International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 06 | Sep-2015 www.irjet.net p-ISSN: 2395-0072

Real time Tweet analysis for event detection & reporting system for Earthquake

Ghansham V.Shendge [1], Mangesh R.Pawar [2], Nikhil D.Patil [3], Pratik R.Pawar [4], Prof: Devdatta B.Bagul [5]

[1 2 3 4 5] *B.E., Computer, BVCOERI, Maharashtra, India*

Abstract - ***Twitter is new social networking trade in currently a day. Real time nature is a very important characteristic of Twitter. We have a tendency to investigate the period of time interaction of events like earthquakes in Twitter associate degreed propose a rule to observe tweets and to sight a target event. To sight a target event, we have a tendency to devise a classifier of tweets supported options like the keywords in a very tweet, the amount of words, and their context. Later, we have a tendency to turn out a probabilistic spatiotemporal model for the target event that may notice the middle of the event location. We have a tendency to regard every Twitter user as a sensing element and apply particle filtering, that square measure wide used for location estimation. The particle filter works higher than different comparable ways for estimating the locations of target events. Projected model is provided which might notice the Centre of the event location. The twitter user's square measure thought to be sensors and apply particle filter, in the main used for detection the situation. Due to the various earthquakes and also the sizable amount of twitter users throughout the country, we will sight associate degree earthquake with high chance simply by watching tweets. As Hadoop is used for processing data huge in size, we will use this framework. And it is mandatory to process or train system by maximum of data in very short time span. In existing system it requires too much time to process and we must develop such system which is very much responsive and very less turnaround time for reporting & alarming. So in our framework we are utilizing Naive Bayes algorithm for training. So we are going to implement machine learning concepts too. And it is very huge point of our system. Our system detects earthquakes promptly and notification is delivered abundant quicker than JMA broadcast announcements.***

Key Words: *Twitter, Event detection, Social sensing element, Location estimation, Earthquake*

1. Introduction

Twitter is classified as a microblogging service. Microblogging may be a variety of blogging that allows users to send transient text updates or micromedia like images or audio clips. Microblogging services aside from Twitter embody Tumbler, Friend Feed, Jaiku, identi.ca, and others [14].3 Our study, which is S.Anand & K. Narayana International Journal of rising Engineering analysis and Technology ninety-seven based on the period nature of 1 social networking service, is applicable to alternative small blogging services, however we tend to specifically examine Twitter during this study thanks to its quality and knowledge volume [15]. An important characteristic that's common among small blogging services is their period nature. Though web log users usually update their blogs once each many days, Twitter users write tweets many times during a single day. Users will savvy alternative users do and infrequently what they're brooding about currently, users repeatedly come back to the location and check to check what people do. Many necessary instances exemplify their period nature: within the case of a very sturdy earthquake in Haiti, several footages were transmitted through Twitter [16]. Folks were thereby able to apprehend the circumstances of injury in Haiti straight off. In another instance, once associate aeroplane crash-landed on the Hudson River in the big apple, the primary reports were revealed through Twitter and tumbler [17].

In such a way, varied update leads to varied reports associated with events. They embrace social events like parties, baseball games, and presidential campaigns. They additionally embrace black events like storms, fires, traffic jams, riots, significant precipitation, and earthquakes. Actually, Twitter is employed for varied period of time notifications like that necessary for facilitate throughout a

large-scale hearth emergency or live traffic updates. Adam Ostrow, the Editor in Chief at Mashable, a social media news web log, wrote in his web log regarding the attention-grabbing development of period of time media4:

This column fine signifies the inspiration of our study. The analysis question of our study is, "can we incline to determine such occasion prevalence in period of time by watching tweets?" This paper presents associate investigation of the period of time nature of Twitter that's designed to determine whether or not we are able to extract valid data from it. We tend to propose an occurrence notification system that monitors tweets and delivers notification promptly victimization information from the investigation. during this analysis, we tend to take 3 steps: 1st, we tend to crawl varied tweets associated with target events First, to get tweets on the target event exactly, we have a tendency to apply linguistics analysis of a tweet. for instance, users may create tweets like "Earthquake!" or "Now it's trembling," that tremor or trembling may be keywords, however users may also create tweets like "I am attending AN Tremor Conference," or "Somebody is shaking hands with my boss." we have a tendency to prepare the coaching knowledge and devise a Tweets Classifier employing logistic regression (KLR) supported options like keywords in a very tweet, the quantity of words, and also the context of target-event words. When doing thus, we have a tendency to acquire a Variation space time model of an instance. We have a tendency to then create a vital assumption: every Twitter user is considered a detector and every tweet as sensory data. These virtual sensors, that we have a tendency to designate as social sensors, square measure of a large selection and have numerous characteristics: some sensors square measure terribly active; others don't seem to be. A detector may well be inoperable or defective typically, as once a user is sleeping, or busy doing one thing else. Consequently, social sensors square measure terribly creaking compared to standard physical sensors. relating to every Twitter user as a detector, the event-detection downside will be reduced to at least one of object detection and placement estimation ubiquitous/ pervasive computing surroundings during which we've various location sensors: a user features a mobile device or a full of life badge in surroundings wherever sensors square measure placed. Through infrared communication or a LAN signal, the user location is calculable as providing location-based services like navigation and deposit guides [9], [10]. We have a tendency to apply particle filters, that square measure wide used for location estimation in ubiquitous/pervasive computing [11]. As AN application, we have a tendency to develop earthquake coverage system victimization Japanese tweets. Japan has various earthquakes. Twitter users square measure equally various and Earthquake Coverage System Development by Tweet Analysis International Journal of rising Engineering analysis and Technology ninety-eight

Geographically spread throughout the country

2. LITERATURE SURVEY

Twitter is a noteworthy example of the foremost recent kind of social media. Varied researchers have examined Twitter. Relating to similar analysis to it conferred during this paper, some researchers have tried topic detection victimization Twitter. Cataldi et al. projected a unique technique to discover rising topics employing a keyword-based topic graph. They succeeded in detective work news keywords that area unit fashionable in Twitter. As an example, (a volcano in Iceland) and Samaranch (the previous President of IOC, World Health Organization died in Apr 2010). Marc et al. divided more and more fashionable keywords on Twitter into patterns of assorted type's victimization Kyrgyzstani monetary unit, thus signifying that Tweet users add to the discussion of those trends. Other than the studies introduced in Section one and these studies, many others are done. We have a tendency to classify studies coping with Twitter or knowledge on Twitter into 3 teams. First, some researchers specifically examine the network structure of Twitter and investigate Twitter network options of assorted types. Java et al. analyzed Twitter as early as 2007. They delineated the common system of Tweet operators and examined the motivations of Twitter users [2]. Haewoon et al. crawled a massive quantity of Twitter knowledge, analyzed the Twitter follower-following topology and hierarchic users by Page Rank [4]. Huberman et al. analyzed quite three hundred thousand users. They found that the relation between friends (defined as someone to whom a user has directed posts victimization associate "@" symbol) is that the key to understanding interaction in Twitter [3]. Second, some agents have observed features of Twitter as common media. Recently, Boyd et al. have continued their investigation of retweet activity, that is that the Twitter-equivalent of e-mail forwarding, by that users post messages that were

originally denote by others [5]. Tumasjan et al. crawled several tweets touching on the election in Germany and tried to predict the results of the election: those political parties would win the election. O'Connor extracts vox populi from Twitter victimization sentiment analysis and reports the chance of employing a projected technique rather than polls. Third, some studies elucidate the advantages of novel applications of Twitter: Ebner and Schiefner establish a microblogging community and study the way to use Twitter as a tool for mobile e-learning. The combination of the linguistics net and microblogging was delineated in a very previous report within which a distributed design is projected and also the contents area unit aggregative. We choose earthquakes in Japan as target events, supported the preliminary investigations. We tend to make a case for them during this section. First, we decide earthquakes as target events for the subsequent reasons:

1. Unstable observations are conducted worldwide, that facilitates acquisition of earthquake data, that additionally makes it straightforward to validate the accuracy of our event detection methodology; and

2. it's quite substantive and valuable to discover earthquakes in earthquake-prone regions.

Second, we decide Japan because the place supported the subsequent investigation. It's apparent that the sole intersection of the 2 maps, those regions with several earthquakes and enormous Twitter users, is Japan. Alternative regions like country, Turkey, Iran, Italy, and Pacific coastal U.S.A. cities like la and city additionally roughly meet, however their various densities are a lot of below that in Japan. Several earthquake events occur in Japan and plenty of Twitter users observe earthquakes in Japan, which implies that social sensors are distributed throughout the country.

We gift a quick summary of Twitter in Japan: The Japanese version of Twitter was launched on Gregorian calendar month 2008. In February 2008, Japan was the No. a pair of country with relation to Twitter traffic.5 At the time of this writing, Japan has the second largest range of tweets (18 % of all tweets are announce from Japan) within the world. Therefore, we decide earthquakes in Japan as a target event as a result of the high density of Twitter users and earthquakes in Japan.

2.1 System Overview

We tend to are getting to propose an occurrence notification system. An occurrence watching system monitors tweets and delivers notification promptly mistreatment investigation results. We tend to propose a system that's supported investigation of tweets i.e. real time investigation. During this analysis, we tend to take 3 steps:

1) We analyze no of tweets associated with target events;

2) We got to style such a probabilistic module to research and extract events from those tweets and predict locations of events with category verifying as positive and negative class.

3) Finally developed coverage method that excerpts tremors from Tweet and shows a message to registered users.

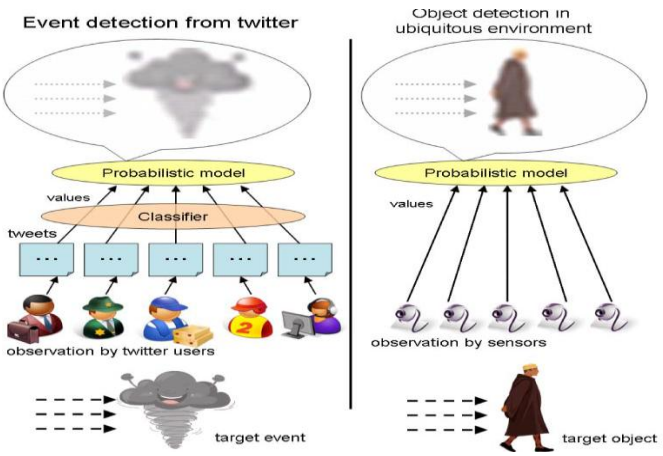

Fig -1: System Architecture [13]

Diagram description:

1) Tweet search API window collects tweets regarding events I large scale.

2) We crawl no of tweets using tweeter crawler to find out useful Tweets and scripted to processing.

3) Processed twitter distinguished between "+ class and - class" by using algorithm.

4) From positive class we find out event detection and location using Hadoop framework training algorithm.

5) Lastly we improve an actual time tweeter operator's method to report real time event detection and analysis of earthquake reporting

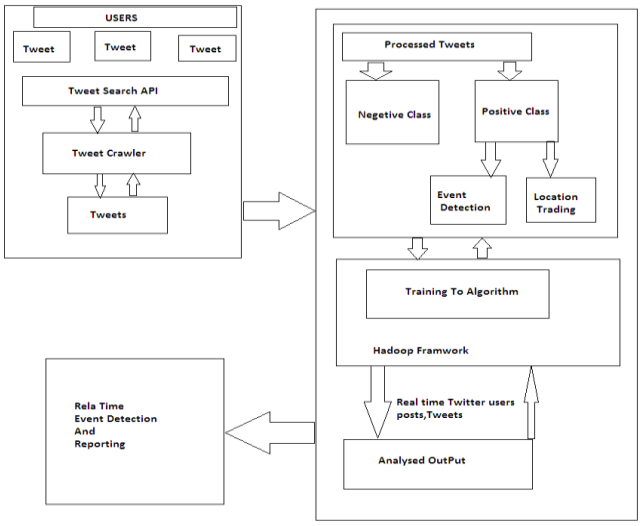

Fig -2: System Overview

2.2 Methods

For event detection and placement estimation, we tend to use probabilistic models. During this section, we tend to 1st describe event detection from time-series information. Then we tend to describe the situation estimation of a target event.

1) Temporal Model

Each tweet has its own post time. Once a target event happens, however the sensors discover the event, we tend to describe the temporal model of event detection. First, we tend to examine the particular information. The several quantities of tweets for a target event: Associate in nursing earthquake. It's apparent that spikes occur within the variety of tweets. Everyone corresponds to an incident} occurrence. Specifically concerning Associate in nursing earthquake, quite ten earthquakes occurred throughout the amount.

2) Spatial Model

Each tweet is related to a location. We tend to describe a technique which will estimate the situation of an occasion from device readings. To resolve the matter, many ways of Bayesian filters square measure planned like Kalman filters, multi-hypothesis following, grid-based and topological approaches, and particle filters. For this study, we tend to use particle filters, each of that square measure wide employed in location estimation.

A) Particle Filters

A particle filter could be a probabilistic approximation algorithmic rule implementing a Bayes filter, and a member of the family of successive Monte Carlo strategies.

B) Consideration of sensing element Geographic Distribution. We should take into account the sensing element geographic distribution to treat readings of social sensors additional exactly in location estimation by physical sensors, those sensors area unit situated equally in several cases. We will treat sensing element readings equally in such things. Actually, social sensors aren't placed equally in several cases as a result of social media user's area unit targeted in urban areas. In Japan, most users board capital of Japan. Therefore, we should always incorporate the geographic distribution of social sensors into abstraction models

C) Techniques to hurry up the method

As represented during this paper, we wish to estimate location of events quickly as shortly as potential as a result of one objective of this analysis is to develop a period earthquake detection system. Therefore, we tend to should decrease the time quality of strategies used for location estimation.

3) Information Diffusion associated with a period Event

Some info associated with an occasion diffuses through Twitter. For instance, if a user detects associate earthquake and makes a tweet regarding the earthquake, then a fan of that user would possibly create tweets that. This characteristic is very important as a result of, in our model; sensors won't be reciprocally freelance, which might have associate unsought result in terms of event detection.

For event detection and placement estimation, we tend to use probabilistic models. From time-series information, we 1st describe event detection. Then we tend to describe the placement estimation of a target event. Each tweet has its own post time. Once a target event happens, however do the sensors observe the event? We tend to describe the temporal model of event detection. First, we tend to

examine the particular information. Everything corresponds to prevalence occurrence. Specifically relating to associate earthquake, over ten earthquakes occurred throughout the amount.

3. PROBLEM DEFINITION

The reference systems used reduced corpus datasets that do not scale to larger amounts of data. The performance, effectiveness and durability of those systems was not being designed to handle big amounts of data but today's online social network service volume of data (creates massive unstructured text data streams) make them obsolete systems. Performing real-time event detection using Twitter requires dealing and mining massive un-structured text data stream that has messages continuously approaching at sky-high data rates. Given this, the approach to deal with this specific problem involves providing solutions that are able to mine continuously, high-volume of open-ended data streams as they arrive. Considering that those sources of data are coming from social network users it is expected that information collected using metrics of networks analysis (nodes, connections and relations, distributions, clusters and communities) could improve the quality of the solution of the algorithm. Apart from, the data in online social network services is also dynamic; messages and arriving at very high data rate. Computation of such vast amount of data needs necessarily technology that has a *highly scalable* storage platform and performs distributed concurrent parallel execution of database. Time and Cost Effectiveness is an issue. Online social network text streams seem to be the ideal source to perform real-time event detection as they are very much Cost Effective.

4. EXISTING SYSTEM

The Existing system, called Toretter is presently working in Japan for Earthquake detection using Twitter has been operated since August 8; 2010.Users can see the detection of past earthquakes. Also they can register to receive notices of future earthquake detections. It alerts users for imminent earthquake. It is hoped that a user receives alert before the earthquake actually affects the area. We assess various conditions under which alarms might be sent to choose better framework for our suggested system.

We set alarm conditions as Ntweet (positive tweets) come in 10 minute. We evaluate those methods by

Precision= Nearthquake/Nalarms
And Recall = Nearthquake / Allearthquake
All earthquake (Nearthquake: Number of earthquakes detected correctly, Nalarms: number of distress signal, Allearthquake: number of tremors that occurred).
We must change the use of this condition vigorously to increase the accuracy of the system, particularly in terms of the repetition and intensity of earthquake.

United States Centers for Disease Control and Prevention uses Twitter as a tool to gather timely health and safety information and encourages the strategic use of Twitter for effectively and inexpensively reach partners for emergency threat and to update individuals about emergency preparedness for health and safety concerns since 2009 swine flu pandemic.

5. PROPOSED SYSTEM

We are going to design the system called 'CrisisCall' is kind of alarming or reporting system based on Hadoop Framework to process huge amount of tweeter data related with 'Earthquake' like calamities. Actually it is one of much needed project started by seeing hazards to people in Nepal. Recently in July-August 2015 Nepal faced very big natural calamity due to Earthquake. Many people lost their lives. By seeing such huge hazard we are proposing such system which will make reports or alarms to social network users on their accounts or public announcing on web by analyzing huge amount of Tweets, Posts related with.

'CrisisCall' is prepared by not only focusing on Earthquake, but also events like Storm, Heavy Rainfall, Flood etc. can be monitor by us. One of best thing in this project is we are using Hadoop Framework which was not present in existing system. Means we can process very huge data within very short time period.

Our Systems' flow will starts from Tweets crawling. First we collect as many tweets as possible from Tweeter Database or else we crawl the web for Tweets. Then we will process them as Positive or Negative Tweets using Navie Bayes Algorithm. So we have to make them sorted as per feeling or sentiment in tweet. In such scenarios we will use sentiment analysis concepts do make type sorting of tweets. Also from such tweets we have detect event and location too. So it is one of big task while developing. As tweeter data gives us locations as time of tweet we can

definitely predict event with respect to time & location too.

Processed data of Messenger is used for preparing the algorithm, so that next time if such data came for prediction it will processed directly. For processing we will use Logistic Regression algorithm over Hadoop Map Reduce framework. When real time Tweeter users tweet on web our system detects event & location & does reporting & alarming automatically. 'CrisisCall' is best prototype for Harmful event detection & location tracking for any kind. What just change is we need to change database for training the system & evaluation.

5.1 Algorithm

1) Developing and Preparing Datasets
We are trying to come up with a function that can predict for future inputs based on the practice it has acquired through the past inputs and their outputs (training set).
2) Logistic Regression is - coming up with a possibility function that can give's a chance, for an input to belong to any one of the various classes'(Classification)
3) Consider binary classification problem as each tweet is either positive (y = 1) or negative (y = 0). These are the 2 parameters here. Our goal is to come up with a possible function that takes in an input X (Number of tweets) and return 'what is the possibility of this tweet to be positive'.
4) This probability function is the Sigmoid Function and which is:

$$\frac{1}{(1+e)^{(-z)}}$$

Since, probability of any occurrence is [0, 1] (between 0 and 1, including both), this task absolutely fit to be used as a probability function for logistic regression.
z = transpose (theta) * X
X =Number of Tweet
Theta=?
X = 0.9 and it gave probability for it to be positive = 0.3, which means it has more possibility of starting gently.
But clearly from our training set this is definitely wrong as for X = 0.9, Y = 1 i.e. malignant.
5) In general the error function in logistic regression is given by:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \cos t(h_\theta(x^{(i)}), y^{(i)})$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1-h_\theta(x^{(i)}))\right]$$

Here m = no of elements in training set,
Y is commonly 1 or 0 and h(x) is solely the 'Sigmoid function' we spoke above.
Since sigmoid is a function of theta (specified in z above), therefore J is a set of theta.
6) Now we minimize J over theta and search out those values of theta for which our error function in minimized.
7) Once we have theta, our probabilistic result (sigmoid) is ready and we can apply it to any size of data and it will give us its.

6. CONCLUSION

We have a tendency to investigate the period nature of Twitter, devoting specific attention to event detection. Linguistics analyses were applied to tweets to category verifies them into a positive and a negative class. We have a tendency to regard every Twitter user as a device, and set the matter as detection of a happening supported sensory observations. Location estimation strategies like particle filtering area unit are used to estimate the locations of events. As associate degree application, we have a tendency to developed associate degree earthquake coverage system, which could be a novel approach to advice folks promptly of associate degree earthquake event

7. ACKNOWLEDGMENTS

This work was supported by Pune University and BVCOERI, Nashik. We are very much thankful that he gave opportunity to complete this work in time to us. We would also like to thanks our Prof. C. K.Patil Principal BVCOE&RI. Prof. H. D.Sonawne H.O.D Computer Department, for providing their valuable support and time throughout engineering.

REFERENCES

[1] M. Sarah, C. Abdur, H. Gregor, L. Ben, and M. Roger, "*Twitter and the Micro-Messaging Revolution,*" technical report, O'Reilly Radar, 2008.
[2] Java, X. Song, T. Finin, and B. Tseng, "*Why We Twitter: Understanding Microblogging Usage and Communities,*" Proc. Ninth WebKDD and First SNA-KDD

Workshop Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07), pp. 56-65, 2007.

[3] B. Huberman, D. Romero, and F. Wu, "*Social Networks that Matter: Twitter Under the Microscope,"* ArXiv E-Prints, http://arxiv.org/abs/0812.1045, Dec. 2008.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "*What is Twitter, A Social Network or A News Media?"* Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 591-600, 2010.

[5] G.L. Danah Boyd and S. Golder, "*Tweet, Tweet, and Retweet: Conversational Aspects of Retweeting on Twitter,"* Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS-43), 2010.

[6] Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe, *"Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment,"* Proc. Fourth Int'l AAAI Conf. Weblogs and Social Media (ICWSM), 2010.

[7] P. Galagan, *"Twitter as a Learning Tool. Really,"* ASTD Learning Circuits, p. 13, 2009.

[8] K. Borau, C. Ullrich, J. Feng, and R. Shen, *"Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence,"* Proc. Eighth Int'l Conf. Advances in Web Based Learning (ICWL '09), pp. 78-87, 2009.

[9] J. Hightower and G. Borriello, *"Location Systems for Ubiquitous Computing,"* Computer, vol. 34, no. 8, pp. 57-66, 2001.

[10] M. Weiser, *"The Computer for the Twenty-First Century,* "Scientific Am., vol. 265, no. 3, pp. 94-104, 1991.

[11] V. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "*Bayesian Filtering for Location Estimation*," IEEE Pervasive Computing, vol. 2, no. 3, pp. 24-33, July-Sept. 2003.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "*Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors,"* Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 851-860, 2010.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "*Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development,"* Proc. Int'l Conf. World Wide Web (WWW '10), pp. 9526, Vol 25 April 2013.

[14] *Microblogging – Wikipedia, the free encyclopedia.*

[15] S.Anand, K.Narayan," *Earthquake Reporting System Development by Tweet Analysis*" IJEERT Volume 2, Issue 4, July 2014, PP 96-106 ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

[16] *Twitter search for Haiti survivors Channel 4 15 Jan, 2010.*

[17] Krūms, Jānis (January 15, 2009). *"There's a plane in the Hudson. I'm on the ferry going to pick up the people. Crazy*"