Predictive Analytics

On Social Unrest due to Election Violence

Sonal Singh
Department of Computer Science
and Engineering
University at Buffalo
Buffalo NY USA
ssingh65@buffalo.com

ABSTRACT

People participate in demonstrations when they want their voice to be heard and to exercise their democratic rights. Studies have shown that the protests in the period before an upcoming election can increase or decrease a candidate's vote by enough to change the final outcome. The Indian anti-corruption movement, commencing in 2011 and followed by a series of protests and demonstrations, concluded with the electoral debut Aam Aadmi Party emerging as the second largest party in the 2013 Delhi legislative Assembly Election. Sometimes, due to a large number of participants, these protests can also cause election violence which can be violent, destructive and hence costly in various ways.

Therefore, it is important to predict such protests in advance to safeguard against such damages. The goal of the project is to predict upcoming election related social unrest events in the countries: India, Indonesia and Thailand. This kind of analysis will also help the social scientists and politicians to delve deeper into the motivating factors and sentiments of the public related to a protest. The project will use text mining and web mining to gather data from the web. It will then train and use Machine Learning models to make the predictions.

CCS CONCEPTS

• Information systems~Data extraction and integration • Information systems~Sentiment analysis • Information systems~Content analysis and feature selection • Information systems~Presentation of retrieval results • Computing methodologies~Supervised learning by classification • Computing methodologies~Classification and regression trees • Computing methodologies~Ensemble methods • Computing methodologies~Feature selection

KEYWORDS

Predictive Analytics, Machine Learning, Sentiment Analysis, Web Mining, Ensemble Learning, Text Classification, Web mining

ACM Reference format:

Sonal Singh, 2019. Predictive Analytics: On Social Unrest due to election violence. *University at Buffalo, Buffalo, NY, USA, 6 pages.*

1 1 Introduction

To predict whether political violence will occur at a location or not, it is important to know the current political environment in that location. As people, if we want to know the current political situation of our city, we watch the news. Ideally, news contains all the important political happenings of a location. In a peaceful period of time, news will contain content that mostly leans positive. During times of heightened political tension, the news of that location will reflect the same. Thus, news is good measure of the political environment of a location.

Predicting the future is hard but there is one observation that we can use to try to make predictions. Political violence does not mostly erupt out of the blue. It is the result of an ongoing political turbulence. This turbulence is captured by the news for that location. If we can use the news data as a measure for the political turbulence, we can train a classifier which can take news for a location as input and predict whether this turbulence will lead to a violent event or not.

Intuition/Hypothesis. All these events create significant chatter on the internet. There is chatter that happens after the event has occurred and there is chatter that happens leading up to the event. Trending news gives us a good signal about what the political climate is like for a location and day.

2 1 Data Source

Building a good dataset for our machine learning models to work on is one of the important steps in supervised learning by classification. The following are the publicly available sources from where data was ingested:

- **ACLED**: Armed Conflict Location and Event Dataset (ACLED), a disaggregated conflict collection, analysis and crisis mapping project, collates and analyses data on political violence and protests in over 50 developing countries. The

project, however, will focus on the data pertaining to three countries: **India, Indonesia and Thailand**. For this project, we are only concerned with events that can be categorized as **Riots, Protests or Violence against civilians**. [1]

- **News API**: Trending news can be retrieved from https://newsapi.org/. [2] They provide a very good API and have great results. Again, for the purpose of this project, we will be retrieving news from India, Indonesia and Thailand.

2.1 1.1 Data Processing

Every event in ACLED will be treated as the ground truth. To get the political climate before that event occurred, we are using News API to get news on a particular date from the region where the event occurred.

This news is then processed so that we can get the general sentiment of the articles. To do that, we will:

- Translate news if not in English (for news retrieved from Thailand and Indonesia) using Google Translation API
- Run the news headlines through a sentiment analyzer using Google Natural Language API
- Store the sentiment scores with their respective events in a database

More on the data processing section will be covered in the [Implementation] section in this paper. The dataset will be then transformed using the Pandas library in Python based on the features we will be using for our model.

3 1 Implementation

This section is divided into three parts for clarity.

3.1 1.1 Granularity

The predictions made by the system are made two days in advance and city wise. Following are the variations of granularity per country:

• India: Top 15 cities where social unrest events take place according to ACLED.

• Indonesia: Top 6 cities where social unrest events take place according to ACLED.

• Thailand: Top 6 cities where social unrest events take place according to ACLED.

3.2 1.1 Methodology

3.2.1 Training Phase of the ML models

We start by collecting news of each day from every city. Using ACLED, we record the occurrence of social unrest events 2 days after this news was published. If there was a social unrest event recorded in ACLED, we classify this news as a signal to Violence happening. If there was no such event, we classify the news as signal to no violence in the city. Sentiment score of each of the news is also recorded in the data processing phase. This score is used as the second feature in training our ML models.

3.2.2 Testing Phase of the ML models

To predict if social unrest event is about to occur, the system collects all the news from various websites of that particular day from a city. It preprocesses the news text into digestible, clean text. It then feeds this text to the previously trained ML model to classify whether this news is forecasting a social unrest event or not.

3.3 1.1 Solution Architecture

3.3.1 News Scraper

This module gets the trending news about a particular location and a particular day by making calls to the News API. The News API gives data from past 30 days.

The locations for which the news is retrieved is decided by the top cities selected from each of the three countries.

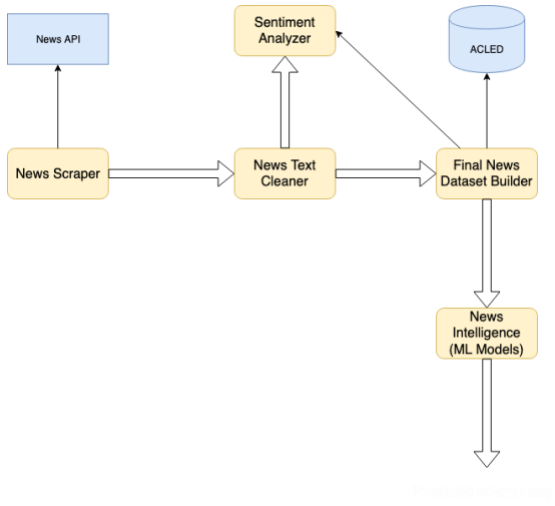


Figure 1: Training the model

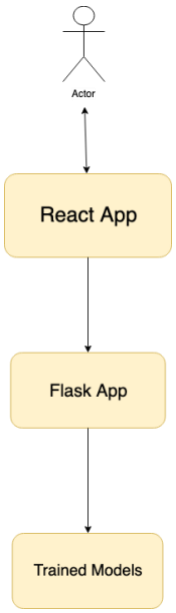


Figure 2: Deploying the model

3.3.2 News Text Cleaner

The news scraped in section 3.3.1 requires preprocessing since we need to transform it into something our ML models can digest. This module performs the following operations on the raw data ingested using the News API:

- Extracts and stores the title of each news object in the JSON files.
- If the text is not in English (when news is from Indonesia/Thailand), it translates it into English using the Google Translate API.
- Removes special characters, punctuations and new lines.
- Saves only the extracted and cleaned text in new files.

3.3.3 Sentiment Analyzer

This module calculates the sentiment scores of news headlines of each news article scraped. The Google Natural Language API used to get of the lines of text. Two values of sentiment are stored: sentiment score and sentiment magnitude. The sentiment score is then multiplied with the sentiment magnitude and 100 to find a total weighted score. At this point, each news article has a sentiment score lying between -1 to +1. To find the final sentiment of the news at a particular location on a particular day, the average of all the sentiment scores of news headlines from that day and location is calculated and stored in the database.

3.3.4 Dataset Builder

Further, we need to transform our data into a dataset which has clearly defined inputs and its corresponding output classes. This dataset will be fed into our ML models as the training and testing sets.

Final dataset includes the following fields:

- News: From a single location on a single day
- Label (Violence/ No_Violence): Denotes whether any social unrest event happened two days after the News were published
- Date_Location: Stores the corresponding date and location of the News
- Sentiment score: the sentiment score calculated in section 3.3.4

For labeling, a News is labelled as Violence if there is an entry in ACLED from two days after the news were published, else No_Violence.

	news	labels	src
0	deputy sp killed in encounter in jammu and ka...	VIOLENCE	2019-02-26__Jammu
1	senior police officer , indian army jawan ki...	NO_VIOLENCE	2019-02-26__Srinagar
2	toxic moonshine kills 133 people and leaves h...	VIOLENCE	2019-02-26__Delhi-New Delhi
3	ludhiana land scam: ask minister ashu to resi...	VIOLENCE	2019-02-26__Ludhiana
4	raja-maharaja government in punjab , #39;mah...	NO_VIOLENCE	2019-02-26__Bathinda
5	assam hooch tragedy toll rises to 124 at leas...	VIOLENCE	2019-02-26__Guwahati
6		VIOLENCE	2019-02-26__Imphal
7	ysrcp chief jagan in london to bring hawala m...	VIOLENCE	2019-02-26__Hyderabad
8	cheque bounce cases: five days later , farme...	NO_VIOLENCE	2019-02-26__Patiala
9	india vs pakistan: sourav ganguly clarifies s...	NO_VIOLENCE	2019-02-26__Kolkata
10	government will teach terrorists a lesson: am...	NO_VIOLENCE	2019-02-26__Amritsar
11	architects to organise chandigarh urban festi...	NO_VIOLENCE	2019-02-26__Chandigarh
12	jewellery , documents go up in smoke - the n...	NO_VIOLENCE	2019-02-26__Bengaluru
13	iit bhubaneswar summer internship: applicatio...	NO_VIOLENCE	2019-02-26__Bhubaneswar
14	pm modi to kick off rs 75 , 000 crore kisan s...	NO_VIOLENCE	2019-02-26__Lucknow
15	can modi govt hold timely elections in j&k ? ...	NO_VIOLENCE	2019-02-27__Jammu
16	can modi govt hold timely elections in j&k ? ...	VIOLENCE	2019-02-27__Srinagar
17	pnr enquiry to complaint status: all in a sin...	VIOLENCE	2019-02-27__Delhi-New Delhi
18	delhi: at rally by sanitation workers , dema...	VIOLENCE	2019-02-27__Ludhiana

Figure 3: Sample from final dataset

	news	violence_label	src	sentiment_score
count	905	905	905	905.000000
unique	905	2	905	NaN
top	regrettable that nonneutral person appointed o...	NO_VIOLENCE	2019-03-14__Patiala	NaN
freq	1	493	1	NaN
mean	NaN	NaN	NaN	0.577265
std	NaN	NaN	NaN	8.767309
min	NaN	NaN	NaN	-64.000002
25%	NaN	NaN	NaN	-3.666667
50%	NaN	NaN	NaN	0.000000
75%	NaN	NaN	NaN	3.916667
max	NaN	NaN	NaN	64.666665

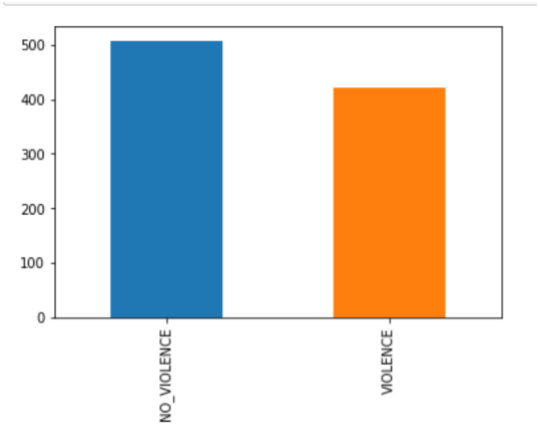


Figure 4: Overview of Labeled dataset

3.3.5 Dataset Builder

The final dataset created in Step 3 is split into training and testing set in the ratio 4:1. This module trains and tests three Machine learning models:

- Gaussian Naive Bayes
- SGD Classifier
- Logistic Regression
- K-neighbors Classifier
- Random Forest Classifier
- Voting Classifier

Each of these models are evaluated and the corresponding accuracy is returned. We will observe the results in the next section. Features used for classification: News Headlines(matrix of tf-idf features) and Sentiment score

3.3.6 Prediction

In this last step, the system makes predictions of whether or not violence will take place, **after 2 days from today**, in the top cities chosen from countries India, Indonesia and Thailand. The ML model with the best accuracy out of the three is chosen to predict the results each time.

Sample predictions from March 28th, 2019 can be seen in the image below.

	Date	City	Prediction
0	30 March 2019	Jammu	VIOLENCE
1	30 March 2019	Srinagar	VIOLENCE
2	30 March 2019	Delhi-New Delhi	VIOLENCE
3	30 March 2019	Ludhiana	NO_VIOLENCE
4	30 March 2019	Bathinda	NO_VIOLENCE
5	30 March 2019	Guwahati	NO_VIOLENCE
6	30 March 2019	Imphal	NO_VIOLENCE
7	30 March 2019	Hyderabad	NO_VIOLENCE
8	30 March 2019	Patiala	VIOLENCE
9	30 March 2019	Kolkata	NO_VIOLENCE
10	30 March 2019	Amritsar	NO_VIOLENCE
11	30 March 2019	Chandigarh	NO_VIOLENCE
12	30 March 2019	Bengaluru	NO_VIOLENCE
13	30 March 2019	Bhubaneswar	NO_VIOLENCE
14	30 March 2019	Lucknow	NO_VIOLENCE

Figure 5: Social Unrest Predictions in top 15 cities in India

4 1 Evaluation

The **sklearn** library in Python provides us with classification reports for the trained ML models. The report consists of the following metrics that are calculated for each of the classes present in the classification:

- Accuracy: Accuracy of the predictions by the model on the testing set.
- Precision: It is the ability of a classifier not to label an instance positive that is actually negative.
- Recall: It is the ability of a classifier to find all positive instances.
- F1-score: gives you the harmonic mean of precision and recall.
- Support: The support is the number of samples of the true response that lie in that class.
- Micro Average: Calculates metrics globally by counting the total true positives, false negatives and false positives.
- Macro Average: Calculates metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

The scores corresponding to every class tells us the accuracy of the classifier in classifying the data points in that particular class compared to all other classes.

```
Gaussian NB
accuracy 0.5195530726256983
              precision    recall  f1-score   support

 NO_VIOLENCE       0.51      0.52      0.51        87
    VIOLENCE       0.53      0.52      0.53        92

   micro avg       0.52      0.52      0.52       179
   macro avg       0.52      0.52      0.52       179
weighted avg       0.52      0.52      0.52       179
```

```
SGD Classifier
accuracy 0.5083798882681564
              precision    recall  f1-score   support

 NO_VIOLENCE       0.49      0.49      0.49        87
    VIOLENCE       0.52      0.52      0.52        92

   micro avg       0.51      0.51      0.51       179
   macro avg       0.51      0.51      0.51       179
weighted avg       0.51      0.51      0.51       179
```

Logistic Regression
accuracy 0.5977653631284916

	precision	recall	f1-score	support
NO_VIOLENCE	0.57	0.69	0.62	87
VIOLENCE	0.64	0.51	0.57	92
micro avg	0.60	0.60	0.60	179
macro avg	0.60	0.60	0.60	179
weighted avg	0.60	0.60	0.59	179

KNeighborsClassifier
accuracy 0.5810055865921788

	precision	recall	f1-score	support
NO_VIOLENCE	0.57	0.53	0.55	87
VIOLENCE	0.59	0.63	0.61	92
micro avg	0.58	0.58	0.58	179
macro avg	0.58	0.58	0.58	179
weighted avg	0.58	0.58	0.58	179

accuracy 0.6256983240223464

	precision	recall	f1-score	support
NO_VIOLENCE	0.54	0.84	0.66	87
VIOLENCE	0.69	0.34	0.45	92
micro avg	0.58	0.58	0.58	179
macro avg	0.62	0.59	0.56	179
weighted avg	0.62	0.58	0.55	179

Voting Classifier
accuracy 0.6256983240223464

	precision	recall	f1-score	support
NO_VIOLENCE	0.59	0.72	0.65	87
VIOLENCE	0.67	0.53	0.59	92
micro avg	0.63	0.63	0.63	179
macro avg	0.63	0.63	0.62	179
weighted avg	0.63	0.63	0.62	179

Figure 6: Classification Reports of the 6 Machine Learning models trained.

For our system, higher values of recall will help us achieve more correct predictions. After running the first 5 models, I observed an accuracy of 55%-60%. To get higher values of accuracy and recall, I decided to go for the **Ensemble Learning** "Wisdom of the crowd". Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. From the classification reports, we observe that the voting classifier had better values for the evaluation metrics.

5 1 Results

For better analytics, I plotted the graphs of the metrics for both violence and Non-violence predictions.

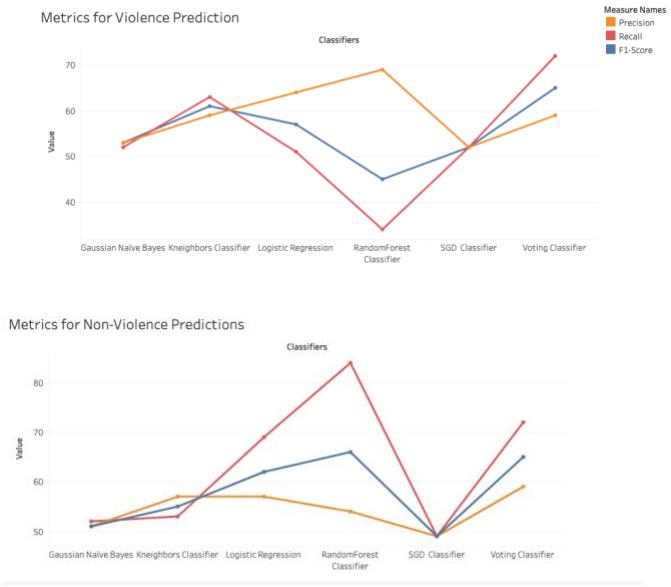
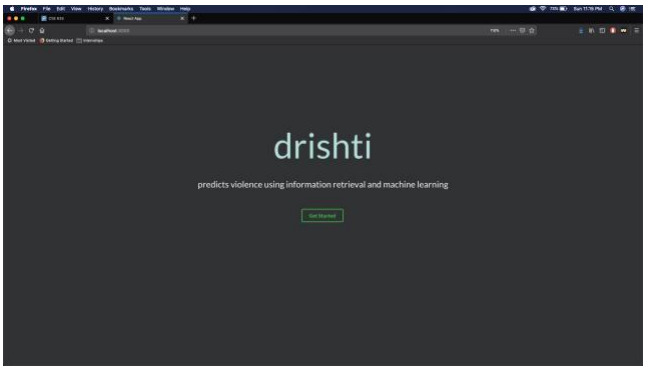
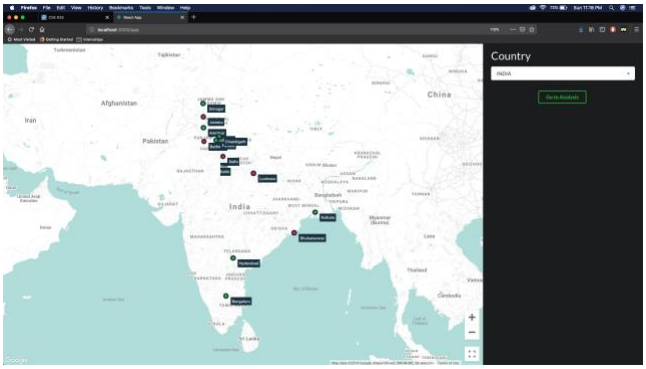


Figure 7: Value vs ML Model for precision, recall and f1-score.

We can see that even though random forest classifier tends to perform better for non-violence predictions, it performs poorly when it comes to predictions related to violence. Whereas, voting classifier works best for both the cases.

The final predictions are shown on a map of the countries India, Thailand and Indonesia as shown in the screenshots below.





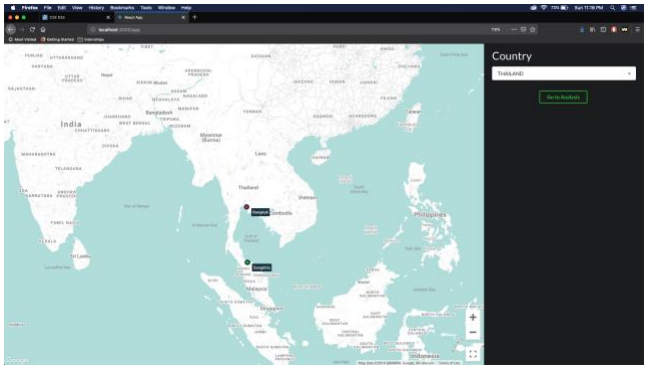


Figure 8: Screenshots of working model

6 1 Conclusion

Predicting an event in the future is an extremely difficult task as there are too many factors behind the occurrence of an event. A naïve model will give the probability of occurrence of an event as 50-50 but we have built a model which gives us more precise results. We are also utilizing 'wisdom of the crowd' by taking into account the multiple machine learning models to improve our model's prediction accuracy. The precision of the model can be further increased by considering sentiment from the population of that location using sentiment analysis on social media.

ACKNOWLEDGMENTS

I would like to thank our professor Dr. Rohini Srihari for her consistent guidance throughout the project. I also appreciate the help provided by our TA Lu Meng whenever I needed it. My friend Yash Choukse, student of Northeastern University, also helped me in finding solutions to problems where I found myself stuck. Thank you, all.

REFERENCES

[1] https://www.acleddata.com/
[2] https://newsapi.org/
[3] https://www.acm.org/publications/proceedings-template.
[4] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, Lakshminarayanan Subramanian, Predicting Socio-Economic Indicators using News Events, New York, USA, https://www.kdd.org/kdd2016/papers/files/rpp0206-chakrabortyA.pdf
[5] Chris Perry, Machine Learning and Conflict Prediction: A Use Case, https://stabilityjournal.org/articles/10.5334/sta.cr/print/