Real Time event occurrence system with particular opinion using Mapreduce Framework for Excavating Twitter Data

Miss.H.A.Patil, Prof. S..A.Joshi
Student, HOD
Computer Network, Flora Institute Of Technolgy,Pune,India

***Abstract*—As per some recent studies, public opinions expressed in social media may be correlated with various social issues. To find out what actually can be discovered in social media data, we need data mining. But traditional data mining has many limitations as size of datasets. So, new data mining approaches that can handle massive amount of data have recently been referred to as big data algorithms. This project proposes a data algorithm to handling Twitter data mining. Computationally, the speed of execution can be shown to increase significantly despite increases in data set size. However gap for communication is filled by the technology through 'social networking' sites. This project proposed a new system that delivers large database of Social Networking Site (SNS) called 'Twitter'. Here processing the tweet involves extraction of metadata of tweet, geocoding the physical address in a tweet, analyzing the sentiment of content in the tweet text and extracting the significant and key phrases from a text. After all the Information Extracted and NER (Named Entity Recognition) text analysis from tweet, are stored into a MongoDB database, as it is more scalable and more flexible among others of NoSQL databases. Here Object Oriented programming and Design patterns are used in implementation of this system, with proper testing and validation are performed at three levels, both normal and performance test results are evaluated to achieve a sophisticated system. Real time data analysis is key feature which we have implemented along with Machine learning approaches like Naïve Bayes theorem.**

***Index Terms* —social media, data mining; big data algorithm, Twitter, Mapreduce, Machine learning, Data Analysis**

- **Introduction**

Twitter is classified as a micro blogging service. Micro blogging may be a variety of blogging that allows users to send transient text updates or macromedia like images or audio clips. There is no doubt that over the last decade, with the rapid development of Web 2.0 technologies and communications engineering, the popularity and expansion of social media has been quite astounding [2]. The large amount of data that is generated from social media could potentially provide some new insights into the market and into consumer behavior [3]. The same applies to other sociological issues such as politics, the environment, the entertainment industry, the stock market, etc. [4]. The key function of data mining is to extract knowledge from data, and as the social media, is like a vast untouched land full of valuable data, there is an obvious incentive to use data mining techniques on that land.For instance, as a typical example of social media, Twitter is a micro-blogging application that allows interested users follow and comment on other users' thoughts or some events in their lives, in real time [5]. As one of the most popular social media, millions of users post over 140 million tweets every day. This situation means that Twitter is a corpus that holds valuable data as a form of collective knowledge, and recently it has attracted much attention from researchers in various fields. Through evaluating past works on social media data mining, we identify the most compelling research problem that need to be solved for social media data mining is, to accurately learn opinions that could uncover sentiments expressed in large scale ambiguous and unstructured contents posted to social media, with fast processing capacity to handle the big data in social media.

- **LITERATURE SURVEY**

Micro blogging may be a variety of blogging that allows users to send transient text updates or macromedia like images or audio clips. There is no doubt that over the last decade, with the rapid development of Web 2.0 technologies and communications engineering, the popularity and expansion of social media has been quite astounding [2]. The large amount of data that is generated from social media could potentially provide some new insights into the market and into consumer behavior [3]. The same applies to other sociological issues such as politics, the environment, the entertainment industry, the stock market, etc. [4]. The key function of data mining is to extract knowledge from data, and as the social media, is like a vast untouched land full of valuable data, there is an obvious incentive to use data mining techniques on that land.For instance, as a typical example of social media, Twitter is a micro-blogging application that allows interested users follow and comment on other users' thoughts or some events in their lives, in real time [5]. As one of the most popular social media, millions of users post over 140 million tweets every day. This situation means that Twitter is a corpus that holds valuable data as a form of collective knowledge, and recently it has attracted much attention from researchers in various fields. Through evaluating past works on social media data mining, we identify the most compelling research problem that need to be solved for social media data mining is, to accurately learn opinions that could uncover sentiments expressed in large scale ambiguous and unstructured contents posted to social media, with fast processing capacity to handle the big data in social media.

- **MOTIVATION**

This document evaluates the available API's to get access data from the twitter, and implementation of suitable procedure to build database of social network data (twitter). To make it useful for visualization of twitter data, which is efficient and effective in utilization and maintenance Also examined and compared existing gazetteers and Entity extraction libraries. For a task of implementing NER (Names Entity Recognition) to extract annotation specific to the defined patterns and formats after proper analysis of input. Sentiment analysis have insight to identify the positive and negative sense in the text, the evaluation focuses mainly on the behavior aspects and words or phrases that means the human emotions. This work simplified to the process to sentiment analysis after proper review on contemporary analysis to classify sentiment over text.

III.**System Requirement & Specification**

Hardware Requirements

- Following are the number of hardware requirements:-
- Processor : Pentium IV or above , 2 GHz or higher.
  - 8 GB RAM
  - Min 80 GB HDD

Software Specification

- Operating System: Windows 7/8 32 bit
- Software Development Platform: Eclipse
- Framework: Hadoop 1.X,
- File System: HDFS
- Browser: Mozilla Firefox
- Os : Ubuntu 14.04 LTS
- Techniques: Java,jsp.
- Scripting Language: JavaScript.

.

VI. SYSTEM IMPLEMENTATION

This is investigated that 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. Also it is important 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

SYSTEM ARCHITECTURE

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:

- We analyze no of tweets associated with target events;
- 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.
- Finally developed coverage method that excerpts tremors from Tweet and shows a message to registered users.

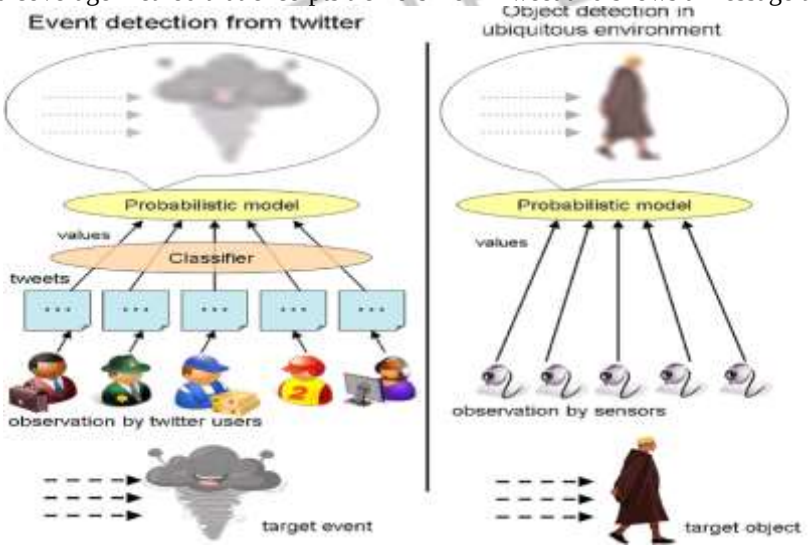

Figure : System Architecture

Diagram Description

- Tweet search API window collects tweets regarding events I large scale.
- We crawl no of tweets using tweeter crawler to find out useful Tweets and scripted to processing.

- Processed twitter distinguished between "+ class and - class" by using algorithm.
- From positive class we find out event detection and location using Hadoop framework training algorithm.
- Lastly we improve an actual time tweeter operator's method to report real time event detection and analysis of earthquake reporting.

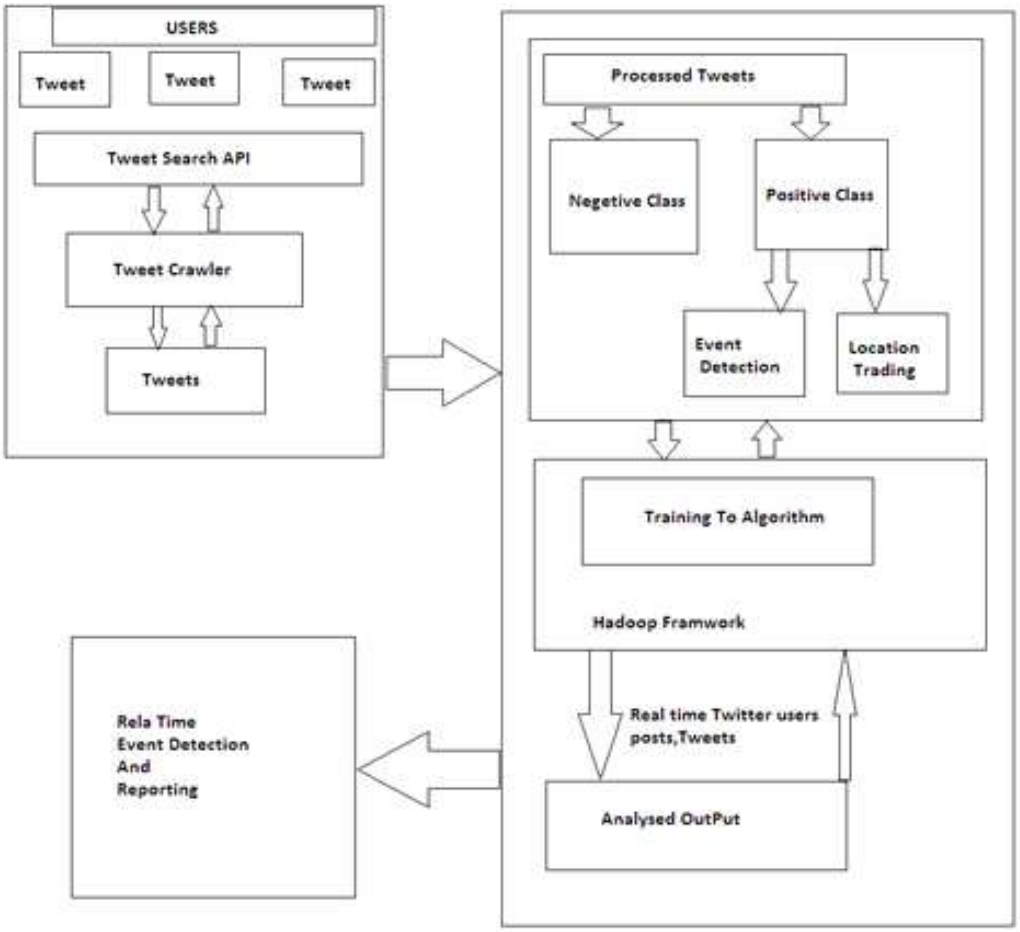

Figure : System Overview

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.

- 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.
- 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
- 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.

METHODOLOGY FOR DEVELOPMENT

This project is an investigation of the real-time nature of Twitter that is designed to ascertain whether we can extract valid information from it. We propose an event notification system that monitors tweets and delivers notification promptly using knowledge from the investigation. In this research, we take three steps: first, we crawl numerous tweets related to target events; second, we propose probabilistic models to extract events from those tweets and estimate locations of events; finally, we developed an earthquake reporting system that extracts earthquakes from Twitter and sends a message to registered users as shown in figure.

Modules

Tweet Collection Module

In this module, we develop our system by posting tweets by the users. It is necessary to collect tweets referring to an earthquake from Twitter. This process includes two steps: crawling tweets from Twitter and filtering out tweets that do not refer to the earthquake. For crawling and filtering tweets, we recommend using script programming languages.

Crawling Tweets from Twitter Module

To collect tweets or some user information from Twitter, one must use the Twitter Application Programmers Interface (API). Twitter API is a group of commands that are necessary to extract data from Twitter. Twitter has APIs of three kinds: Search API, REST API, and Streaming API. In this section, we introduce Search API and Streaming API, which are necessary to crawl tweets from Twitter. We explain REST API later because REST API is necessary to extract location information from Twitter information. Additionally, it is known that Twitter API specifications are subject to change. When using Twitter API, it is necessary to know the latest details and requirements. They are obtainable from Twitter API documentation

Twitter Search API Module

The Twitter Search API extracts tweets from Twitter, including search keywords or those fitting other retrieval conditions, in chronological order. It is possible to use language, date, location and other conditions as retrieval conditions.

- Some points must be considered when using Twitter Search API:

It is possible to collect tweets posted only during the prior five days. It is not possible to search tweets posted six days ago.

It is only possible to collect the latest 1500 tweets at one time. (Technically speaking, it is possible to access one page with a request and track pages back to the 15th page. One page includes 100 tweets at most. Therefore it is possible to acquire the latest 1500 tweets at one time.)

Filtering Tweets using Machine Learning Module

We collected data from tweets including keywords related to earthquakes, such as earthquake, shake. Those tweets include not only tweets that users posted immediately after they felt earthquakes, but also tweets that users posted shortly after they heard earthquake news, or perhaps they misinterpreted some sense of shaking from a large truck passing nearby.

When the seismic activity reached its peak, the graph of tweets invariably showed a peak. However, when the graph of tweet counts showed a peak, the seismic activity did not necessarily show a peak. Some "false-positive" peaks of the graph of tweet counts arise from mistakes by users or some news related to earthquakes. Therefore, we must filter tweets to extract those posted immediately after the earthquake. We designate tweets posted by users who felt earthquakes as positive tweets, and other tweets as negative tweets.

Here, we describe the creation of a classifier to categorize crawled tweets into positive tweets and negative tweets, using Support Vector Machine: a supervised learning method.

Semantic Analysis on Tweets Module

Semantic Analysis on Tweet Search tweets including keywords related to a target event Example: In the case of earthquakes “shaking”, “earthquake” Classify tweets into a positive class or a negative class Example: “Earthquake right now!!” ---Positive “Someone is shaking hands with my boss” --- negative Create a classifier

Semantic Analysis on Tweet Create classifier for tweets use Support Vector Machine (SVM) Features (Example: I am in Japan, earthquake right now!) Statistical features (7 words, the 5th word) the number of words in a tweet message and the position of the query within a tweet Keyword features (I, am, in, Japan, earthquake, right, now) the words in a tweet Word context features (Japan, right) the words before and after the query word

Earthquake Reporting System Module

- In this module, the users will be altered if the earthquake occurs based on their location and the tweets. Effectiveness of alerts of this system Alert E-mails urges users to prepare for the earthquake if they are received by a user shortly before the earthquake actually arrives.

V.RESULTS

The following results are observed after dynamic data analysis.

FIG:Event Occurrence Representation

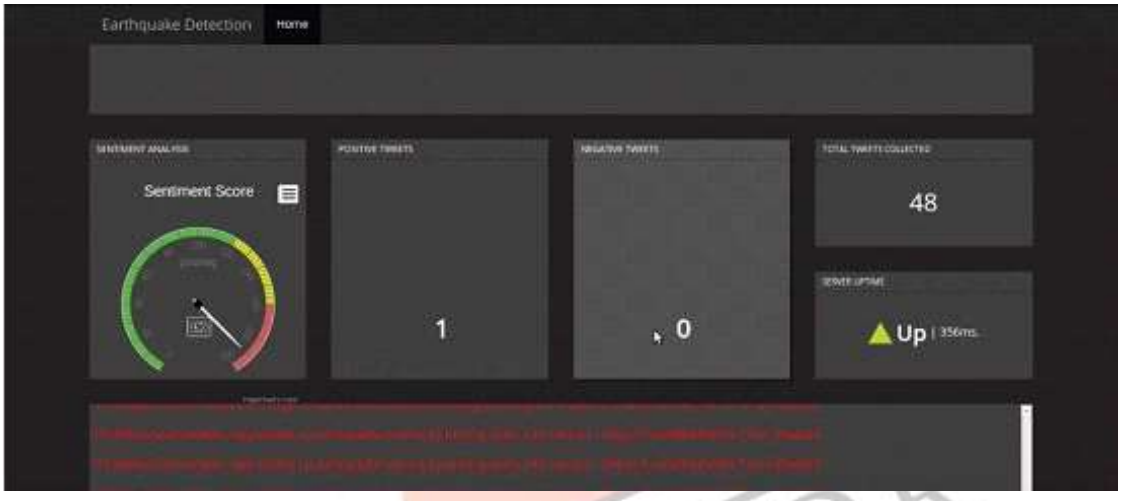

FIG:Tweet Analysis At Real Time

VI. CONCLUSION

This paper aims to serve a processed twitter tweet database to frontend third party visualization applications leads to locate & detect real time event like Earthquake. Text analysis focused on processing the tweets to extract information from the raw data of tweet, which can benefit the frontend application in projecting more information to the user, in terms of usability and exploring text-engineered data. The algorithms like MapReduce are used for sentiment analysis on top of hadoop while machine learning concepts also implemented for supervised learning for datasets which are coming as processed output. Here we focused on Earthquake related datasets which after preprocessing and sorting as positive & negative given to training algorithm as input while real time tweets coming from different users related such event will detect event & location. And at same time as per our proposed system module it will report the event. We have given name to this project as 'CrisisCall' and it will suits this architecture while our system will do very fast processing on huge datasets at very short time as we used hadoop framework for it. And it will definitely very helpful to fastest detection of events like happed in Nepal.

VII. FUTURE SCOPE

In future work for various datasets, there will be connectivity between this system and Disaster management team of nation. Also one important part is tweets coming over twitter platform are not of fixed format, so by applying machine learning algorithms system can find out such patterns or formats with respect to region from where security threats may evolved or it will create us issue. This system need to implement cloud architecture as scope of project besides current project satisfies this situation.

VIII. REFERENCE

[1] LI Bing and Keith C.C. Chan,"A Paralleled Big Data Algorithm with MapReduce Framework for Mining Twitter Data" 2014 IEEE Fourth International Conference on Big Data and Cloud Computing.

[2] A. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media", Business Horizons, vol. 53 (1), pp. 59-68, 2009.

[3] F. Wang, K. Carley, D. Zeng and W. Mao, "Social Computing: From Social Informatics to Social Intelligence". IEEE Intelligent Systems, vol. 22 (2), pp. 79-83, 2007.

[4] B. Ulicny, M. Kokar and C. Matheus, "Metrics for monitoring a social political blogosphere: A Malaysian case study". IEEE Internet Computing, vol. 14 (2), pp. 34-44, 2010.

[5] E. Schonfeld, "Mining the thought stream", 2009. Retrieved May 7, 2014

[6] M. Joshi, D. Das, K. Gimpel and N. Smith, "Movie Reviews and Revenues: An Experiment in Text Regression". Proceedings of HLT' 10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 293-296, 2010

[7] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market. Journal of Computational Science", vol. 2 (1), pp. 1-8, 2010.

[8] F. Zhu, Huan Sun and X. Yan. "Network mining and analysis for social applications." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[9] J. Li, et al. "Mining Trajectory Data and Geotagged Data in Social Media for Road Map Inference." Transactions in GIS, 2014.

[10] S. Baccianella, E. Andrea and S. Fabrizio, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," LREC, vol. 10, 2010.

[11] P. E. Greenwood and M. S. Nikulin, "A guide to chi-squared testing," John Wiley & Sons, 1996.

[12] B. Li, C. C. Chan,. "A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data". Proceedings of the 7th IEEE/ACM International Conference on Utility and Cloud Computing, workshop on Big Data and Social Networking Management and Security,London, UK, 2014.

[13] C. C. Chan, K. C. Wong and K. Y. Chiu, "Learning Sequential Patterns for Probabilistic Inductive Prediction," IEEE TSMC, vol 24(10), pp. 1532-1547, 1994.

[14] B. Li, C. C. Chan., "A Fast Big Data Collection System using0 MapReduce Framework". Proceedings of the 3rd IEEE International

[15] M. Sarah, C. Abdur, H. Gregor, L. Ben, and M. Roger, "Twitter and the Micro-Messaging Revolution," technical report, O'Reilly Radar, 2008.

[16] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Micro blogging 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.

[17] 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.

[18] 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.

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

[20] A. 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.