Mid-Term Report : Disruptive Event Detection using Machine Learning Algorithms

(TECH WARRIORS)
Anagha Sarmalkar
Srishtee Marotkar
Srishti Tiwari

➔ BRIEF PROBLEM DESCRIPTION

Disruptive Events: Recurring violent incidents taking place in a city/town which can be classified together to tap the trends over the years. Higher the violence rating, more is the place unsafe to live/travel.

We will be analyzing the crime data set of India and determine the possibility of various crimes (disruptive events) occuring in a given district, in a given state. For touching all possible crimes, we have narrowed down our demographic to the country of India.

Classification dataset will be used to train model using Naive Bayes and Random Forest Machine Learning algorithms. Naive Bayes Model works best with constant data whereas Random Forest works well with large and real time data. As we are using Static data for project we will be using Naive Bayes algorithm for predicting disruptive event. But the fact is Crime data should be read in real-time because according to 2015 crime clock statistics released by FBI [1] every second there are huge change in crime count, therefore though we do not have real-time data we are considering to create real-time environment by passing data in batches (this approach needs more research which we are planning to do after dataset is in place) and use Random Forest algorithm.

From This Project we are expecting to predict disruptive events from crime data in India and comparing Naive Bayes and Random Forest machine Learning algorithm performances with static and artificially created real-time environment respectively.

➔ LITERATURE SURVEY

We placed major emphasis on data gathering and analysis for this project report.

- ◆ Analysis of Time Series Data.
  Since we were dealing with the crimes in India over the years, we realized it was important to capture the trends in crime data. It was important to come up with the crime rate, normalize it and then smoothen it using median smoothing. The main advantage of using median as compared to moving average smoothing is that its results are less biased by outliers. [2]

➔ ACCOMPLISHED MILESTONES FROM THE ORIGINAL PLAN

We have accomplished **Dataset preparation and preprocessing**. Following are the steps:

1. **Searching Dataset** : We have picked up 'Crime in India' dataset from Kaggle. This datasets gives us total count of crime from 2001-2010 for different types of crimes. [4][5]

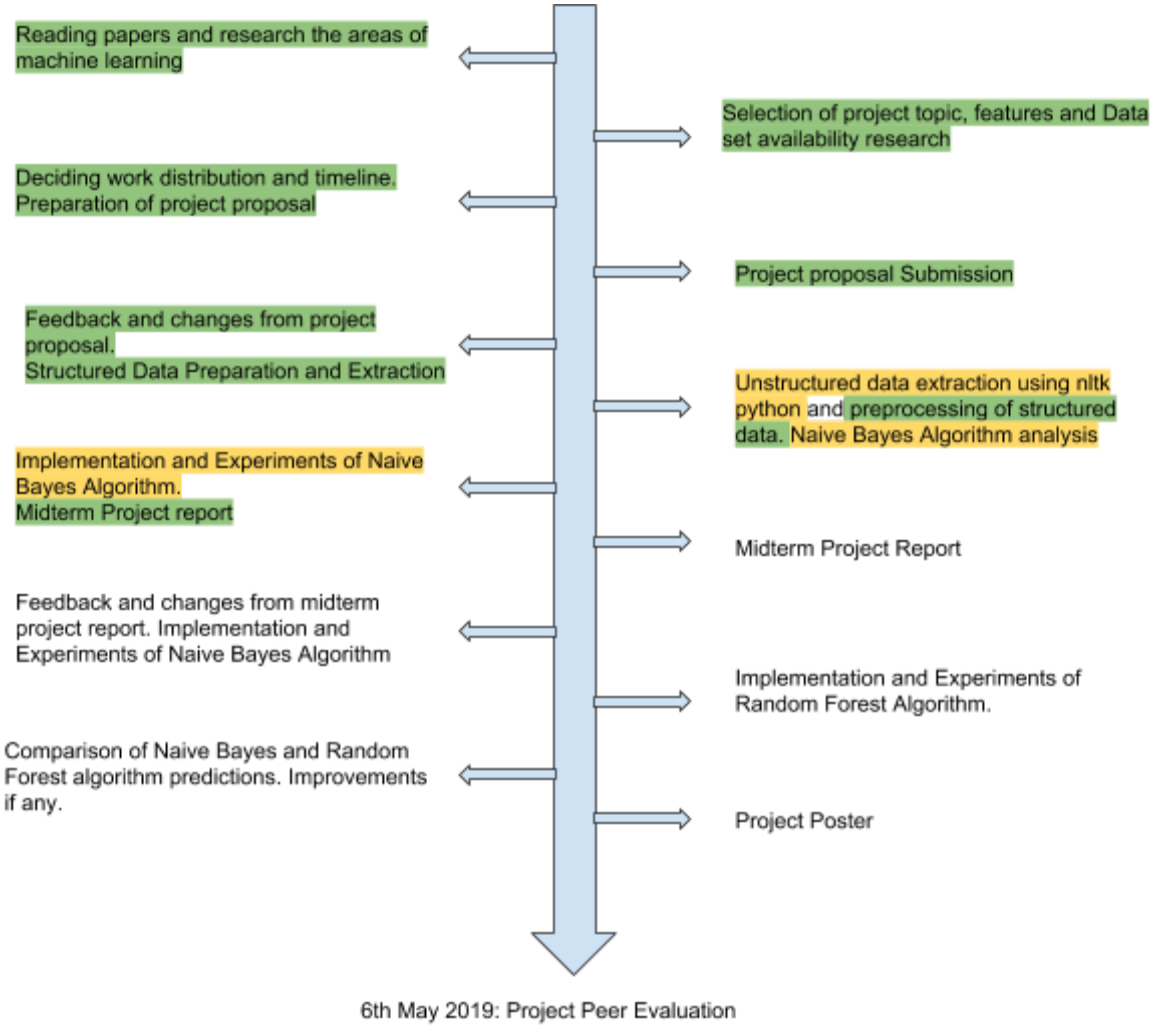
2. **Preparation of Dataset** : We have started by picking up data for 2001 -2003. Then Basic preprocessing is done for State and District Column using Label Encoders. [6]

3. **Calculating Crime Rate and Violence Rate** : This part includes calculation Crime Rate, normalizing Crime Rate and calculating Violence Rate. Detailed Description is given in **APPROACHES** section.

➔ DIFFICULTIES OR PROBLEMS EXPERIENCED

1. The data preparation part is turning out to be time consuming. Searching Crime dataset was easy but it took time to drill population data which is required to calculate crime rate formula explained in below Section.
2. We are working on calculating 'Violence Rate' from 'Normalized Crime Rate'. Problem faced to come up with approach to normalize 'Normalized Crime Rate' (which is calculated over state and district from 2001-2010) to fit all pairs of State and District for all years. (This is in progress)
3. As one of our teammate (Madhuri) withdrew from course, we had to change project timeline and task distribution.

➔ MODIFIED PLAN/ TIMELINE FOR REST WORK



Abc :: Activities in progress

Abc :: Activities completed

➔ DISTRIBUTION OF WORK

Sr.no.	Tasks	Anagha	Srishee	Srishti
1.	Reading Research papers,articles and general analysis for project idea.	✓	✓	✓
2.	Selection of project topics, features and data availability research		✓	✓
3.	Deciding work distribution and estimate. Preparation of project proposal	✓	✓	✓
4.	Structured data preparation and extraction	✓		
5.	Unstructured data extraction using nltk python and preprocessing of structured data. Naive Bayes Algorithm analysis		✓	✓
6.	Implementation and Experiments of Naive Bayes Algorithm. Preparation of Midterm Project report	✓		✓
7.	Preparation of Midterm Project report And submission	✓	✓	✓
8.	Feedback and changes from midterm project report. Analysis of Random Forest Algorithm	✓	✓	✓
9.	Implementation and Experiments of Random Forest Algorithm.		✓	✓
10.	Comparison of Naive Bayes and Random Forest algorithm predictions. Improvements if any.	✓	✓	✓
11.	Project Poster preparation	✓	✓	✓
12.	Structuring of Final project report	✓	✓	✓

➔ (DETAILED) APPROACHES

- ◆ Data Gathering:
  We gathered crimes that have been committed in the country of India and have been reported. (penalized under IPC: Indian Penal Code)
  This data includes a wide range of criminal offences and has been collected over the span of 10+ years.

- ◆ Data Cleaning:
  There was some inconsistency with the individual values of crimes committed per location and the column “TOTAL IPC CRIMES” which seems to be the summation of all the individual crimes. For the sake of being on the safer side, we will calculate the total crimes committed and disregard this column.

- ◆ Exploratory Data Analysis :
  - Crime Rate :
    A crime rate describes the number of crimes reported to law enforcement agencies per 100,000 total population. A crime rate is calculated by dividing the number of reported crimes by the total population; the result is multiplied by 100,000. For example, in 2010 there were 58,100 robberies in California and the population was 38,826,898. This equals a robbery crime rate of 149.6 per 100,000 general population.

    $$\frac{58,100}{38,826,898} = 0.0014964 \times 100,000 = 149.6$$

    We analyzed the data and calculated the rate of every crime that happened for every year [3].

  - Population per year per state per district.
    We calculated the population per year for calculating the crime rate by interpolation using the population from the census. In India census is calculated after every 10 years so we have the census data for year 2001 and 2011. This data comprises of the population count of every state and every district in that state. We will select the states and districts depending upon the ones present in the crime dataset available to us. [7] [8]

  - Normalized Crime Rate
    The ‘Normalized Crime Rate’ is calculated by normalizing the data in the crime rate using min-max scaling to fit the crime rate in a certain range.

- Violence rating
  The violence rating is calculated by choosing the median value of the crime rate number as the threshold value. We have chosen the median value. Choosing the crime rate number greater than this value will have a violence rating of 1 and less than this value will have a violence rating of 0. VR = 1 will indicate more chances of acts of violence taking place in the particular location. VR = 0 will indicate the opposite.

- Total dataset
  Thus, we get a total data of (states*districts*years) number of rows and (no. of crimes + crime_total + no. of years + population per year + crime rate + Normalized Crime Rate + violence rating) number of columns.

➔ DIFFERENCE/NOVELTY

The research on this topic is ongoing and a few papers have been published. From the papers we studied, we observed that lot of them have used clustering algorithms to detect crime patterns based on regions. We are trying to use Crime dataset and formulate crime rate and violence rate. Violence Rate will be classifier, here violence rate can help us predict the expected disruptive event beforehand to take necessary actions to avoid harsh resulting activities. We are planning to train model using Baye navie and Random Forest Classification Machine Learning Algorithm.

➔ REFLECTIONS (RESPONSE TO THE INSTRUCTOR'S FEEDBACK)

- **Problem Statement: "***good but wouldn't it be too small number of riot cases in data? how are you going to handle the imbalance?"*
  Solution: Riot cases are very less and there is no concrete source from where we can get that information. After taking your feedback, we have shifted our focus to a broader problem statement which includes calculating violence rating of a given place.

- **Steps & Approaches:** *"good intro to data but method is not clearly described"*
  Solution: We have described the methods for data preparation in this midterm report.

- **Steps & Approaches -- Dataset -- Data Collection -- Unstructured Data:** *“riot happens in past. current data might not help. Twitter api won't provide you to search over long history.”*
  Solution: Addressed in first answer.

- **Survey on the area:** *“good summary but it is missing how those are related to your works. pros /cons? limitations to come up with your method? ....”*
  Solution: Through Introduction and Approaches section we have tried answering this question.

➔ REFERENCES

[1]https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/resource-pages/crime-clock
[2]http://www.statsoft.com/Textbook/Time-Series-Analysis#trend
[3]https://oag.ca.gov/sites/all/files/agweb/pdfs/cjsc/prof10/formulas.pdf
[4]https://github.com/smarotka/Disruptive-Event-Predictor/blob/master/Datasets/crime-in-india/crime/01_District_wise_crimes_committed_IPC_2001_2012.csv
[5]https://www.kaggle.com/rajanand/crime-in-india
[6]https://github.com/smarotka/Disruptive-Event-Predictor/blob/master/crime_dataset_prep.ipynb
[7] https://www.kaggle.com/danofer/india-census
[8] https://www.kaggle.com/bazuka/census2001
[9] https://ieeexplore.ieee.org/abstract/document/4053200
[10] https://dl.acm.org/citation.cfm?id=3123282
[11] https://journals.sagepub.com/doi/pdf/10.1177/0165551517698564
[12]https://www.researchgate.net/publication/315871444_Can_We_Predict_a_Riot_Disruptive_Event_Detection_Using_Twitter
[13] https://www.sciencedaily.com/releases/2017/06/170626093522.htm
[14]https://www.slideshare.net/apnic/machine-learning-approaches-for-crime-pattern-detection
[15]https://oag.ca.gov/sites/all/files/agweb/pdfs/cjsc/prof10/formulas.pdf

➔ GITHUB REPO

https://github.com/smarotka/Disruptive-Event-Predictor.git