Analysis of Terrorist Attacks

Brundha P
Computer Science and Engineering
PES University
Bangalore, India
brundha1801@gmail.com

Charushree A
Computer Science and Engineering
PES University
Bangalore, India
charushree.a@gmail.com

Debaditya Ray
Computer Science and Engineering
PES University
Bangalore, India
debadityaray00@gmail.com

***Abstract*—Terrorism has stricken fear in the hearts of many people across states and nations. To predict future patterns and gain past insights, in this report we aim to analyse the GTD dataset for various terrorist attacks and try to incorporate meaningful inferences from it along with bringing new novel patterns and relations undiscovered till now using various machine learning models like decision trees, random forest classifier,k-NN etc after applying some initial cleaning and exploratory analysis on the data using classic visualising techniques like wordcloud, bar charts and more.**

***Index Terms*—GTD, attacks, visualisation, wordcloud, terrorist**

I. Introduction

With the complete invasion of Afghanistan by the Taliban, the whole world has been on its toes. Global terror threats have shook nations and there is a constant need for surveillance all over the planet. Terrorism is one of the major issues plaguing countries and therefore military budgets of all developing and developed countries have increased significantly in the past years with many undergoing debts to finance defense operations.

Randy Borum's report in 2004 identifies terrorism as ”acts of violence intentionally perpetrated on civilian non combatants with the goal of furthering some ideological, religious or political objective.” [1] Annually lots of lives of soldiers and civilians are lost due to terrorist attacks and many places have become hubs or red zones due to their attack frequencies. It is a growing problem across nations and various countries are adopting new security and surveillance measures to tackle this issue.

In this context, we aim to analyse the GTD [4] database for all the past terrorist attacks which has various attribute values for many terrorist attacks all the way from 1973. We aim to bring meaningful inferences and predictions from the data which would make the tasks easier for the national security analysts and prevent possible terror attacks.

II. Related Work

We first proceed by analysing work already done in this domain. One report involves Social Network Analysis(SNA) and open source data collection involving information retrieval and entity recognition on the sourced data along with the GTD database and uses various visual techniques using pie-charts, maps and networks and mainly tries to find around the links between the various types of organisation, target and attack networks for all attacks targeted in India. [2]Centrality for SNA is used here which can account for the degree, betweeness and closeness measure. Records up to 2014 are taken from GTD and from 2014-16 are taken from online news articles and twitter. From the results, it is found that every group which targeted India for a particular decade was closely linked to every other group and that all terrorist activities revolve around certain concentrated regions. The results were satisfactory but due to the open source nature origin, many data values might appear skewed or incorrect due to propaganda and nationalist news agencies and journalists along with no scope of prediction for future attacks is mentioned anywhere.

Another approach involves using machine learning algorithms such as k-NN and random forest to build classifiers which would group and predict various future terrorism activities based on the GTD databse itself after filtering out columns that have less than 20% values and applying mean imputation on the rest of the missing data. They built a weapon classifier with k value 12 in k-NN algorithm which had an accuracy of 88.74% along with a perpetrator classifier built using random forest algorithm which had an accuracy of 90.45%, and precision of 89.95% along with important visualisations involving attacks and fatalities by years along with countries by attack and total attack types. [3] There is further scope for improvement in this work in the preprocessing stage as well as the classification phases.

Similar approach like [3] has been used in [6] where machine learning algorithms notedly, decision trees and random forest algorithms have been used to build classifiers based on supervised machine learning which would predict geographical areas or regions prone to attacks and even display the probabilities on a map using colour gradients based on the same dataset. The models gave an accuracy of 75.45% and 89.544% for region wise attacks and 79.24%, 90.414% for types of terrorist attacks made on decision tree and random forest algorithms respectively.

III. Dataset

The Global Terrorism Database(GTD) dataset is a publicly available free dataset which can be downloaded from the GTD website and is maintained by the National Consortium for

Identify applicable funding agency here. If none, delete this.

the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland. [5] It includes all the the possible collected data about every terrorist attack all the way from 1970 to end of 2019 nationally and internationally and has over 135 attributes and 201183 entries.On a yearly basis it is updated with new information and additional columns and is used extensively by many researchers and security organisations for the fine level of detail and accuracy of the data.

IV. Data Preparation

A. Data Cleaning and preprocessing

Real world data will always contain some noise, outliers and missing data due to which working directly on the data without any modification might cause some errors in our final results. The data collected here is assumed accurate and consistent due to the fact that it is being maintained by the START organisation on a best-effort accurate basis. There are a lot of values missing for many attributes due to lack of minute information and the data is extremely sparse thereby invoking the need of dimensionality reduction. PCA is not very useful for the initial cleaning stages as manual intervention from experience does a better job.

- In this dataset we first start cleaning by removing all the records which have null values for these attributes-summary(brief description about the attack), latitude, longitude,nkill(number of people killed in the attack) ,propextent(extent of property damage) as these parameters are vital for our study and are the basic attributes which are supposed to be provided and maintained in the dataset by start. This leaves us with around 52660 rows with complete data for these attributes
- For tackling extremely sparse data, a filter was set where attributes with more than 50% null values are removed. Due to this more than 60 columns were removed which had null values touching as high as 99% of the records for almost each of the attributes. The number of attributes were brought down to 63 as a result of this.
- Some irrelevant attributes which were redundant and could be inferred from other attributes or were insignificant were dropped from the dataset.
- Finally those attributes were removed which were highly skewed or biased towards a particular value which was not very effective and useful along with duplicate records too ending up with 52660 rows and 41 columns
- Rest of the null values of all attributes were filled with another category value labelled missing for the mainly categorical variables as performing any mode or median imputation might make the data inconsistent and the null values for numeric variables were replaced with the means.

B. Visualisation

From the initial exploratory data analysis, lot of inferences and meaningful insights could be obtained from it.

1) *Correlation Matrix:*

- natlty1(nationality of target victim) and country where the attack occurred has a value of 0.64 which implies that majority of the attacks have a tendency to target same country individuals.
- weaptype1(Type of weapon used for the attack) and attacktype1(General method of attack like bombing, hijacking, assassination, etc) has a correlation of 0.68. Therefore specific weapons are preferred for specific attacks and given a particular weapon we could predict the attacks that could take place with it.
- Another relation found(a bit obvious though) is that nkill(Total number of people killed) and nwound(Total number of people wounded) have a value of 0.75 showing that if an attack has resulted in the deaths of many people, a huge number of people would end up alive and wounded.

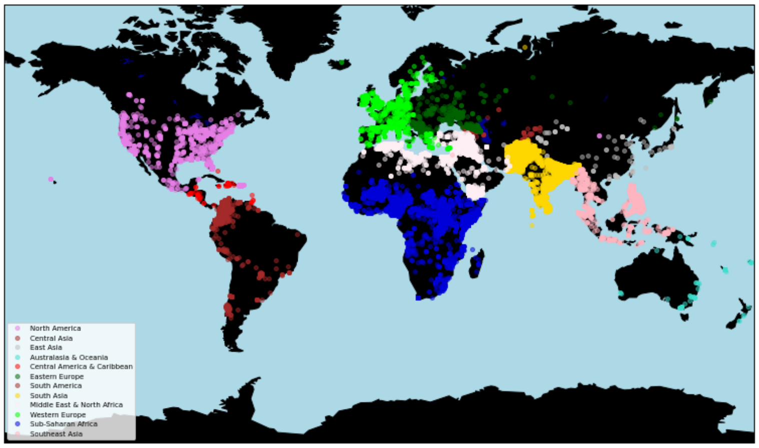

Fig. 1. Global Terrorist Attacks, 1970-2019

2) *Map plot:*

- Using matplotlib library of python, the locations of the terrorist attacks are plotted on a world map based on the latitude and longitude attributes with different colour codes for various regions of the world. The plot shows a clear pattern where terror attacks are prevalent all around the globe except some noticeable places where almost no attacks take place or very rarely like the poles, Russia, Central Australia, Canada and North-Eastern South America. This may be explained by the extremely low population occurring in such places where terror attacks might not have much of an effect in accomplishing the terrorist's demands.

3) *Scatter plot:*

- Similar to the map plot, this time the coordinates of the attacks are plotted on a scatter plot(due to two continuous variables) and from this we can narrow down on the location of the attacks at a higher level. The bulk of the attacks take place between latitudes 0 and 40 and longitudes 0 and 100 thereby signifying majority of

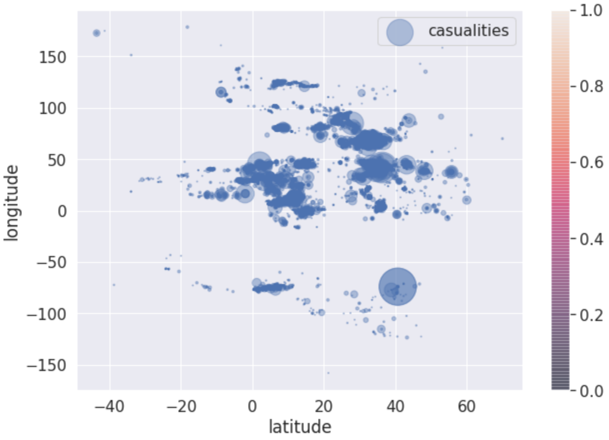

Fig. 2.

attacks of the world are concentrated near Africa, Western Europe, Middle East and South-East Asia.

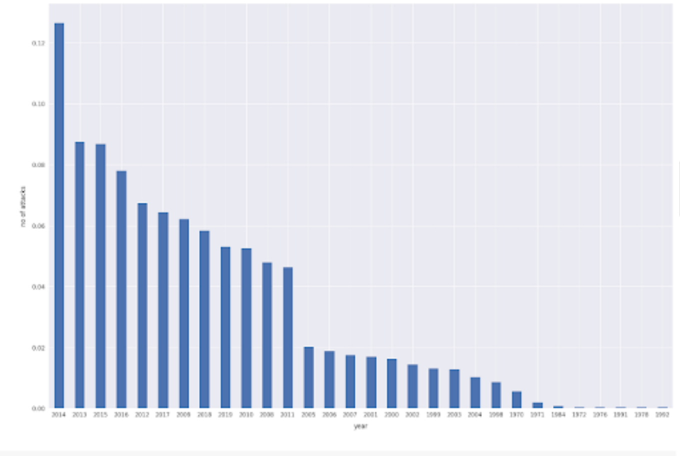

Fig. 3. Number of attacks vs years

4) Bar graph:

- A bar chart is visualised with the number of attacks on Y axis and the corresponding years on the X axis. The number of attacks shows a clear growing trend over the years with the last decade having shown the maximum number of attacks with the highest number of attacks reported in 2014(possibly due to the Gaza conflict).
- Trying to find out the main groups behind the attacks from the bar graph(due to categorical variables), we discover that majority of attacks are not claimed by any organisation or are attacks with unknown perpetrator groups. The most notable ones leading in terror attacks are the Taliban, ISIL(Islamic State of Iraq and the Levant) and Boko Haram groups which are known by all.

5) WordCloud:

- A wordcloud is an ingenious way of analysing and representing strings whose appearance is most prominent and frequent. While constructing a wordcloud of the types of attack which take place, the words assault and armed are most observable and since there is a value for the attribute attacktype1 known as ”Armed Assault”, it is the preferred mode of attack for terrorists which is relatively easy to carry out along with slight emphasis on infrastructure attacks and bombing ones which require a good amount of funding.

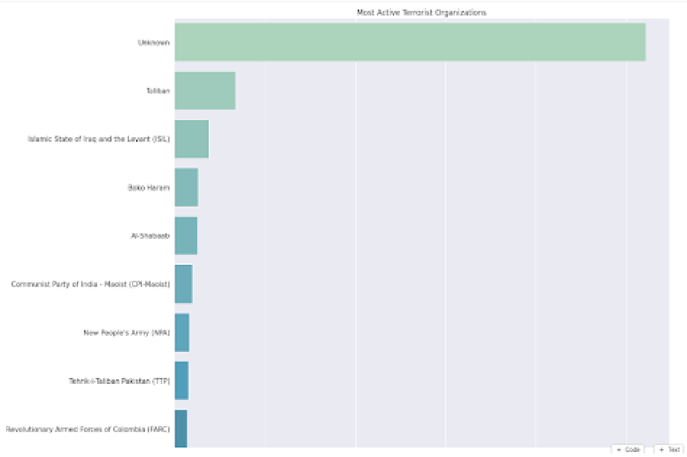

Fig. 4. Most active terrorist organisations

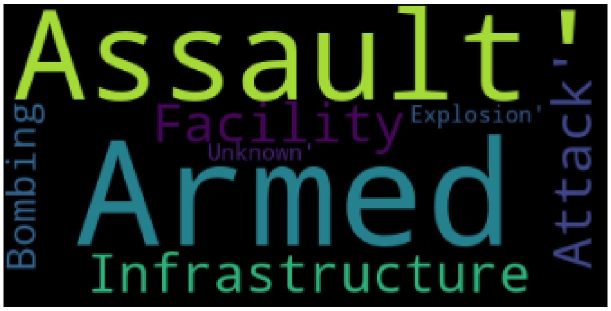

Fig. 5. Wordcloud of attack types

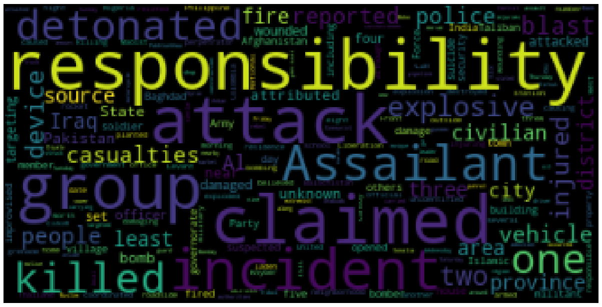

Fig. 6. Wordcloud of attack summaries

- Another wordcloud on the attack summaries reveal the keyword responsibility. This shows that all attack reports with high priority tend to find the group or persons associated with the attack. Other common words which are common and come in conjuction with responsibility are attack, group, claimed, incident and more.

V. Problem Statement

In this enormous dataset with multiple records and attributes various inferences and models could be toyed around and built for it using various algorithms such as regression, decision trees, random forest, k-NN and so on. However, our minds are revolving around addressing 3 proposed questions or problems which we feel are troubling experts.

- Predict the regions which are susceptible to attacks in the nearby future and the probability of the neighbouring regions of attacked areas being under a significant threat.
- Given an attack, figure out the group mostly responsible for it.
- For an attack made by a specific group, estimate the amount of property and human damage caused by the event.

Bringing a solid answer to the above questions can help the police prepare, estimate and tackle future attacks. We feel these are unique problems not repeated in any other journals or professional reports and will provide great contribution to the work done by many researchers.

We plan to use k-NN and regression for tackling the first issue along with a complex regression model for the second problem. The third one calls for a multi regression model. All the approaches mentioned here are still in their initial stages and slight modifications may arise along the way due to the nature of the dataset. The task is little bit challenging due to the complexity however it should be accomplished with ease and we hope to bring out excellent results from the modelling and analysis.

References

[1] Randy Borum, The Psychology of Terrorism, University of South Florida, 2004.

[2] Lavanya Venkatagiri Hegde, Nerella Sreelakshmi and Kavi Mahesh.,"Visual Analytics of Terrorism Data", 2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM),2016.

[3] S. Kalaiarasi, Ankit Mehta, Devyash Bordia, Sanskar.,"Using Global Terrorism Database (GTD) and Machine Learning Algorithms to Predict Terrorism and Threat", IJEAT,ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019 .

[4] Start, "Global Terrorism Database." [Online]. Available: https://www.start.umd.edu/gtd/.

[5] Start, Global Terrorism Database Codebook, no. October. 2019

[6] Enrique Lee Huamaní, Alva Mantari Alicia and Avid Roman-Gonzalez.,"Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020.