iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46798

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Anatomizing Terrorism

Bhavana Yadav K[1], Charushree A[2], Shreya Vishwas[3]
[1, 2, 3]*Computer Science and EngineeringPES University Bangalore, India*

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 pat- terns and relations undiscovered till now using various machine learning models like random forest classifier,bagging,k-NN,ANN after applying some initial cleaning and exploratory analysis on the data using classic visualising techniques like wordcloud, heat maps,geographical graphs and more.***
Index Terms: GTD, weapon, k-NN, accuracy, random forest

I. INTRODUCTION AND APPROACH

With the mass shooting in Iraq and 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. Terrorismis 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 combat- ants 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. Itis a growing problem across nations and various countries are adopting new security and surveillance measures to tackle thisissue.

In this context, we aim to analyse the GTD [4] databaseand happiness database. 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 the Study of Terrorism and Responsesto Terrorism (START) at the University of Maryland. [5]It includes all the the possible collected data about everyterrorist attack all the way from 1970 to end of 2020 nationallyand 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 ofdetail and accuracy of the data. The World Happiness dataset is a landmark survey of the state of global happiness . The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis,national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations.

We aim to bring meaningful inferences and predictions fromthe datasets which would make the tasks easier for the nationalsecurity analysts and prevent possible terror attacks. We also specifically try to solve two issues here.

1) Given an attack, figure out the attack type.
2) For an attack made by a specific group, estimate theamount of property and human damage caused by the event.
3) Predicting the success rate of an attack given country andits attack type

It is observed that when an attack takes place, the perpetrators are unknown unless someone claims responsibility for theattack and therefore predicting the attack type based on the attack parameters can help the defense forces counteract followup attacks with quicker and effective action since they would be more well informed of their techniques. Similarly, we would like to estimate the typical amount of damage a terror group could create if they attacked which would be useful for government entities so that they can prepare their financing and personnel accordingly for conflict areas where attacks are frequent.

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 retrievaland entity recognition on the sourced data along with theGTD database and uses various visual techniques using pie- charts, maps and networks and mainly tries to find aroundthe 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 onlinenews articles and twitter. From the results, it is found that everygroup which targeted India for a particular decade was closelylinked 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 datavalues might appear skewed or incorrect due to propaganda and nationalist news agencies.

Another approach involves using machine learning algo- rithms such as k-NN and random forest to build classifiers which would group and predict various future terrorism activ- ities based on the GTD databse itself after filtering out columnsthat 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 accuracyof 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 wellas the classification phases.

Similar approach like [3] has been used in [6] where machine learning algorithms notedly, decision trees and ran- dom 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 displaythe 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. METHODS

A. *Data Cleaning And Preprocessing*

Real world data will always contain some noise, outliers andmissing data due to which working directly on the data withoutany modification might cause some errors in our final results. The data collected here is assumed accurate and consistentdue to the fact that it is being maintained by the START organisation on a best-effort accurate basis. There are a lotof 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 fromexperience does a better job.

1) In this dataset we first start cleaning by removing allthe 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 param- eters 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

2) 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 attributeswere brought down to 63 as a result of this.

3) Some irrelevant attributes which were redundant and could be inferred from other attributes or were insignifi- cant were dropped from the dataset.

4) 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 recordstoo ending up with 52660 rows and 41 columns

5) 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. *Building a Model*

For tackling problem one of finding the attack type(bombing,assassination,etc)based on the the attribute of the attack, we develop and compare 3 machine learning and 1 deep learning models-k-NN,random forest classifier,bagging and ANN. Meanwhile estimating the damage done by a group is solved only using random forest classifier.

1) *k-NN:* K-Nearest Neighbour is a supervised lazy learn- ing algorithm which classifies or regresses the test instances based on the target classes and values of the k most similar instances or neighbours to it. It measures the similarity of the nodes by comparing the distances between them where variousdistance measures such as Euclidean,Minkowski, Manhattan, etc are used depending on the nature of the objects.

The most crucial part of k-NN is finding the correct k value to minimise the bias-variance measure of the classifier with the elbow method being the most prominent technique in choosingk with the minimum classification error.

2) *Random Forest and Bagging:* Random forest and bag- ging are ensemble learning techniques which employ multiple decision trees. A decision tree composes of 3 parts: root node,decision and leaf nodes. The training dataset is divided into branches based on the attribute values where attributesfor a node are selected depending upon the entropy valueobtained on splitting the dataset at that node. This is continuedtill leaf nodes are obtained. In bagging several decision trees are trained in parallel on different subsets of the data andtheir combined predicted results is used in classifying a test instance. Random forest classifier is an extension of bagging where at each node split only a portion of the features are selected at random for consideration and the attribute with lowest entropy is selected. It is an improvement of bagging since it prevents overfitting and classifies better with different test instances.

For classifying the perpetrator, the k value was obtained by the elbow method by comparing the classification errors of the test instances with k value in the range of 1 to 9. The valueof k=3 gave the minimum error of 0.18 and was chosen forthe model. The Random Forest model for the first problemwas built with 10 estimators where the value 10 was obtained

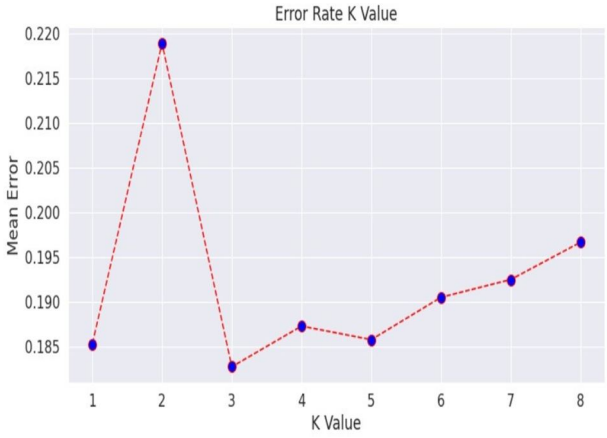


Fig. 1. The elbow method

by trail and error by observing the results that gave optimal accuracy. The bagging technique was devised with the base estimator involving decision trees and a max cap of tree lengthas 10 .

3) *Ann:* the deeplearning model takes input dimension 2 with relu activation layer, and it is passed to the hidden layer with 8 neural nodes and output layer with sigmoid activation

, model is compiled with adam optimizer with accuracy as evaluation metrics. The respective model is trained with 50 epochs.It has an accuracy of 94.63% From this model the success rate of terror attack can be analyzed which wouldgive an on which group has more power for specific region.

C. *Evaluation*

On the cleaned and processed dataset, we make use of crossvalidation which would estimate the skill of the model on newdata and split the data in a ratio of 80-20 for training and testing for all models for first 2 problems. Mean error is a parametric used to determine the accuracy of the model onthe testing dataset. And for the Deep learning model inputwas split into 80-20 ration and accuracy was the metrics totest the model.

IV. RESULTS

A. *Model Performance*

1) *Predicting The Type Of Attack Based On The Attack Pa- Rameters:* Every model will be analysed and run on the same cross validation set to 0.82 with the Random Forest model andbagging fetching similar values of 0.88 with random forest obtaining slightly better results.
2) *Predicting The Property And Human Damage Extent Of An Attack:* The Random Forest model which was built without any hyper parameter gave a staggering accuracy of 0.94 onthe test instances.
3) *Predicting The Success Rate Of Terror Attack:* The model used in deep learning model because the input size is smallso linear models wont perform that well that is the reason a deep neural network is implemented which gave better result of 94.63% .

B. EDA

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

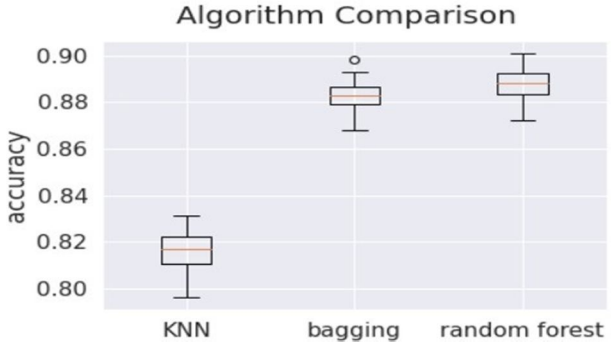


Fig.2. Comparison of different classifiers for attack type

1) Correlation Heatmap

a) natlty1(nationality of target victim) and country where the attack occurred has a value of 0.64 which implies thatmajority of the attacks have a tendency to target same country individuals.

b) weaptype1(Type of weapon used for the attack) and attacktype1(General method of attack like bombing, hi- jacking, 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.

c) 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 showingthat if an attack has resulted in the deaths of many people, a huge number of people would end up alive and wounded.

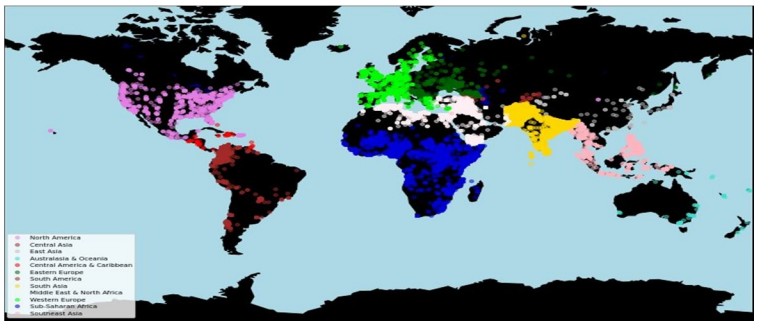


Fig.3. Global Terrorist Attacks, 1970-2019

2) Map Plot

Using matplotlib library of python, the locations of the terrorist attacks are plotted on a world map based onthe 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 aroundthe globe except some noticeable places where almostno attacks take place or very rarely like the poles, Rus- sia, 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.

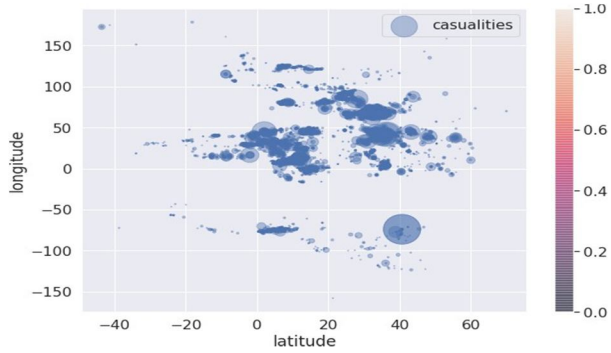


Fig.4

3) Pie Chart

Using matplotlib library of python,pie chart is created with top 15 highly affected countries by human loss .So each country is grouped with number of kills made over the years. From this we can infer that middle east countries are most affected by terror attacks.

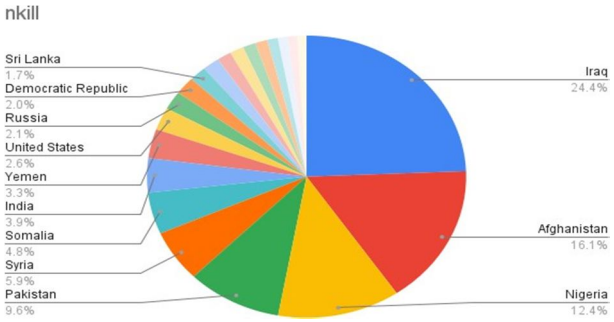
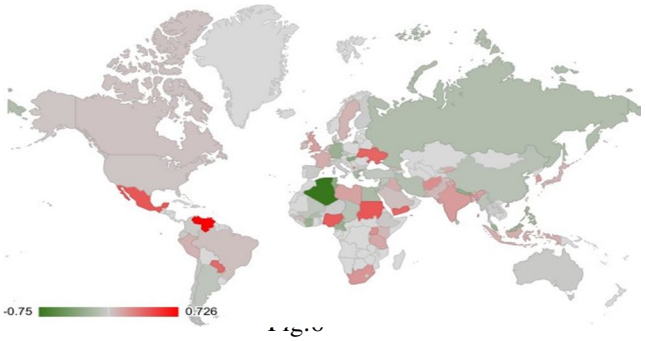


Fig.5

4) Geographical Map

Using matplotlib library of python, the locations of the terrorist attacks are mapped to the happiness difference that is occurred during the years when the terror attack happened 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 happiness index of ones country ,if the happiness index of the country is more its more likelyto get attacked which reduces its GDP and happiness index . So when an attack happens the happiness index of the country reduces drastically and this will affect the progress of ones country.If the happiness index of the country is more then the probability of it getting attacked in near future is really high. We were unable to make a model to predict the probability as the happiness dataset was insufficient to build the model.



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

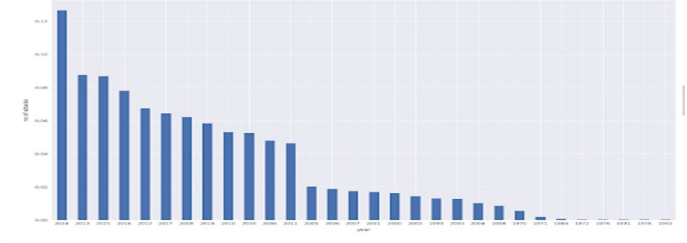


Fig.7 Number of attacks vs years

6) Bar Graph

A bar chart is visualised with the number of attacks onY 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).

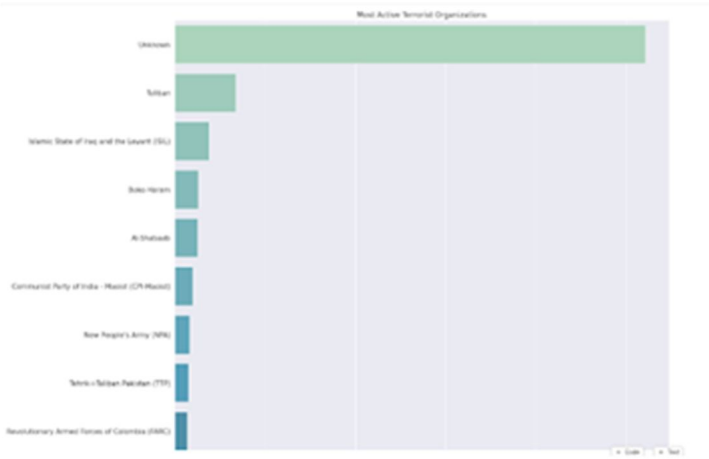


Fig.8 Most active terrorist organisations

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.

7) Word Cloud

A wordcloud is an ingenious way of analysing and representing strings whose appearance is most prominentand 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.

Another wordcloud on the attack summaries reveal the keyword responsibility. This shows that all attack reportswith 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. CONCLUSION

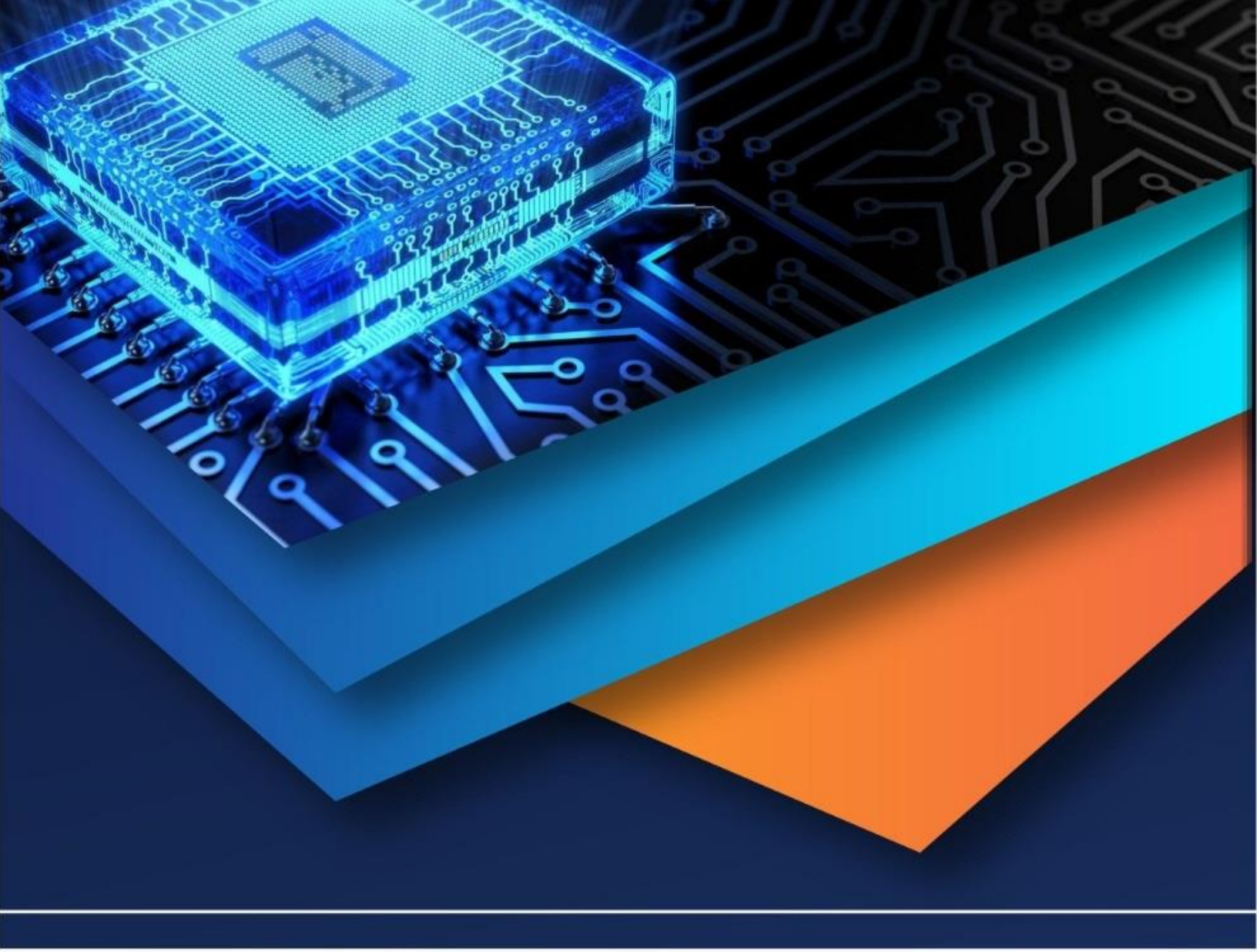
Our analysis of the dataset concludes with the fact that thereis a lot of inferences which can be derived from multiplerecords and attributes present and models could be toyed around and built upon it using various machine learning techniques such as bagging, decision trees, random forest,k-NN, etc.

Fig.9 Wordcloud of attack summaries

Future Improvements is to build a model with sufficient happiness dataset or an alternate approch is to predict the probabiltiy of decrease in happiness index of a country when there is an attack. The deep neural network with ann architecture performed well in creating better success ratecompared to linear models for predicting success rate of an attack. Through our experiment we found out that random forest is better than bagging and K-nn with respect to thisdata set. Our models built can predict information about an attack accurately and therefore would be a valuable asset to professionals and government entities which would help in keeping the terrorists at bay.

REFERENCES

[1] Randy Borum, The Psychology of Terrorism, University of South Florida, 2004.
[2] Lavanya Venkatagiri Hegde, Nerella Sreelakshmi and Kavi Ma- hesh.,"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] Kaggle,World Happiness Report up to 2022,January 2022.
[7] Enrique Lee Huaman´ı, Alva Mantari Alicia and Avid Roman- Gonzalez.,"Machine Learning Techniques to Visualize and Predict Ter- rorist Attacks Worldwide using the Global Terrorism Database", Inter- national Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY