Unauthorized Terror Attack Tracking System using Web Usage Mining

Akash Yadav[1], Minal Chudasama[2], Dipti Patil[3], Ranjita Gaonkar[4]

[1-3]BE Student, Department of Information Technology, MES Pillai College of Engineering, Navi Mumbai, India
[4]Assistant Professor, Department of Computer Engineering, MES Pillai College of Engineering, Navi Mumbai, India
---***---

Abstract - *With the speedy increase in social network applications, people are using these platforms to voice their choices regarding the daily problems. Gathering and Analyzing people's reactions towards shopping for a product, public services, so on are very important. Sentiment Analysis (or Opinion Mining) may be a common dialogue preparing task that aims to discover the feelings behind opinions in texts on varying subjects. In recent years, researchers within the field of sentiment analysis are involved with analyzing opinions on various topics like movies, commercial products and daily social group problems. Twitter is a tremendously popular micro blog on which users might voice their opinions. Opinion investigation of twitter data is a field that has been given a lot of attention over the last decade and involves dissecting "tweets" (comments) and the content of those expressions. Solely this, Paper explores the assorted sentiment analysis applied to twitter data and their outcomes.*

Key Words: **TextBlob, Vader Sentiment, Natural Language Processing, Sentiment, Analysis, Sentiment Analysis, CSV, Tweets, Twitter, Countries, Terrorism, Terror.**

1. INTRODUCTION

Day to day the utilization of internet goes on increasing drastically, comparatively the massive range of techniques are there rising each day on varied dynamic platforms. Use of Internet is expanding however, together with these obstructive minded people are victimizing the net for hurting to the society and people. It is believed that the detection of terrorist's activities on the Internet may stop more terrorist attacks. Several terrorists or terrorist teams' square measure created to use such techniques, applications, and web to unfurl the fear. They traditionally attract the youths to be entangled in such activities. It is necessary to notice such attempts as to stop such hazardous things. Terrorists utilize social sites. It's a massive challenge to notice such venturesome terrorist attacks. It's similar to eavesdropping.

Sentiment analysis is also called "opinion mining" or "Emotion Artificial Intelligence" and alludes to the utilization of Natural Language Processing (NLP), text mining, computational linguistics, and bio measurements to methodically acknowledge, extricate, evaluate, and examine emotional states and subjective information. Sentiment analysis mostly involves the voice in client materials; for example, surveys and reviews on the Web and web-based social networks.

2. LITERATURE SURVEY

1. Naseema Begum A, Hanu Rakavi S, Mohanambal R, Aswathy R, "Detection of online spread of terrorism using web data mining", [1]

The application that is to be developed will stop the spread of terrorism by preventing the radicalization of the people through online websites and the other media available on the internet. People are denied access to these websites and these media because the terrorist organizations are using the internet as the media to promote and spread terrorism and hence the terrorism is avoided. Data mining technique is utilized to determine and define the pattern from the available collection of the websites and the data from the websites mined are the huge volume of data sources, the results obtained are widely used [1]. Web mining is also similar to data mining since it involves the text mining methods that are used to scan the data and also extracting the useful pattern from unstructured data

2. Abdullah Alsaeedi, Mohammad Zubair Khan,"A Study on Sentiment Analysis Techniques of Twitter Data" [2].

The diverse techniques for Twitter sentiment analysis methods were discussed, including machine learning, ensemble approaches and dictionary (lexicon) based approaches. Also, hybrid and ensemble Twitter sentiment analysis techniques were explored. Research outcomes demonstrated that machine learning techniques; for example, the SVM and MNB produced the greatest precision, especially when multiple features were included. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive. Machine learning algorithms, such as The Naive Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram models were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve classification accuracy approximately.

3. T. Suvarna Kumari, Narsaiah Putta, "Web Data Mining to Detect Online Spread of Terrorism" [3].

This paper talks about a method for mental oppressor area by using Web traffic substance and using data mining techniques because the survey information is seen here. The suggested approach, called ATDS, knows psychic oppressors

standard result (profile) by adding a knowledge mining estimate to the writing of Web objectives related to dread. The resulting picture including records is employed by the framework to execute ongoing customer revelation linked to being crowded with functions out of fear monger. Examination of the Receiver Operator Characteristic (ROC) shows that this scheme can defeat a framework of intrusion recognition centered on going. Thusly, fear related activities are often perceived using data mining frameworks and log access and handling.

4. Harshali P. Patil, Dr. Mohammad Atique "Sentiment Analysis for Social Media: A Survey", [4].

Masses of users share their feelings on social media, making it a valuable platform for trailing and exploring public sentiment. Social media is one in all the largest platforms where massive instant messages are revealed everyday which makes it a perfect supply for capturing the opinions towards varied curious topics, like product, goods or celebrities, etc. the most goal of this paper is to administer associate overview of latest updates in sentiment analysis and classification strategies and it includes the temporary and brief discussions on the challenges of sentiment analysis that the work wants to be done. We tend to additionally find that the majority of the works done area unit based on machine learning methodology instead of the lexicon based methodology. Here, detailed survey on different techniques used in Sentiment Analysis is carried out to understand the level of work.

5. Ramesh Yevale, "Unauthorized Terror Attack Tracking Using Web Usage Mining", [6].

The purpose of the project is that the project is incredibly helpful to search out the terror related activities. By this project, users of the project will come across to understand the suspicious spoken conversation, and therefore will notice the supply of the terror and can be helpful to attenuate the fear activities. The project has terribly wide scope within the security of the national assets. The project is incredibly helpful to search out terror activities. It's chiefly helpful for the Military and CBI agents. By mistreatment they'll come back to understand the various terror communication and therefore the accidents will be avoided. Blogs, social networking sites, largely hosted by terrorist sympathizers is avoided by mistreatment net usage mining.

3. SYSTEM DESIGN

The website that we built starts with a home page that highlights the key features of the website, leading users to the other pages which provide them with different features. The three main features of our website are namely , Global Search, Country Search and Upload Dataset. These services then lead the user to the further functioning of the system.
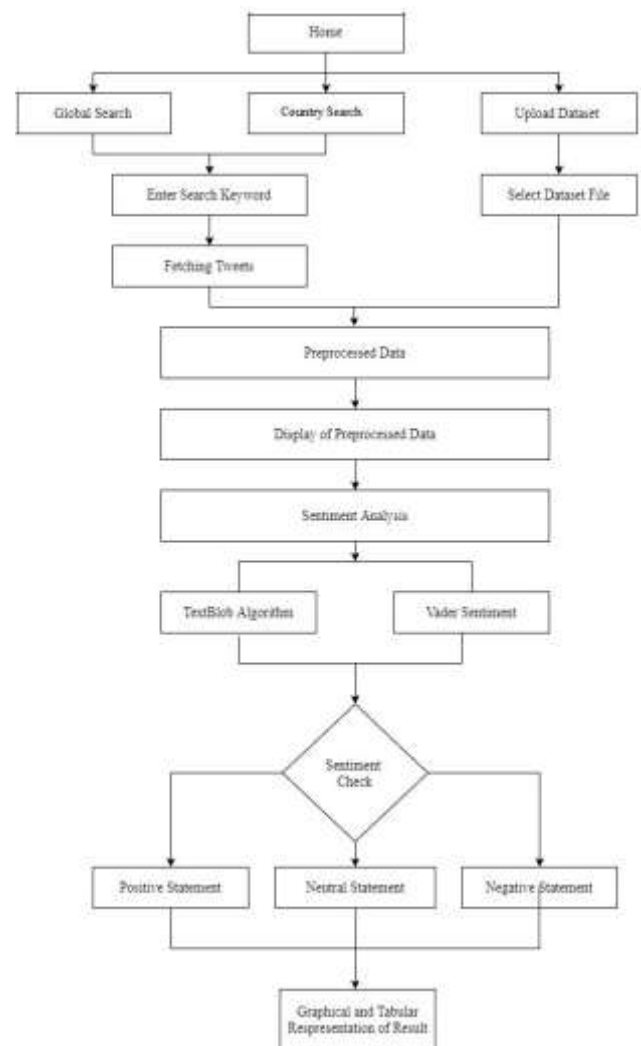

Fig 1:- Proposed System

3.1 Sentiment Analysis

Sentiment analysis [7] is to analyze the collection of these personal preference expressions, also known as emotion mining and opinion extraction. According to the size of the text, the sentiment analysis can be divided into three categories: 1) sentiment analysis based on the words. It creates a sentiment lexicon to separate positive words and negative words by hand or automatically, and then uses the lexicon to analyze the data based on the frequency of the word or the sentiment value which is set when building the lexicon. 2) sentiment analysis based on the sentences. It mainly analyzes the relevant factors of the sentiment tendency such as the semantic expression and sentence representation of sentences and so on, combined with the polarity of the modifier and the negative words. 3) sentiment analysis based on the document. Basically, it's similar to the second category, because the statement can be seen as short documents. In this paper, sentiment analysis involves different techniques such as the establishment of sentiment lexicon, the sentiment analysis based on the words and the label of the sentiment value such as positive, negative or

neutral. The sentiment analysis of the word is the basis of the sentiment analysis of the text, and the premise of sentence and chapter emotional analysis.

3.2 TextBlob in Python

TextBlob is a python library and offers an easy API to access its methods and perform basic Natural Language Processing tasks. TextBlob is a Python (2 and 3) library for processing textual data. It provides an easy API for dividing NLP tasks like part-of-speech tagging, phrase extraction, sentiment analysis, classification, translation, and more. TextBlob and Naive Bayes Classifier are used for this process.

3.3 Naïve Bayes Classifier

Bayesian classification provides practical learning algorithms and previous knowledge and observed data are often combined. Bayesian Classification provides a helpful perspective for understanding and evaluating several learning algorithms. It calculates explicit possibilities for hypothesis and it's sturdy to noise in input file.

1. Consider a training data set D consists of documents which belongs to different classes say class A and B.
2. Prior probability of both classes A and B is calculated as shown Class A=number of objects of class A / total number of objects. Class B=number of objects of class B / total number of objects.
3. Now calculate the total number of word frequencies of both classes A and B i.e., ni na = the total number of word frequency of class A. nb =the total number of word frequency of class B.
4. Calculate the conditional probability of keyword occurrence for given class P(word1 / class A) = word count / ni(A) P(word1 / class B) = word count / ni(B) P(word2 / class A) = word count / ni(A) P(word2 / class B) = wordcount / ni P(wordn / class B) = wordcount / ni (B)
5. Uniform distributions are to be performed in order to avoid zero frequency problem.
6. Now a new document M is classified based on calculating the probability for both classes A and B P (M/W). a) Find P(A / W) = P(A) * P(word1/class A)* P(word2/ class A)......* P(wordn / class A). b) Find P(B / W) = P(B) * P(word1/class B)*P(word2/ class B)......* P(wordn / class B).

7. After calculating probability for both classes A and B the class with higher probability is the one the new document M assigned.

3.4 VADER Sentiment

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is absolutely open-sourced. VADER uses a combination of a sentiment lexicon is a list of lexical options (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not solely tells regarding the Positivity and Negativity score however additionally tells about how positive or negative a

sentiment is Manually making (much less, validating) a comprehensive sentiment lexicon is a labor intensive and generally error prone method, therefore there is no surprise that several opinion mining researchers and practitioners rely heavily on existing lexicons as primary resources.

The Python code for the rule-based sentiment analysis engine. Implements the grammatical and syntactical rules represented within the paper, incorporating empirically derived quantifications for the impact of every rule on the perceived intensity of sentiment in sentence-level text. Significantly, these heuristics transcend what would usually be captured in an exceedingly typical bag-of-words model. They incorporate word-order sensitive relationships between terms. For instance, degree modifiers (also referred to as intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity.

The algorithm goes as such:

1. Count the words in the sentence.
2. Generate the score for the all the words in the sentence individually.
3. Calculate the sum of all the word scores.
4. Generate a compound score which shows if the sentence is positive / negative / neutral.

The compound score is computed by summing the valence score of every word within the lexicon, adjusted consistent the rules and so normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the foremost helpful metric if you wish one one-dimensional live of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is correct.

It is also useful for researchers who would like to set standardized thresholds for classifying sentences as positive, neutral, or negative. Typical threshold values (used in the literature cited on this page) are:

1.Positive sentiment: compound score >= 0.05
2.Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
3. Negative sentiment: compound score <= -0.05
The pos, neu, and neg scores are ratios for proportions of text that fall in each category (so these should all add up to be 1... or close to it with float operation). These are the most useful metrics if you want multidimensional measures of sentiment for a given sentence

4. IMPLEMENTATION DETAILS

4.1 GUI

This is the overall UI of the application. The website is made by using HTML, CSS and CSS libraries like Bootstrap, Flask and JavaScript. Users will interact with this part of the application only, which is the website. On the Home Page, the user has the option to pick one of 3 given ways to provide data, The Global Search, Country Search and the Upload Dataset.

4.2 Global Search

Global Search is the first option to download Tweets from Twitter. The user downloads the tweet by entering the search Keyword in the form provided. This query will then be passed to the Python function called download tweets. This function will use the search Keyword provided to download data from Twitter, without Retweets. After the tweets are downloaded a CSV file will be created. This CSV file will be preprocessed again and a final CSV file will be created which will be later displayed on the preprocessed table page.

After the CSV file is created download tweets () function will redirect the user to the table page which displays the preprocessed data in a tabular form and a link in the form of a button to the next page. When the user clicks on the Evaluate Sentiment button it runs the tweets graph() function which performs sentiment analysis on the preprocessed data using both Vader Sentiment and Text Blob techniques and then writes the twitter username, tweet and the sentiment values for each tweet to the CSV files and redirects the user to the next page.

The next page consists of a bar graph generated from the sentiment values using JavaScript and the following JavaScript libraries: c3.js, papa parse and d3.js, a table and a pie chart. The bar graph and the table is displayed on the page with an option to download the table. A pie chart is also generated from the statistics of tweets downloaded and the stats are displayed alongside the bar graph and the table. The user can also view a summary of the tweets from 5 preset major countries relevant to the search query provided by the user. This summary page has a bar graph which shows number of positive, neutral and negative tweets for each country and a pie chart with statistics and a summary table.

4.3 Country Search

The country search feature is similar to the global search feature but instead provides tweets from 5 countries and displayed one at a time and the user can select which country tweets to view. When the user enters the search query the query is passed to the python function country tweets() and this function downloads the tweets for 5 preset countries and saves the tweets to the CSV files for each country and then another function country sentiment() runs which performs sentiment analysis on the preprocessed data using both Vader Sentiment and Text Blob techniques and then writes the twitter username, tweet and the sentiment values for each tweet to the CSV files and redirects the user to the next page.

The next page consists of a bar graph generated from the sentiment values, a table and a pie chart. The bar graph and the table is displayed on the page with an option to download the table. A pie chart is also generated from the statistics of tweets downloaded and the stats are displayed alongside the bar graph and the table. The user can choose the country from which they want to view the data from the drop down menu.

4.4 Upload Dataset

The upload dataset feature allows the user to provide their own dataset and it is preprocessed for the sentiment visualizer which displays the sentiments for the uploaded dataset. The user can upload their dataset in a CSV format using the upload form and when the user hits upload the python function processing() runs and the CSV file is preprocessed and the required fields are extracted from the provided dataset and placed in the appropriate fields in the preprocessed CSV file. Then sentiment analysis is performed on the preprocessed data and the sentiment values are written to a new CSV file. The user is then redirected to the next page.

The next page consists of a bar graph which shows the sentiment of each tweet alongside the table and the pie chart which also consists of the statistics showing the number and percentage of positive, neutral and negative tweets in the provided dataset. The user can also download the CSV file from this page.
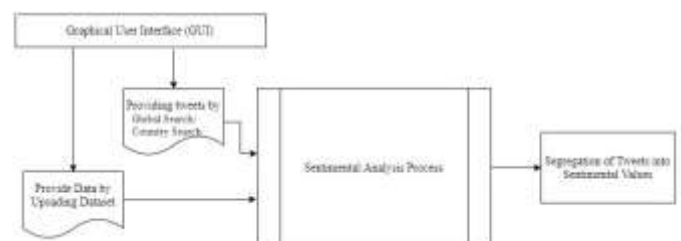

Fig 2:- System Functional Module

5. PROJECT INPUT AND OUTPUT SCREENSHOTS

5.1 Home Page

Our website provides the services to perform Sentimental Analysis on user generated data. The process of sentimental analysis helps us define the general idea of the nature of the statement on social media in our case tweets. There are 3 services to do, Global search, Country search and Update Dataset**.**

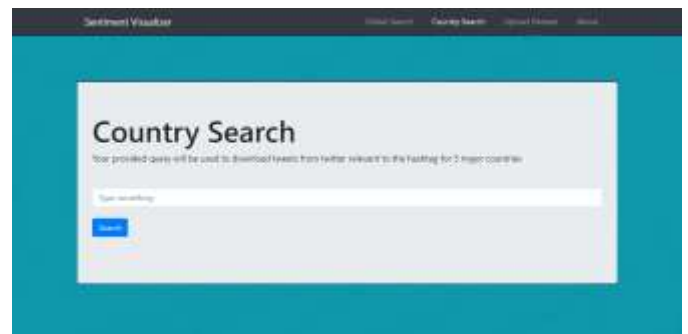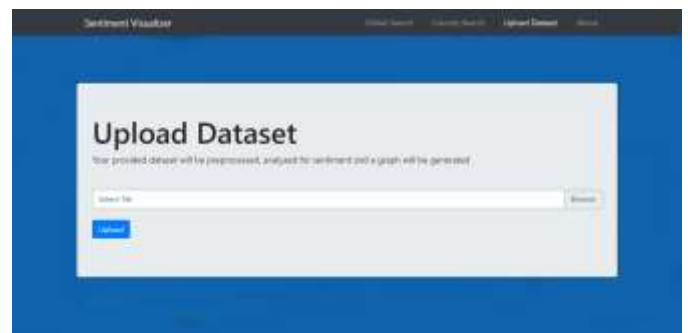Fig 3:- Home Page (1)

Fig 4:- Home Page (2)

Fig 5:- Home Page (3)

5.2 Global Search

When enacted on this Global Search page by the user; this module allows the user to enter search any term, related to which the user wishes to obtain data or in our case Tweets. The user enters the search query, which results in a hash tag search on twitter using the Python library "tweepy".
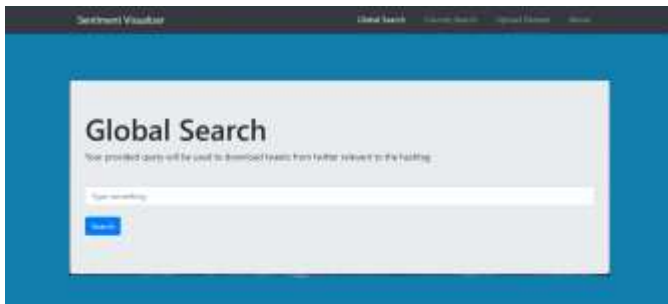

Fig 6:- Global Search Page

5.3 Country Search

This module provides user with the option to download tweets related to any term but for 5 specific predefined major countries. This feature is very useful for research purposes and makes the data more refined and country specific. The user can select to view one country at a time from the dropdown menu.

Fig 7:- Country Search

5.4 Upload Dataset

In Upload Dataset, there is an option for the user to provide the data in the form of CSV file. The further process is then done similarly, on the provided data CSV file by the user.

Fig 8:- Upload Dataset

5.5 Preprocessed Data Table

The resultant tweets are gathered and downloaded one by one into a CSV file in a raw format. This CSV file is a collection of preprocessed data. The output CSV file consists of two fields: the twitter username and the corresponding tweet. Only the required fields are selected ignoring the duplicate data and this new set of data is then written into another CSV file. This creates our preprocessed data. From this data we can perform Sentimental Analysis.

Fig 9:- Preprocessed Data Table

5.6 Graph Generation

This process altogether creates a clean version of the sentiments generated from the content of the tweets making it visual friendly by creating a graphical format of the two algorithms used namely, TextBlob and VADER sentiment

which displays the data over the course of time. The graph displays, The Positive sentiment, which are the tweets having sentimental value 15, the Negative sentiment, which are tweets with tweets with value 5 and Neutral sentiment, tweets having sentimental value 10 and a pie chart which displays all the distribution of positive, neutral and negative sentiment percentage in the dataset. Similarly, we can view the result for Country specific search.

Fig 10:- Results – Graph

Fig 11:- Results – Pie chart

Fig 12:- Results- Graph (Country Search)

5.7 Download Result Table

The expected result generated is also represented in a tabular form for better understanding and convenience and can be downloaded for further use by the user. The table contains fields like Username, Tweet and the resultant sentiment.

Fig 13:- Result Table

Fig 14:- Downloaded Result Table

5.8 Countries Overview

The result could also be further divided in to country specific data and graph as shown in the figure 10, Countries Overview. When this function runs it downloads tweets for the search term provided by the user for 5 different countries. The data is distributed into 5 pre-defined major countries.

Fig 15:- Countries Overview

6. CONCLUSIONS

In conclusion, we have developed an application that performs Sentiment analysis on tweets posted by the users round the world and conjointly in specific countries and so this displays the tabular information within the style of interactive graphs. The sentiments are calculated using Natural Language Processing (NLP) techniques such as TextBlob and Vader sentiment that are very accurate. We need to examine the assorted Data Mining and Web Mining technologies and see how they can be adapted for counter-terrorism, for future scope.

REFERENCES

[1] Naseema Begum A, Hanu Rakavi S, Mohanambal R, Aswathy R, "Detection of online spread of terrorism using web data mining", IEEE 2019.

[2] Abdullah Alsaeedi, Mohammad Zubair Khan,"A Study on Sentiment Analysis Techniques of Twitter Data", IEEE 2019.

[3] T. Suvarna Kumari, Narsaiah Putta, "Web Data Mining to Detect Online Spread of Terrorism", IEEE July 2019.

[4] Sonali Vighne, Priyanka Trimbake, Anjali Musmade, Ashwini Merukar, Sandip Pandit, "An Approach to Detect Terror Related Activities on Net", IEEE 2016.

[5] Harshali P. Patil, Dr. Mohammad Atique "Sentiment Analysis for Social Media: A Survey", IEEE 2015.

[6] Ramesh Yevale, "Unauthorized Terror Attack Tracking Using Web Usage Mining", IEEE 2014.

[7] Sentiment analysis, In Wikipedia. Retrieved November 2, 2017,from https://en.wikipedia.org/wiki/Sentiment_analysis