IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

A Diagnosis System Framework for the Time-series analysis of the Terrorism attacks Worldwide

Dharanija G[1*], B. Chandana Priya[2], B. Manasa Sai [3], G.V. Vishnu Vardhan Reddy[4], Sujatha K[5]

[1,2,3,4] School of Computing and Information Technology, REVA University, Bangalore, India
[5]School of Computing and Information, REVA University, Bangalore, India

Corresponding Author: dharanijagudur@gmail.com, Tel.: +91 9035994460

DOI: https://doi.org/10.26438/ijcse/v7si14.1822 | Available online at: www.ijcseonline.org

***Abstract*—** Social media(twitter) is easily conveyed organization for the enrolled people that may fuse content, photos, chronicles and hyperlinks. Individuals post whereabouts, opinions and information to help or against social media. The most terrified subject is terrorist strikes happening far and wide. Terrorist exploits the web-based life to consistently impart utilizing code signs or to build their backhanded proximity. The words with the hash sign related with them are broke down, get the evaluation of the twitter posts. This paper displays a methodology for sentiment analysis on terrorist related posts and to deal with the slants with their geolocations. Machine learning procedures like KNN (K-Nearest Neighbor), Random Forest are connected and the information is prepared utilizing Exploratory Data Analysis. The results are looked at and exhibited.

***Keywords*—**Sentiment Analysis; Exploratory Data Analysis; KNN; Random Forest; Geolocations;

I. INTRODUCTION

Presently days in India, there are numerous information bunches who share their information features as short messages in smaller scale blogging administrations, for example, Twitter. Writers of these messages expound on their life, share assessments on assortment of points and talk about current issues. In perspective on association of messages and a basic accessibility of little scale blogging stages, Internet customers will as a rule move from traditional specific tools, (for instance, standard sites) to scaled down scale blogging organizations.

As an ever-increasing number of clients post about items and administrations they use, or express their formal and informal views, become profitable wellsprings of individuals' conclusions and assessments. We utilize information set framed contain gathered texts from social media(twitter). It contains an expansive number of extremely mini tweets made by clients of the smaller scale blogging stage. Substance of the tweets shifts from individual considerations to open explanations. The tweets will be ordered by the framework into a gathering: war terrorist-wrongdoing.

In this examination, the tweets identified with terrorist activities are recognized from Twitter with their definite area, and conclusion investigation is performed on them. Initially, the Exploratory data analysis is utilized to prepare the information. At that point the AI calculations, for example, KNN, Random Forest are connected for order of the slants of the posts. The outcomes are thought about and introduce.

The rest of the bit of this exploration paper is composed as pursues: Section II gives the foundation research to the examination of the social media(twitter) information; Section III gives the design to the way toward doing sentiment analysis. Segment IV talks about the strategies for order of the estimations and the simulation results. Segment V introduces the end and future work.

II. RELATED WORK

Govind A Ali dissected the terrorist related texts both in English and in Arabic in his proposal [1]. To discover the every now and again showing up words in the posts he utilized the k-means Nearest Neighbor calculation and has recognized, checked three client id that have terrorist related texts by utilizing system chart. Enghin Omer identified terrorist type texts in his postulation using hash sign words [2]. Information set were collected of three distinctive types: First is that bolsters terrorist, second is enemies of terrorist, and last one is that terrorist is unable to find irregular texts. He utilized Naïve Bayes, Ada Boost and bolster vector machine (SVM) for the grouping. Matthew Rowe et. al. [3] examined the reaction of individuals after terrorist assaults. They inspected the conduct of individuals amid pre-and post-terrorist assaults from tweets and approved by utilizing dictionary basis methodology.
Abhishek Barve said KNN furnishes exactness of 100% contrasted and SVM and Random Forest which gives the

precision of 74.20 and 74.69 individually [4]. Agarwal et. al. [5] introduced a down to earth approach for analyzing sentiments of tweets utilizing polarity where content is grouped into positive, negative and impartial. To classify the posts three models were used: tree part based and highlight based methodology, unigram. On unigram approach they used Support Vector Machine (SVM), and accomplished 71.35% normal precision. By using AI calculations Alec go et. al. investigated a novel way to deal with order assessment of texts [6]. A polarity-based methodology was utilized to discover the assumption of tweets utilizing an emoji word reference. By utilizing Naïve Bayes, Support Vector machines and Maximum Entropy for unigram include extraction they accomplished a precision of over 80%.

III. METHODOLOGY

A. Preprocessing

In this module, the tweets which are imported to database from the twitter API, these tweets comprise of superfluous words, whitespaces, hyperlinks and exceptional characters. First, we have to do sifting process by expelling every single superfluous word, whitespaces, hyperlinks and uncommon characters.

B. Self Learning and word Standardization System

In this module, first we have to introduce the word reference (first emphasis dictionary). In the lexicon by and large we have to instate the positive, negative unbiased and things. Every single huge datum and data mining ventures dependent on the prepared information, without prepared information (introduction of words). So, instatement of the prepared information is essential. In oneself learning framework, we are doing word standardization, here we are not considering past, present and future status of the words, just we are thinking about the word.

I. Exploratory Data Analysis:

In insights, exploratory data analysis is a way to deal with breaking down informational indexes to condense their principle qualities, frequently with visual strategies. A factual model can be utilized or not, however in a general sense exploratory data analysis is for seeing what the data can tell us past the formal showing or hypothesis testing task.

Imputation: A threshold of greater than three standard deviations is used to identify attributes with outliers. Since the mean is not robust and is affected by outliers, the median recommended for imputation.

C. Sentiment Analysis Module

In this module, preprocessed tweets are brought from the database one by one. First, we need check one by one catchphrase whether that watchword is thing are not, if thing we will expel it from the specific tweet. After that the rest of the catchphrases checked with estimation type, regardless of whether that watchwords have certain feeling or negative slant or impartial conclusion. The rest of the catchphrases in the tweet which does not has a place with any of the assumption will be allocated transitory supposition dependent on the more tally of positive, negative and nonpartisan. In the second cycle if the rest of the word crosses the edge of positive, negative or unbiased, that watchword forever included as extension in the lexicon. At last Sentiment of the tweet is identified dependent on the positive, negative and impartial words in the specific tweet.

I. K-Nearest Neighbor Classifier:

One of the different classifiers, 'KNN classifier' is a case-based learning algorithm which depends on a separation or comparability work for different sets of perception, for example, the Euclidean distance function. It is tried for some applications as a result of its effectiveness, non-parametric and easy to execution properties. However, under this technique, the grouping time is extremely long and it is hard to find the optimal value of K. For the most part, the best option of k to be chosen relies upon the data. Likewise, the impact of noise on the classification is diminished by the bigger estimations of k yet make limits between boundaries and few classes. By utilizing different heuristic systems, a good 'k' can be chosen. So as to beat the above said drawback, alter conventional KNN with various K esteems for various classes instead of fixed an incentive for all classes.

KNN calculation is utilized to characterize cases dependent on closest preparing models in the edge space. KNN algorithm is known as lazy learning algorithm in which function is approximated locally and calculations are delayed until classification. A greater part of cases is utilized for grouping process. Object is classified into the specific class which has most extreme number of closest examples.

II. Random Forest:

A Random forest incorporates an additional degree of randomness to pressing. Albeit each tree is created using another bootstrap trial of the dataset, the procedure by which the gathering trees ate constructed is improved. A random forest indicator is a gathering of individual classification tree indicators. For each discernment, every individual tree vote for one class and the forest predicts the class that has most of votes. The client needs to determine the quantity of randomly selected factors (mtry) to be looked over for the best split at each point(node).

Among all variables in standard trees node utilizes the best part, in a random forest the node is part using the best among a subset of pointers randomly selected at that node. The greatest tree possible is created and isn't pruned. The root node of each tree in the forest contains a bootstrap test from

the main data as the planning set. In the planning set observations are not present, are implied as "out-of-bag" discernments.

IV. RESULTS AND DISCUSSION

Here, in this research python programming is utilized, as it is cross platform compatible, gives large number of packages, better diagrams and exact strategies. At first, tweets related to terrorist related hashtags are recovered utilizing twitter API. This twitter API gives the cutoff of recovering such short tweets for a given client in XML document position. Each XML record could add mini texts up to 200 without a moment's delay. By at that point, the essential cleaning of these tweets is finished utilizing Regular Expression, tokenizing and making vocabularies. Regular Expression is finished utilizing an inbuilt function in python. By then the polarity-based thoughts are settled. Secondary cleaning of the tweets helps in structure corpus and frequently happening words. Python gives the text mining package (tm), which gives works that fuses stacking the substance into corpus and for cleaning the tweets information utilizing the cutoff named "tm_map". The frequency of words as 25 were made by the word cloud. The sentiment classification is done by applying KNN, Random Forest and Exploratory data analysis on the evaluation of the data. The execution estimations are arranged and are examined. Moreover, the through and through zone is removed using the geocoding API.

In Exploratory data analysis, imputation is performed on the data and the mean and standard deviation is calculated after replacing the missing data.

	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	max
nperpcap	120439.0	0.111276	1.984861	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	406.0	406.0
nkill	120439.0	2.506157	12.142033	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	2.0	5.0	1570.0	1570.0
nkillus	120439.0	0.035595	5.512533	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1360.0	1360.0
nkillter	120439.0	0.469449	4.075478	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	500.0	500.0
nwound	120439.0	3.498149	38.268746	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	3.0	7.0	8191.0	8191.0
nwoundus	120439.0	0.016946	0.811726	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	151.0	151.0
nwoundte	120439.0	0.098647	1.438473	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200.0	200.0

Figure1: Summary statistics after imputation

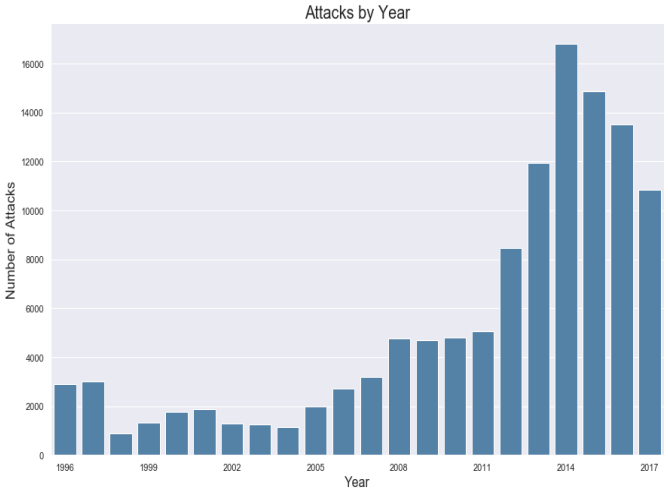

Figure2: Number of attacks by year

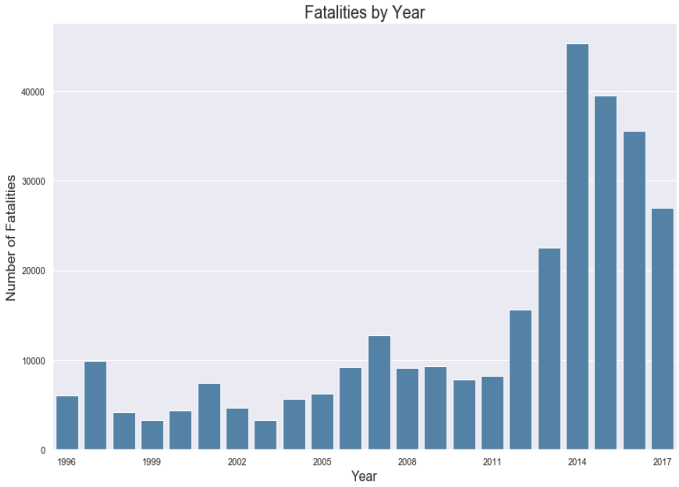

Figure3: Number of Fatalities by year

Figure4: Attacks by Geographical region

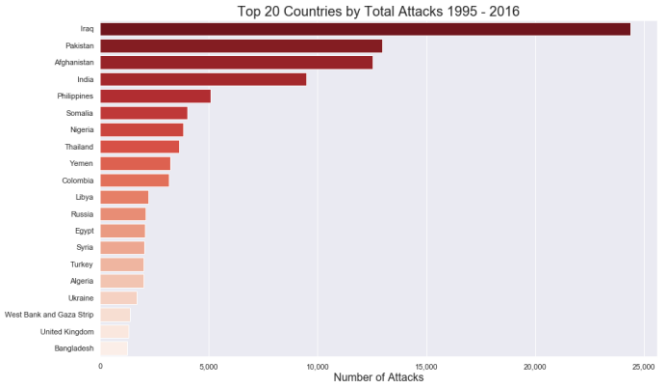

Figure5: Total number of attacks by country from 1995 to 2016

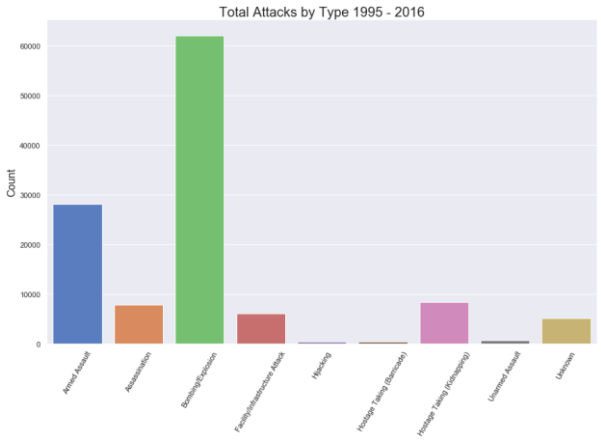

Figure6: total number of attacks by attack type from 1995 to 2016

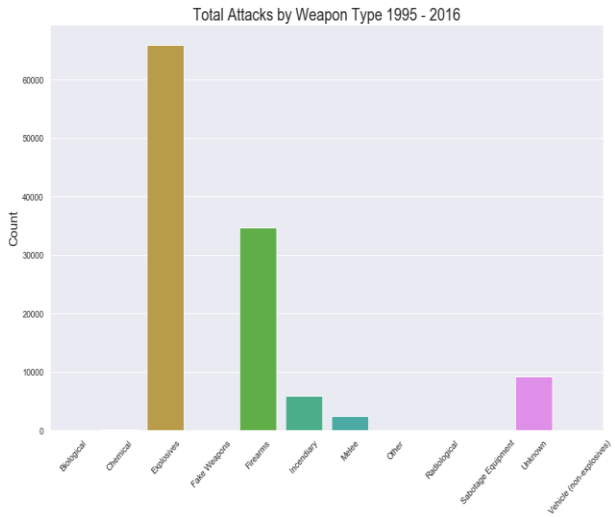

Figure7: total number of attacks by weapon type from 1995 to 2016

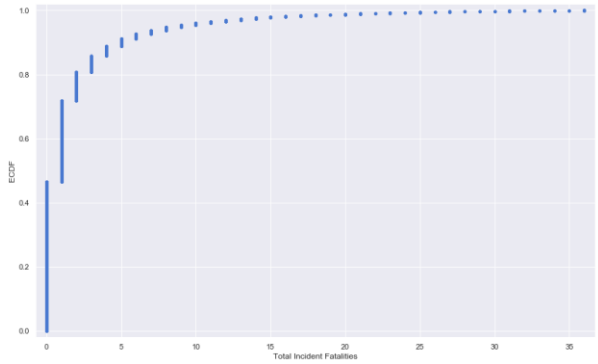

Figure8: Fatalities Empirical Cumulative Distribution

V. CONCLUSION AND FUTURE SCOPE

In this result, request of the terrorist related data from twitter and the desire for its estimation are done. A two-advance cleaning process, slant examination, habitually happening words, datamining calculations, and geolocation are joined. System was readied using KNN, Random Forest and Exploratory data analysis. The ampleness of the preparation framework can be measure using review and accuracy. accuracy is the probability of recouping relevant short messages. Review is the probability of the criticalness of recouped short messages. The consonant measure (F-measure) was used to get a singular motivation for review and accuracy. The weighted F - measure (Fβ measure) was used as accuracy was ought to have been highlighting in current situation. The system gives best results to Accident, Development-Government, Climate-Disaster, Entertainment, Health, Education, Sports, War-Terrorist Crime, Politics and Economy-business social occasions. KNN gives precision of 88.05 and Random Forest gives the exactness of 89.

As a future work, gathering more information from various dialects and applying supposition examination on n-grams with increasingly precise geolocation's API should be possible. Consolidating more lexicons could help in giving better outcomes to supposition examination. Also, breaking down more tweets could prompt the terrorist accounts and can give other private and unlawful information.

REFERENCES

[1] Ali, Govand A., "Identifying Terrorist Affiliations through Social Network Analysis Using Data Mining Techniques,". M.S Theses, Dept. Information Technology, Valparaiso Univ., Indiana, USA, 2016.

[2] Enghin Omer, "Using Machine Learning to Identify Jihadist Messages on Twitter,". M.S Theses, Dept. Information Technology, Uppsala Univ., Sweden, 2015.

[3] M. Rowe and H. Saif, "Mining pro-ISIS radicalisation signals from social media users," in Proceedings of the 10th International Conference on Web and Social Media, 2016.

[4] Abhishek Barve, "Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages,"

International Research Journal of Engineering and Technology (IRJET), Vidyalankar Institute of Technology, India, 2018.

[5] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R, "Sentiment Analysis of Twitter Data," in Proc of ACL HLT Conf, 2011.

[6] Go, A., Bhayani, R., and Huang, L, "Twitter Sentiment Classification using Distant Supervision," Technical Report, Stanford Digital Library Technologies Project, 2009.

[7] Juan DU, Zhi an Yi. "A New KNN Categorization Algorithm for Harmful Information Filtering", 2012 IEEE.

[8] Mateusz Budnik, Iwona Pozniak-Koszalka, Leszek Koszalka, "The Usage of the k-Nearest Neighbour Classifier with Classifier Ensemble", 12th International Conference on Computational Science and Its Applications, 2012 IEEE.

[9] Mohammad Abdul Wajeed, T. Adilakshmi," Semi- Supervised Text Classification Using Enhanced KNN".2011 IEEE.

[10] Lijun Wang, Xiqing Zhao," Improved Knn Classification Algorithms Research in Text Categorization". 2012 IEEE.

[11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon," What is Twitter, a Social. or a News Media?". International World Wide Web Conference Committee (IW3C2), April 26–30, 2010.

[12] G. Kesavaraj, Dr.S. Sukumaran, "A Study on Classification Techniques in Data Mining", IEEE 4th ICCCNT - 2013 July 4 - 6, 2013.

