

# A Probabilistic Model of Real Time Event Detection and Reporting

Dharnasi Rama Devi<sup>1</sup> · Ms. Naga Aswani Puduru<sup>2</sup>

<sup>1,2</sup>*Department of Computer Science and Engineering,  
Aurora's Technological and Research Institute, uppal,  
Hyderbad, Andhra Pradesh*

**ABSTRACT**--A probabilistic model provides a way to detect multiple instances of real time events and to estimate the location of targeted event like earthquakes, typhoons, traffic jams. For this, two models have been proposed named temporal and spatial models to detect real time events and estimate the targeted event locations respectively by dealing with sensor reading appropriately. our work is based on the twitter-which is used to deal with sensor reading appropriately real time events and particularly for location estimation. An important characteristic of twitter is its real-time nature. We investigate the real-time interaction of events such as earthquakes in twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, we device a classifier of tweets based on features such as the keywords in a tweet, the number of words and their context, by using MI-SVM . We regard each twitter user as a sensor and apply particle filtering .which are widely used for location estimation. The particle filter works better than other other comparable methods for estimating the locations of target events. In a temporal model, each tweet has its own post time, when target event occurs. Depends on quantities of tweets, target event behavior can be analyzed. This distribution is apparently an exponential distribution i.e. 0.34 on average, on which data fit very well. In spatial model, each tweet, that is associated with a location and by using probabilistic approximation algorithm called" particle filter", can estimate the location of event occurred from sensor readings. Particle filtering works better than other comparable methods for estimating the location of events. Finally, we develop a probabilistic model for real time event detection and reporting system using event detection algorithm.

**Keywords**--Social networking, data mining, SVM, tweet analysis.

## I. INTRODUCTION

A probabilistic model is a statistical analysis tool that estimates, on the basis of past (historical) data, the probability of an event occurring again. Different statistical tools are available, some of them are simple, some are complicated, and often very specific for certain purposes. In analytical work, comparison of data, or sets of data has been done to quantify accuracy (bias) and precision. Fortunately, with a few simple statistical tools : the "t-test", the "F-test", and regression analysis , can quantify the accuracy and precision . Clearly, statistics are a tool, not an aim. Simple inspection of data, without statistical treatment, by an experienced and dedicated analyst may be just as useful as statistical figures on the desk of the disinterested. The value of statistics lies with organizing and simplifying data, to permit some objective estimate showing that an analysis is under control or that a change has occurred. Equally important is that the results of these statistical procedures are recorded and can be retrieved.

Our work is on detection of real time events like earthquakes , using online social network ,particularly by use of micro blogging service. Twitter has been examined by some researchers , and notified it as ,is one of the micro blogging service. An important characteristic of Micro blogging service is of its real time nature. So , twitter has been consider for our research , to detect the real time event ,depending on the tweets posted. Tweet is a status update message that posts to our friends and colleagues , are socially connected around the world

For instance using Twitter people comment on various ongoing and real time events. elections are talked about over Twitter. The tweets made by various people can have useful information when mined. Such knowledge can help in making well informed decisions pertaining to advertisements, marketing, promoting sales campaigns and so on. The tweet analysis in the real time also helps to know the important topics on which people are focusing. The process of acquiring business intelligence from the tweets of Twitter became active research area of late.

Many researchers tired to build applications based on Twitter. That way research on Twitter became very important as provides insights to make businesses grow faster. Micro blogging is attributed to tweets concept in Twitter. This kind of blogging enables people to send text updates to online users besides expressing their responses to various real time events. From the twitter tweets and studying them we understood that it is possible to analyze tweets and gain business intelligence. Recently Sakaki *et al.* [1] presented an approach to analyze tweets for real time event detection. Inspired by this research, in this paper, we propose a framework that allows tweet analysis. We also built a probabilistic model that demonstrates the proof of concept. Our application can detect multiple instances of events from analyzing the tweets in spatiotemporal domains. Our empirical results revealed that the proposed application is very useful to detect multiple instances of real time event and notify people so as to make necessary steps in the aftermath of events.

The remainder of this paper is structured as follows. Section II presents review of literature on the closely related works. Section III presents proposed system to detect multiple instances of real time events through tweet analysis. Section IV presents the probabilistic spatiotemporal model that demonstrates how the application works. Section V presents experimental results while section VI concludes the paper besides giving directions for future work.

## II. RELATED WORK

Social media has been active for many years and the users of it are growing exponentially. This section provides review of literature on related works such as real time event detection. Key word based topic search was proposed by Cataldi et al. [2] in order to identify emerging topics with respect to news and keywords. Twitter network features were investigated by some researchers [3]. Later on Haewoon et al. [4] crawled Twitter data and applied Page Rank algorithm on it. Then Huber man et al. [5] analyzed 3 laky Twitter users to discover interactions among friends. Characteristics of Twitter as social media were investigated by some researchers [6]. With respect to the election of Germany Tumasjan et al. [7] analyzed Twitter tweets to predict winners in elections. Sentiment analysis concept was used in [8] for knowing public opinion. Some researchers investigated Twitter to relate it with mobile e-Learning [9]. In [10] some investigation is made to know the relationships between micro blogging and semantic web. Many applications came into existence in order to examine Twitter data and analyze the content for various benefits such as marketing, advertising and so on.

The Twitter and its spatial aspects were studied by Backstop et al. [11] which is close to the study made in [1]. Both the researches [11] and [1] are close the present research in this paper. Another research was carried out on blogging for event detection through discovery of spatiotemporal patterns [12]. Photographs from Flickr were used in [13] to map them with world map. The place and event semantics were explored in [14]. In the field of social media location estimation related studies were found. Estimation of the location of an object under study has many real time utilities. To achieve this different type soft GPS and infrared badges were used [15]. Another technique used for location estimation is the particle filters as explored in [16]. The particle filter approach proved to be efficient.

Recently Sakaki et al. [1] presented an approach to detect real time events by analyzing tweets. Their research focuses on probabilistic spatiotemporal model in order to locate the event place and event recognition. Particle filtering is used in their search for exact location estimation. As there were numerous earthquakes happening in the world, they study was focusing on them. In this paper we did related work but our approach can detect multiple instances of the events at a time. This helps to have better results when it comes to notifying users so as to let them understand about real time events. This is achieved by analyzing tweets. As the tweeting became part of social media and that is the means to generate voluminous data over Internet, the tweet analysis was given importance in our paper too.

## III. SYSTEM ARCHITECTURE

In this paper we proposed a framework for , the detection of multiple instances of events and the location of such events by analyzing tweets posted on Twitter , describes the event has been occurred . Recent study on this kind of research was done in [1]. However it locates single instance of an event . We use the concepts and ideas from

those authors and present similar kind of :framework that detects and locates multiple instances of an event with some modifications. The framework proposed is as shown in Figure 2.

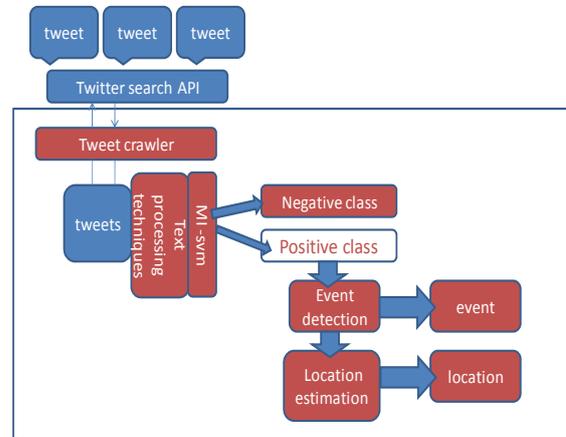


Fig.2

As can be seen in Figure 2, it is evident that the framework has various modules for performing tweet analysis and finally showing multiple instances of events along with location map. The functional requirements of the system are logically divided into the following modules such as Tweet Crawler, Classification, Event Detection and Location Estimation. Tweet Crawler Module is responsible to connect to the social networking site - Twitter and obtain tweets based on the keywords given. To achieve this it makes use of Twitter search API provided by Twitter for public use. Depends on the search word “earthquake”, it is possible to gather the tweets that consisting the search word ”earthquake”. This should be happens by categorizing the tweets into three groups of features i.e. statistical, keyword, word context features. In our application, keyword based feature has been adopted. Depends on this keyword based search, we should consider the tweets that are more relevant to the search word “earthquake”. Next, by adopting the tweets that are most relevant to the search word, those tweets should be processed. In this tweets processing, which are in natural language, we use text processing techniques[17] like stop –word removal and stemming. After processing these tweets , we apply classification algorithm that separates the tweets into positive class or negative class.

In text classification, this F- test has most important to quantify the accuracy on different data sets. More F-measure, will gives most accurate data. So, Classification Module is responsible to take tweets obtained by crawler module and apply classification algorithm of data mining. To make classification of tweets whether they are related to event i.e either positive or not, Multi Instance – Support Vector Machine has been used [13]. Due the success of the Support Vector Machine (SVM) algorithm and the various positive theoretical results behind it, maximum

margin methods have become extremely popular. So, SVM have been proposed for the MIL problem.

SVM is used to find a hyperplane in the input space that separates the training data points with as big a margin as possible. The classifier is defined by the hyperplane normal  $w$  and the offset  $b$ ,  $h = \{w, b\}$ . The margin, is defined as the smallest distance between a positive or negative point and this hyperplane. For example , we assume that the margin is at least 1, and shrink the size of the hyperplane normal  $w$ . Data may actually not be separable, we also include slack variables  $\epsilon_i$  for each point  $x_i$ . The closest points to the hyperplane are called support vectors.

Algorithm for text classification is

```

initialize  $y_i = YI$  for  $i \in I$ 
REPEAT
  compute SVM solution  $w; b$  for data set
  with imputed labels
  compute outputs  $f_i = (w; x_i) + b$  for all  $x_i$ 
  in positive bags
  set  $y_i = \text{sgn}(f_i)$  for every  $i \in I, YI = 1$ 
  FOR (every positive bag  $BI$ )
  IF ( $\sum_{i \in BI} (1 + y_i)/2 == 0$ )
  compute  $I^* = \text{argmax}_{i \in BI} f_i$ 
  set  $y_{i^*} = 1$ 
  END
  END
  WHILE (imputed labels have changed)
  OUTPUT ( $w; b$ )
    
```

After the classification of tweets into positive class, these positive tweets will help us in detection of an event.

*Temporal model* For this we make use of temporal model, which describes the time series of data. Each tweet has its own post time , depending upon the quantity of tweets about the target event, have to calculate the probability density function . for our application , it is an exponential distribution , because of chance of getting more than 10 earthquake events my happens within particular period. So this exponential distribution should be

$$f(t;\lambda) = \lambda e^{-\lambda t} \text{ where } t > 0 \text{ and } \lambda > 0$$

here,  $\lambda$  can be the probability of post a tweet from time  $t$  to  $\Delta t$ . Assumes ,  $n$  sensors, produce positive signals and the probability of all  $n$  sensors returning a false alarm is  $p_f^n$ . Therefore, the probability of event occurrence can be estimated as  $1 - p_f^n$

so, the probability of an event occurrence at time  $t$  is

$$p_{\text{occur}}(t) = 1 - p_f^n$$

We can calculate the probability of event occurrence if we set  $\lambda = 0.34$  and  $p_f = 0.35$ .

*Location Estimation Module*, after detecting events, the location estimation module is responsible for analyzing spatial and temporal information present in the tweets and estimates the locations of the real-time event. To locate the location of the event occurred , “particle filtering “ algorithm has been proposed[7].

The algorithm is presented below.

```

Algorithm particle_filter ( $X_{t-1}, u_t, z_t$ ):
 $\bar{X}_t = X_t = \emptyset$ 
for  $m = 1$  to  $M$  do
  sample  $x_t^{[m]} \sim p(x_t | u_t, x_{t-1}^{[m]})$ 
   $w_t^{[m]} = p(z_t | x_t^{[m]})$ 
 $\bar{X}_t = \bar{X}_t + \langle x_t^{[m]}, w_t^{[m]} \rangle$ 
endfor
for  $m = 1$  to  $M$  do
  draw  $i$  with probability  $\propto w_t^{[i]}$ 
  add  $x_t^{[i]}$  to  $X_t$ 
end for
return  $X_t$ 
    
```

“Social sensors” are used in the process of detecting a target event and the location from the tweets analyzed. In [1] alerts are generated for a single instance of target event like earthquakes. With this system, users are sent alert messages in the form of tweets, about earthquakes in other places once event is detected. And administrator confirms that earthquake and report to all users in this site. With this alert public can take preventive steps to protect their lives and properties. In proposed system, we generate alerts for multiple instance of target events like earthquakes, typhoons, traffic jams. With this system, user sends messages about real time events in their places, once event is detected .It alerts public, so that the public and government can take preventative steps to protect their lives and property.

#### IV. EXPERIMENTAL RESULTS

Experiments are made with the proposed application that demonstrates the concept of multiple instances of real time event detection though tweet analysis. Our experiments are mainly focused on detection of multiple instances of real time event like earthquake.

The **proposed algorithm** is as the following:

1. Given a set of queries  $Q$  for a target event.
2. Put a query  $Q$  using search API every  $s$  seconds and obtain tweets  $T$ .
3. Find the relevant tweet from  $T$ , by using text processing techniques. And then apply the multiple instance learning algorithm for classification to obtain positive or negative class of tweets.
4. If the enough number of positive tweets comes ( $p_{\text{occur}}$  in(1) exceeds 0.99 under the condition: 10 tweets in 10 minutes;  $\lambda = 0.34$ ;  $p_f = 0.35$ ;) then proceed to step 5.
5. For each tweet, we obtain the latitude and the longitude  $l_t$  by
  - 1) Using the associated GPS location,
  - 2) Making a query to Google Map for the registered location for user  $u_t$ .
 Set  $l_t = \text{null}$  if neither functions.
6. Calculate the estimated location of the event from  $l_t$ ;  $t \in T$  using normal particle filtering.
7. Send alert e-mails to registered users.

Put Number of days, the tweets are utilized for experiments The results reveal that the application is able to detect multiple instances of real time events and identify locations besides sending notifications to people concerned.

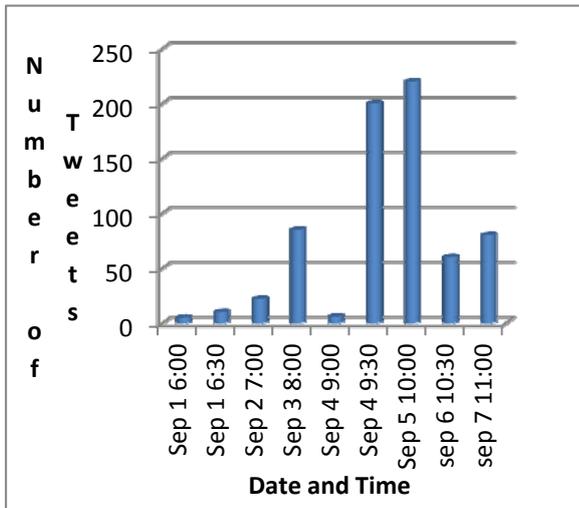


Fig-3

As can be seen in Figure 3, it is evident that, the real time tweets are taken for experiments. Especially observations are made on earthquake events and the statistics are plotted. The horizontal axis represents date and time of tweets while the vertical axis represents the number of tweets that are related to earthquakes.

**V. CONCLUSIONS AND FUTURE WORK**

In this paper we studied Twitter as social medium. The tweets are considered to have insights into real time events such as earthquakes, tsunami and so on. Analyzing such tweets can help in detecting real time events and finding the location where the events occurred. We proposed a framework that helps in real time detection of such events. Unlike prior works we consider detecting multiple instances of events so as to make the application more useful. Location estimation methods are used in order to detect events and the results are presented in the form of Google maps so as to visualize the location of events. We built a prototype application that demonstrates the proof of concept. The empirical results are encouraging. In future we would like to investigate advanced algorithms ,useful for expansion of query in detection of real time event detection and estimate the accurate location of an event.

**REFERENCES**

[1] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development"

[2] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation," Proc. 10th Int'l Workshop Multimedia Data Mining(MDMKDD '10), pp. 1-10, 2010.

[3] A. 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.

[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] 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.

[6] 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.

[7] Miodrag Bolic "Theory and implementation of particle filters "

[8] B. O'Connor, R. Balasubramanian, B.R. Routledge, and N.A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," Proc. Int'l AAAI Conf. Weblogs and Social Media, 2010

[9] M. Ebner and M. Schiefner, "Microblogging - More than Fun?" Proc. IADIS Mobile Learning Conf., pp. 155-159, 2008.

[10] A. Passant, T. Hastrup, U. Bojars, and J. Breslin, "Microblogging: A Semantic Web and Distributed Approach," Proc. Fourth Workshop Scripting for the Semantic Web (SFSW '08), <http://data.semanticweb.org/workshop/scripting/2008/paper/11>, 2008.

[11] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial Variation in Search Engine Queries," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 357-366, 2008.

[12] Q. Mei, C. Liu, H. Su, and C. Zhai, "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 533-542, 2006.

[13] Stuart Andrews, Ioannis Tsochantaridis and Thomas Hofmann "Support Vector Machines for Multiple Instance Learning"

[14] T. Rattenbury, N. Good, and M. Naaman, "Towards Automatic Extraction of Event and Place Semantics from Flickr Tags," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 103-110, 2007

[15] 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.

[16] J. Hightower and G. Borriello, "Particle Filters for Location Estimation in Ubiquitous Computing: A Case Study," Proc. Int'l Conf. Ubiquitous Computing (UbiComp '04), pp. 88-106, 2004

[17] Mita K. Dalal and Mukesh A. Zaveri "Automatic Text Classification: A Technical Review" Machine Learning, International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011

**AUTHOR BIOGRAPHIES**



Dharnasi Rama Devi ,pursuing M.Tech in computer science and engineering branch from Aurora's Technological and Research Institute,uppal, Hyderabad, Andhra Pradesh. Her Areas of interest are data mining and artificial intelligence



Ms. Naga Aswani Pudur working as a Assistant professor in the Department of Computer Science and Engineering in Aurora's Technological and Research Institute with a teaching experience of 1year and worked as a Software Test Engineer .She had received her M.Tech in Computer Science (Software Engineering) from JNTU Hyderabad. Her areas of interest include Cloud Computing, Network Security and her specialization in Cloud Computing.