Final Report
Terrorist Attack Analysis and Prediction using GTD

Srikar Chundury
01FB15ECS306
Dept. of Computer Science
PES University
Bangalore, India 560085
Email: chundurysrikar@gmail.com

Srinivas Shekhar
01FB15ECS308
Dept. of Computer Science
PES University
Bangalore, India 560085
Email: srinivasshekar01@gmail.com

Swati N H
01FB15ECS320
Dept. of Computer Science
PES University
Bangalore, India 560085
Email: swatinagaraj2@gmail.com

***Abstract*—Recently, worldwide, there has been an urgent need to deal with terrorism. Understanding and analyzing various aspects of global terrorism as a big picture creates an untactful awareness among those interested to contribute solutions to this international problem. Our analysis through visualizations will include a time series analysis of growth of terrorism over the years, categorizing the types of terrorist attacks, identifying the most vulnerable countries in different categories of attacks, discovering the weapons used for deadly attacks, learning the extent of different types of attacks, to name a few. We plan to build reasonably accurate models to predict the next attacks features such as an approximate range of victim count, extent of damage, responsible terrorist group, to name a few. We seek to make the world a better place by providing meaningful analysis.**

I. INTRODUCTION

To be able to understand what is the terrorism, it has become necessary to learn its process, its significance, and how it is developing into our society, what is its roots, how it has evolved until present time, and how can be stopped. Terrorism is a form of violence that has been used since the first century to change public opinions, and to influence political decisions.

Terrorism adversely affects many aspects of the victim country like life-loss , economy , tourism , market , mental health ,exchange rate behavior , education etc.It is every countrys responsibility to minimize such events.Is it possible to turn technology around and use it to not only catch terrorists but predict and potentially stop terror attacks before they happen? However , it is important for us to be very cautious as the models built even with good accuracies may fail in the worst case scenario. This paper is a sincere effort to help the society from the dangers of terrorism.

We are focusing on to give an estimate of the number of terrorist incidents in the coming future years so that government will take appropriate actions/measures to reduce the number of incidents.Also we are visualizing the past terrorist attacks to understand the pattern and also to know which countries are prone to attacks,which country is free from any attack(not completely though).We are keen to solve the problem of predicting the terrorist group that is responsible for the attack.

II. BRIEF SUMMARY OF THE LITERATURE SURVEY

The literature survey covers significantly about three research papers published in IEEE journals.Out of the three papers ,we found this[1] paper to be similar to our approach of solving the problem.
In this paper,a Terrorist group prediction model (TGPM) was built, and it analyses the trend of nature of attacks followed by different terrorist group over a period of time. Based on the different features like target victim analysis, extent of damage done to public property, nature of attack, weapons used, location of attack, and so on. Patterns can be learnt by TGPM to predict which group may have initiated the given attack. CLOPE algorithm, a partition based clustering approach was implemented to build TGPM. CLOPE algorithm was employed because it works well with categorical attributes and missing values.
This paper primarily concluded that historical data can be used to predict the responsible terrorist group in the given attack by employing data mining prediction models such as TGPM and clustering based techniques.

III. PROBLEM STATEMENT

1) Prediction of Fatalities , Damage and the Cause of an attack
2) Predicting the Group cluster responsible for an attack i.e Suspected list of group responsible
3) An overall overview of Global Terrorism using visualizations

A. Visualizing the dataset :

- Year wise and Decade wise Attacks
- Month wise Attacks
- Weapon type wise Attacks
- Year wise Weapon of preference Analysis

- Terrorist group wise Analysis (in a given year interval)
- Type of Attack vs Target of Attack Analysis
- Most Destructive Weapon
- Top Bomb-prone countries
- Region of Attack
- Vulnerable countries
- Visualizing Growth of terrorism region-wise in a period of 50 years
- Geo-referential attack type Visualization
- Geo-referential successful attack visualization
- Geo-referential Number of terrorist incidents heatmap
- Multiple chart analyzing frequency of attack and victim count

B. Data Source :

- Global Terrorism Database (GTD)[1] is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism.It contains features that describes the terrorist attacks from 1970 till 2016.
- It has information about more than 1,70,000 terrorist attacks from 1970-2016 (May).
- Huge variety of information distributed across 135 features.
- Handling NA's , Nan's in GTD is a challenging task for data analysts.

C. Constraints :

- The dataset provides the mapping from non-numeric to numeric entries for most of the features.
- Although , it doesn't provide such a mapping for "gname" (i.e. group attacked). Hence, a new feature called "gno" (i.e. group-number) is added for mathematical modeling purposes.
- The dataset has not been maintained for a brief period of time in between 1980-1990.
- Filling the unknown values has been done differently in different places. Example : "unknown" in "property" is filled with "-9" whereas that in "nperps" is filled with "-99".
- A set of features that were added after 1997 have no information for the years before 1970 and there are a few features that have been removed after a certain year , that have no information since-forth.An attempt has been made to fill these gaps using supplemental collection efforts.Hence, the dataset contains many missing values.

D. Assumptions :

- The population of the country is not taken into account while predicting the "nkill" feature.
- The Economic status of a country is not considered for Target wise Analysis.
- NA is coerced to Unknown though there is a slim difference.

IV. PROPOSED WORK - OUR APPROACH

Refer Figure 1.1

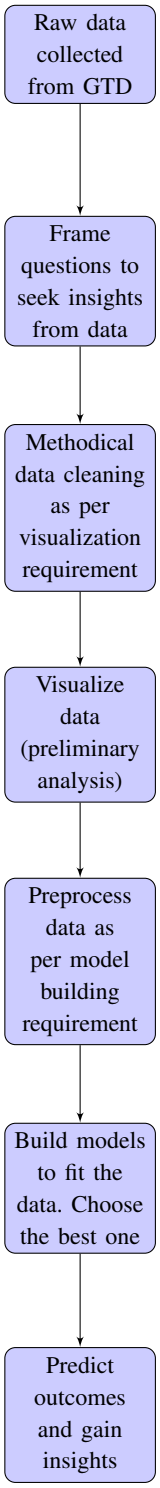


Figure 1.1

V. DETAILED EXPLANATION OF EACH STEP

A. Data Cleaning

GTD dataset is a collection of 135 attributes with many missing values. Some of the attributes have been added along the years. So the challenge lies in understanding the

importance of each attribute and filling in the missing values meaningfully for each attribute. Different kinds of visualizations demand different aspects of data cleansing. Since unknown values, not available and missing values depict atypical meanings in different attribute scenarios, especially from global terrorism perspective, we cannot just remove them. We carefully imputed missing values with appropriate replacement such as means, zeroes, constants and so on.

To illustrate all the three variations of imputations, examples are as follows. To fill missing values in Number of kills(nkill) attribute, we computed the means of Number of kills per weapon, and replaced respective missing values with mean of weapon. To fill missing values in Number of perpetrators(nperp) column, we filled a constant 11, because that was a minimum standard value as mentioned in the codebook published for the dataset. In case of ransom amount(ransomamt), missing values were filled with 0 when no ransom was demanded.Some attributes had to be converted to numeric form to perform mathematical computations on them.

Preprocessing of data involved Principal Component Analysis to pick the most significant Principal Components as there were many attributes.

B. Data Visualization

Visualizations form the foundations of preliminary data analysis. Initially to get a sense of the data, we plotted basic visualizations.As a analyst,it makes us understand what the data actually means i.e. to get the significance of the data. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier through data visualization.

- Sub-setting based on different features helps us analyze the dataset with respect to that feature alone.Ex - Year wise analysis , month wise analysis , target wise analysis , attack type wise analysis , primary weapon type wise analysis and so on..
- Separating out every 10 years and analyzing all the above described features on that 10 year interval - Decade based analysis.
- Suicide ?
    1) Analyzing suicides with Type of Attack
    2) Analyzing suicides with Target of Attack
    3) Analyzing suicides with Type of weapon used
- Decade vs Weapon of choice Analysis
- Decade vs Type of Attack Analysis
- Most Destructive Weapon based on Number of fatalities , Property damage
- Vulnerable Countries
- Region of Attack

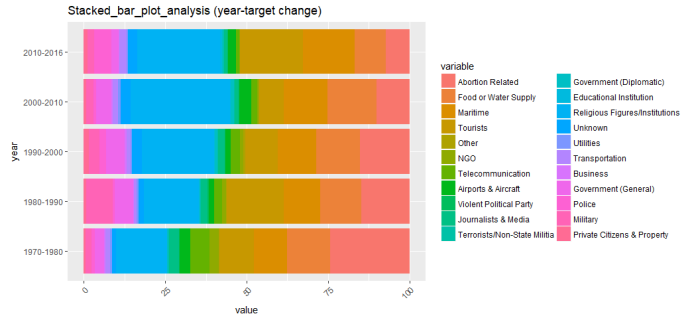
— To get an overview of growth of terrorism across the globe, we did a region wise time series victim count analysis by plotting animated bubble plots. In order to emphasize more on geo referential attributes of the dataset, scatter maps and choropleths were plotted. Scatter map of Terrorist Incidents differentiated by the attack types was geographically plotted. Scatter map of successful Terrorist incident was visualized to geographically locate higher percentage of successful attacks. Choropleth of Frequency of attack was georeferenced to find the highest intensity of frequency of attack across the globe. A Multiple chart to compare Number of lives lost(line chart) and Frequency of attack (As bar chart), with time as x axis and y axis accordingly for the two graphs was plotted to inference correlations if any.Scatter mapbox plot of countries that are prone to attacks was plotted to indicate vulnerable countries.Scatter mapbox plot of region of attack that represents the region that is more prone to attacks.

C. Model Building and Prediction

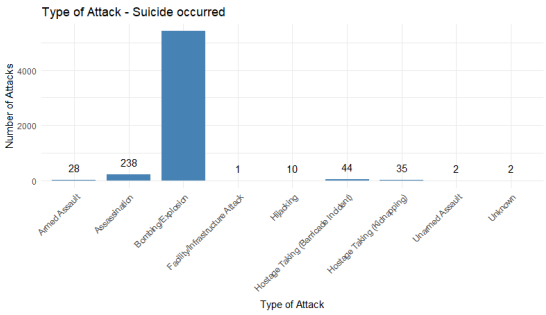
Given an incident we built a model to predict whether it is a terrorist attack. If it did not fall into a terrorist activity, we predicted which other category does it fall under. If the incident were indeed a terrorist attack, we aimed to predict the terrorist group responsible for the attack. Since there are nearly 3454 terrorist groups, it is difficult to achieve higher accuracy of prediction. To overcome this challenge, we clustered the terrorist groups using k-means clustering technique into 50 clusters, and predicted which cluster which was responsible for the attack using decision trees. A remarkable accuracy was achieved. We could also predict the extent of property damage to a reasonably good level given the details of weapons used, attack type and the terrorist cluster group. Predicting the loss of life during the incident was fairly accurate.Also predicting the number of terrorist attacks in the upcoming years using AR[Auto Regression],ARIMA [Auto Regression Integrated Moving Average] and ETS [Error Trend Seasonality] models.Since 1993's data is missing we built the model from 1994 to 2016 and used the same model for prediction.

VI. EXPERIMENTS AND RESULTS

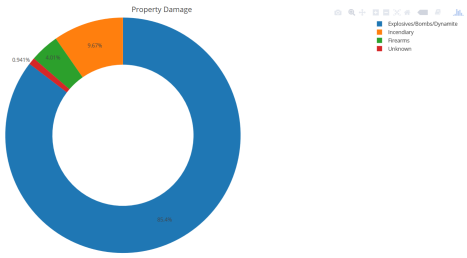
A. Basic Visualizations

- 

  The Stacked Bar Plot analysis for Decade vs Target of attack clearly indicates the groups of society that are always attacked in all the decades.We can also see that the decade 2000-2010 primarily targeted Religious Figures/Institutions.



As expected suicides in a terrorist attack mostly occur in a bombing/explosion incident, next followed by assassination category.



We can see that 85%(maximum) property is damaged because of Explosives , other weapons damage property only to a certain extent.

This visualization tells us the countries that are more prone to attacks. As in the plot,we can see that India,Iran,Iraq,Pakistan,Afghanistan are more prone to attacks.On the other side Australia,New Zealand,Russia are more heavenly countries[since these countries are less prone to attacks].

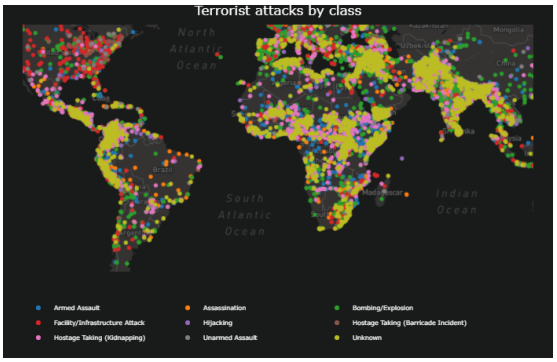


Figure 1.5

Few interesting inferences drawn from Fig 1.5 are the following : North America witnessed mostly Facility/Infrastructure attacks. Hostage taking situations are more likely in Middle East, Pakistan. In India, kidnapping incidents are mostly observed in the North eastern region and Assam, West Bengal. Bombing/Explosion mostly witnessed in Europe, Pakistan.

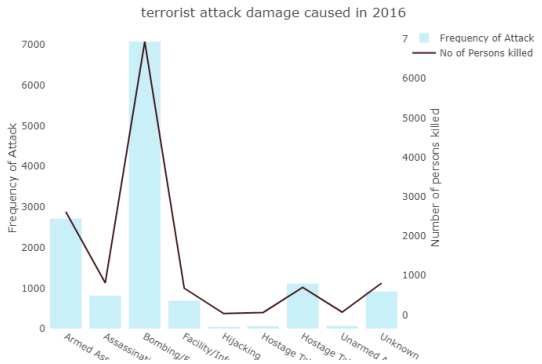
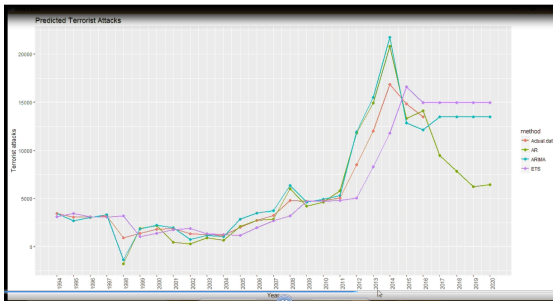


Figure 1.6

A Multiple chart to compare Number of lives lost(line chart) and property damage (As bar chart), with time as x axis and y axis accordingly for the two graphs in the Fig 1.6 Some of the inferences drawn are Facility/Infrastructure attack, Hijacking, Hostage taking(barricade incidents), Unarmed assaults, Assassination type of attacks have a higher Victim count per attack ratio. Unpredicted incidents like these have to be dealt with cautiously. Bombing, Armed assault, Hostage taking(Kidnapping) have approximately balanced Victim count/attack ratio.

B. Predictions

- Estimating Extent of property damage
  1) Naive Bayes Classifier : Accuracy = 54.46%
  2) kNN classifier (n=7) : Accuracy = 61.03%
- Logistic Regression to predict the safe return of hostages Accuracy : 73.01%



To forecast the number of attacks in the upcoming years ,we used AR[Auto Regression], ARIMA[Auto Regression Integrated Moving Average], ETS[Error Trend Seasonality] models. Refer figure 1.2 for the results that we have achieved using all the three models.

```
> pred1<-subset(final_pred,final_pred$Year==2016)
> pred1
    Year       method    value
23  2016           AR 14116.03
50  2016          ETS 14945.57
77  2016        ARIMA 12124.00
104 2016 Actual.data 13488.00
> pred1<-subset(final_pred,final_pred$Year==2017)
> pred1
    Year       method     value
24  2017           AR  9490.956
51  2017          ETS 14945.736
78  2017        ARIMA 13488.000
105 2017 Actual.data        NA
>
```

Figure 1.2

As per the results obtained, AR and ARIMA model are closer to the actual data in the year 2016.So,using ARIMA model ,the number of terrorist attacks is approximately 13K and using AR model ,the number of terrorist attacks is approximately 9K in the year 2017.[chosen ARIMA model because RMSE,ME are less compared to other models]

By building a rule based decision tree model, a few specific prediction questions were answered with a reasonably good accuracy.

1) Given an incident, predicting whether it is a terrorist activity : Accuracy = 95.32%.
2) If the incident does not categorize as terrorism, then predicting the nature of this violent activity : Accuracy = 96.86%.
3) Predicting the group responsible for the terrorist attack : Accuracy = 92.34%.
   Here, since there were 3454 terrorist groups in total, we clustered similar terrorist groups into clusters using K-means clustering technique and predicted the cluster responsible for the attack.
4) Predicting the possible victim count given features like attack type, weapon details and so on : Accuracy = 62.26%.
5) Predicting the extent of property damage given the nature of attack. Accuracy = 88.65%.

VII. CONCLUSION

Unless we are not united against terrorism ,it is certain that the number of terrorist incidents will strongly grow exponentially .As Vladimir Putin rightly said ”Terrorism has once again shown it is prepared deliberately to stop at nothing in creating human victims. An end must be put to this. As never before, it is vital to unite forces of the entire world community against terror”.Terrorism is a threat to peace and security,prosperity and people. As our model suggested that there is always an increase in number of terrorist attacks ,it is important for us to take precautionary measures/actions to avoid such heinous crimes.Finally we are ending the report with the quote by Malala Yousafzai, ”Terrorism will spill over if you don’t speak up.”

VIII. REFERENCES

[1] Pilley PH, Sikchi SS (2014) Review of Group Prediction Model for Counter Terrorism Using CLOPE Algorithm. J Def Manag 4: 115. doi:10.4172/2167- 0374.1000115

[2]National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2016). Global Terrorism Database [Data file]. Retrieved from https://www.start.umd.edu/gtd

IX. CONTRIBUTION OF EACH TEAM MEMBER

1) Srikar Chundury
   - Basic Visualizations
     - Year , Decade , Primary weapon , Country , Target , Attack Type , Suicide etc..
     - Stacked Bar Plot Analysis.
     - Donut Chart.
   - A part of data cleaning - interpolating NA’s with weapon specific means.
   - Naive Bayes Model to estimate the extent of property damage.
   - Knn Model to estimate the extent of property damage.
   - Logistic regression to predict the safe return of hostages.
2) Srinivas Shekar
   - Basic Visualizations
     - Geo referential scatter plot of Vulnerable countries
     - Geo referential scatter plot of Region Mostly attacked
   - Forecasting the number of terrorist attacks in the upcoming years.
     - Initially built ETS model but because of high error[RMSE,ME] when compared to actual data,also built AR,ARIMA model.
     - A part of data preprocessing -since the dataset didn’t have 1993’s data so interpolated the previous year’s data to fill in the values for the year 1993.

3) Swati N H
   - Basic Visualization
     - Geo referential time series animation heat map of frequency of terrorist attacks across globe spanning 50 years.
     - Geo referential time series animation bubble chart of Victim count in several regions across the world spanning 50 years.
     - Geo referential scatter plot of terrorist attacks by type of Attack.
     - Geo referential scatter plot of terrorist attacks by their attack success.
     - Multiple chart measuring lethality of an attack with features such as Victim count and Frequency of attack.
   - A part of Data Cleaning and preprocessing - identifying the necessary complete columns to build models, PCA.
   - Built rule based Decision tree prediction model.
   - K-means clustering technique to cluster similar terrorist groups.