A SURVEY ON FORECAST OF ONLINE SPREAD OF TERRORISM USING TWITTER TEXT CONTENT

Ms. Poonam Sahibani[1], Mr Ketan Rathod[2], Ms. Ankita Padhiyar[3]
[1,3]Computer Department, Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, India
[2]Computer Department, Babria Institute of Technology, Vadodara, Gujarat, India

Abstract: Online social networking sites such as Twitter, Facebook and Google+ among others are the great way to be in touch with our friends and relatives. Nonetheless, the power of social networking sites can be exploited for the purpose of illegal activities such as planning terrorism related activities or spreading some sorts of hate message in the society. So if there is systems which can identify such activities and the persons involved, then that system will prove to be a boon for the law enforcement people. So to solve this problem, the data mining techniques gives the better, accurate and affordable solution to predict the terrorism related activities on the web. In data mining, we can extract words from the tweets and by using classification method and machine learning approach we can predict tweets that are terror related or not.

Keywords: Terrorism, ISIS, social media radical content, text mining, natural language processing, user cluster

I. INTRODUCTION

Today, terrorist groups are well formed with so many resources; social media such as twitter is one of them. Many radical groups use their online twitter accounts to spread their propaganda. Terrorist groups use their network organizations and strategies to put audio video or text content [4]. So, it has become very important to control terrorism and stop their spread before certain amount of time [1]. As observed some previous work, they have proposed some efficient algorithm or designed a system to capture radical cluster or radical content on twitter.

II. RELATED WORKS AND LITERATURE SURVEY

In [4], authors have used machine learning approach to classify tweets. They have build classifier on the basis of three dataset.

Dataset	Description
TW-PRO	Tweets that are pro-ISIS
TW-RAND	Randomly collected tweets
TW-CON	Tweets that are against ISIS

Table I : dataset used for experiment in [4] .

Author have focused on the English hash tag which are #IS, #ISLAMICSTATE, #LOVEISIS, #ALLEYESONISIS and other hash tags [4]. Authors have used three different classes of features.

A. STYLOMETRIC FEATURE

Stylometric feature contains most frequent used words like state, Islamic, not, do, kill , support, abu, allah, people and al.

Function words	Frequency of various function words	293
Frequent words	Frequency of major frequently used words	173
Punctuation	Frequency of characters,, . , [,] , ! , ? , &	13
Hash tags	Frequency of most frequent letter bigrams	100
Letter bigrams	Frequency of most frequent letter bigrams	133
Word bigrams	Frequency of most frequent words bigrams	99

Table II The list of words that have been used in [4].

Stylometric features also include punctuation, letter bigrams, word bigrams and the most frequently used hash tags [4].

B. TIME BASED FEATURE

Time based feature contain detailed description about when tweet is posted . The following attributes are specified in [4]:
• Hour Of Day: Hour1, Hour2, . . . , Hour24,
• Period Of Day: Morning, Afternoon, Evening, Night, Mid Night.
• Day: Sunday, Monday, . . . , Saturday
•Type Of Day: WeekDay, WeekEnd.

C. SENTIMENT BASED FEATURE

Sentiment analysis determines the attitude of text towards a specific topic. The analysis of sentiment was done using natural language processing, the values the sentiment can take are: very negative, negative, neutral, positive, very positive. Authors have used three different classifiers AdaBoost, Naïve Bayes and SVM using all features on all datasets [4]. They have finally got result that AdaBoost performs slightly better than both Naïve Bayes and SVM.

In paper[1], authors had used DOM Tree concept to extracting text data from web pages and smartly designed web mining algorithms to mine textual information on web pages and detect their relevancy to terrorism. In this way they may judge web pages and check if they may be promoting terrorism [1]. Firstly, they had a crone thread which accepts web page links from web crawler then apply parsing on that links using DOM Tree [1]. After applying parsing they had done clustering of extracted pages using k-means clustering technique. Then segmentation is applied on clustered data and lastly pattern matching is done for text data and object recognition is done for image objects. If any link is affected or promotes terrorism then they had stored it in stack for further processing otherwise go back to crone thread and same process had followed for next website or web link.

In paper [2], auther said that their goal is not to change the minds of the terrorists, nor do they intend to engage in unproductive debate with extremists online. Rather, their project seeks to identify the most impactful narratives that potentially facilitate radicalization of the online audience. To achieve this, they utilize big data analytics and employ a variety of computational methods (semantic web, Natural Language Processing (NLP), and crowdsourcing) to accomplish the following objectives[2]: (1) develop a new computer ontological model, Active Narrative Capturing of Online Religious Extremism, that detects, captures, and measures predominant violent Islamist Extremist Narratives[2] (2) identify predominant themes and violent IENs used to recruit youth online; and (3) measure the impact factor of the disseminated IENs. predominant IENs systematically.

In paper [3], author said that the social network provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the pattern observed in these structures.Similarly terrorist networks have an identical structure like social networks where each node is directly/indirectly linked to the terrorist organization. In this paper author had representated some standard measures of social networks and their connection to social influence measures.

In paper [4], author had used machine learning approach to automatically identify jihadist message. They have used three kinds of dataset like proISIS, Randomly collected tweets, tweets that are against ISIS. They have used three different classes of feature : 1) Stylometric feature 2) Time based feature 3) Sentiment based feature. Bases on these features and by using machine learning approach they have concluded results. One of the major problems with classification is that in most cases the data is manually labeled by analysts as either jihadist on non-jihadist (they used the terms radical or non-radical). they avoided this and related issues such as analyst disagreement by working with data labeled from incorporated hashtags and using networks of known jihadists to assure radical content. One of the drawbacks with thier method is that many of the features are dependent of the dataset.

In paper [5] , author said that terrorist groups use the Web as their infrastructure for various purposes. One example is the forming of new local cells that may later become active and perform acts of terror. They had developed own system as name as The Advanced Terrorist Detection System (ATDS), is aimed at tracking down online access to abnormal content, which may include terrorist-generated sites, by analyzing the content of information accessed by the Web users. ATDS operates in two modes: the training mode and the detection mode. In the training mode, ATDS determines the typical interests of a pre specified group of users by processing the Web pages accessed by these users over time. In the detection mode, ATDS performs real-time monitoring of the Web traffic generated by the monitored group, analyzes the content of the accessed Web pages, and issues an alarm if the accessed information is not within the typical interests of that group and similar to the terrorist interests. An experimental version of ATDS was implemented and evaluated in a local network environment. The results suggest that when optimally tuned the system can reach high detection rates of up to 100% in case of continuous access to a series of terrorist Web pages.

In paper [6], authors said that suspended users and their affect on analyses performed on the social structure of a particular social media platform, Twitter. They noted that suspended users are a subset of all malicious users in social media; hence, their results may be underestimating the impact of such users. Their study focuses on three key points. First, they seek to understand the impact that removing these suspended users has on the structure of the social networks that can be extracted via Twitter when two users mention each other. Second, they considered how these suspended users impact our understanding of the topical focus of a collection of users. Finally, they performed a clustering analysis on the set of all suspended users in addition to a subset of non-suspended users to better understand the different types of suspended users in their data and the differing roles they might play in the social environment.

Author in [7], featured that 90% of fear monger exercises completed on the web are composed through long range interpersonal communication locales. They have assigned a framework which isn't completely automated to accomplish this framework requires subordinate client association [7]. Their proposed framework gains contribution through a settled database of email [7]. This email is drawn from the Enron email dataset. They proposed a framework plan which distinguishes the bunch of individuals or radical gatherings in long range informal communication locales, whose conduct are suspicious [7]. They have additionally centered around finding the group of clients who are examining about same point, which is finished by discovering likenesses in message which are being traded among web based life clients. The structure of creator's proposed framework is gathering of five sub framework [7].Which are following:

- Online data monitoring system and Database
- Suspicious message identification using NLP/Keyword system
- Latent semantic analysis (LSA) system
- Suspicious users identification system
- Visual representation of suspicious users.

Author in [8] connected their Algorithm on the extended information from facebook Operation [8]. The calculation distinguished dynamic frameworks or hubs that can initiates different hubs in the gathering effortlessly on account of them position among different hubs in the system [8]. Distinguishing the dynamic hubs is finished by utilizing mix of various mainstream centrality measures on the hubs in the gathering [8].

The Dark Web Forum Portal (DWFP) keeps up an accumulation of 29 online jihadist discussions, which as of now contains 14,297,961 messages and 1,553,122 strings from 362,495 authors [9]. They had talked about just two dark web gatherings Islamic Network and Islamic Awaking Forum and their information has been recovered from dark web-based interface [9]. By utilizing dark web analysis, security organizations can execute cautiousness and information accumulation for CT since dim networks are huge wellspring of data. This investigation will help security organizations to distinguish and keep away from fear based oppressor dangers. Dark web analysis will reveal the shrouded examples and come to an obvious conclusion in data space. Along these lines, Dark web investigation can be utilized for identifying and maintaining a strategic distance from dread dangers or radical movement.

III. RESEARCH GAP

In [4] a considerable lot of the features are subject to the dataset. We can utilize the two information ward and information free highlights and assess the outcome. In [1] author have utilized DOM tree to extricate data from the web to do web mining. In [2], [3], [7] and [8] they have concentrated on to discover cluster of people where suspicious action found. In [5] author have made claim framework, yet not utilized explicit calculation for precise outcomes. In [9] they have utilized dark forum portal and analyzed fear monger activities.

REFERENCES

[1] Ms. Pooja S. Kade1, Prof. N.M. Dhande, “ A Paper on Web Data Segmentation for Terrorism Detection using Named Entity Recognition Technique” presented at International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056,Volume: 04 Issue: 01 | Jan -2017,

[2] Budak Arpinar & Ugur Kursuncu and Dilshod Achilov, “Social Media Analytics to Identify and Counter Islamist Extremism: Systematic Detection, Evaluation, and Challenging of Extremist Narratives Online” presented at International Conference on Collaboration Technologies and Systems. 978-1-5090-2300-4/16 2016 IEEE.

[3] Surajit Dasgupta, Chandan Prakash, “Intelligent Detection of Influential Nodes in Networks” presented at International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) , 978-1-4673-9939-5/16, 2016-IEEE,

[4] Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, Nico Prucha, “Detecting Jihadist Messages on Twitter” presented at European Intelligence and Security Informatics Conference, 978-1-4799-8657-6/15, 2015 IEEE.

[5] Sonali Vighne, Priyanka Trimbake, Anjali Musmade, Ashwini Merukar, Sandip Pandit, “An Approach to Detect Terror Related Activities on Net” presented at IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-1 2016.

[6] Wei Wei Carnegie, Kenneth Joseph, Huan Liu, Kathleen M. Carley,” The Fragility of Twitter Social Networks Against Suspended Users” International Conference on Advances in Social Networks Analysis and Mining, 2015 IEEE/ACM.

[7] Sharath Kumar A and Sanjay Singh, ” Detection of User Cluster with Suspicious Activity in Online Social Networking Sites” Second International Conference on Advanced Computing, Networking and Security, 978-0-7695-5127-2/13,2013 IEEE,.

[8] Ala Berzinji, Frzand Sherko Abdullah, Ali Hayder kakei, “Analysis of Terrorist Groups on Facebook”, 978-0-7695-5062-6/13,IEEE.

[9] Abhishek sachan, “Countering Terrorism through Dark Web Analysis” ICCCNT'12, 26th _28th July 2012, Coimbatore, India, IEEE-20180.