Disruptive Event Detection using Classification Machine Learning Algorithm

Team Members (Group 15)

Anagha Sarmalkar, Madhuri Pawle, Srishtee Marotkar, Srishti Tiwari

Problem Statement

If there is a protest happening somewhere, can it turn into a riot?
In our project we will be trying to predict riots based on attributes such as crime rate, number of deaths, number of participants, region, issues, target etc. using machine learning algorithms.

Overview

Polarization of political point of views is happening due to social media. Social media has become a very good medium to bring together the people with similar ideologies. While this is a good thing, things can go south spirally and cause protests, violence, riots, etc. It also can provide easy access to weapons, drugs etc. The information gathering from social media can help predicting the violent disturbances before the situation worsens and predicting it in early stage can help alert the security forces to control any such violences.

Using regional crime statistics information available on various dataset providers we will detect crime patterns. Studying these patterns will help us in extraction of attributes. These attributes will play important role in predicting crime in particular region by calculating crime rate and violence rate. Based on these parameters we will have output as binary classification indicating violent disturbance predictor (0 for no and 1 for yes). After training our model, we are targeting that when we test the model it should correctly predict the upcoming violent disturbance event in specific region which can help us in taking necessary measures to avoid major consequences like riots, mass protest etc.

Steps & Approaches

1. Dataset

In this project we wish to predict whether a protest can become a riot using various data sets. We will need to extract relevant attributes from various datasets for our project. This will require a lot of data preprocessing. We have segregated a few data sources as listed below:

Data Collection

I. Structured Data

- We will be collecting structured Dataset from Social Conflict Analysis database from University of Texas Austin https://www.strausscenter.org/scad.html
- We will also use www.knoema.com for getting crime-rate, issues, targets etc.

II. Unstructured Data

- We have multiple sources to collect unstructured data such as: Facebook, Twitter, youtube, news api etc.
- We plan to use search engine api for getting the news data (any of the search engine api)
- We also plan to use Twitter api to collect real-time data
- Using the above unstructured data we will make it structured using natural language toolkit and positive and negative words collection.

2. Classification Machine Learning algorithm for violent disturbance detection

- Algorithms to be used : Naive Bayes & Random Forest
- After Data preprocessing and cleaning we will be applying above classification algorithms.
- The reason for selecting above two algorithms:

1. **Naive Bayes** : This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

2. **Random Forest** : This algorithm can handle large dataset with high dimensionality and output Importance of Variable

3. Survey on the area

To reach to the idea of Violent Disturbance Prediction,we divided our survey as in two parts.

- Firstly, understand the crime data pattern so to understand the pattern of factors to be considered while calculating violence rate or crime rate which we will use as a basis to classify data into binary output classes.
- Secondly, We studied papers on existing Crime detection machine learning models using clustering algorithms.

Below are the papers and articles we studied:

1. Machine Learning approaches for Detect Crime pattern

This paper highlights the various machine learning approaches that were used to detect crime patterns.It discusses about the different crime detection models currently being used and the techniques employed. It also discusses the current research being done in this field. Data had been collected from major sources like criminal records from law enforcement authorities,social media, IoT devices, newspaper articles, call data records etc. After analyzing these data sources, geographical hotspots have been identified. This paper also discusses the importance of various other seemingly unrelated factors like weather conditions to the occurence of crimes. Several methods of operations have been discussed, which would be helpful pointers for us while developing our model.

Data extraction using supervised learning approach called support vector machine is most optimal amongst all other text classification algorithms. Most of the researches have employed K means clustering for classification and pattern recognition because of its reliability and relativity to the criminal data. But K-means would be productive only if the data is not noisy and the number of clusters are less.

2. Using Machine Learning Algorithms To Analyze Crime Data

This paper aims to analyze how effective and accurate the machine learning algorithms used in data mining analysis are, in predicting violent crimes patterns. The authors use WEKA (Waikato Environment for Knowledge Analysis), an open source data mining software to compare the three algorithms - Linear Regression, Additive Regression and Decision Stump Algorithm, that were used to predict violent crimes like murder, rape, robbery and aggravated assault. The machine learning algorithms were trained using the dataset from neighborhoodscout.com which was provided by the FBI and consisted of crime statistical data of the state Mississippi, for the year 2013 and also the Communities and Crime Unnormalized Dataset provided by the University of California-Irvine repository. After the implementation of the algorithm, WEKA outputs five metrics - Correlation Coefficient, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error, that are used to determine the efficiency and effectiveness of the three algorithms. The algorithm that had the greatest correlation coefficient value also generated the lowest error values among the three algorithms. Based on the comparison, Linear Regression performed the best as it could handle the randomness of the data. On the other hand, Decision Stump Algorithm performed poorly.

3. Disruptive Event Detection Using Twitter

This article focuses on how social media as a medium to detect crime pattern and identify upcoming violent disturbances. People are having freedom to post their view on social medias like Facebook, Twitter and YouTube. Due to real-time textual data which may have informal lingo, irregular and abbreviated word, spelling and grammatical errors makes it difficult to track down smaller event or minor consequences, Therefore this paper focuses on tracking down larger events using supervised learning and then predicting smaller crime pattern that may result using unsupervised approaches like clustering.

Take away from reading, imbibe how real-time news, social media article and post textual data is used as medium to identify or predict future mishaps. This paper plays an important in giving us direction to how we can use crime pattern data to help predict violent disturbances likes, riots, rampage and uproar in specific group of people and how can we provide security beforehand to avoid tumult in specific geographical area.

4. Crime Analysis and Prediction Using Data Mining

This paper demonstrate the use of clustering algorithms for data mining to detect crime patterns to speed up process of solving crimes. Paper focus more towards main kind of crimes and not to depth of criminal justice. Using clustering referring to specific geographical area and lost of crime happening there. With real-time crime data from Sheriff's department they identifies the type of suspect and type of victim of a crime. Suspect of crime is either identified or unidentified but most of the time victim is identifiable as in most cases person reporting crime is victim, thus large number of crimes making the job for crime detectives easier.

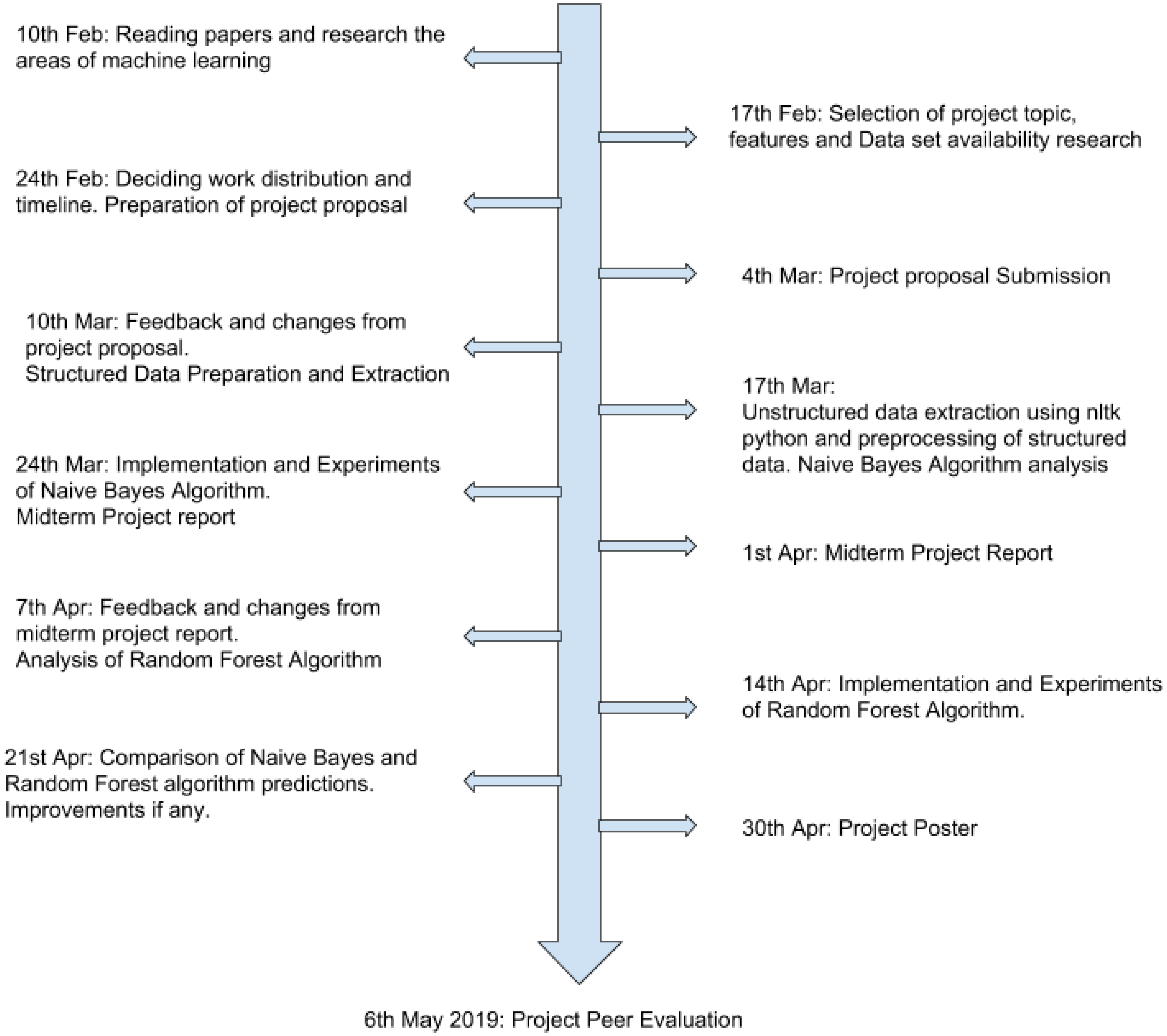
Take away from this reading, While analyzing crime datasets, the type of victim and suspect, then identifying there count in geospatial area will play an important role to identify hotspot for specific type of crime pattern.

5. Crime Pattern Detection Using Data Mining

Unlike paper no. 4, this paper speaks that it is difficult to identify the type of victim and suspect of committing it but more on place that has high probability of its occurrence. The results of this paper has showed that providing security in particular area has proved to be reducing factor for crime. Drawback for this technique is that it rely on only existing crime type and data available, there is no way to predict places for crime happening in future. Based on this characteristics they have used Bayes theorem which shows 90% of accuracy.

Take away from this paper, gives an insight of how important it is to understand the pattern of crime data and understand that crime pattern changes over time so we need to device a model which reads data from real-time system and our model learns from real time data.

Weekly Timeline for workplan



10th Feb: Reading papers and research the areas of machine learning

17th Feb: Selection of project topic, features and Data set availability research

24th Feb: Deciding work distribution and timeline. Preparation of project proposal

4th Mar: Project proposal Submission

10th Mar: Feedback and changes from project proposal.
Structured Data Preparation and Extraction

17th Mar:
Unstructured data extraction using nltk python and preprocessing of structured data. Naive Bayes Algorithm analysis

24th Mar: Implementation and Experiments of Naive Bayes Algorithm.
Midterm Project report

1st Apr: Midterm Project Report

7th Apr: Feedback and changes from midterm project report.
Analysis of Random Forest Algorithm

14th Apr: Implementation and Experiments of Random Forest Algorithm.

21st Apr: Comparison of Naive Bayes and Random Forest algorithm predictions.
Improvements if any.

30th Apr: Project Poster

6th May 2019: Project Peer Evaluation

Distribution of Work

Sr.no.	Tasks	Anagha	Madhuri	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	✓	✓	✓	✓

Question we want to answer during the project

- How a machine Learning algorithm can be used to solve complex social problem whose data is hard to structure.
- Whether the attributes selected to determine crime pattern in specific region can accurately predict violent disturbances before the situation worsens.
- As crime patterns varies with real-time data by using Naive Bayes and Random Forest Algorithms we can compare and study how much data is required by both of these algorithms.

Expectation of what we will learn from the project

- Gathering different crime datasets, study them and extracts attributes which plays vital role in predicting crime rate or violence rate.
- Performing data pre-processing techniques, to make the data ready for analysis.
- This project will give us insight of how to compare and choose best machine learning algorithms based on data availability.
- We will get to learn different phase of developing Machine learning model from Data collection, data pattern identification, data preprocessing, implementing classification algorithms, experiments and compare prediction results.

Is our idea novel?

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 predict the whether any crime pattern will result into mass violence like riots, man made accidents etc. We will attempt to combine crime statistics and real time data to achieve this. From the readings we observed that the accuracy would increase significantly if more number of features are included in the context of event discovery such as social network features (community influence detection), visual features (images and video), and semantic features. We will try to incorporate this into our model and improve its efficiency to predict beforehand the violent disturbance rampage like riots, mass protests etc. thus, taking necessary actions to prevent situation from worsening.

Future Scope

1. The social media data extraction can be improved by adding multiple sources of information and ways to process information can be highly enhanced.
2. This project can also be extended to extract violence information from images and videos.

References

1. Crime Pattern Detection Using Data Mining https://ieeexplore.ieee.org/abstract/document/4053200
2. Protest Activity Detection and Perceived Violence Estimation from Social Media Images https://dl.acm.org/citation.cfm?id=3123282
3. A survey on real-time event detection from the Twitter data stream https://journals.sagepub.com/doi/pdf/10.1177/0165551517698564
4. Can We Predict a Riot? Disruptive Event Detection Using Twitter https://www.researchgate.net/publication/315871444_Can_We_Predict_a_Riot_Disruptive_Event_Detection_Using_Twitter
5. Twitter-monitoring system detects riots far quicker than police reports https://www.sciencedaily.com/releases/2017/06/170626093522.htm
6. Machine Learning Approaches for Crime Pattern Detection https://www.slideshare.net/apnic/machine-learning-approaches-for-crime-pattern-detection