Terrorism in India: Integrated Approach Using Machine Learning models And Data Mining Analysis

Navanshu Khare,
Indore,India

ABSTRACT: Terrorism is becoming one of the most tedious problems to deal with and a grave threat to mankind and it is evident that there has been enormous growth in terrorist attacks in recent years. This paper is to enhance counter-terrorism by predicting the most accurate terrorist group responsible for an attack that caused massive losses of lives and property and to analyzes and enumerates the losses occurred, trends in attack frequency and places more prone to it, by considering the attack responsibilities taken as evaluation class. This paper analyses Data set provided by Global Terrorism Database which is a publicly available database and contains information on terrorist event far and wide from 1970 through 2017 comprising 156,772 reported attacks in India by modelling the behavior of terrorist groups using machine learning algorithms like Random Forest, Naive Bayes, K-nearest neighbor and Neural Network and then combining them to form Ensemble Classifier using . It becomes challenging due to reason that dataset is unbalanced and needs meticulous preprocessing. Ensemble classifier gave the highest possible accuracy of 93.6% to find the terrorist groups responsible for various attacks and important observations regarding the terrorism pattern are deduced. The results of the evaluation are classified based on accuracy, precision, recall and F1 score. Random forest and Naïve bayes also performed excellently, neural network performance can be classed as decent and satisfying while performance of k-nn needs more tuning.

Keywords—Terrorism, Data mining, Random Forest, Naïve Bayes, K-nearest- Neighbor, Neural Network, Ensemble approach

I. INTRODUCTION

Since the past two decades Terrorism has increasingly become a major security concern for the Countries and continuous growth in terrorism is not putting rest to it. Terrorism clearly has a very real and direct impact on human rights, with devastating consequences for the enjoyment of the right to life, liberty and physical integrity of victims. In addition to these individual costs, terrorism can destabilize Governments, undermine civil society, jeopardize peace and security, and threaten social and economic development. The Global Terrorism Index measures the direct and indirect impact of terrorism, including its effects on lives lost, injuries, property damage and the psychological aftereffects. It is a composite score that ranks countries according to the impact of terrorism from 0 (no impact) to 10 (highest impact) and India Scores a whooping score of 7.57 in 2019 and Stand 7 in the List of out of 138 Countries.

The objectives of this research include classifying the terrorism data using various machine learning algorithms, construction of classification models with highly accurate algorithms and then combined to make majority vote base ensemble. Use these models to predict the terrorist groups responsible for attacks in various parts of India and using terrorism specific domain knowledge to analyze and extract macro-level conclusions about the pattern of terrorist to narrow down the terrorist risk to the target. The study has been evaluated and tested by using the real dataset taken from the Global Terrorism Database (GTD) managed by the National Consortium for the study of Terrorism and Responses to Terrorism (START), at the University of Maryland and is considered to be the most comprehensive data globally covering over more than 170,000 reported terrorist incidents.

Prediction is difficult problem to solve mainly because of low incidence of terrorism, and the tendency of terrorist tactics to evolve fairly rapidly, makes it difficult to build good predictive models. There are multiple different roles in terrorism, and multiple pathways to fulfilling those roles. This means that it is impossible to list definitive indicators of involvement or exclusion from terrorist activities. The data is filtered after preprocessing by means of several filters such as linear and non-linear. The final and the major pre-processing is ended by solving the class imbalance problem in the dataset. It is a situation where observations that belong to one class are significantly lower when compared to the observations of other classes. This will make the test results biased towards the classes with higher observations. Here data set is highly biased towards Maoists with around 1165 occurrences against the classes having frequency of 2.

From the experiments the result that has been concluded is that Ensemble classifier, Random forest and Naïve bayes also performed excellently, neural network performed decent and satisfying while performance of k-nn needs more tuning. Among all classifiers Ensemble model showed the best result. Multiple Observations are deducted on analyzing the data. Terrorism in India is continuously growing since 2008 with Maoist most active terror groups in India. Terrorism is highest in Jammu Kashmir where terrorist mostly target civilians and police.

II. RELATED WORK

This section highlights background related work for the terrorist group prediction model and various algorithms to tackle this issue.

In [11] The author Pawan H. Pilley and S. S Sikchi [24] have used the CLOPE algorithm to perform analysis and reviewed the terrorist group prediction model. Historical data is used to detect the terrorist group and a relationship is determined between terrorist group and the attacks occurred before. CLOPE clustering algorithm is used to make the clusters of the data that is particularly used for the categorical features. It is concluded through analysis that CLOPE algorithm is quite efficient in prediction.

In [15] This paper reviewed Ensemble-based classifiers in his research presented the idea of combining different classifier to build ensemble classifiers for enhanced performance. Innumerable techniques can be used for combining classifiers one of which is majority voting. In majority voting technique classification of an unlabeled instance is performed according to the class that obtains the highest number of votes.

In [18] The Author I. Rizwan and A. Masrah [1] compare two different classification algorithms that is to say, Decision Tree and Naïve Bayes for forecasting "Crime Category" for different states in UAA. Tenfold cross validation was applied to the input dataset in the experiment, separately for both Naïve Bayes and Decision Tree to evaluate the accuracy of the classifiers which showed that DT algorithm out performed NB algorithm and achieved 83.951% accuracy in predicting "crime Category".

In [19]. The aim of this research is to offer two diverse approaches to handle the missing data as well as provide a detailed comparative study of the used classification algorithms and evaluate the achieved results via two different test options. With the help of Weka, experiments are performed on real-life data to conclude the final evaluation based on four different performance measures. Results from mode imputation approach showed that SVM, is more accurate than NB and KNN, ID3 has the lowest classification accuracy although it performs well in other measures, and in Litwise deletion approach, KNN outclassed the other classifiers in its accuracy, but the overall performance of SVM is satisfactory than other classifiers.

In [22] This research proposes a different ensemble framework for the classification and prediction of the terrorist group that consists of four base classifiers, naïve bayes (NB), K nearest neighbour (KNN), Iterative Dichotomiser 3 (ID3) and decision stump (DS). To combine these classifiers, Majority vote based ensemble technique is used and then the results of individual base classifiers are compared with the majority vote classifier and it is observed through experiments that ensemble approach achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifiers.

This research is to identify the various groups responsible for the terror attacks in India using the GTD database by applying Random Forest, Naïve Bayes, K-Nearest Neighbor and Neural Network and to combine them to form an ensemble model to evaluate their Accuracy, precision, recall and F1 score to state the best possible model to provide a support in fighting against terrorism.

III. BACKGROUND

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network. One of the most frequently overlooked challenges of predictive modeling is acquiring the right data to use when developing algorithms and process it.

A. Random Forest

Random Forest is a classification and regression algorithm with collection of decision trees using approach of divide-and-conquer to improve performance. Random forest trains multiple decision trees and then ensemble them to create "forest". A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. When classifying a new unknown data point, each decision tree will then test the observation and vote on which class it believes the observation to be. By majority vote, the random forest will output the most likely classification. Each tree is trained independently of the others therefore data instances training is extremely fast and is also effective against large data sets. however, since main principle behind random forest is to group weak learners to form a string learner, some of interpret ability of single tree is lost thus increasing computational complexity exponentially.

$$H = -\sum p(x)\log p(x)$$

Figure 1: The entropy

equation **B. Naïve Bayes**

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. The central feature of Naive Bayesian model is that it is easy to build, with no complex iterative parameter estimation making it particularly useful for huge datasets., it recast the dataset into a frequency table and then later it creates likelihood table by finding the probabilities. Lastly it calculates the posterior probability for each class using the Naïve Bayes equation predicting the class with greatest posterior probability. It assumes the presence of a peculiar feature in a class is unlinked to the presence of any other feature. Naive Bayes models can be considerably faster than many of the other refined methods. Naïve Bayes also alleviate problems stemming from the curse of dimensionality by decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution.

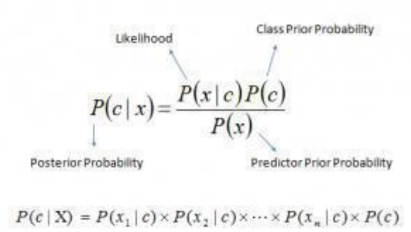
Likelihood
Class Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X)=P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$

Figure 2: Bayes

Theorem **C. Neural Network**

A neural network is a series of organized layers that are made up of interconnected nodes which predict the outcome by deriving a pattern by communicating with each other. Patterns are presented via input layer which then communicates with one or more hidden layer where adaptive weights between neurons are tuned by a learning algorithm that learns from observed data in order to improve the model. In addition to the learning algorithm itself, one must choose an appropriate cost function. optimization techniques such as gradient descent or stochastic gradient descent are used *to* determine the best optimal solution and values for all of the tunable model parameters, with neuron path adaptive weights being the primary target, along with algorithm tuning parameters such as th*e* learning rate.Since Neural networks can adapt to changing input, so the network generates the best possible result in the output Layer.

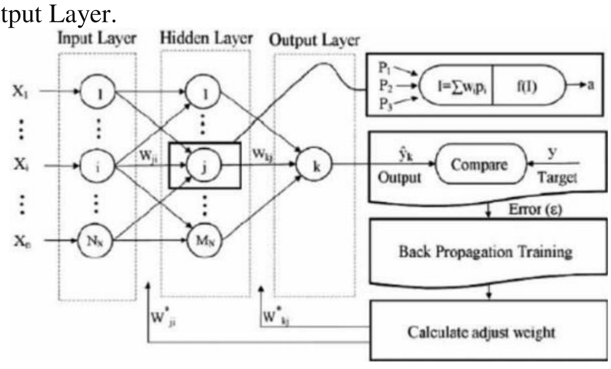

Figure 3: Architecture of Neural

Network **D. K-Nearest-Neighbor**

The k-nearest neighbors (KNN) algorithm is one of the simplest similarity-based artificial learning algorithms, offering interesting performance in some contexts. The input data that is used as training data-sets of k-NN algorithm is plotted against multi-dimensional element, which is assigned into different segments which are then characterized based on the order of training data-sets. The class of the new instance is then determined according to the most frequent class among the k nearest neighbors. The choice of the value of k must be chosen a priori low value of K will make predictions less stable and inversely high value of K will increase the risk of error, various techniques have been proposed to select it such as cross-validation and heuristics. The performance of KNN also depends largely on the measure used to calculate the distances between the instances. To select the K that's right for the data, KNN algorithm is run several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm's ability to accurately make predictions when it's given data it hasn't seen before.

$$d(\mathbf{p},\mathbf{q})=d(\mathbf{q},\mathbf{p})=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$
$$=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Figure 4: Euclidian Distance formula

IV. METHODOLOGY

An orderly procedure is followed to ensure accurate and fitting results which guide the results in a correct manner. The dataset is taken from the source and then pre- processed to remove any discrepancies and inconsistencies, which is then used to build the classifier from which the results are inferred.

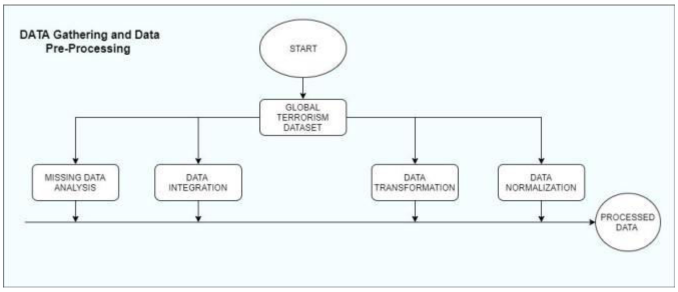

Figure 5: Data gathering and Data Pre-processing

Data Collection Data for this particular research is obtained from the Global Terrorism Database (GTD) managed by the National Consortium for the study of Terrorism and Responses to Terrorism (START), at the University of Maryland and is considered to be the most comprehensive data globally covering over more than 170,000 reported terrorist incidents. This data set attributes around 137 variables which showcase the details like type of attack, time,

place, causalities, tactics, perpetrators and terrorist group responsible etc. collected from over the period of 1970-2016. This particular research covers the Terrorist Incidents related to India and 18 Variables are filtered to predict the terrorist group responsible. The classifier can predict good results only if pre-processing is done on the dataset in the most efficient way. Data Preprocessing is a transformation of a raw unstructured data to understandable and optimized data. Data is cleaned by treating missing values and removing noisy data.

Data Transformation by Normalization and are used for data transformation and lastly data reduction techniques like data aggregation and dimensionality reduction are applied to obtain the incidents related to Indian soil and from entire dataset with 137 columns, 18 rows are extracted. These attributes are selected on the basis of their relevance to the predicted attribute (terrorist group). Aggregation of Fields nkill, nwound and nhost to create a new casualties field is done Since this research focuses on the prediction of the group responsible for the terror attacks, those rows with empty fields of group responsible were removed as they are useless records.

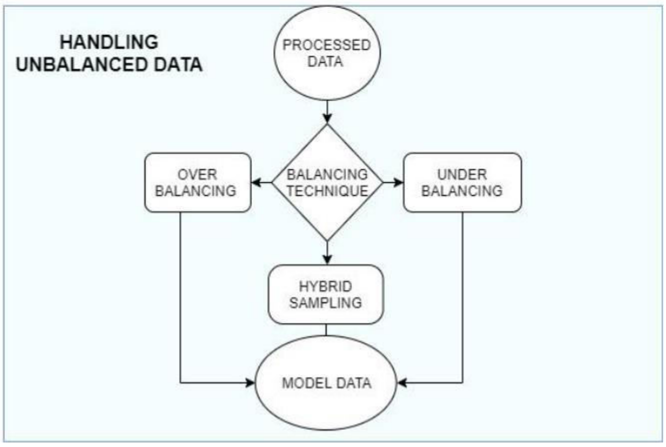

Figure 6: Handling unbalanced data

One major impedance was that the dataset was highly imbalanced that is some classes like Maoists have very high frequency of occurrence. This limitation was overcome by balancing the data using Overbalancing, Under Balancing and Hybrid Balancing by SMOTE. Over sampling works by replicating the minority class records whereas under sampling excludes some records of the majority classes so that the discrepancy between the minority and majority classes lessens and hybrid sampling is a combination of both under sampling and oversampling. In the current dataset oversampling and hybrid sampling are applied to solve the class imbalance problem. The SMOTE (Synthetic Minority Over-Sampling Technique) algorithm is used to perform the hybrid sampling procedures.

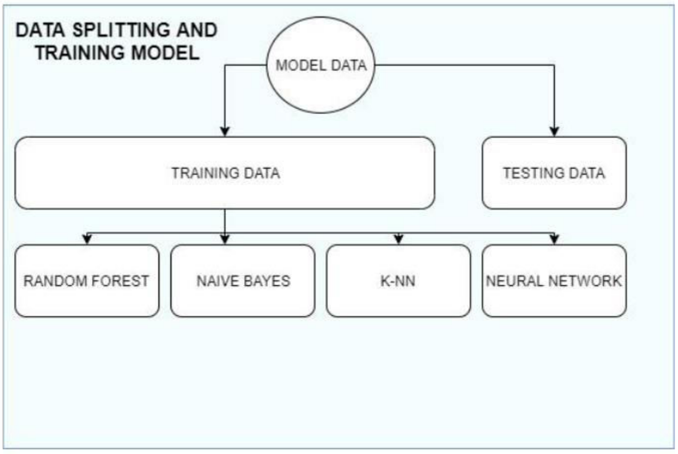

Figure 7: Data splitting and training

This data is then separated into two part, one is training dataset and another one is test data set. Now K fold cross validation is done that is the original sample is arbitrarily partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k −1 subsamples are used as training data, first with training data, Individual classifier like Naive Bayes, Random Forest, K nearest neighbor and Neural Network are created and then in second step these individual classifiers are combined together to create a new ensemble method.

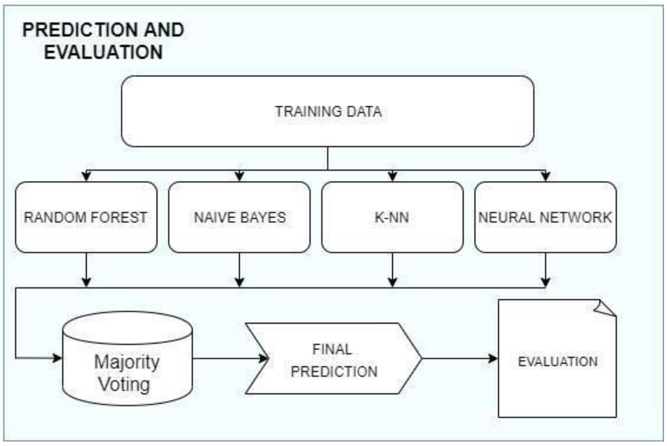

Figure 8: Prediction and evaluation

The performance of Classifiers is evaluated based on accuracy, precision, recall and F1 Score.

True positive (TP): Prediction is positive and group is responsible, we want that.
True negative (TN): Prediction is negative and group is not responsible, we want that too.
False positive (FP): Prediction is positive and group is not responsible, false alarm, bad.

False negative (FN): Prediction is negative and group is responsible, the worst.

Accuracy can be said as a degree to which the outcome of a measurement, calculation, or specification conforms to the true value or a standard. Accuracy is a beneficial metric when all the classes are equally important but not true when working with a class-imbalanced data set. Precision-Recall is a suitable degree of success prediction when the classes are imbalanced. The metric our intuition tells us we should maximize is known in statistics as recall, or the ability of a model to find all the relevant cases within a dataset. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant. F1 score is a singular metric that pools recall and precision using the harmonic mean and

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precison = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,Score = \frac{2 * Precison * Recall}{Precison + Recall}$$

V. RESULTS AND OBSERVATIONS

METRICS				
CLSSIFIER	Accuracy	Precision	Recall	F1 Score
Random Forest	*0.903*	*0.901*	*0.946*	*0.922*
Naïve Bayes	*0.872*	*0.887*	*0.932*	*0.908*
K-Nearest Neighbor	*0.829*	*0.786*	*0.833*	*0.808*
Neural Network	*0.846*	*0.843*	*0.877*	*0.859*
Ensemble Approach	***0.936***	***0.929***	***0.964***	***0.946***

Table 1: Performance Matrices

Terrorism in India is continuously increasing with around total 618 attacks in 2008 and highest 976 attacks in 2016 with an average of around 266 attacks per year.

India has faced around 4168 terrorist attacks in the 5-year span of 2013-2017

CPI Maoist is the most active group in India and activities of CPI increased rapidly after 2007 and in 2010 CPI was on its peak.

Sikh Extremist is second most distressing terror group in India. Sikh Extremists were active from 1882 to 1993 but with time become less active.

LeT has caused greater no. of casualties in less no. of attacks and Sikh Extremist has killed the most.

Unknown group refers mostly to the terrorists which cross border from Pakistan to Jammu and Kashmir to enter India but no terrorist group in Pakistan takes responsibility for such attacks to avoid any suspicion over their collaborations with Pakistan

Jammu and Kashmir is the most affected state by terrorism with around 2438 terror attacks followed by Chhattisgarh with around 912 attacks.

Terrorist mainly target Civilians and property followed by Police force.

The intent of Naxalites and Maoists is to paralyze the state machinery and cause public damage. These groups mainly target state officers and state property

VI. CONCLUSION

This paper addresses the problem of terrorism faced by India by using data mining and machine learning techniques. After pre-processing the data meticulously, class imbalance problem of the dataset is solved using Under Sampling and Oversampling Techniques. This pre-processed data is used to create classification models and later combined to create an ensemble model and the models are analyzed using various parameters like Accuracy, recall, precision, f1-score. Ensemble classifier gave the highest possible result to find the terrorist groups responsible for various attacks if the perpetrator is unknown. Random Forest and Naïve also performed well in identifying the responsible terror group. Neural network was satisfying but K-nn needs more tuning. This helps the anti-terrorist organizations to reduce the list of possible suspects and help them act rapidly to find and catch the real suspect. The study also offers observations from the results which can be used efficiently to narrow down the terrorist risk faced by the states and shows the culprit groups with the highest frequent attacks and casualties. It can also be established that biggest threat to security of India is from home grown terror groups such as Naxalites and Maoists (they maybe be funded by foreign elements). Terrorism through foreign groups such as Let, JeM are contained in Jammu and Kashmir and have not spread through much of India but are serious concern in Jammu Kashmir. In future, the research aims to study in detail the about the distinct terrorist groups and their terrorism pattern in India, also relations between various terrorist group can be analyzed by different algorithms and methods like deep learning models.

VII. References

[1] A R. Kalpana and K. L. Bansal.: A comparative study of data mining tools. In International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, (2014).

[2] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi and M.A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers", 2007 International Conference on Convergence Information Technology, IEEE DOI 10.1109/ICCIT.2007.148.

[3] Goteng GL, Tao X. Cloud computing intelligent data-driven model: Connecting the dots to combat global terrorism. IEEE International Congress on Big Data (BigData Congress). 2016. p. 446-53

[4] F. Bolz et al. The Counterterrorism Handbook: Tactics, Procedures, and Techniques, CRC Press, (2001).

[5] S. Yingbing and L. Qingsun, "KNN algorithm based on feature weighted", Hanna University of Science and Technology,2008, Vol. 26 No. 4,352-355.

[6] Han., J., Kamber, M., Pei., J.: Data mining concepts and techniques. 3rd edition, Morgan Kaufmann (2012)

[7] Mahmood T, Rohail K. Analyzing terrorist incidents to support counter-terrorism - Events and methods. International Conference of Robotics and Artificial Intelligence; 2012. p. 149-56. Crossref

[8] Ghada M. Tolan and Omar S. Soliman.: An Experimental Study of Classification Algorithms for Terrorism Prediction. International Journal of Knowledge Engineering, Vol. 1, pp. 107-112. DOI: 10.7763/IJKE.2015.V1.18 (2015)

[9] Kathleen McKendrick : Artificial Intelligence Prediction and Counterterrorism, International Security Department | August 2019

[10] S. Neelamegam and E. Ramaraj, "Classification algorithm in data mining: An overview," International Journal of P2P Network Trends and Technology (IJPTT), vol. 4, p. 369, Sep. 2013.

[11] P. H. Pilley and S. S. Sikchi, "Review of group prediction model for counter terrorism using CLOPE algorithm," International Journal of Advance Research in Computer Science and Management Studies, vol. 2, issue I, ISSN: 2321-7782, January 2014

[12] S.Ozekes and O. Osman, "Classification and prediction in data mining with neural networks," Journal of Electrical and Electronics Engineering, vol. 3, no. 1, pp. 707-712, 2003

[13]] J. Baylis, S. Smith, and P. Owens. The globalization of world politics: an introduction to international relations. Oxford: Oxford University Press, (2017). [Print]

[14] S. Ejaz Hussain, "Terrorism in pakistan: Incident Patterns, Terrorists' Characteristics, and The Impact of Terrorist Arrests on Terrorism" (2010).Publicly accessible Penn Dissertations.Paper 13

[15] C. D. Amato, ,D. Malerba , F. Esposito and M. Monopoli, "Extending The K-Nearest Neighbour Classification Algorithm To Symbolic Objects", Atti del Convegno Intermedio della Societ à Italiana diStatistica"Analisi Statistica Multivariata per le scienze eco2nomico2sociali,le scienze naturali e la tecnologia". Italia :Napoli ,2003.

[16] F. Bolz et al. The Counterterrorism Handbook: Tactics, Procedures, and Techniques, CRC Press, (2001).

[17] "Classification, LDA" in Data Mining and Analysis. Stanford University. [Online].

[18] I. Rizwan, A. Masrah, A. M. Aida, H. Payam, and K. Nasim, "An experimental study of classification algorithms for crime prediction," Indian Journal of Science and Technology, vol.6, March 2013.

[19] Ghada M. Tolan and Omar S. Soliman ,An Experimental Study of Classification Algorithms for Terrorism Prediction, International Journal of Knowledge Engineering, Vol. 1, No. 2, September 2015

[20] F. Ding, Q. Ge, D. Jiang, J. Fu, and M. Hao. Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. Plos One, 12(6), (2017)

[21] K. Dalacoura. Islamist Terrorism and Democracy in the Middle East, Cambridge: Cambridge University Press, (2011).

[22] Faryal Gohar1, Wasi Haider Butt2, Usman Qamar3, Terrorist Group Prediction Using Data Classification Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition, Kuala Lumpur, Malaysia, 2014

First A. Author: Navanshu Khare is a Software Engineer at Tata Consultancy Services .He completed his B.Tech in Computer Science from SRM University, Kattankulathur. His areas of interest includes Data Science and Neural Networks.