Disruptive Event Detection using Classification Machine Learning Algorithms

Anagha Sarmalkar, Srishtee Marotkar and Srishti Tiwari

Under the guidance of Dr. Minwoo Jake Lee, Assistant Professor, Department of Computer Science

University of North Carolina at Charlotte

1. Abstract

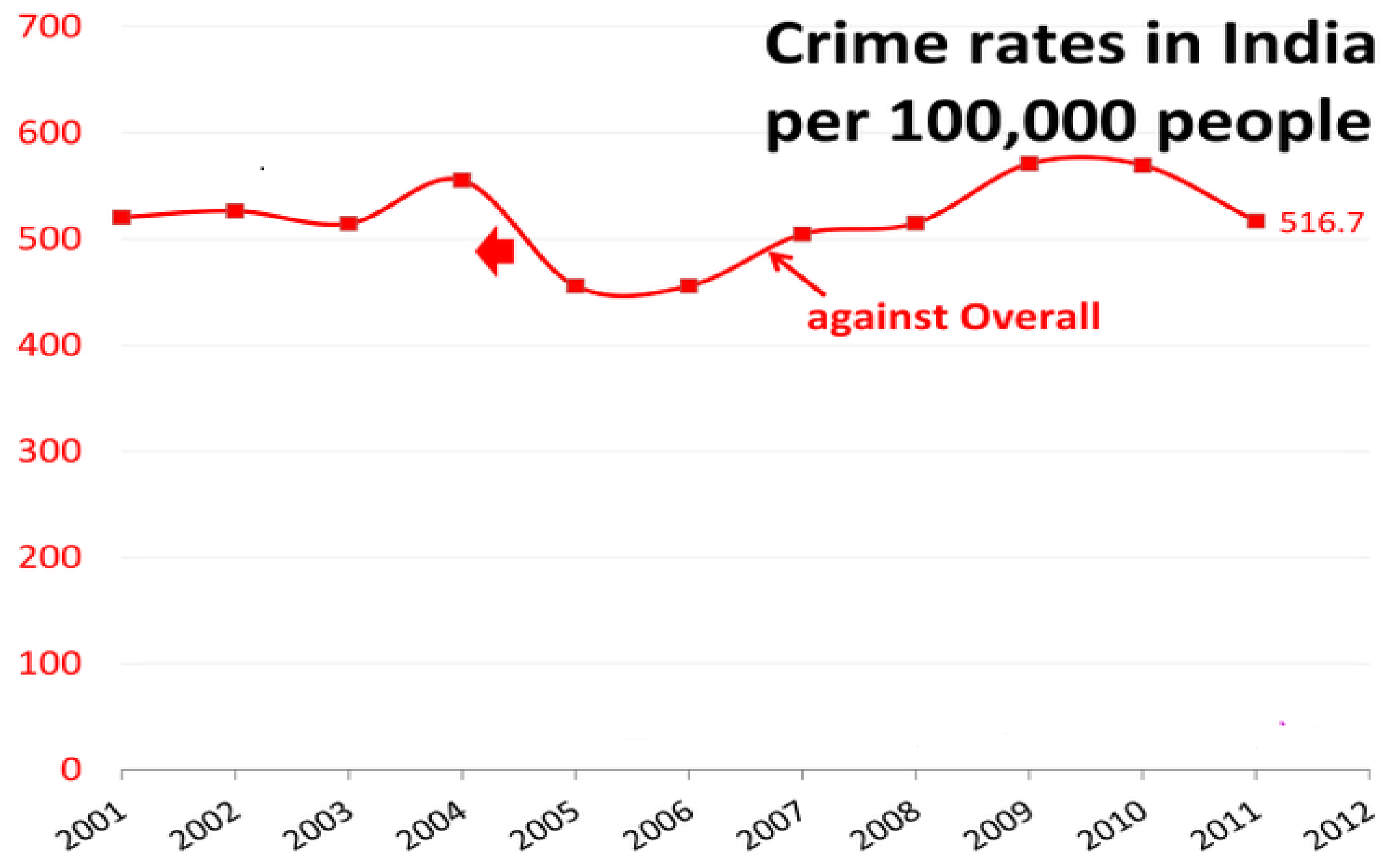
Recurring violent incidents taking place in a city/town can be classified together to tap the trends over the years. Higher the violence rating, more are the chances of people come out in rage which leads to incitement to protest. And protest that turn violent may lead to riots, mass protest etc. There is huge data of criminal activities available to us which can be used to get data-driven insights to help know what is coming. It is extremely important to utilize this knowledge to prevent/control any such activities even before they occur.

Considering this as an inspiration we leverage the vast dataset of criminal activities (reported in India) and employ machine learning models for determining if disruptive events may occur in a certain area.

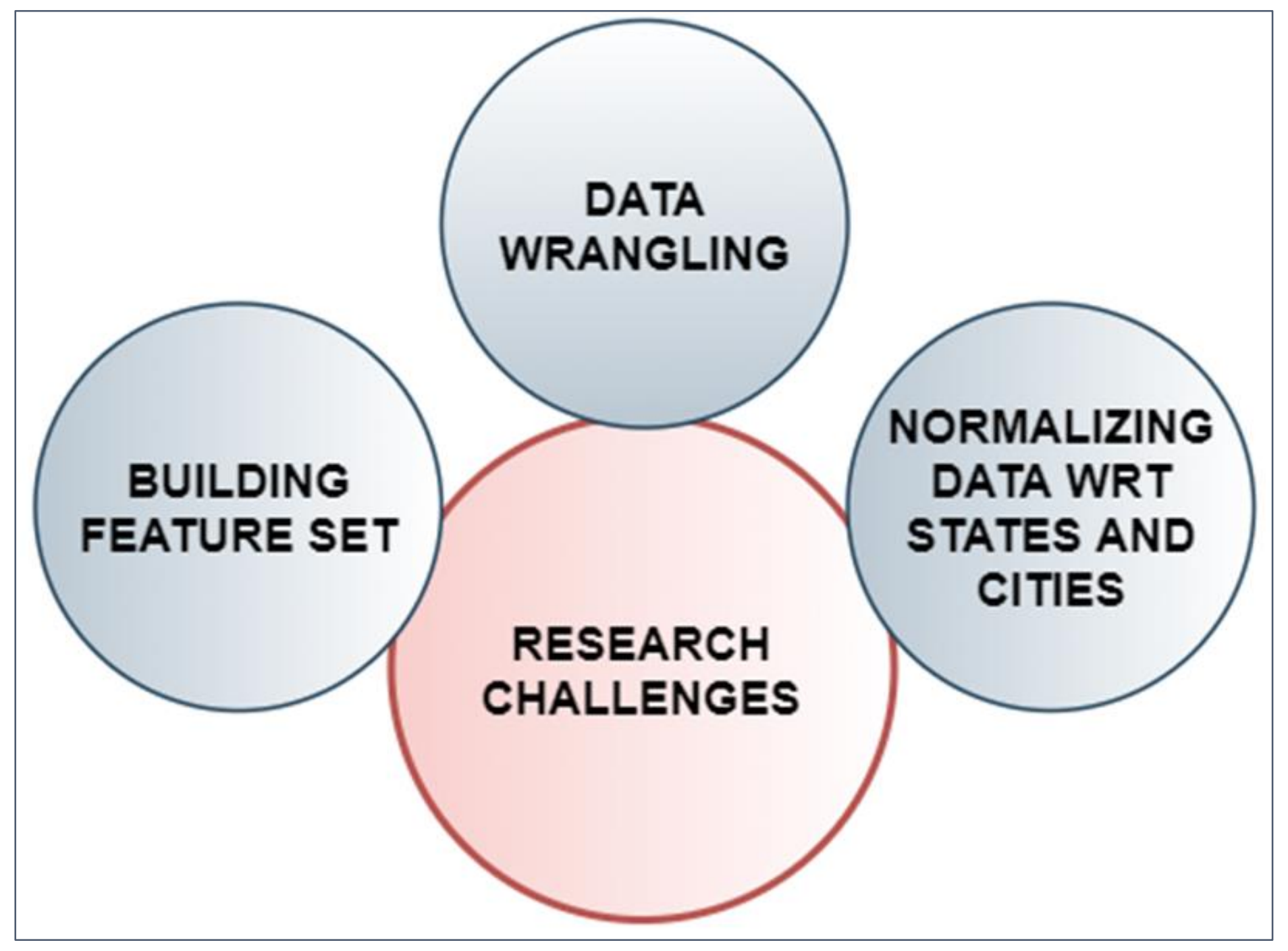
Using regional crime statistics information available on various dataset providers we detect crime patterns across regions in India. Studying these patterns help us in extraction of attributes. These attributes plays an important role in predicting crime rate in particular region and crime pattern over ten years helps in rating any region as ‘0’ or ‘1’ which indicates the occurrence of a disruptive event occurrence as ‘Yes or ‘No’.

2. Motivation

- Many papers have been focusing on predicting crimes that may happen in specific region using clustering machine learning algorithms.
- It is equally important to notice that high crime rate leads to disruptive events like riots, mass protest, employee strikes etc. Thus our motivation is to predict upcoming disruptive event beforehand which can alert the police department to take necessary actions before situation worsens.

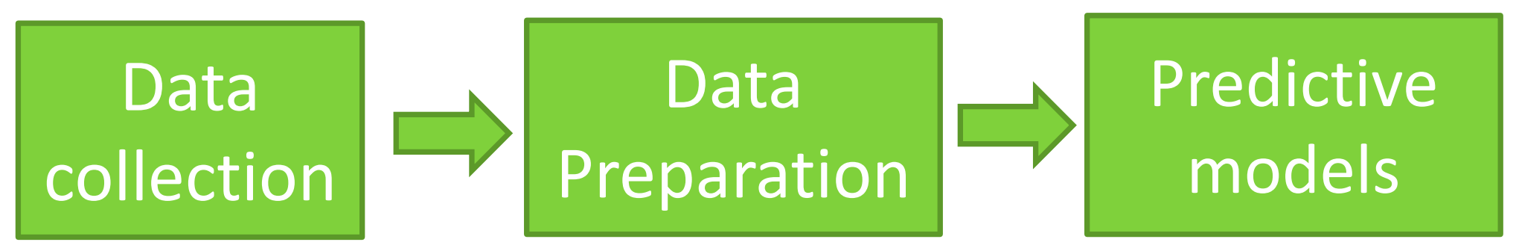


3. Challenges



Research Challenges

- In todays world the rate at which crime is happening and rate at which it is reported and recorded has made doors open to huge data availability.
- Having bulk data is not enough, calculating Violence rating by segregating relevant data that contributes towards accurate crime rate is main challenge.



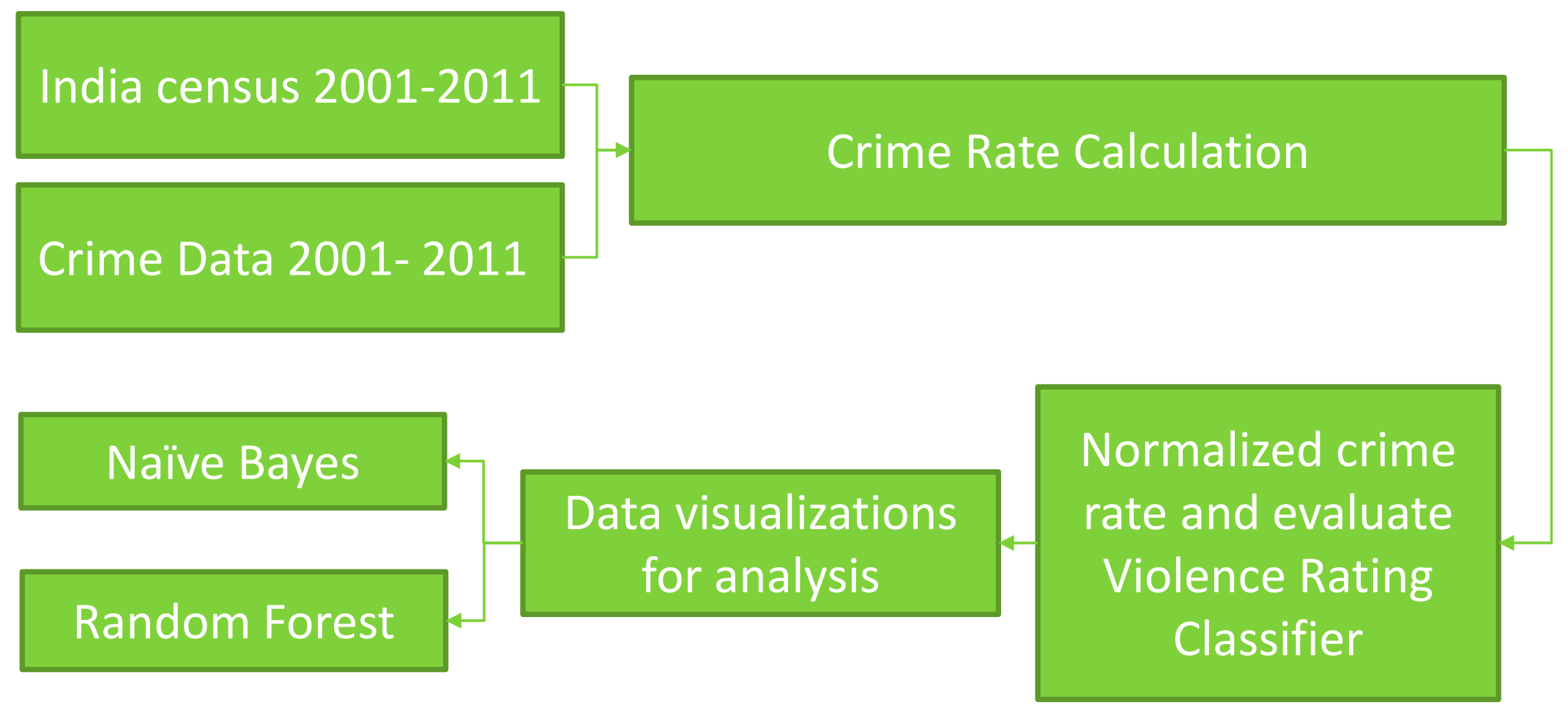
Data collection → Data Preparation → Predictive models

Problem

- Crime Dataset of India and Indian census data from 2001 to 2011 were two different datasets from different resources.
- Main challenge was to mismatch in categorical attributes in both datasets like ‘State’ and ‘District’ of both the datasets, which is key for merging datasets.
- Identifying the robust machine learning algorithm which is capable of handling randomness in crime data which occurs because of live reporting and needs quick attention.

4. Solution

- Crime Rate in any region is total crime reported by total population. Thus, Total Crime in India is merged with Indian census from 2001-2011.
- To suppress outliers Crime Rate is normalized to fit between 0 and 1.



India census 2001-2011

Crime Data 2001- 2011

Crime Rate Calculation

Normalized crime rate and evaluate Violence Rating Classifier

Data visualizations for analysis

Naïve Bayes

Random Forest

Flow of Violence Rating Dataset Preparation

- High Crime Rate is one of the major reason for disruptive events occurrence. Considering this as base to calculate Violence Rating , the median of Normalized Crime Rate is fixed as a threshold over which there are high chances of disruptive events. Thus all the normalized Crime Rate over threshold gives ‘1’ and rest is ‘0’ value as Violence Rating classifier.

Choice of Algorithms

- **Naïve Bayes:** We selected naive bayes because our training set is small, therefore high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN), since the latter will overfit. However as the training set grows low bias/high variance datasets will give better results (they have lower asymptotic error), high bias classifiers won't be powerful enough to provide accurate models. Naive Bayes converges quicker than discriminative models like logistic regression, hence lesser training data is required.

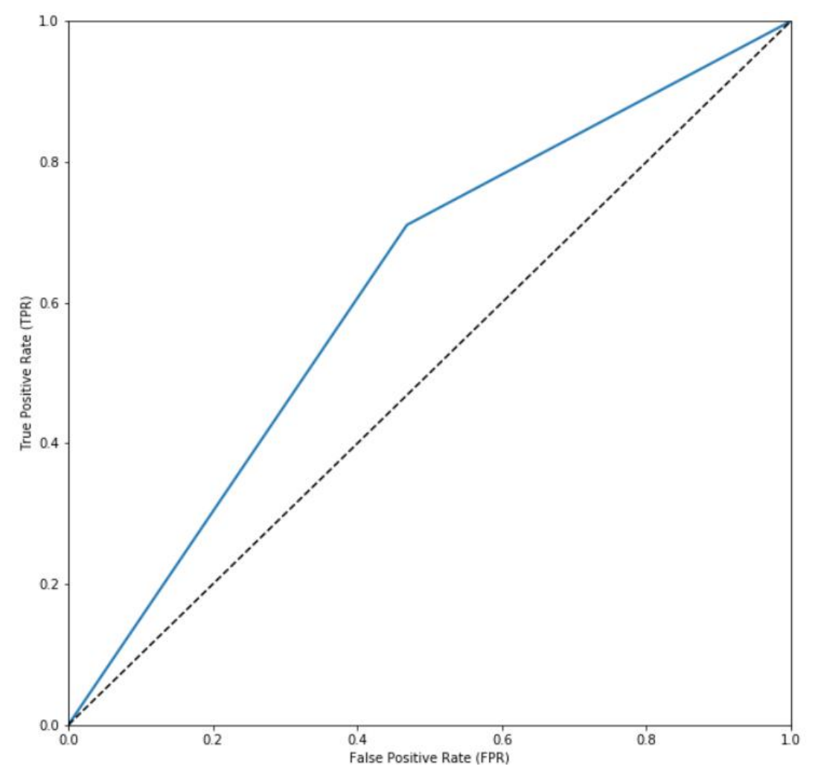
- **Random forest:** Random forests are among the most popular machine learning methods because of relatively good accuracy, robustness and ease of use. They can handle the problem of missing values and maintain the accuracy of a large proportion of data. The model is quite generalized since the presence of more number of trees will not allow overfitting in the model. Since our dataset had high dimensionality we decided to use it.

However, better data often beats better algorithms, and designing good features goes a long way. Hence we equally focused on making our cleaned and processed dataset as thoroughly as possible.

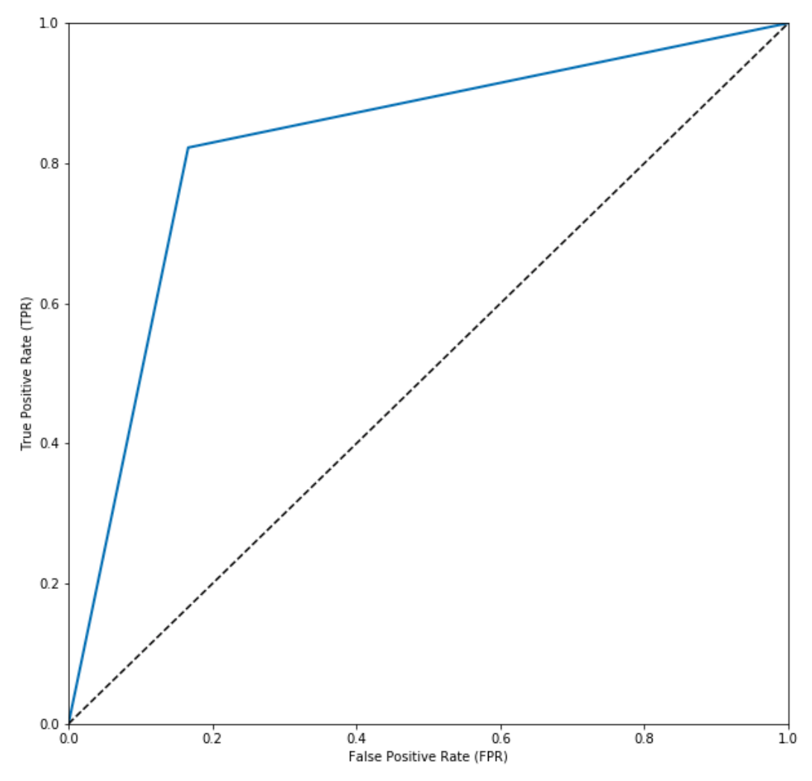
5. Evaluation

	Naïve Bayes	Random Forest
Accuracy	0.56	0.83
Precision	0.71	0.82
Recall	0.19	0.85
Specificity	0.91	0.81
F1 Score	348	134
ROC	0.62	0.83

Evaluation metrics comparison of Naïve Bayes and Random Forest Classifier



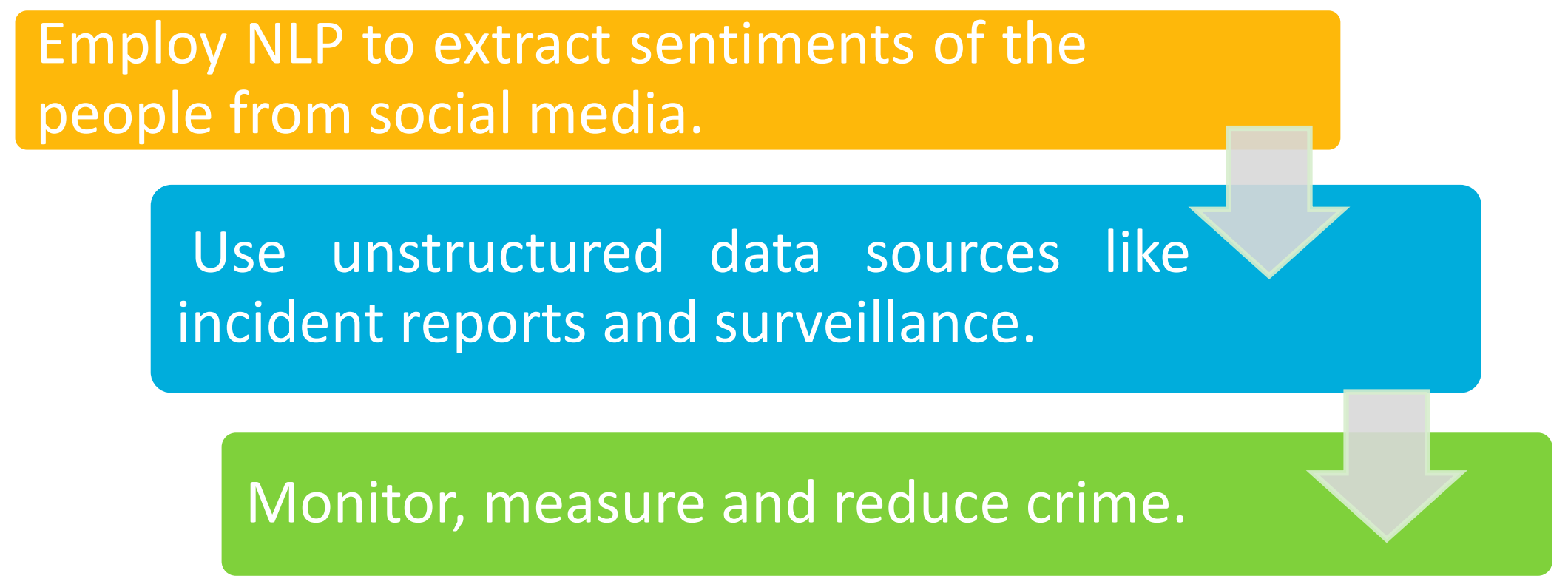
ROC Curve for Naïve Bayes



ROC Curve for Random Forest

- Comparing the result of both classification algorithms, the random forest classifier shows better accuracy.
- ROC curve for random forest lies farther from that of purely random classifier compared to Naive Bayes.
- Random forest can get the relative feature importance, which helps in selecting the most contributing crime leading to disruptive events.

Future work



Employ NLP to extract sentiments of the people from social media.

Use unstructured data sources like incident reports and surveillance.

Monitor, measure and reduce crime.

6. References

[1] More information about this project, visit: https://github.com/smarotka/Disruptive-Event-Predictor

[2] Crime dataset for India: https://www.kaggle.com/rajanand/crime-in-india

[3] Census dataset for India (2011): https://www.kaggle.com/danofer/india-census

[4] Census dataset for India (2001): https://www.kaggle.com/bazuka/census2001

RESEARCH POSTER PRESENTATION DESIGN © 2011
www.PosterPresentations.com