.12%**
- There is still some room for improvement

Crime Severity Survey



Crime Density - Heat Map of Delhi



Heat Map

- CS of every Article should decay over time
- Cluster the articles based on their location
- Summed these decayed CS of Articles for every location

$\lambda = \ln 2 / \text{Half Life}$

Half Life = 180 days

$$\text{Decayed cs} = cs * e^{(\lambda * t)}$$

CS = Crime Score t = Age of the Article

Future Work

- Improve on Heat Map
- How to normalize Crime Score of Locations
- Generate ground truth for Duplicate News

Questions and Answers

Thanks!