Event Detection Techniques and Performance Analysis using Twitter Data

[1]Bhakti Kulkarni, [2]Neeta Dimble

[1]Student, [2]Professior
Computer Engineering
Flora Institute of Technology, Pune, India

Abstract : A public safety event is that which needs to be detected early as it is danger for area at which it is occurred. It needs fast response and quick recovery. When event is occurred, its correct information is necessary to be collected quickly to provide the response. Now a day's most of the places are checked with the surveillance cameras. But they need continuous attention and it may not be possible. So the social media users who are present at occurred event can play important role to provide data related to a public safety event. Twitter is one of the best social platforms these days. Twitter can provide real time data of particular public safety event. System investigates the real-time data of events. Various algorithms are proposed to investigate tweets and detect a target event. Keywords in a tweet and their contexts are used to detect event related information. Huge data stored on twitter. To get this data for study purpose there are different techniques. Twitter data will be useful for public safety events. Subsequently, various classification, machine learning techniques and natural language process techniques (NLP) will be applied. In this paper, using twitter data how spatial and temporal distance will be found is discussed. Different techniques for data extractions are discussed. How performance will be measured by different parameters is discussed in this paper.

IndexTerms **- event detection, social media, Twitter data, spatial and temporal distance**

I. INTRODUCTION

Nowadays, number of events occurring publicly around the world is increase tremendously. The large, real-time use of internet makes the social media as a most important information source of a public safety event. With the improvements and developments in communication technology, mobile users, twitter users work as reporter detection and analysis of a public safety event.

Social media users post information of a public safety event .The social media is a most effective platform. So, Social media works as the information resource of the public safety event. At last, internet users provide the online opinion, attention, sentiment of the physical space to the public safety event. Social sensors are the users who generate data of the event.

For the detection of public safety event, data mining will be used. Information is extracted from data to detect the occurred event. Spatial and temporal distance of social sensors is found by using data mining.

Public safety event's information is extracted from the dataset by performing fetching, elaborating, classifying datasets generated by social sensors. Social sensors are the users of social media providing information of the event.

II. RELATED WORK

M. Vijay Kumar [1] target detection of events. Every event consists of two values of space and time. When event is occurred, area of event may be filled with user who are twitting about events, products, location in space and time. Event detection system finds various events such as earthquakes, traffic jams, typhoons about which user tweets on twitter. Sports events, exhibitions, accidents, and political campaigns, large parties are the examples of events that are detected in this system. Natural safety events such as storms, heavy rains, tornadoes are also detected very correctly.

Haisheng Li [2] recognized a spatial and temporal distribution of sensors in the affected area. Here visual analysis of weibo is done. Sensors participate in event and create comments and retweet. Algorithms are also designed to generate retweet relation from downloaded data. Main two modules are data preprocessing and visualization. Visualization is basic information of spatial and temporal distribution.

Neela Avudaiappan [3] detected emergent keywords related to events by monitoring them. Author also summarizes semantic graphs which are detected from documents which are always streaming. Author detects minimum weight set cover of particular event using algorithm for summarization of emergent events. Then keywords are ranked. The system is demonstrated on the live twitter data generated by users related to public safety events.

Zheng Xu [4] recognized the method for finding steps for storytelling of public safety events. By using weibo post related to public emergency events Method is proposed. Method mines the multiple information related to event as well as storytelling of event correctly. Three stages of this method are irrelevant weibo post filtering, mining multi-modal information and storytelling generation.

Takeshi Sakaki [5] investigated the real-time nature of twitter and an event notification system. This system investigates tweets and notifies promptly about event. Semantic analysis is applied to tweets about public safety event. Support vector machine is used to classify trained data using 4 features keywords in a tweet, the number of words, and the context of words.

III. METHODOLOGY

3.1 System to get tweets related to target with accuracy

The proposed model is extended by modeling the data to cope with the spatial temporal analysis of Twits. System approaches to perform analysis using various NLP techniques and special temporal distance finding techniques. For the detection of event, analysis of data is performed. Method of analysis of tweets for detection of target event is presented in Figure 1. Data of tweets is taken as input to the analysis system. These tweets and user data are accessed through Twitter API. Different steps are performed on this twitter data. Classification of data is performed according to different events. These separated tweets are again used for NLTK and ML processing. Both these methods are used for event detection and finding location of event.

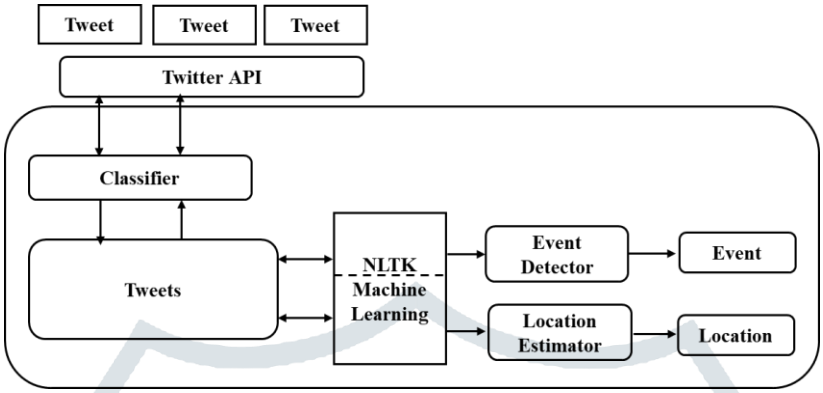

Figure1. System to get tweets related to target with accuracy

Keywords are an important elements for detection of any event. For every event, there can be 3-4 exact keywords which can describe the event. Here without considering sentiment or emotions of user, only keywords are used as detectors. Only semantic information is extract from the keywords. Any user who is twitting about event if contains say area, exact situation, loss of human lives, loss of property, then that tweet is ranked higher. By various semantic analysis methods of natural language toolkit (NLTK) and machine learning, tweets will be ranked according to content. Then emergency event will be detected using noun words with real and clear meaning .Every Twitter user has its own identification information and extra information like location, time of tweet, device used, check-ins etc. For more accurate detection of event, check-ins, location of tweet are extracted. User who is or may get affected by emergency event publishes information of event. He informs other people that " Be alert, prepared. There is dangerous situation". So location of such user is also important. This check-ins will be extracted from HTML page. Location information will be mined with longitude and latitude of user [4].

3.2 System Architecture

In following Figure 2. Shows system architecture. Social Sensors also known as users of social media generate data for event. By using Twitter developer account, Twitter API is generated. Twitter API streams live training datasets. Every time live data is used. This data is taken as input for processing. Here using twitter dataset for the detection of event, status update messages are fetched from dataset. They are converted in the structured format by using various techniques. Classification of structured data and calculation of spatial temporal distance are performed and public safety event is detected.

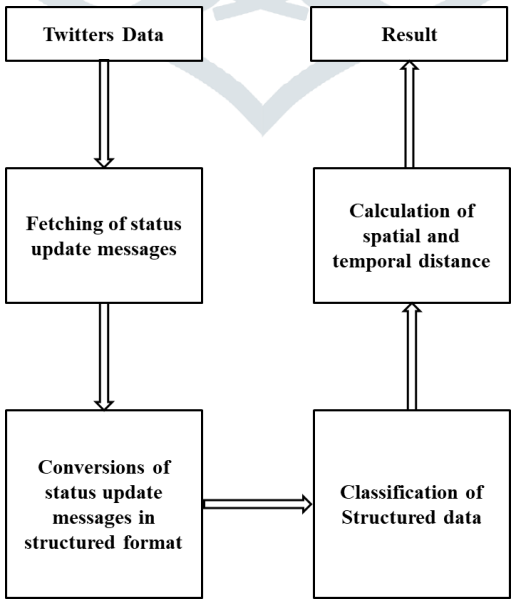

Figure 2. System Architecture

3.3 NLP, Spatial and Temporal mining Techniques

By using Natural language processing (NLP) human language is converted to computer program. Basic NLP tasks such as tokenization and parsing are used for syntactical parsing. Stemming, part-of-speech tagging are for context extraction process.

By using these steps language detection and identification of semantic relationships is done. Analysis of Natural language processing is done using different techniques. Few are discussed here.

3.3.1 Pattern Matching

Interpretation of data input without changing structure and meaning of words is done in pattern matching. Here words interpretations are matched against data inputs. Large number of input patterns are this pattern matching, in this way analysis is done with deep level. Solution to this problem can be hierarchical pattern matching. Here sub phrases are used to perform pattern matching. Another way to reduce the number of patterns is to match with semantic primitives instead of words.

3.3.2 Syntactically driven Parsing

Analysis of syntax of input given is performed in this step. Inputs of data are used for syntactic analysis. Here internal representations are created using different syntax. And then these internal representations are used that can be understood by computers. So basically, it is important parsing step to convert data into computer language.

3.3.3 Semantic Grammars

In case of syntactically driven parsing, resources are wasted while generating parsers which are syntactically correct and nonsensical. More deeper and efficient level of analysis will be done by eliminating the production of meaningless parses. Here grammars are set so that only meaningful parses are generated. There are many similarities between Natural language analysis based on semantic grammar and syntactically driven parsing. Only one difference is that in semantic grammar, the categories that are used defined semantically and syntactically

3.4 Spatial and temporal distance for public safety events

Raw data has different relations like distance, direction, shape, occurrence time, duration. These relations have to be extracted from that raw data. Required inferences are drawn from this raw data.

3.4.1 Spatial distance

The exact physical distance between twitter user and actual event occurred are found using longitude, and latitude of both the user and event area. If the event such as graduation ceremony takes place in same state then considered as low level. If it is takes place in other state then it is processed as high level. When twitter user twits about safety event, distance is checked between user and event area. If distance is far away from user then that person may not give or tweet correctly about event. If distance is near then user will tweet correctly about event [2].

When user tweets about event, twitter will save location of tweet. Then this location will be extracted from twitter data. Other than location word, different information can be used like the check-in information of real location of the user.

3.4.2 Temporal Distance

Temporal distance refers to distance in time. Each tweet has its own post time. Temporal distance is the time to taken from one place to other. It means how long is taken from location where the event happen to the tweet location [2].

According to Richardson and smith(1993) to make the model more effective and efficient the selection criteria for the shares in the period are: Shares with no missing values in the period, Shares with adjusted $R^2 < 0$ or F significant (p-value) >0.05of the first pass regression of the excess returns on the market risk premium are excluded. And Shares are grouped by alphabetic order into group of 30 individual securities (Roll and Ross, 1980).

3.5 Algorithms

Following algorithm shows the spatio-temporal distribution of users participating in the event detection. [3] By using this algorithm, system can find location and time of event occurred.

Algorithm I: The algorithm of spatio temporal distribution of user

Input: Tweets table (W), comments (C), user info (I).

Output: Spatio-temporal distribution of users.

Procedure:

1: Timeline finding. Get and store time of tweet from W, C and get the first and last point of time, break time interval into T − 1 time interval between T time points;

2: Extract coordinates. Extract latitude and longitude for user's current location from I;

3: Statistics of data. For each point of time, calculate the users publishing tweets, made opinion from different locations respectively.

In the following algorithm, every sensor with particular data is compared with every other sensor [6]. If similarity is found in then compared sensor is discarded. If very few or no similarities are found then compared sensor is added to sensor set. For new sensor which is not present in the sensor set, semantics and spatial temporal distances are found. Finally Social Sensor set which contains distinct social sensors with datasets is generated. The algorithm is shown as follows:

Algorithm I: Social Incidents Detection

Input: Incident, I; Detector Set, DS; Dataset, D

Output: Social Detector Set, SDS

```
i:   For every di 1 D
ii:        SDS Empty
iii:       if Equal (di, I. semantic) > b
iv:                SDS (recognize identities from dsi)
v:                 For each dsi 1 DS
vi:                        dsi (find semantic out of dsi)
vii:                       dsi (find spatial out of dsi)
viii                       dsi (find temporal out of dsi)
ix:                        if dsi 1 DS
x:                              upgrade SDS dsi
xi:                        else
xii:                            add SDS dsi
xiii:                      end if
xiv:               end for
xv:        end if
xvi:  end for
xvii: Go back SDS
```

3.6 Data Preprocessing

3.6.1 Tokenization

Tokenization has two types one is word tokenizers and another one is sentence tokenizers. Tokenization performed on corpora and lexicon. Corpora is nothing but body of text and lexicon is words and their means.

3.6.2 Stop words

A stop words are commonly used words which are ignored by search engine. Stop words are like the, a, an, in etc.

3.6.3 Stemming

It is the process of words where words are reduced to their root by removing their unnecessary part.

3.6.4 Part of Speech Tagging

In part of speech there is process of classifying in their part of speech and labeled them accordingly. Parts of speech are divided into mainly 8 parts.

3.6.5 Chunking

Chunking in natural language processing is generally used for users certain language pattern. Chunking is used for to reach higher meaning of particular searching.

IV. RESULTS AND DISCUSSION

For proposed work, tweets related dataset is used. Real-time dataset is used for proposed work. Twitter developer account is used for fetching the real-time tweets. Tweet id, tweet text, latitude and longitude of users are the parameters of the dataset. Nepal earthquake related tweets are used for the measuring the accuracy. For measuring accuracy different classifier are used. Nepal earthquake dataset have total 10378 entries present out of that 50 -50 percent entries are used for training and testing purpose.

Table 1: Performance of different classifier

Classifier	Accuracy in percentages
SVM	79.59
NB	84.7
KNN	90.57
LR	77.26
XgBoost	65.0
RF	56.07
DT	99.56
ANN	98.3

V. ACKNOWLEDGMENT

I would like to thank all the authors for inspiring research oriented studies by their publications. I am thankful to all the authors whose papers are referred for preparing this article.

REFERENCES

[1] M. Vijay Kumar, "Real-Time Event Recognition and Earthquake Reporting System Development by Using Tweet Analysis" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol 3, pp. 459-464, 2018

[2] Haisheng Li, Xunge Liang, Xuan Song, Qiang Cai, "Visual Analysis of Spatio-temporal Distribution and Retweet Relation in Weibo Event" IEEE International Conference on Big Data and Smart Computing, pp.9-16, 2018

[3] Neela Avudaiappan, Alexander Herzog, Sneha Kadam, Yuheng Du, Jason Thatcher, Ilya Safro, "Detecting and Summarizing Emergent Events in Microblogs and Social Media Streams by Dynamic Centralities" IEEE International Conference on Big Data, pp. 1627-1634, 2017

[4] Zheng Xu, Hui Zhang, Yunhuai Liu and Lin Mei, "Crowd sensing based Multi-Modal Storytelling of Urban Emergency Events using Social Media" 9th EAI International Conference on Mobile Multimedia Communications, pp.214-219, 2016

[5] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" 19th International Conference on world wild web, pp. 851-860, 2010

[6] Farzindar Atefeh, Wael Khreich, "A Survey of Techniques for Event Detection in Twitter", International Journal of Computational Intelligence, Vol. 31, pp. 132-164, 2015

[7] Doaa Mohey El-Din Mohamed, Mohamed Hamed Nasr El-din "Performance Analysis for Sentiment Techniques Evaluation Perspectives" International Conference on Advanced Intelligent Systems and Informatics, pp.448-457, 2017

[8] BiswaRanjanSamal, Anil Kumar Behera, Mrutyunjaya Panda, "Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis", IEEE 3rd International Conference on Sensing, Signal Processing and Security, pp. 128-133, 2017