Using Global Terrorism Database (GTD) and Machine Learning Algorithms to Predict Terrorism and Threat

S. Kalaiarasi, Ankit Mehta, Devyash Bordia, Sanskar

Abstract: ***It is evident that there has been enormous growth in terrorist attacks in recent years. The idea of online terrorism has also been growing its roots in the internet world. These types of activities have been growing along with the growth in internet technology. These types of events include social media threats such as hate speeches and comments provoking terror on social media platforms such as twitter, Facebook, etc. These activities must be prevented before it makes an impact. In this paper, we will make various classifiers that will group and predict various terrorism activities using k-NN algorithm and random forest algorithm. The purpose of this project is to use Global Terrorism Database as a dataset to detect terrorism. We will be using GTD which stands for Global Terrorism Database which is a publicly available database which contains information on terrorist event far and wide from 1970 through 2017 to train a machine learning-based intelligent system to predict any future events that could bring threat to the society.***

Keywords: Data Mining; Global Terrorism Database (GTD); Social Media Threats;

I. INTRODUCTION

There has been a huge growth in Internet users in the past decade. Technological advancement has not only benefited the society but also has given rise to various problems in the society. One of such threats is the growth in cyber terrorism. Due to exponential increase in Internet users it is evident that people have started using this technological advancement for carrying out unlawful activities. Purposeful use of computers, web and various other systems for hurting others or carrying out un-lawful activities or personal motive can be called as cyber terrorism. Like every coin has two sides, technological advancement has also its advantage and disadvantage. Nowadays we are exposed to much more vulnerability and risk on the web on the hands of predator or cyber-terrorists.

Estonia, a Baltic nation which is continually developing as far as innovation, turned into a battleground for cyber terror in April, 2007 after debates with respect to the expulsion of a WWII soviet statue situated in Estonia's capital Tallinn.

It is very necessary to identify the fact that cyberterrorism has had very disastrous effect in our society. We intend to use various Machine Learning algorithms to analyze, predict and categorize various terrorist activities using various algorithms like Natural Language Processing (NLP), sentiment analysis, k-NN algorithm and random forest algorithm to train a system by feeding it with data from Global Terrorism Database and various threats related text on Social Media.

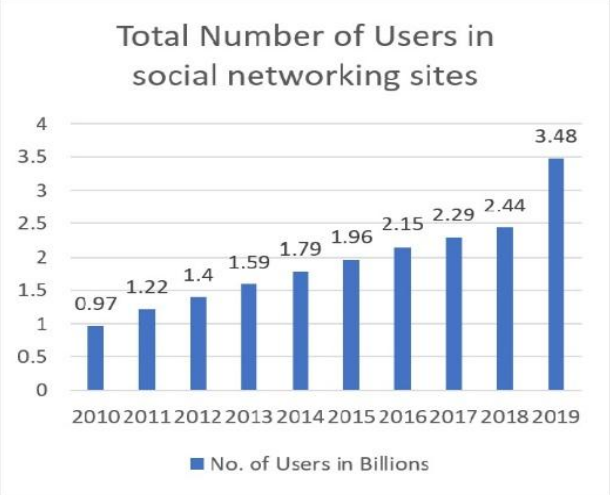

Fig.1. Bar-diagram depicting number of social media users

II. RELATED WORK

This paper intends to deal with the following details: -

a. To Study and investigate various techniques and work done used to identify and channel vicious and hurtful solitary substance online on the web.
b. To play out a comprehensive investigation and contemplating of these systems and discover basic patterns.

The underlying advance is to begin by thinking of a few meanings of withdrawn conduct, and after that quest up for some ideas on terrorist conduct.

Studying asocial practice: Antisocial conduct, incorporates exercises, for example, "trolling, cyberbullying, grieving and criminal activities", they have generally been utilized and discussed. Research on Online people group to recognize introverted conduct has significantly led with subjective measures instead of quantitative.

These examinations for the most part included the various kinds of social defaming which happens and persuasive explanation for performing these types of exercises and various systems that are used by many people in light of these kinds of activities.

Techniques to distinguish antisocial conduct: There are a few similar kinds of studies done on asocial practice. In any case many had been completed in this domain. Endeavors had been additionally done in the field of recognition along with avoidance of substantial indications identified with reserved messages in networks.

A. Antisocial Content Identification

The Fundamental goal is to discover untruthful, exploitative, bigot and fake substance on the online networks. Influencing the attitude of an individual or to coast misleading information to the network is the primary undertaking. Examining this information will result in the wellspring of abhor advertiser or reserved information.

B. Obtaining Criminal Information

Searching as well as investigating for criminal information and evidence can give concealed examples along with obscure patterns utilized by every lawbreaker. In spite of the fact that it is documented as well as investigated by concerned authorities that manages the wrongdoings and offenders to keep away from its abuse.

C. Cyber Infrastructure Shielding

It has the procedure to ensure digital foundation if there should be an occurrence of digital dangers and digital assaults. These assaults could be a foundational withdrawn action. It can also be an endeavor for risking the safety measures.

There can be numerous strategies as well as procedures for examining an individual's brain however techniques haven't been created for natural identification of asocial practices on the internet. Information as well as content available on interpersonal organization is either not structured or not classified information which turns into significant obstacle for locating reasonable technique for natural identification of anti-social practice.

III. DESIGN AND METHODOLOGY

Following approach were used for accomplishing the required target:

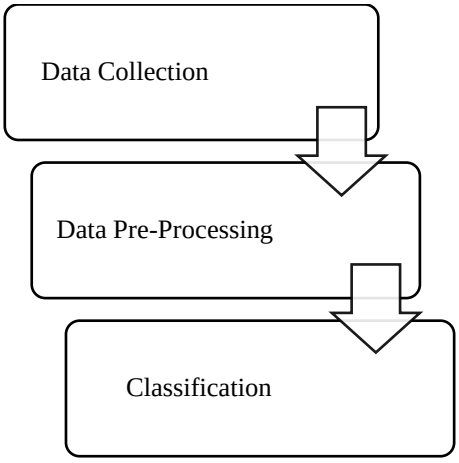

Fig.2. Structure Map

A. Data Set

The Global Terrorism Database, is an organization which is financed by (START) which stands for National Consortium for the Study of Terrorism and Responses to Terrorism, is a dataset of terrorist events ranging from a period length since 1970. The GTD holds more than 170 thousand assaults by terrorists having more than 40 factors for every occurrence, that made this dataset as "right now finest exhaustive unsorted information dependent upon psychological militant occasions on the planet. In perspective on its effectiveness and exhaustiveness, GTD information has highlighted in various scholastic papers as exact investigation for the examination on different parts of fear-based oppression, from current patterns to sorts of psychological oppression.

B. Data Preprocessing

The GTD incorporates in excess of 170,000 instances of terrorist events far and wide from 1970 to 2017. For every occurrence, data is accessible on the area and date of the comparing episode, summing up to 132 characteristics. Be that as it may, the information was gathered from various information assets, which would bring about information inconsistence. Regarding the present problem, we are setting the limit of offer proportion to twenty percent, that indicates those traits which have been documented inside over twenty percent of the all-out occurrences are investigated. Ensuing measurable investigation, more than 50 properties were chosen but left out information yet resides inside portion's record. For taking care of the issue, the Mean Imputation (MI) technique is utilized, trading left out information using every single recognized estimation of the trait of the set in which the left-out quality has a place.

Terrorism can be categorized into:

a. Politics Related Terrorism: Terror attacks carried because of political dispute between parties.
b. Protesting Terrorism: The terror attacks carried out in protest of ruling party or because of unsatisfactory decision against a particular group or region.
c. Religious Terrorism: Terror attacks carried because of religious disputes between various religious groups.

d. Underworld Terrorism: Terror attacks carried out by organized criminal groups like mafia using dangerous and unlawful methods to disrupt society for money related advantage.

Table-I: Various categories of Terrorism

TYPE/DATA	SAMPLE
Political	99,530
Protesting	20,678
Religious	24,243
Underworld	27,290

C. Classification

Appointing items to one of a few preset classes is the fundamental concept of classification. Two broadly utilized classifiers have been considered, k-Nearest Neighbor (k-NN) and Random Forest Algorithm.

i. K-Nearest Neighbour Classification Approach

k-NN falls under automatic learning algorithm category. K-NN algorithm is considered to be an instance-based algorithm. This algorithm is one of the notable methodologies in the field of example acknowledgment. The k-NN algorithm is normally and broadly utilized in categorizing text, and classifying text. k-NN algorithm is one of the widely used algorithm for mining purpose and is popularly used in classification of text and is effectively adaptable to enormous applications. Since k-NN algorithm can have multiple classifiers and class labels, it is preferred in multi-model classes.

The input data that is used as training data-sets of k-NN algorithm is plotted against multi-dimensional element, which is portioned into different areas which are then characterized based on the order of training data-sets.

Fig.5 demonstrates a diagram of locale region of a k-NN classifier which includes three unique classifications, in particular w1, w2 and w3. These classifications are related with the preparation information and circles, triangles as well as square-blocks are marks of information point's w1, w2 and w3, separately. In the following figures, 'X' means the information preparing information for grouping.

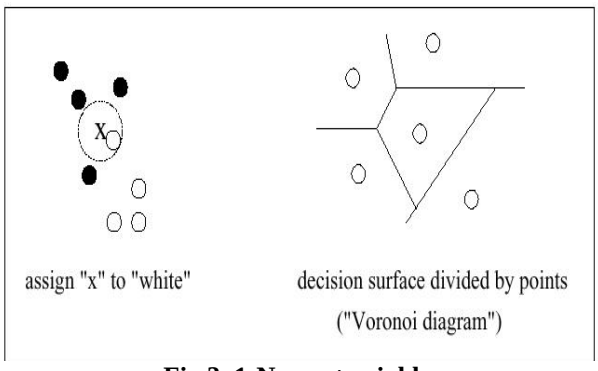

Fig.3. 1-Nearest neighbor

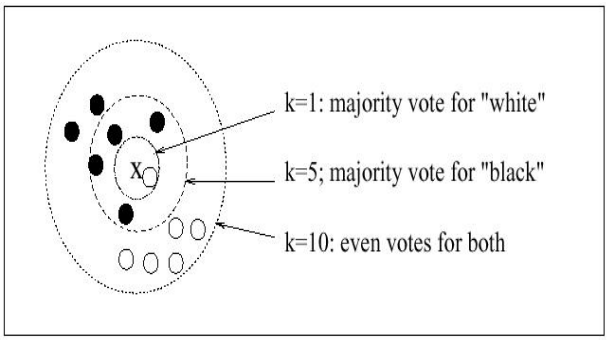

Fig.4. k number of Nearest Neighbors

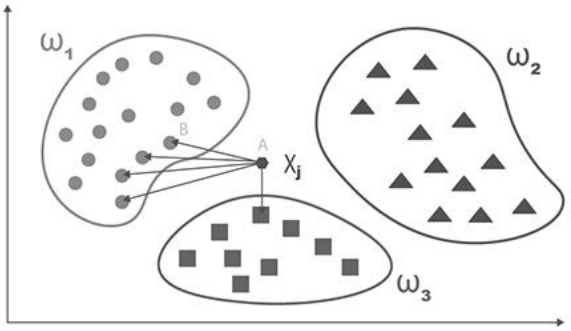

Fig.5. Three-dimension region space of a k-NN classifier

ii. Random Forest

Random Selection as well as bagging are the two machine learning methods that RF merges together to combine their benefits. In bagging forecast is made by selecting the maximum number of votes. This is done by training each and every tree, using bootstrap test. While Random-Feature selection looks for the best part in every node over irregular portion of the highlights. It is a renowned coordinated learning algorithm by taking decision tree as the fundamental classifier. It has demonstrated its accomplishment in applications like email spam filtering, voice classification, and picture classification and text classifier. To order another document from the information vector, it passes the information vector through every one of the trees of the forests with each tree giving a result, i.e. a classification, which is named as "votes" for that specific class. Like the election results, the last result would be the class that has the most votes. The most important features of Random Forest algorithms are enlisted below:

a. Random Forest is un-affected by noise as well as outliers.
b. Random Forest gives better result and is faster than boosting as well as bagging
c. Incase of huge data sets RF is considered to be quite efficient.
d. In comparison to other readily available classification algorithms, RF is considered more precise and hence superior.
e. In case of missing data, Random Forest Algorithm is considered to be more productive for evaluating purpose.

IV. RESULTS

In this paper we have explored various classifiers on the basis of their accuracy and speed. We finally went with k-NN algorithm for Weapon Classifier and Random Forest algorithm for Perpetrator Classifier. k-NN algorithm was best suitable for Weapon Classifier and similarly Random Forest Algorithm stood out to be best for Perpetrator classifier on our multiple attempts using various algorithms. The data-sets generated using GTD was divided into the ratio of 8:2 for training and testing our model respectively.

A. Visualization

We have visualized various contents from the Global Terrorism Database, to better understand the Datasets we are provided with. We have visualized various attributes from the year 1996 to 2017 that were enlisted in the Global Terrorism Database. Following are the diagrams generated from the GTD after visualization of the contents based on attacks by year, fatalities by year, countries by total attack, and attacks by type. The diagrams below have been plotted using "catplot()".

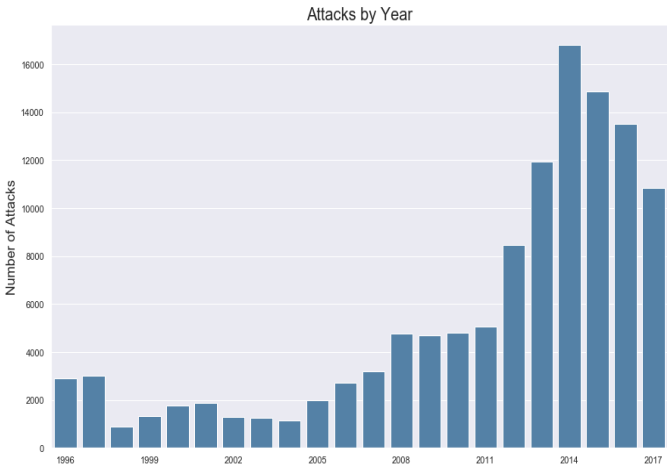

Fig.6. Attacks by Year

The above bar diagram shows us the number of attacks carried out every year from the year 1996 to 2017. We can conclude that the attacks were increasing from the year 2011 to 2014 and then it started to gradually drop.

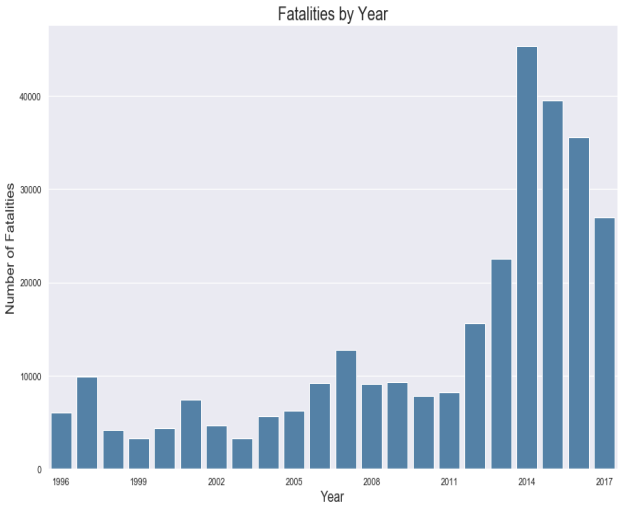

Fig.7. Fatalities by Year

The above bar diagram shows the number of fatalities that happened every year from 1996 to 2017.It is evident that the most fatalities happened between the years 2011 to 2017.

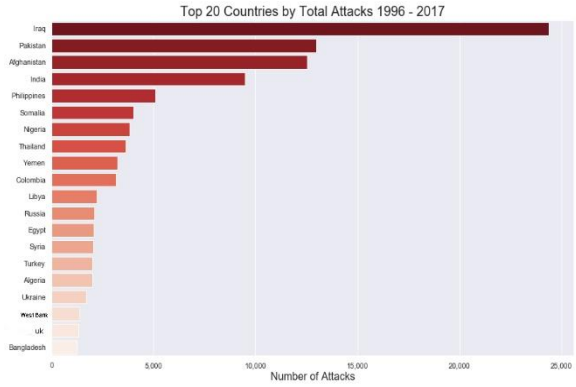

Fig.8. Top 20 Countries by Total Attack

The above diagram lists the countries based on the most number of terrorist acts acted upon. It is evident that Iran, Pakistan, Afghanistan and India are the top 4 countries where most number of attacks have happened from the year 1996-2017.

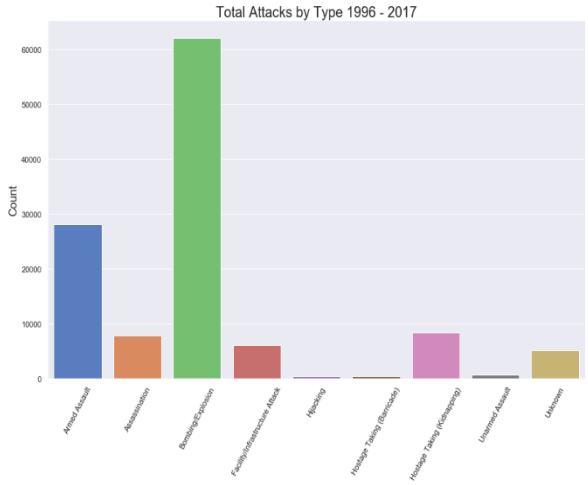

Fig.9. Total Attacks Type

The above bar diagram depicts the total attacks by type. It is evident that bombing and explosion followed by Armed Assault are the most preferred attack types with a total count of more than 90,000 attacks from the year 1996-2017.

Fig.10. Terrorism Hotspots

The World map in Fig.10 shows the hottest Terrorist locations, showing active terrorist encounters in South and East Asia.

B. Weapon Classifier

The weapon classifier was built using k-NN algorithm that classifies the attacks based on types of weapons. We have chosen 'K' by trial and error strategy for which we obtained the ideal outcome. We have calculated the neighbors from the GTD where k=12.

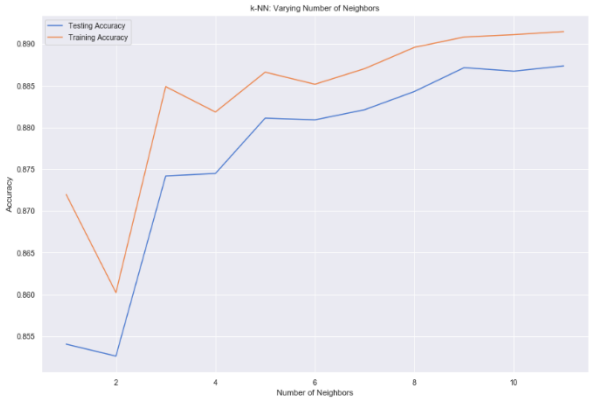

Fig.11. k-NN Varying Number of Neighbors

The above graph was generated using k-NN algorithm by trial and error. The graph shows the accuracy for 12 different clustering and we finally got the most effective clustered dataset when k was equal to 12. Thus, from the majority of the 12 attributes we predicted the weapons that could be used. The accuracy that we got using the k-NN algorithm was 88.74%.

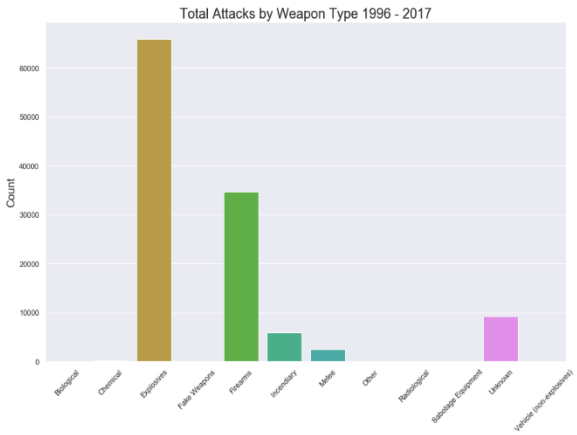

Fig.12. Total Attacks by Weapon Type

The above graph classifies the attacks by weapon types from the year 1996-2017. It is evident that explosion and firearms are the most preferred weapon types.

C. Perpetrator Classifier

The perpetrator classifier classifies various groups or organization that carries out illegal and influences terrorism. We used Random Forest Algorithm for creating the Perpetrator Classifier. We grouped and listed various groups responsible for various attacks from 1996 to 2017 using random forest algorithm. A list of major perpetrator groups was displayed, along with the accuracy, precision, recall and f1after the training of the model.

```
# Display the number of attacks by group
major_groups['gname'].value_counts()
```

```
Out[7]: Unknown                                        62459
        Taliban                                         7454
        Islamic State of Iraq and the Levant (ISIL)     5584
        Al-Shabaab                                      3274
        Boko Haram                                      2408
        Communist Party of India - Maoist (CPI-Maoist)  1876
        New People's Army (NPA)                         1800
        Maoists                                         1616
        Revolutionary Armed Forces of Colombia (FARC)   1493
        Tehrik-i-Taliban Pakistan (TTP)                 1349
        Kurdistan Workers' Party (PKK)                  1302
        Houthi extremists (Ansar Allah)                 1021
        Al-Qaida in the Arabian Peninsula (AQAP)        1011
        Liberation Tigers of Tamil Eelam (LTTE)          843
        Al-Qaida in Iraq                                 635
        National Liberation Army of Colombia (ELN)       625
        Donetsk People's Republic                        623
        Muslim extremists                                603
        Abu Sayyaf Group (ASG)                           511
        Fulani extremists                                509
```

Fig.13. Output of Perpetrator Class (Attacks by Group)

Performance Metrics

Following four types of performance metrics were generated using Random Forest Algorithm. The formulas for each of the performance metrics along with the result obtained from our model are enlisted below.

a. Accuracy score :

$$accuracy = \frac{Tr.Pos. + Tr.Neg.}{Tr.Pos. + Fa.Neg. + Tr.Neg. + Fa.Pos.}$$

Accuracy: 0.9045279383429673

b. Precision:

$$precision = \frac{Tr.Pos.}{Tr.Pos. + Fa.Pos.}$$

Precision: 0.8995287972392408

c. Recall:

$$recall = \frac{Tr.Pos.}{Tr.Pos. + Fa.Neg.}$$

Recall: 0.9045279383429673

d. F1:

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

F1: 0.894189908871535

Where,
Tr.Pos. = True-Positive
Tr.Neg. = True-Negative
Fa.Neg. = False-Negative
Fa.Pos. = False-Positive

Hence, we got accuracy of 90.45%, and precision of 89.95% from our model using the Random Forest Algorithm for creating Perpetrator Classifier.

V. DISCUSSION AND CONCLUSION

Terrorism has become a huge threat over the world. Various Machine learning system, artificial intelligence and

Data-Analytics have provided us with a system to help the investigator and anti-terrorist or counter-terrorist squad to rapidly decide the most probable perpetrator responsible of a particular terrorist attack. We have demonstrated how the methods like k-Nearest algorithm and Random Forest can help us predict the responsible perpetrator precisely eight out of ten times. 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. In future we further mean to attempt different algorithms and methods like deep learning models and package classifiers to further improvise the accuracy of result and hence successfully predict the perpetrator with more precision and accuracy. Besides this in future we also intend to use web-scrapping methods and sentimental analysis to study various posts and comments on social media sites for hatred speech and text and further filter them and create a classifier to merge the current project with the social media texts.

REFERENCES

1. Crime Data Mining, Threat Analysis and Prediction. Maryam Farsi, Alireza Daneshkhah, Amin Hosseinian Far. (2018)
2. Using Fuzzy Sets for Detecting Cyber Terrorism and Extremism in the Text. Vahide Nida Uzel , Esra Saraç Eşsiz ,Selma Ayşe Özel. (2018)
3. Psychological and Behavioural examinations of online terrorism. Sheryl Prentice, Paul J. Taylor. (2018)
4. Complex Networks for Terrorist Target Prediction. Gian Maria Campedelli, Iain Cruickshank, Kathleen M. Carley.(2018)
5. Prediction of terrorist attacks based on GA-BP neural network. Qinghao Li, Zonghua Zhang, Zhen Shen. (2019)
6. Events classification and operation states considering terrorism in security analysis. A. Torres ; C. Tranchita
7. Text Classification Techniques Used to Faciliate Cyber Terrorism Investigation.•David Allister Simanjuntak ; Heru Purnomo Ipung ; Charles li ; Anto Satriyo Nugroho.2010
8. Terrorism analytics: Learning to predict the perpetrator.Disha Talreja ; Jeevan Nagaraj ; N J Varsha ; Kavi Mahesh.2017
9. Positing the problem: enhancing classification of extremist web content through textual analysis. George R. S. Weir ; Emanuel Dos Santos ; Barry Cartwright ; Richard Frank. 2016
10. Development of a Framework for Analyzing Terrorism Actions via Twitter Lists .Kuljeet Kaur. 2016
11. Anti Social Comment Classification based on k-NN Algorithm ,Nidhi Chandra , Sunil Kumar Khatri , Subhranil Som , 2017
12. An International Study on the Risk of Cyber Terrorism . Suhannia Ponnusamy, Geetha A. Rubasundram , 2019
13. applications of artificial intelligence techniques to combating cyber crimes: a review , Selma Dilek , Hüseyin Çakır and Mustafa Aydın. , 2015
14. Crime Pattern Detection Using Data Mining Shyam Varan Nath, 2006.
15. Lexicon-Based Methods for Sentiment Analysis Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, 2011
16. Sentiment Analysis of Twitter Data - Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau
17. Research on Prediction Method of Terrorist Attack Based on Random Subspace. Author(s) Luo Zijuan ; Ding Shuai. 2017.

AUTHORS PROFILE

Mrs. S. Kalaiarasi completed her Bachelor of Engineering in the stream of Computer Science and Engineering from Muthayammal Engineering College. She completed her Master degree M.E.-CSE from Sathyabama University in the year 2011. She is pursuing her Ph.D. in SSE, Saveetha Institue of Science and Medical Science. She is working as Assistant Professor in SRM institute of science and technology and life member of ISTE. She has got good publication records.

Ankit Mehta is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He has participated in various national level technical fests and events. He has played vital role in organizing technical fests in SRM IST. He is also a member of various technical and mathematical clubs of the Institute. He is currently working on various Deep-Learning Projects.

Devyash Bordia is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He is also a member of various technical clubs of SRM IST. He is currently working on a ML project on creating a classifier that studies Cancer.

Sanskar is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He is also member of various technical clubs of SRMIST. He has participated in many national level technical events and fests. He is currently working on various web design and development projects.