IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1754)
Available online at: https://www.ijariit.com

Prediction of future terrorist activities using Machine Learning algorithm

Santhosh Kumar S.
santhosh.kumar.s@kssem.edu.in
KS School of Engineering and Management, Bengaluru, Karnataka

Akshatha M. N.
akshathamnshroff@gmail.com
KS School of Engineering and Management, Bengaluru, Karnataka

Arun Malle
arunmalle998@gmail.com
KS School of Engineering and Management, Bengaluru, Karnataka

Aishwarya Bhat
ashuu1131@gmail.com
KS School of Engineering and Management, Bengaluru, Karnataka

H. Sai Rohit
sairohit.sairohit.bobby5@gmail.com
KS School of Engineering and Management, Bengaluru, Karnataka

ABSTRACT

The exponential increase in internet users, has led people to start using the technology-based development for carrying out activities against the law. According to the studies made from previous knowledge and experiences it can be found out that, cyber terrorism puts out stress and anxiety, increases feelings of intrusion, threat and hardens political frame of mind. Therefore, it is important to identify the fact that cyber terrorism has had very awful and shocking effect in our society. Thus, With the use of a Machine learning model, it helps us to predict the possible future terrorist activities by detecting the instigating text messages in social network platforms. The proposed system aims to develop a method for of prediction of future terrorist activities using machine learning algorithm that will help the system in detecting potential terrorist attacks and in providing safeguards and strengthening defense for required areas consequently lessens the loss of life and property. Along with that it also provides, Fast detection to illegal activity, improving information accuracy. It has the ability to distinguish between instigating (terrorism related) messages and non-instigation (non-terrorism based) messages. This method also helps in identifying the user responsible for sending such messages and also recognize the specific messages and also shows the best machine learning algorithm that best fits and that can be used for prediction based on the accuracy analysis of the algorithms.

Keywords: *Decision Tree, RFA (Random Forest Algorithm), NBC (Naïve Bayes Classifier), K-NN algorithm, Training phase, Testing phase*

1. INTRODUCTION

There has been a immense increase in internet users in the past decennary, technology based development has not only helped the society but also has given rise to various problems in the society one of such threats is the growth in cyber terrorism. Terrorism is a global issue that has drawn considerable attention, especially after the events of 9/11 in the USA and 26/11 in India and other parts of the world. However, the ability to dynamically identify and even speculate a probable terrorist risk is critically important for government agencies to react in a timely manner. According to Global Terrorism Database more than 98,773 terrorist attacks were reported between year 2001 and year 2016, which resulted in approximately 238,808 deaths. By Using earlier techniques of prediction models, it was difficult to capture the varying effects and complex interactions of terrorist attack predictors. Since the earlier method of prediction was not efficient, this led to the introduction of machine learning techniques which is an analytic trend that continues to be used in the present day. It is very necessary to identify the fact that cyber terrorism has had disastrous effect in our society. This project intends to use various machine learning algorithms to analyze, predict and categorize various terrorist activities using algorithms like k-nn algorithm, random forest algorithm to train a system by feeding it with data from dataset and various threats related text on social media.

A. Scope

The proposed system detects and predicts the nature of the message that is whether it is terrorism related instigating messages or non-terrorism messages. It also predicts the person responsible for sending such messages. It also gives the accuracy outcome of the machine learning algorithms used during the training phase. The system provides the facility to access the latest tweets of the specific person as well. The concerned authorities can initiate necessary action by passing the information to the authorized person so that necessary actions can be initiated and effective methods can be done followed to stop this

B. Motivation

The main intent of machine learning is to build a model that performs well on both the training set and the test set. Once a machine learning model is built, there are a number of ways to fine-tune the complexity of the model. Due to exponential increase in internet users, it is evident that people have started using this technological advancement for carrying out unlawful activities. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

C. Organization

The contents of the paper are organized as follows. Section II explains about related work. Section III explains the proposed system along with hardware and software requirements. In section IV experimental results are presented. Section V is the conclusion of paper.

2. RELATED WORK

In recent years a lot of work has been done in the field of “Sentiment Analysis on Twitter " by various researchers around the globe. In its early stage it was intended for binary classification which assigns opinions to bipolar classes such as positive or negative only.

In [1] authors are using various customer reviews on two different restaurants. This approach incorporates different kinds of algorithms to improve the accuracy and segregate the data of the Twitter; this methodology uses the unsupervised algorithms for the high accuracy.

In this paper [2] it is about extracting live twitter data regarding any topic and converting it into structured form from unstructured form. Opinions are extracted from the text data and the same is assigned against each tweet.

This Paper [3] predicts the products to its users based on previous tweet date. Spring XD tool was used for fetching real time tweet data and using HIVE language to get sentiments about any personality. Data is fetched from tweets and separated into words, then these words are compared with the dictionary available ignoring the emoticons as they are not part of the dictionary.

[4] Uses a dictionary-based approach to classify the reviews accurately as positive, negative and neutral and implemented by using the Support Vector Machine (SVM). Both user and the product owner can identify the product quality for each review based on the generated graph of product video.

[5] made use of Dictionary Based approach, Support Vector Machine (SVM) and Naïve Bayes, algorithm for prediction of Indian election held in 2016 in India on Hindi twitter to determine the result as positive, negative and neutral for political parties of India.

In [6] develops sentiment analysis approaches embedded in public Arabic tweets and Facebook comments. They used supervised machine learning algorithms such as Support Vector Machine (SVM) and Naïve Bayes, and they used binary model (BM) and TF-IDF to see the effect of several terms weighting functions on the accuracy of sentiment analysis.

Machine (SVM) and Naïve Bayes, and they used binary model (BM) and TF-IDF to see the effect of several terms weighting functions on the accuracy of sentiment analysis.

(BM) and TF-IDF to see the effect of several terms weighting functions on the accuracy of sentiment analysis.

Machine (SVM) and Naïve Bayes, and they used binary model (BM) and TF-IDF to see the effect of several terms weighting functions on the accuracy of sentiment analysis.

Sentiment Analysis is nothing but the use of machine learning and Natural Language Processing to identify, extract and categories the text documents. Sentiment analysis is also called opinion mining. Generally, sentiment analysis is done at document based and sentence based, in document based, gives the positive or negative opinion in the whole document as a single entity. And in sentence based, analyzing each sentence to be positive, negative or neutral opinion in the document [7].

product owner can identify the product quality for each review based on the generated graph of product video.

[5] made use of Dictionary Based approach, Support Vector Machine (SVM) and Naïve Bayes, algorithm for prediction of Indian election held in 2016 in India on Hindi twitter to determine the result as positive, negative and neutral for political parties of India.

In [6] develops sentiment analysis approaches embedded in public Arabic tweets and Facebook comments. They used supervised machine learning algorithms such as Support Vector Machine (SVM) and Naïve Bayes, and they used binary model (BM) and TF-IDF to see the effect of several terms weighting functions on the accuracy of sentiment analysis.

Sentiment Analysis is nothing but the use of machine learning and Natural Language Processing to identify, extract and categories the text documents. Sentiment analysis is also called opinion mining. Generally, sentiment analysis is done at document based and sentence based, in document based, gives the positive or negative opinion in the whole document as a single entity. And in sentence based, analyzing each sentence to be positive, negative or neutral opinion in the document [7].

3. THE PROPOSED SYSTEM

The proposed system detects illegal activities on social media platform such as twitter by recognizing instigating text messages/tweets shared on social networks and reporting to the intelligence agencies and security services for further action. This system will help in detecting potential terrorist attacks which will help in preventing deaths and theft. Fast detection to illegal activity, improved information accuracy, understanding machine learning and its motivation, paves a way to bring smart approach in solving high level problems. The system does the analysis by following 3 processes: Data collection, Data pre-processing and Classification.

A. Data Collection

Data collection schemes are proposed for astounding Twitter terminal to obtain chronological tweets. Rigorous works in sentiment analysis use a public information streaming platform known as Twitter Standard Search API, which is an interface that has potential for gaining insight in consecutive order for no longer than a week.

B. Pre-Processing

The dataset was cleansed to obtain grade and the impact in number format from given words. The developed facts field is properly configured to turn it from words to data type. Every blank space and blank values are returned with nullified for words columns and 0 for quantity fields. Once the cleaning process is done, the next phase is eradication of abbreviation such as „Sr. to senior", deleting punctuations and replace the numbers with text. Additionally, the stop words are removed.

Stop word removal in natural language processing is nothing but eliminating the words or phrase which do not support classification, here we remove most widely used words from the sentences even the web crawler does not use for indexing like „the, a" etc. Stemming is defined as method for shrinking the text to its root word. For example: „loud", „loudly", „loudness" can be reduced to make one base word „loud". After pre-processing is finished, data is set to be classified.

The below diagram represents system architecture of the project and overall model of a machine learning algorithm model works. It is divided into 3 stages: Training, Testing & Prediction.

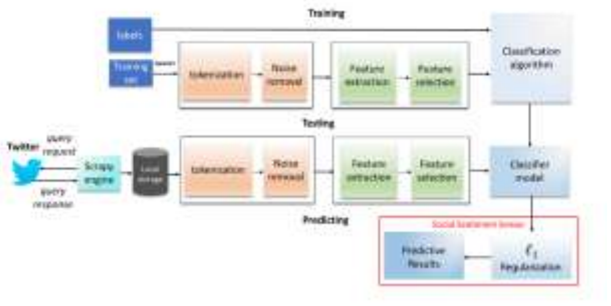

Fig 1: system architecture

I. *Labels & Training set:*

Identification of raw data. Here x is the input variable and y are the output variable. Training sets are data sets used for the implementation.

II. *Tokenization:*

is a process of turning meaningful data into string of character called tokens so it serves as a reference to original data. It is used to perform basic prepossessing such as removing punctuations marks, converting all words into lower case.

III. *Noise Removal.:*

Noise removal is the removal of duplicate data, incomplete data, repeated data and meaningless data from the dataset.

IV. *Feature Extraction & Selection.:*

Feature extraction is the process of removal of redundant data so that processing can be done easily. It is basically formatting the number of features as per requirements of algorithm. In feature selection fitting of irrelevant data is done but it keeps the original features.

V. *Classification Algorithm & Classifier model:*

The pre-processed dataset is later categorized into given number of classes. It is done to identify to which category new data will fall under. Classifier model is a set supervised machine learning model which are trained to analyses if text is instigating or not. The 3 supervised model used are Random Forest, Decision Tree and Naïve Bayes.

Random Forest: defined as a machine learning technique which has potential for conducting regression and classification job with the help of many decision trees and also analytical process known as bagging. Bagging together with boosting are said to be the most well-favored techniques mainly focus in handling of high variance and high bias. Random Forest uses two key concepts that defines the name *random*:

1. Random sampling of training observations when building trees.
2. Random subsets of features for splitting nodes.

Random Forest process steps:

Step 1: Illustrative examples are got hold from training data so that every data point has a uniform probability of getting accepted and every sample will have exact size as the native training dataset.

For E=example consider following data:
x= 0.1,0.5,0.4,0.8,0.6, y=0.1,0.2,0.15,0.11,0.13 where x is an independent variable with 5 data points and y is dependent variable. Now bootstrap samples are taken with replacement from the above data set. N **estimators** is set to 3 (no of tree in random forest), then:

The first tree will have a bootstrap sample of size 5 (same as the original dataset), assuming it to be: x1= {0.5,0.1,0.1,0.6,0.6} likewise x2= {0.4,0.8,0.6,0.8,0.1}

x3= {0.1,0.5,0.4,0.8,0.8}

Step 2: this model is instructed at each bootstrap sample drawn in the previous step, and a prediction is accounted for every sample.

Step 3: the ensemble prediction is obtained by doing an average of all the predictions of the above trees that produces the final prediction.

Recall from analysis that for large n,

$$\left(1-\frac{1}{n}\right)^n \approx \frac{1}{e} \approx .368.$$

Naive Bayes:

Step 1: storage servers are scanned for the obtaining necessary data for mining process from database, cloud, excel sheet etc.

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p] The probability for every attribute is computed. For each class(n) we should apply the formulae.

Step 3: Compute using the below formula.

P (attributevalue (ai) / subjectvalue (vj)) = (n_c + mp) / (n + m)

Where:

n = the number of training examples for which v = vj
n_c = number of examples for which v = vj and a = ai
p = a priori estimate for P(aijvj)
m = the equivalent sample size

For each class, here we multiple the results of each attribute with p and final results are used for classification.

Step 5: now values are compared and then attribute values are classified to one of the predefined sets of class.

Decision Trees: are the best implement for supporting the selection of best course of action. It supports very high productive structure in which one can provide choices and explore the possible result of selecting those choices.

A decision tree is graph which consists of nodes which represents the place where we choose an attribute and question; here edges represent the solution to the question; and the leaves represent the actual outcome or class label. Decision trees are used widely in non-linear decision making with simple linear decision surface.

Decision Tree has the major challenge that is to identify the attribute for the root node at each level. This process is called as attribute selection. The two popular attribute selection measures are:
Information Gain
Gini Index

Information Gain: Information gain is a measure of change in entropy.

Entropy: Entropy is the estimation of uncertainty of a random variable, it provides traits of the impurity in an arbitrary collection of examples. The larger the entropy more the information content.

Gini Index

- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- That is an attribute with lower Gini index should must be recommended for use.
- Formula for the calculating Gini Index is given below.

$$GiniIndex = 1 - \sum_j p_j^2$$

I. *Query request and response:* Scrapy engine sends a request to the twitter database to gain data in order to manipulate. In response data is sent. The web crawling tasks are done for document scraping in an automated and efficient manner. The collected insight is then processed by Scrapy.
II. *Scrappy Engine:* The received web information is next processed by scrappy engine using python web scraping framework that will extract and process large amount of data from twitter.
III. *Regularization:* L1 & L2 is used to reduce the errors to prevent overfitting.

Work Flow: the general workflow of the Machine learning process is shown below.

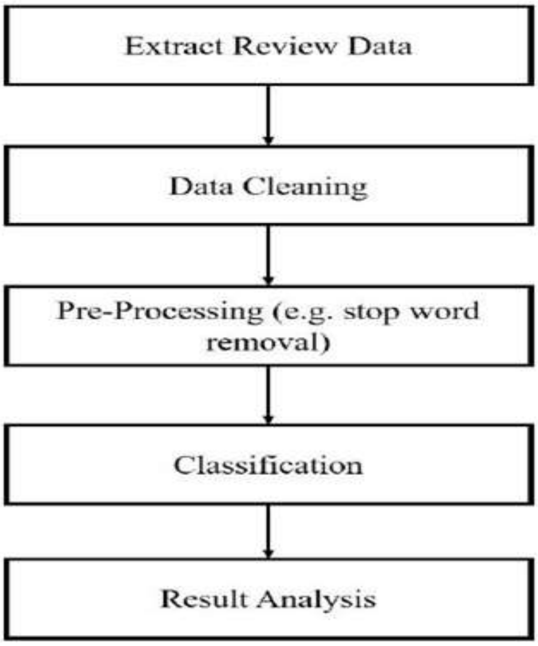

Fig 2: work flow diagram

1. **Data absorption:** initially data is being loaded from a file and is saved in the memory.
2. **Data modification**: data which has been loaded previously is now being transitioned, cleared and normalized which can fit to the algorithm.
3. **Model Training**: a model is developed using the finalized algorithm.
4. **Model Testing**: here the model which was developed in the previous step is used for testing data set and the outcome produced will be utilized for developing a new model, which will consider previous model.
5. **Model Deployment**: finally, the best model is chosen.

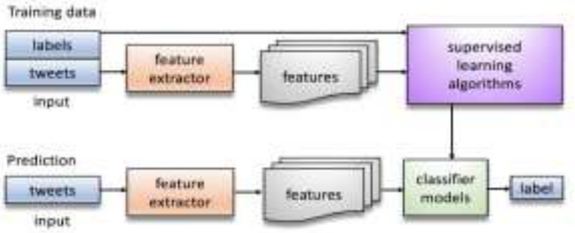

Fig 3: General workflow process

4. EXPERIMENT AND RESULT

The prediction of future terrorist activities is carried out by using machine learning algorithm such as Decision Tree Algorithm (DTA), Random Forest Algorithm (RFA) and Naïve Bayes Classifier (NBC). The project is divided into three phases, which are training, testing and prediction phase.

In training phase, the dataset is imported as a csv file (Fig 4 Shows the sample of dataset) which is used further process for analysis.

In the prediction page (Fig 5) the username is entered and it provides the results by differentiating the tweets between tweets may be related to terrorism and tweets may be related to non-terrorism (Fig 6). Once detected, the user's name of the individual will be forwarded to anti-terrorism agencies for further actions.

Fig 4: Datasets

Fig 5: Prediction Page

Terrorism Prediction

Fig 6: Predicting Page

5. CONCLUSION

The physical way of going through each tweet usually consumes a lot of time, the mission of our structure is to minimize the complexity and time-consuming process of prediction of future terrorist activities by automating the analyzing process of tweets. Our system uses machine learning algorithm's and provides high accuracy hence strengthening the defense system. Also, there can be improvisation done in our system such as providing live tweets analyzing method and adding a multi-language detection system.

6. REFERENCES

[1] El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019). Sentiment Analysis of Twitter Data. 2019 International Conference on Computer and Information Sciences (ICCIS).

[2] Aslam, A., Qamar, U., Khan, R. A., Saqib, P., Ahmad, A., & Qadeer, A. (2019). Opinion Mining Using Live Twitter Data. 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).

[3] Fernandes, R., & D'Souza, R. (2016). Analysis of product Twitter data though opinion mining. 2016 IEEE Annual India Conference (INDICON).

[4] Abinaya. R, Aishwaryaa. P, Baavana. S, ³Automatic Sentiment Analysis of User reviews´, IEEE International conference on Technological Innovations in ICT for Agriculture and Rural Development, 2016, pp 158-162.

[5] Paul Sharma, Teng-Sheng Moh, ³Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter´ , IEEE International conference on Big Data, 2016, pp.1966-1970.

[6] R. M. Duwairi and I.Qarqaz, "A framework for Arabic sentiment analysis using supervised classification" , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016.

[7] A. kaushik, S. Naithani, ³A Study on Sentiment Analysis: Methods and Tools´ International journal of Science and Research, vol. 4, 2015, pp 287-291.