Terrorism: Data Analysis & Predictive Modelling

Shivprasad Kathane
Course: DS203-Programming for Data Science
Centre for Machine Intelligence and Data Sciences
Indian Institute of Technology Bombay
Mumbai, India

Ayush Sarraf
Course: DS203-Programming for Data Science
Centre for Machine Intelligence and Data Sciences
Indian Institute of Technology Bombay
Mumbai, India

***Abstract*—The volume of terrorism in recent times has been rising alarmingly. Terrorist activities cause huge loss of lives & damages to property and have deep-threatening impacts on the minds of the people, hence a comprehensive analysis of past occurrences is needed in order to be better prepared for future ones. This report serves to aid counter-terrorism as we explore the The Global Terrorism Database (GTD) to examine different trends in the occurrence of terrorist activities using data analytics practices and provide a base for predictive modelling using Machine Learning (ML) after aggregating data from other sources. Existing analyses revolve around extracting patterns from GTD and predicting information on terrorist events. Here, comprehensive data visualisation is done in order to understand temporal, regional and other variations in GTD variables, classifier model is built to predict whether an attack would be successful or not based on input values for these variables and ML Regression models are built using political, economic and social factors that may influence terrorism. The analysis is useful for policy making so as to assess threat of terrorism based on the situation of country at a particular time and take actions accordingly.**

***Index Terms*—Terrorism; Global Terrorism Database (GTD); Machine Learning; Data Mining; Factors influencing terrorism; Regression**

I. Introduction

Terrorism means intentional, indiscriminate and illegal use of power to create and exploit fear through violence or a threat of violence in order to attain some political, economic, religious, social or legal objectives by non-state individuals or terrorist organisations.

Terrorism has become a huge threat to both humanity and society in recent times. Terrorist attacks lead to loss of innocent lives and damage to public and private property apart from mental trauma, fear, anxiety, uncertainty, frustration among people affected by it. Hence counter-terrorism measures are employed by security forces to prevent occurrence of such attacks. An important basis for such measures turns out to be data on the occurrence of such activities, which is analysed in order to be better prepared for future such occurrences.

To start with, enhanced security combat requires knowledge on the popular weapons used in these attacks. Which cities, countries and regions are prone to terrorism governs deployment of security forces. Identification of most active terrorist organisations, most targeted sections of society, popular modes of attack serves as a critical requirement for planning purposes. Further, addressing how these have varied with time in the past provides better temporal insights. All of this information is easily extracted using standard data visualisation techniques. As part of exploratory data analysis, a more elegant analysis and presentation is provided in this report with focus on ranking and time-variation of **GTD** variables and additional specific analysis for India. 2 classification models: decision tree and random forest are implemented to predict the success of an attack on the basis of these variables.

The **Global Terrorism Database (GTD)**, a comprehensive database on terrorist attacks has been popularly mentioned in literature and utilised for data analysis and machine learning in the study of terrorism. Review of existing literature utilising **GTD** primarily reveals machine learning models for prediction of attacks and attackers **[1]**, deep learning models for prediction of future terrorist activities **[2]**, correlation factors influencing terrorist attacks **[3]**. With the growing potential of AI, the existing work has been of significant relevance in understanding the behaviour of terrorist activities and developing counter-terrorism techniques.

Generally, a problem such as terrorism ultimately finds its roots in political, economic or social issues. Hence, assuming these factors influence terrorism, then the political, economic and social indicators of a country may as well in combination be a good indicator of the level of terrorism present in that country. The degree to which this is represented by the data forms the basis of our work inspired from **[3]**. The ultimate desirable goal for counter-terrorism can be to develop real-time location based threat perception and formulate anti-terrorism policies for the nation based on the current situation in the country.

Our work aims to aid the anti-terrorism policies along with the ongoing reported research based on AI. On the data front which is currently annual national in nature, we sought the objective to be able to predict the number of terrorist attacks and casualties in a country in a year given data on political, social and economic fronts reported for that country in that year. We aggregated data from multiple sources like **GTD, Penn World Tables, Database of Political Institutions, Polity5 project, ourworldindata.org, etc**. for many countries across the years and built regression models using popular machine learning techniques. Five models: Linear and Lasso

Regression, K-nearest neighbour, Random Forest and Neural Network with varying data input were evaluated and compared on the basis of Mean Absolute Error as error metric.

II. RELATED WORK

- **Alaa S. Alsaedi et al [1]** used three machine learning (regression) algorithms: K-nearest neighbor (KNN), Naive Bayes (NB) and Random Forest (RF) to acquire valuable information about the predicted attacks and attackers exclusively based on data of Global Terrorism Database with RF performing the best.
- **M. Irfan Uddin et al [2]** used five different deep neural network (classification) models to predict information on future terrorist activity like whether it will succeed or not, whether it would be a suicide, the type of weapon used, the mode of attack and the region of attack. One of their DNN models demonstrated more than 95 percent accuracy compared to other techniques in machine learning.
- **Shuo She et al [3]** used a negative binomial regression model based on data of 49 Belt and Road (BR) countries from multiple sources from 1999 to 2014. They explored the impact of political, economic and social factors on the number of terrorist attacks and resulting casualties and found on the basis of their model coefficients that "trade deficit caused by sluggish export growth, structurally unreasonable but enormous military expenditure, the level of democracy and fragile political structure are the main causes of frequent terrorist attacks" in those 49 countries under the China's proposed programme.

Our work based on this idea of national indicators as factors correlating with terrorism includes all countries and utilises popular ML regression models to predict same outputs with varying input data for time period from 1970 to 2017. Thus our independent analysis expands the time space domain of input data and builds model for predictive application rather than just determining the significant causes for terrorism.

III. DATASETS

The **Global Terrorism Database** is a publicly available database which documents more than 190,000 international and domestic terrorist attacks that occurred worldwide since 1970 till 2018 in its latest version and is maintained by the **National Consortium for the Study of Terrorism and Responses to Terrorism (START)** at the University of Maryland. The **GTD codebook** was an important resource for variable description. Chief variables used for analysis and their description is as follows:

- **iyear:** Year in which the terrorist attack took place
- **country-txt:** Country in which the attack took place
- **provstate:** Attacked-State (*relevant for India*)
- **city:** Name of city where attack took place
- **attacktype1-txt:** Principal mode of attack
- **success:** Successful or not
- **suicide:** Suicide attack or not
- **weaptype1-txt:** principal type of weapon used in attack

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	cou
0	197000000001	1970	7	2	NaN	0	NaT	58	
1	197000000002	1970	0	0	NaN	0	NaT	130	
2	197001000001	1970	1	0	NaN	0	NaT	160	
3	197001000002	1970	1	0	NaN	0	NaT	78	
4	197001000003	1970	1	0	NaN	0	NaT	101	
...	...	...	...	...	...	...	...	...	

Fig. 1. Overview of the GTD

- **targtype1-txt:** major target group attacked
- **gname:** name of terrorist organisation
- **nkill:** total number of fatalities
- **nwound:** total number of people injured

There weren't any issues involved in visualising data of **GTD**. For ML purposes, the data was grouped by country and year respectively using the **'groupby'** function. The count for each year gave the number of attacks for that year by country and the number of casualties was taken by summing values in columns **'nkill'** and **'nwound'**.

There are several open access datasets on political, social and economic (PSE) indicators with many indicators themselves. They were chosen on the basis of maximum possible common time period and their comprehensive quality. World Bank data is a good source for many variables but has a time series (horizontal stacked data) with one dataset catering to a single variable. So on grounds of incompatibility & incomprehensiveness with respect to others, it was left untouched.

country	year	flag	fragment	democ	autoc	p
Afghanistan	1800	0	NaN	1	7	
Afghanistan	1801	0	NaN	1	7	
Afghanistan	1802	0	NaN	1	7	
Afghanistan	1803	0	NaN	1	7	
Afghanistan	1804	0	NaN	1	7	

Fig. 2. Overview of the Polity5 Project Dataset

The **Polity5 Project Dataset** records political regime characteristics and transitions from 1800 to 2018. It has assigned scores for some parameters of a political system. Three variables were used as features for primary model:

- **democ:** A state's level of democracy is measured by its democracy score (0-10), which is made up of four components: the competitiveness of executive recruitment, the openness of executive recruitment, constraints

on the chief executive and the competitiveness of political participation
- **autoc:** Level of autocracy assigned scores in similar manner
- **durable:** Political Stability Score denoting the length of years during which the same party rules a country

In contrast the **Database on Political Institutions (DPI)** has absolute data on many political variables. It isn't included in the primary model.

Penn World Tables (PWT) has a dataset on economic and social indicators from 1950 to 2017. 6 variables were used as features for the primary model:

sumption	HCI	XR	CP	%emp	%urb	nattacks	ncasual
8.579590	1.950753	7.000000	0.209760	0.391950	0.34024	1	0.0
5.688477	2.516159	7.559302	0.183802	0.406555	0.36428	1	0.0
9.219727	2.515733	9.914172	0.190120	0.405223	0.36700	1	0.0
7.018555	2.515308	32.276627	0.245260	0.327174	0.37249	3	1.0
8.912109	2.514457	94.623337	0.194902	0.342752	0.38354	2	1.0

Fig. 3. Overview of the primary dataframe used for ML

- **rgdpna:** Real GDP at constant 2011 national prices (in mil. 2011USD)
- **rconna:** Real consumption at constant 2011 national prices (in mil. 2011USD)
- **xr:** Exchange rate, national currency/USD (market+estimated)
- **hc:** Human capital index, based on years of schooling returns to education
- **emp:** Calculated as proportion of population engaged in workforce
- **pl-c:** Price level of household consumption relative to that of USA in 2011

Data on population in urban and rural areas for each year for each country from 1960 to 2017 was acquired from **ourworldindata.org** and urbanisation level was taken as percentage of population in urban areas and used as a feature in the primary model.

For all data aggregation purposes data-frames were merged using inner join.

All missing or absurd values were replaced with the padding method and remaining were dropped so that null or incorrect values do not enter ML training.

IV. ANALYSIS PIPELINE

A brief step-wise pipeline is as follows:

- Data was first collected from Global Terrorism Database
- Data visualisation was performed
  - Popularity-based ranking of variable value was done
  - Time-variation of these variables was studied
- 2 success-classification models were built
- Data on political socio-economic factors was collected
- All data was transformed into country-year wise format
- Data from all sources was merged and cleaned
- 10 seemingly relevant indicators were chosen as ML input
- 5 regression models were developed using above inputs
- 4 regression models were implemented on the datasets
- The best model was employed on the final dataset
- Regression models are compared using MAE
- Necessary stats and graphs are displayed along the way
- Observations and derived conclusions are recorded

Execution details for machine learning are as follows:

- Column values of all data frames were standardised before being used for training.
- Each dataset was split into train, validation and test set in ratio 0.7, 0.15, 0.15.
- Sklearn library was used for all machine learning models.
- Linear regression was directly employed while optimum hyperparameter alpha for lasso regression was found using a 'for' loop.
- Neural network with a single hidden layer in the form of Multi Layer Perceptron Regression was used only on the primary dataframe.
- Optimum hyperparameters for KNN, MLP(NN) and SVR were obtained using Random Search Cross-Validation algorithm.
- Mean absolute error is reported in terms of average of MAE for the 2 outputs

V. RESULTS

Some results are evident from data visualisation.

- There has been a steep rise in the terrorist activities in South-Asia, Middle East and Africa regions in recent times.

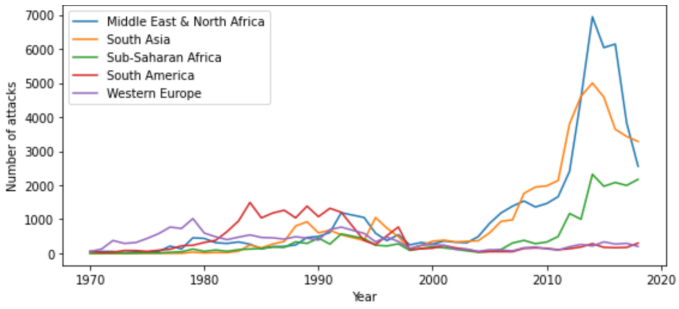

Fig. 4. Regional Variation of Terrorist attacks over time

- Taliban and ISIL have been the most active terrorist groups hence, a close watch on their activities is necessary.
- Explosions and explosives have been the dominant mode of attack type of weapon and have resulted in huge loss of lives and damage to property of common citizens.
- In India, there has been an alarming rise in the number of attacks on police in recent times with the former state of Jammu and Kashmir being the worst affected.

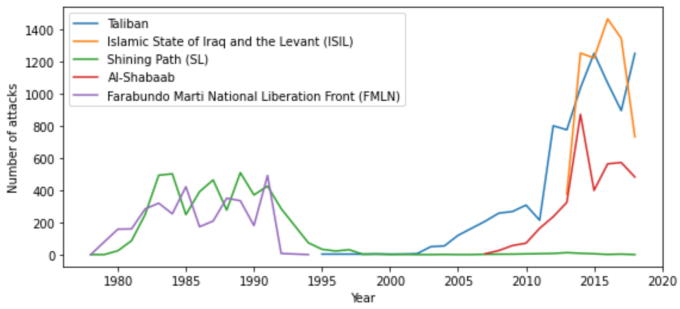

Fig. 5. Group-wise Variation of Terrorist attacks over time

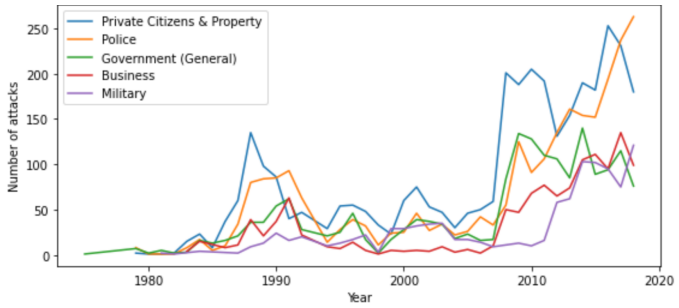

Fig. 6. Target-wise Variation of Terrorist attacks over time in India

- India witnessed its single most deadly terror attack as an explosion in Mumbai local trains in 2006 followed by the 2008 series of attacks in the same city.
- Hence, identifying threat perceptions and formulating anti-terrorism policy has been critical since then and utilisng the power of ML can be helpful.

Correlation heat map reveals:

- Highest correlation of -0.14 between percentage employment and number of attacks and fatalities.

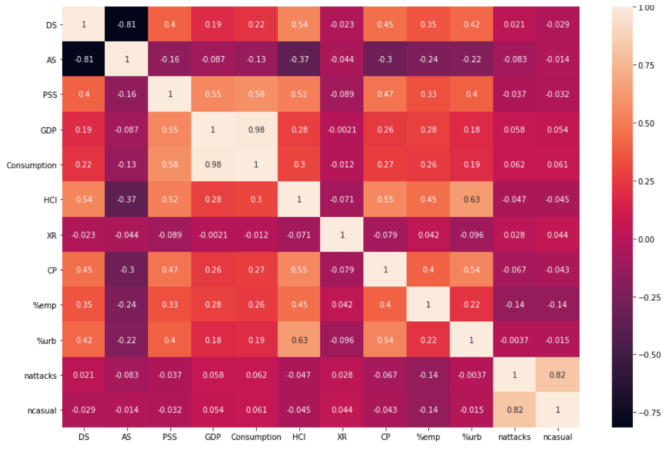

Fig. 7. Heat-Map showcasing correlations between variables

- Positive correlation between terrorism and each of GDP, consumption, exchange rate
- Negative correlation between terrorism and each of human capital index, political stability, autocracy, consumer price levels, urbanisation
- Correlation is disputed for democracy level among attacks and casualties
- In general, terrorism is found to be low where there is more education, employment, urbanisation, political stability, economic strength

Machine Learning Results:
85 percent weighted average accuracy was obtained on a decision tree classifier used to predict the success of a terrorist attack.
Mean absolute error is reported in terms of average of MAE for the 2 outputs

Results for Machine Learning models using 10 features from 3 datasets:

Machine Learning Model	Best Validation MAE
Linear Regression	236.58
Lasso Regression	210.26
Random Forest Regression	97
Neural Network (MLP) Regression	198.13
KNN Regression	88.13

Table 1: ML model MAE

KNN gave lowest validation error and on test set yielded MAE=121.77

Results for Machine Learning models using all variables of PWT as features:

Machine Learning Model	Best Validation MAE
Linear Regression	297.5
Lasso Regression	238.64
Random Forest Regression	140.46
KNN Regression	114.84

Table 2: ML model MAE for PWT

KNN gave lowest validation error and on test set yielded MAE=111.83

Results for Machine Learning models using all variables of DPI as features:

Machine Learning Model	Best Validation MAE
Linear Regression	212.02
Lasso Regression	169.3
Random Forest Regression	92.81
KNN Regression	75.39

Table 3: ML model MAE for DPI

KNN gave lowest validation error and on test set yielded MAE=86.52

As KNN performed best on all the 3 cases. It was directly employed on the super dataframe with all variables from PWT and DPI with test set size of 0.2 and no validation set. It yielded MAE=99.54 on the test set in this case.

VI. DISCUSSION

Apart from essential qualitative conclusions and findings from data visualisation and literature review, the results for the machine learning part are significant. K-Nearest Neighbours Regression performed the best across all dataset combinations used with lowest Mean Absolute Error on test set being 86.52 on DPI data. The primary dataset formed using 10 features from 3 datasets and 2 labelled outputs from GTD serves as an example for future relevant work. KNN fared the best on this data with an MAE=121.77 on the test set. Although this error is large compared to the mean number of attacks (50) it still is much better compared to the maximum value of number of casualties (30769) and given that this was a fresh analysis with no such previous instance found by us in existing literature for comparison purposes, it is some good value to begin with upon which others can improve.

To conclude, our work demonstrates effective data mining practice for analysing terrorism as a function of national factors and the ML modelling serves as a base for predictive applications and policy making. Taking 'data-driven' decisions is key in this information-heavy world which is true not just for business purposes but also in areas like security as here to identify threat perceptions from terrorism. Developing problem specific algorithms and choosing relevant indicators can be the extension to this work, done as part of a short term course project.

ACKNOWLEDGEMENT

We would like to express our gratitude to our TA **Arijit Jain**, who with utmost patience helped us get through the dilemma of choosing a relevant topic for the analysis. We would further like to thank our classmates, whose proficient discussions across topics inspired us to take up this analysis and played a key role in the completion of the project.

REFERENCES

[1] A. S. Alsaedi, A. S. Almobarak and S. T. Alharbi, "Mining the Global Terrorism Dataset using Machine Learning Algorithms," 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2019, pp. 1-7.

[2] M. Irfan Uddin, Nazir Zada, Furqan Aziz, Yousaf Saeed, Asim Zeb, Syed Atif Ali Shah, Mahmoud Ahmad Al-Khasawneh and Marwan Mahmoud, "Prediction of future terrorist activities using Deep Neural Networks," Wiley Hindawi Complexity, vol. 2020.

[3] She, S., Wang, Q. Weimann-Saks, D.,"Correlation factors influencing terrorist attacks: political, social or economic? A study of terrorist events in 49 "Belt and Road" countries," Qual Quant 54, 125–146, 2020.

[4] Global Terrorism Database (GTD), START, University of Maryland

[5] Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" American Economic Review, 105(10), 3150-3182.

[6] Polity5 dataset version 2018, Polity Project, Center for Systemic Peace

[7] Hannah Ritchie (2018) - "Urbanization". Retrieved from: "https://ourworldindata.org/urbanization" [Online Resource]

[8] Database of Political Institutions (DPI)