IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Prediction of Online Spread of Terrorism on Twitter Using Naïve Bayes And SVM Classification

A.L. Munani[1*], B.A. Tanawala[2], P.B. Swadas[3]

[1]Computer Department, B & B Institute of Technology, V.V.Nagar, Gujarat, India
[2] Computer Department, BVM Engineering College, V.V.Nagar, Gujarat, India
[3] Computer Department, BVM Engineering College, V.V.Nagar, Gujarat, India

[*]*Corresponding Author: anilmunani@gmail.com, Tel.: +91-9998202023*

Available online at: www.ijcseonline.org

Accepted: 15/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Online networking is quickly getting to be one of the mediums of decision for understanding the social beat of an area. To survey this social beat it is basic to have an exact appraisal of who is saying what in web-based social networking. Fear monger bunches like al-quida, Indian mujahedeen, ISIS and other psychological oppressor bunches are spreading their purposeful publicity utilizing web or distinctive web-based social networking sites like Facebook, Twitter and Google+. Essential plan to stop or diminish spreading of psychological oppression is to expel these records. To execute this thought needs bunches of human endeavors which incorporate perusing part of data and dissecting contain. So to lessen human endeavors we will make a framework which recognize message given by fear based oppressor amass on twitter. Our framework will arrange tweets and discovers tweets are supporting ISIS gathering or not. We need to fabricate a framework which will give better outcome for analyzers.

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

I. INTRODUCTION

Terrorism has developed its underlying foundations very somewhere down in specific parts of the world. With expanding terrorist activities it has turned out to be vital to check terrorism and stop its spread before a specific time. So as distinguished web is a noteworthy well spring of spreading terrorism through talks and pictures. Fear monger associations utilize web to brain wash people and furthermore advance terrorist activity through provocative site pages that motivate vulnerable individuals to join terrorist associations. So here we propose a proficient web data mining framework to recognize such web properties and banner them consequently for human audit.

Data mining is a strategy used to mine out examples of valuable information from vast informational indexes and make the most utilization of got comes about. Data mining and web mining is utilized together now and again for productive framework improvement. Web mining likewise comprises of content mining approaches that enable us to sweep and concentrate helpful substance from unstructured information. Content mining enables us to recognize examples, catchphrases and important data in unstructured writings. Both Web mining and data mining frameworks are generally utilized for mining from content. Data mining calculations are productive at controlling composed informational collections, while web mining calculations are broadly used to sweep and mine from sloppy and unstructured pages and content information accessible on the web. Sites made in different stages have distinctive information structures and are hard to peruse for a solitary calculation. Since it isn't possible to manufacture an alternate calculation to suit different web advances we have to utilize productive web mining calculations to mine this tremendous measure of web information. Site pages are comprised of HTML (Hypertext markup dialect) in different courses of action and have pictures, recordings and so forth intermixed on a solitary site page.

II. METHODOLOGY

The framework for any vision related classification is same. First step start is preprocessing of twitter dataset like remove special symbols, comma, semicolon etc. Extract all word from dataset. The methodology, step by step is shown in below block diagram.
The algorithm for this methodology is:

1. Input of twitters text content.
2. Preprocessing of tweets.
3. Feature Extraction
4. Prediction of twitter content using Naïve Bayes and SVM Classification.

We are using online and offline dataset both. Twitter providing to create your own application using your twitter account and we can download up to 18000 tweets at a time.

To download tweets we have used hash tags like Indian_Mujahid, #Al-Qaeda, #ISIS and other randomly collected tweets. Also we have large dataset (offline dataset) of twitter. Our dataset is in the form of excel file which contain 65508 tweets. We also have a bag of words which contain terror related words and some sort of code word which are most used by fear mongers. Our twitter dataset is given as input for pre-processing which includes tokenization of tweets, stop word filtering from tweets, and stemming - lemmatization of words.

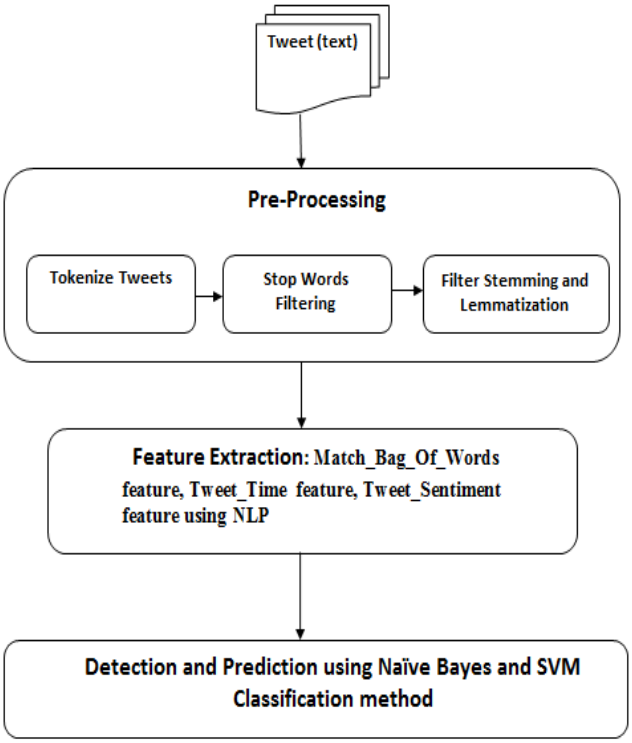

Figure I (Block Diagram)

A. PRE-PROCESSING OF TWEETS

TOKENIZE TWEETS

Tokenization, similar to all preprocessing, is application particular. It relies upon the ultimate objective. The most troublesome and interesting parts of tweets are usernames, hash tags, URLs, and emojis [4]. Distinctive applications will display these components in an unexpected way. In this way the requirement for various tokenizers with various parameters.

STOP WORD FILTERING

Once in a while, some to a great degree basic words which would have all the earmarks of being of little value in choosing archives coordinating a client require are avoided from the vocabulary completely. These words are called stop words. stop word may likewise contain comma, semicolon and hash and other exceptional images. So we need to channel all these sort of stop words to do analysis of tweets.

STEMMING AND LEMMATIZATION

Stemming algorithms work by removing the end/start of the word, considering a rundown of basic prefixes and postfixes that can be found in a curved word. This aimless cutting can be effective in a few events, however not generally, and that is the reason we certify that this approach shows a few impediments.

Lemmatization on the other hand, thinks about the morphological investigation of the words. To do as such, it is important to have detailed dictionaries which the algorithm can look through to interface the frame back to its lemma.

B. FEATURE EXTRACTION

MATCH_BAG_OF_WORDS FEATURE

In this feature, we have analyzed most frequent words like isis, influence, decline, battle, killed, bombed and others. Terrorist also use some sort of code words like bebo, shaadi, kalmashori, changez khan, aaka, eggplant and other related terms. We have compared all these words with over datasets.

TWEET_TIME FEATURE

Tweet_time feature contains details about when tweet is posted. The attributes that we have used are similar to what was used. The attributes are:

Table I (Tweet_time Feature)

Hour Of Day	1st Hour, 2nd Hour,……, 24th Hour
Day	Sunday, Monday,……., Saturday
Month	January, February, …….., December

We have noticed Friday is most active day where they are posting hate kind of messages in public.

TWEET_ SENTIMENTIMENT FEATURE

This feature includes tweet sentiment and tweet emotions. By using this feature we have analyzed the attitude of the user who have posted their tweets on twitter [4]. Sentiment analysis includes sentiments like negative and positive. It also includes emotions based attribute like anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

C. CLASSIFICATION OF TERRORIST MESSAGES

Analyzing messages automatically on twitter is most effective task for government security agencies [4]. It is not possible to read information manually. Data may be available in other formats like images, videos and may be it text contain in other languages like Arabic. So it is required to

automatic classify tweets using computer, it can identify it is radical content or not.

So we use machine learning approach to classify tweets as it is terror related or not. To build classifier we must have suitable dataset to select feature to determine it is terror related or not. We have collected set of tweets from different tweet handlers like IndianMujahid, #Al-queda, #ISIS. We have also randomly collected tweets around 65508 (offline dataset). Our dataset is also containing tweets which are against of terrorism. We only use tweets that were written in english.
As mentioned in the previous section we have used three different dataset and labelled them.

Table II (Twitter Datasets)

DATASET	DESCRIPTION
PRO_TERROR_DATA	TWEET THAT RELATED TO ISIS, AL-QUEDA AND INDIA MUJAHID
RANDOM_DATA	RANDOMLY COLLECTED TWEETS OF VARIOUS TOPICS.
AGAINST_TERROR_DATA	TWEETS THAT ARE AGAINST TO ISIS, AL-QUEDA AND INDIA MUJAHID

All our implementation results were conducted using machine learning tool called R tool. For implementation result we have used total 18000 tweets (Online data) and 65000 tweets (Offline data) from which 80% data are used as training data and 20% data are used as testing data. We have used Naïve Bayes and SVM classification algorithms.

RESULTS AND DISCUSSION

IMPLEMENTATION RESULT

Accuracy may vary as per we are giving dataset size or taking sample of tweets like 1000 tweets or 3000 tweets or up to 18000 tweets. Below tables are showing results of 1000 tweets, 5000 tweets and 18000 tweets respectively.
As shown in tables we have used two different classifiers using all features of dataset. As can be seen in the table SVM performance is better than naive basian classification method.

Table III
(ACCURACY OF MODELS USING 1000 TWEETS)

Classification method	Accuracy of Model
Naïve Bayes	78%
SVM	89%

TABLE IV
(ACCURACY OF MODELS USING 5000 TWEETS)

Classification method	Accuracy of Model
Naïve Bayes	85%
SVM	93%

TABLE V
(ACCURACY OF MODELS USING 18000 TWEETS)

Classification method	Accuracy of Model
Naïve Bayes	90%
SVM	98%

As we have implemented by using frequently used hash tags (which hashtags are frequently used on twitter), time based feature and sentiment based feature, we have plot those features on graphs as shown in figure.

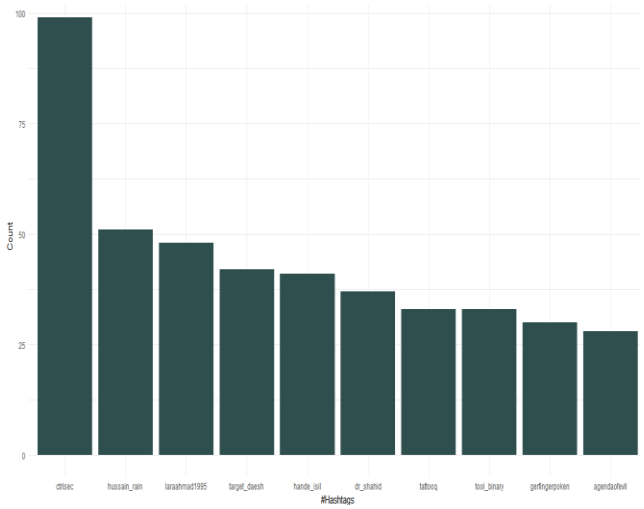

Figure II (Frequently used Hash (#) Tags)

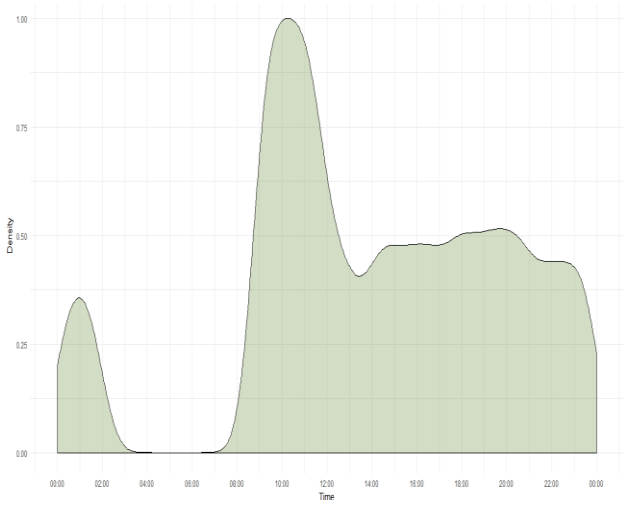

Figure III (Time based feature)

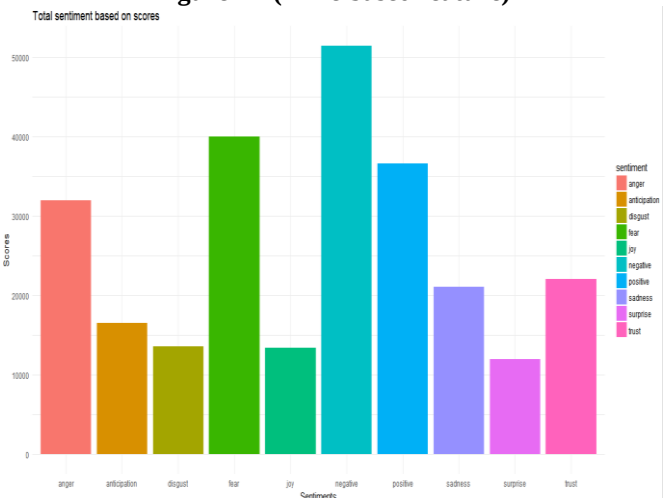

Figure IV (Sentiment based feature)

CONCLUSION AND FUTURE SCOPE

In this work we have used machine learning approach to predict and automatically check tweets are terror related or not. We have majorly focused on online dataset of twitter for classification. Because of we are getting tweets online we have given class label by applying conditions on trained data. After feature extraction we have applied both classification algorithms and compared it. For future work, we can also apply algorithm before data extracted from online.

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, p-ISSN: 2395 – 0072,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 ISBN: 978 – 1 – 4503 – 3854, 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-2018

Authors Profile

Anil L Munani, Completed his M.tech. in computer engineering from B.V.M Engineering College, vallabh vidya nagar, Anand, Gujarat, India in year 2018. He is currently working as Professor in computer engineering department of Bhailalbhai and Bhikhabhai Institute Of Technology, Vallabh Vidyanagar, Anand, Gujarat, India.. He has more than 6 year experience in acadamic. His main research work focuses on Data Mining Algorithms.

Bhavesh A. Tanawala, having work experience of more than 8 years, as an Assistant professor in computer engineering department of B.V.M. Engineering College, Vallabh Vidyanagar, Anand, Gujarat, India. His Current research areas are Data Mining and Data Analytics. He did his masters in Computer Science from IIIT, Pune.

Prashant B. Swadas, having work experience of more than 27 years, as an Associate professor in computer engineering department of B.V.M. Engineering College, Vallabh Vidyanagar, Anand, Gujarat, India. His Current research area is network security. He did his masters in Computer Science from IIT, Delhi.

