Ain Shams Engineering Journal (2018) 9, 365–378

Ain Shams University

Ain Shams Engineering Journal

www.elsevier.com/locate/asej
www.sciencedirect.com

A review on application of data mining techniques to combat natural disasters

Saptarsi Goswami [a], Sanjay Chakraborty [a,*], Sanhita Ghosh [a], Amlan Chakrabarti [b], Basabi Chakraborty [c]

[a] Institute of Engineering and Management, Kolkata, India
[b] A.K.Choudhury School of Information Technology, Kolkata, India
[c] Faculty of Software and Information Science, Iwate Prefectural University, Japan

Received 17 August 2015; revised 26 November 2015; accepted 16 January 2016
Available online 16 April 2016

KEYWORDS

Natural disaster;
Data mining;
Twitter;
India;
Big Data

Abstract Thousands of human lives are lost every year around the globe, apart from significant damage on property, animal life, etc., due to natural disasters (e.g., earthquake, flood, tsunami, hurricane and other storms, landslides, cloudburst, heat wave, forest fire). In this paper, we focus on reviewing the application of data mining and analytical techniques designed so far for (i) prediction, (ii) detection, and (iii) development of appropriate disaster management strategy based on the collected data from disasters. A detailed description of availability of data from geological observatories (seismological, hydrological), satellites, remote sensing and newer sources like social networking sites as twitter is presented. An extensive and in-depth literature study on current techniques for disaster prediction, detection and management has been done and the results are summarized according to various types of disasters. Finally a framework for building a disaster management database for India hosted on open source Big Data platform like Hadoop in a phased manner has been proposed. The study has special focus on India which ranks among top five counties in terms of absolute number of the loss of human life.

1. Introduction

Natural disasters affect human and animal lives and properties all around the globe. In many cases the reasons are not in our control. As noted in [1], for the three decades namely 1970–80 (rank 2nd), 1980–90 (rank 4th), 1990–00 (rank 2nd), India ranks in first 5 countries in terms of absolute number of the loss of human life. It is not only the immediate effect as observed in [2], and exposure to a natural disaster in the past months increases the likelihood of acute illnesses such as diarrhea, fever, and acute respiratory illness in children under

* Corresponding author. Mobile: +91 9038205310.
E-mail addresses: saptarsi.goswami@iemcal.com (S. Goswami), sanjay.chakraborty@iemcal.com (S. Chakraborty), sanhita.ghosh@iemcal.com (S. Ghosh), amlanc@ieee.org (A. Chakrabarti), basabi@iwate-pu.ac.jp (B. Chakraborty).
Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.asej.2016.01.012
2090-4479 © 2016 Ain Shams University. Production and hosting by Elsevier B.V.

5 year by 9–18%. The socioeconomic status of the households has a direct bearing on the magnitude and nature of these effects. The disasters have pronounced effects on business houses as well. As stated in [3] 40% of the companies, which were closed for consecutive 3 days, failed or closed down within a period of 36 months. The disasters are not infrequent as well. Only for earthquake [4], there are as many as 20 earthquakes every year which has a Richter scale reading greater than 7.0. The effects of the disasters are much more pronounced in developing countries like India.

Meteorologist, Geologists, Environmental Scientists, Computer Scientists and Scientists from various other disciplines have put a lot of concerted efforts to predict the time, place and severity of the disasters. Apart from advanced weather forecasting models, data mining models also have been used for the same purpose. Another line of research, has concentrated on disaster management, appropriate flow of information, channelizing the relief work and analysis of needs or concerns of the victims. The sources of the underlying data for such tasks have often been social media and other internet media. Diverse data are also collected on regular basis by satellites, wireless and remote sensors, national meteorological and geological departments, NGOs, various other international, government and private bodies, before, during and after the disaster. The data thus collected qualify to be called 'Big Data' because of the volume, variety and the velocity in which the data are generated.

A brief technical description of some of the major natural disasters is as follows:

- *Earthquake:* A sudden movement of the earth's crust, causing destruction due to violent activities caused due to volcanic action underneath the surface of the earth. 55% of India's landmass are in seismic zone III–V.
- *Landslide:* A sudden collapse of the earth or mass of rock from mountains or cliff due to vibration on the earth's surface. In India the northern sub-Himalayan region and Western Ghats are prone to landslides.
- *Cloudburst:* It is an extreme form of unpredicted rainfall in the form of thunder storm, hail storm and heavy precipitation which is short lived. Unseasonal heavy rainfalls are common in India. A devastating effect of it was the flash flood in North India in 2013 that killed thousands of pilgrims and animals.
- *Storm:* A bad weather in the form of rain or snow caused by strong winds or air currents formed due to unexpected changes in air pressure on the earth's surface. Cyclones are common in various parts of India, especially the coastal regions that leave long lasting and expensive damages to human lives and properties.
- *Flood:* An overflow of huge water masses beyond normal limits over dry land. Every year, millions of human lives, cattle and agricultural crops are destroyed in India due to lack of planning and improper weather forecasting.
- *Tsunami:* High sea waves that are large volumes of displaced water, caused due to an earthquake, volcanic eruption or any other underwater explosions. The 2004 Tsunami that hit parts of the southeastern coast of India had devastating effects on the mainland and Andaman and Nicobar Islands.
- *Volcanic eruption:* It is a sudden, violent discharge of steam, gases, ashes, molten rocks or lava from the surface of the earth that are ejected to heights and spread for several miles. Underwater volcanoes on the islands surrounding the landmass of India are common. However, they have not imposed significant damages to the mainland till now.

The unique contributions of the paper are as follows:

- A comprehensive summary of different data mining techniques applied to various tasks pertaining to the natural disasters.
- A detailed account of various types and sources of data for each category of task and disaster.
- A brief account of disaster management 'status-quo' from Indian context.
- A brief review of suitability of 'Twitter' as a data source.
- A presentation of proposed architecture to streamline disaster management.

The organization of the paper is as follows: In Section 2, natural disasters have been discussed with focus on India; a brief description of the existing disaster management structure is also outlined. In Section 3, the broad categorizations of the tasks that can be achieved with respect to natural disaster are presented in detail. In Section 4, granular levels of tasks are enlisted with respect to the major type of tasks discussed in Section 3. Details of the tasks, data used in the task, data mining methods used, country or region have also been discussed. In Section 5, a structured view of different types of required data and their corresponding sources has been discussed. A short review of twitter and other Internet resources as data source has been discussed, along with their application for natural disaster in Section 6. In Section 7, a process flow and architecture of a disaster management system have been proposed. Section 8 contains, conclusion with the direction of future work.

2. Natural disaster from an Indian context

India is vulnerable to various natural disasters due to its unique geo-climatic condition as a result of its geographical location. This subcontinent is surrounded by water bodies on three sides and the Himalayas on the North. The country has been hit approximately by 8 natural calamities per year and there has been about 5 times increase in frequency of natural disasters in the past three decades. The calamities that affect the country can be categorized as follows: 57% landmass is prone to earthquakes, 12% floods (about 40 million hectares of land is vulnerable to floods) and 8% are prone to cyclones. Table 1 records such disasters for last 15 years.

Asia, tops in terms of number of disaster events among the continents. Close to 60% of the disasters in Asia are originated in South Asia and 40% are originated in India. In the below figure (Fig. 1), the above statistics are displayed.

India is a victim of natural disasters every year and the loss of lives and properties adds up to millions of rupees which this country cannot afford to lose. There are certain reasons for such poor disaster management procedures followed in this country.

Table 1 Natural disasters in India in last 15 years.

Disaster Type	Year	Origin (India)	Tolls
Earthquake	2001	Gujrat	20,000
	1999	Chamoli	150
Cyclones	2012	Tamil Nadu	20
	2011	Tamil Nadu	41
	2010	Andhra Pradesh	32
	2009	West Bengal	100
	1999	Orissa	15,000
Tsunami	2004	Indian Ocean	230,000
Floods	2007	Bihar	41
	2005	Mumbai	5,000
Cloud Burst	2014	Jammu & Kashmir	4,500
	2013	Uttarakhand	5,700
Landslides	2014	Manlin, Pune	28
	1998	Malpa, Manasarovar Yatra	380

- Inadequate early warning system.
- Poor preparation before the disaster occurs.
- Inadequate and slow relief operation.
- Lack of proper administration.
- Slow process of rehabilitation and reconstruction.
- Poor management of finances for relief work.
- Lack of effective help to victims.

The apex body that handles disaster management in India is the National Disaster Management Authority (NDMA) whose Chairman is the Prime Minister himself. Similar authorities are also set up at state and district levels which are respectively headed by the Chief Ministers and Collectors or Zilla Parishad Chairperson. The Natural Urban Renewal Mission has been set up in 70 cities due to the recent unprecedented weather conditions in major metros and megacities. Need of research to predict, prevent and reduce means of losses from a disaster is far from over. Over 100,000 Rural Knowledge Centers (or IT Kiosks) have to be established for meeting the need for spatial E-Governance and therefore offering informed decisions in disaster prone areas to improve the response time and type of relief aid offered to the victims on time. National disaster management structure is depicted in Fig. 2.

3. Broad category of tasks with respect to natural disasters

In this section, broad categories of tasks that can be solved using different types of data have been discussed. We can

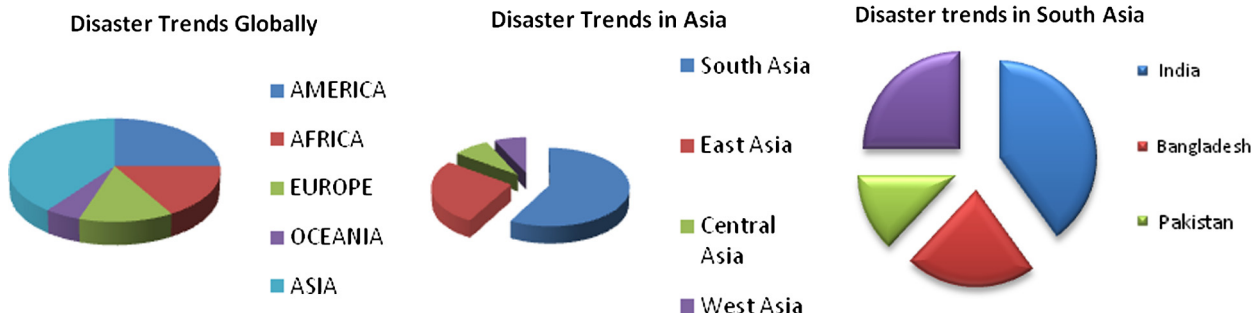

Figure 1 Disaster trends across the globe.

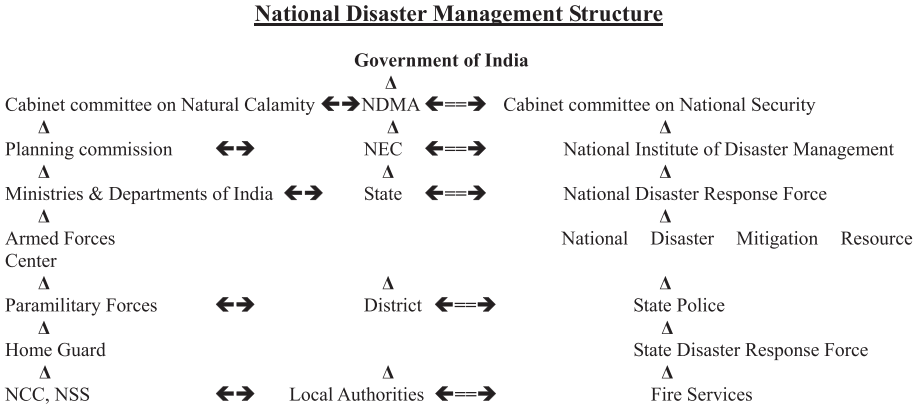

Figure 2 Disaster management structure India.

classify the objectives of the tasks, in the following three major categories:

- *Prediction:* These sets of tasks involve prediction of the natural disaster, disaster prone area and different attributes of a natural disaster that can occur. Basically these tasks involve prediction or forecasting of time, place and magnitude of the disaster.
- *Detection:* These sets of tasks involve detection of the natural disaster promptly after it has occurred. Literature studies indicate that the social sensors in terms of tweets and other social media websites report a natural disaster much faster than the observatories.
- *Disaster management strategies:* These methods deal with identification of different entities that are taking part in combating a disaster so that communication is enhanced, appropriate concern of the affected people is identified and distribution of relief items is optimized.

Another branch of study deals with carrying out the psychological and behavioral changes over affected regions after the disaster.

In many cases, the classification stated above is overlapping. As an example, it can be surely argued that, detecting the natural disaster helps in disaster management strategies. Even the psychological studies can give lot of insight to disaster management strategies. So the classification is based on the direct objective of the task involved. There are some rare cases which fall in a borderline.

Prediction: There is no doubt that this would be the most 'ideal' problem to solve. But very often, this is not a problem that can be solved with available data and techniques. However, it is possible to predict, the areas which are susceptible to a particular type of disaster, let us say, landslide or flood. The prediction techniques have been seen to be of more use for predicting various characteristics of a natural disaster, which has occurred. As an example, the techniques can be used to predict the magnitude of an earthquake, track and intensity of a cyclone, etc. Analysis of various spatial and temporal data is often needed for such tasks. Though handful, another branch of research has focused on using unusual animal behavior to predict a natural disaster.

Detection: Often the meteorological observatories detect the natural disaster, but the news of the detection takes a long time to be communicated to proper authorities with the exact location of the detection.

Disaster management strategies: These sets of tasks are involved in forming appropriate disaster management strategies. An example of such tasks is identifying critical entities for disaster management, identifying proper communication study, and identifying the needs of the disaster affected area. Social media data are very important in these types of tasks.

The aim of disaster management should be the following:

- Minimize casualties.
- Rescue victims on time.
- Offer first aid instantly.
- Evacuate people and animals to safe places.
- Reconstruct the damages immediately.

4. Data, models, tasks

In this section, a summary has been enclosed highlighting the granular level tasks corresponding to the major type of tasks such as prediction, detection and disaster management. An account of model/techniques has been given along with the data used and the country of the disaster/research.

The findings are summarized in Tables 2a–2h. The tables are divided as per the natural disasters; separate tables are presented for earthquake (Table 2a), cloudburst (Table 2b), flood (Table 2c), landslide (Table 2d), volcanic eruption (Table 2e) storm (Table 2f) and tsunami (Table 2g). Table 2h represents generic efforts without focus on a specific type of disaster, which have been covered.

In the tables referred above, different research directions in combating natural disaster have been discussed. This is a multidisciplinary activity needing experts from environmental science, geology, meteorology, social science, computer science, etc. The above list is not exhaustive, but an effort has been made to cover last 10–12 years data in this section. Here are few of our observations:

- Twitter as a source has become important for real time detection and understanding of the need and concern of the affected people. Ten out of the forty papers we reviewed above use twitter as the data source. Interestingly, we did not find any referential work, where twitter has been used in an Indian context.
- It is also observed, though India has a much higher loss in terms of human life and property, and adequate research as in countries such as USA has not been done in India (Fig. 3).
- Level of activity, in the research communities around each disaster type can be roughly estimated by Fig. 4. We have used the number of results from Google Scholar, using the disaster specific keyword. We have restricted the search results using the 'since 2014' filter. The above searches were not limited to the application of data mining.
- The main tasks, where research activities have been going on are prediction and disaster management respectively. As expected, the prediction tasks for each disaster are varied. Apart from the data mining tasks, quite a few papers [4,17,45] focused on building a data warehouse and OLAP structures for better information decimation and consumption. In Table 3, a summary of the prediction tasks that are being researched for each disaster type has been presented.
- We find neural network, SVM, and decision tree have been used extensively as the data mining models. As many of the data are actually time series, time domain techniques as well as frequency domain techniques such as wavelet transformation have been used. Evolutionary techniques such as Genetic Algorithm and Particle Swarm Algorithm have also been used. For newer sources such as twitters, blogs, server logs, and News text processing techniques (topic clustering, LDA) have been applied. Word cloud visualization has been applied on the techniques and it is shown in Fig. 5. Though most of the data obtained are high dimensional in nature, apart from LDA or wavelet transform, we do not see uses of feature selection or dimensionality reduction in the approaches.

Table 2a Earthquake data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Predict magnitude of earthquake [5]	Particle swarm optimization	Seismological data	China
	Focus on abnormal animal behavior, rather than the geophysical indicators. The study has been done mainly in Japan, China and USA. Animals are much more sensitive to the change in electric field precursor to the earthquake [6]	NA	NA	Japan, China, USA
	Predict magnitude of earthquake [7]	Neural network	Seismological data	USA
	Building a data warehouse for earthquakes, for uniformity of data and structure of data for a uniform interchange and better decision making [8]	Ontology, star schema, data warehouse	Seismological data	All over the world
	Predict earthquake based on time series data [9]	Nonlinear time series and fuzzy rules	Seismological data	All over the world
	Predict earthquake from historical data and also propose a grid system for distributed processing and better information interchange [10]	Feature generation and clustering	Seismological & GIS data	USA
Detection	Discover major earthquakes faster than seismological observatories [11]	Text mining	Twitter	USA
	The affected area citizens visit web pages of the Swiss Seismological Service, by doing an IP tracing and volume analysis, the affected regions can be tracked easily [12]	Regular log mining techniques	Web server logs	Switzerland
	To detect earthquake from social sensors i.e. twitter, do a spatial, temporal analysis and send notification much faster than that of the Japan Meteorological Agency (JMA) [13]	Temporal Analysis, Kalman filter	Twitter	Japan
Disaster management	A temporal analysis of peoples need after the earthquake from blogs and social media. This can make the relief operation more effective [14]	Text Mining, latent semantic analysis (LSA), time series	Blogs & social media data	Japan
Behavioral and social analysis	To study general peoples reaction after a natural disaster like an earthquake and how long they take to subside to normal level [15]	Time series, text processing	Twitter	Japan

Table 2b Cloudburst data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Observe different parameters of climate from the earth science data, to find out whether there was enough indication of the Uttarakhand disaster [16]	Anomaly detection, time series	Earth science data	India
	To leverage OLAP structure to store metrological data and analyze them to identify cloudbursts [17]	OLAP cubes, K means clustering	Meteorological data	India
	Real-time newscast and prediction of rainfall in case of extreme weather like cloud burst from Doppler weather radar data [18]	Mesoscale model	Doppler Weather Radar data (DWR)	India

5. Data sources & types

The objective of this section is to give researchers and practitioners a high level overview of the type of data that are useful for analysis and prediction of a natural disaster. Much of the data will qualify to be called 'Big Data', because of all or some of the dimensions of volume, variety and velocity as listed below:

1. Volume (GIS data, meteorological data, social media data).
2. Variety (text, time series, spatial data, GIS images).
3. Velocity (because of the rate in which data are generated as well as because of the speed in which a decision needs to be taken).

We try to answer the following questions in this section:

- What are the different types of data that are useful for each type of disaster?
- What are the sources and format of such data, at national and international level?
- How the data can be accessed? Whether it is freely available or not?

In Table 4, we have listed data types for different disaster types, their corresponding data types, the format of the data and the various agencies that collect or capture the data. The abbreviations of the agencies are used in Table 4, and the details of the agencies with a URL are available in Appendix A.

Table 2c Flood data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Build a model and select appropriate parameters to assess the damage from flood [19]	Decision tree	Hydrological data, remote sensing data, GIS	Germany
	To build a model to find susceptible flood regions based on spatial data [20]	Logistic regression and frequency ratio model	Meteorological data (digital elevation model), river, rainfall data, etc.	Malaysia
	To build a model to predict monsoon flood (1 day ahead). The built system gave better results than existing auto regressive models [21]	Wavelet transform, genetic algorithm, artificial neural net	Hydrological time series data	India
	Build a system for flood forecast for medium- to large-scale African river basins (before 2 weeks) [22]	Probabilistic model & ensemble	Hydrological data	Africa
	Build a flood routing model based on past data [23]	Muskingum flood routing model, Cuckoo Search (for parameter values and calibration)	Hydrological and hydraulic data	All world
Disaster management	A study of tweets during various floods was done to identify key players. The study shows the effect of local authority involvement in successfully tackling a disaster [24]	Text mining methods	Twitter	Australia

Table 2d Landslide data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Build a classifier to identify landscapes to landslide susceptible areas based on soil properties, geomorphological, and groundwater conditions, etc. [25]	Discrete rough set and C4.5 decision tree	Remote sensing data and GIS	Taiwan
	To build a classification model to predict land slide. The various factors considered are rainfall, land use, soil type, slope, etc. [26]	SVM, Naïve Bayes	GIS (rainfall, land use, soil type, slope and its)	India
	A generic note on usefulness of data mining/machine learning models in predicting place and time of a land slide [27]	NA	NA	All over the world
	Build a prediction model based on an inexpensive wireless sensors placed on susceptible regions [28]	Distributes statistical prediction method	Wireless sensor data	India
	To build a model to identify areas of shallow landslide [29]	Spatial distribution	Geomorphologic information and hydrological records	Taiwan
	Predicting landslide based on past data [30]	Back propagation neural network, genetic algorithm, simulated annealing		China

Table 2e Volcanic eruption data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Analysis of multivariate time series data to understand the state of the volcano and potential hazard assessment [31]	Multivariate time series clustering	Geophysical data through monitoring network	Italy
	To monitor and predict trajectories of volcanic ash cloud, to minimize air crash [32]	Not mentioned	Plume height, mass eruption rate, eruption duration, ash distribution with altitude, and grain-size distribution	USA

We would not go into details of the data acquisition methodology, and several literatures are available on the same. However we thought we should briefly mention 'flash flood' technology because of the recentness of the method.

Flash Floods with advanced WSN technology: Nowadays, with the advancement of several technologies, flash flood is going to be introduced with high-tech WSN activity. According to this concept, there is a Wireless sensor device associated with a flash flood technology which is emerged into a certain level of water. If the water level reached to a certain level of threshold value, then the WS device sends a danger signal to receiver station through broadcasting. This technique is planned to be introduced in some areas of Delhi city in India [46] (Fig. 6).

Table 2f Storm data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Prediction	Take the data of current storm and compare with the historical & synthetic storms using storm similarity index (SSI) from the databases to understand the effect. Visualization of the storm path is done on Google Earth. The study was done on two previous storms Katrina and Camille [33]	Data mining techniques	National Hurricane Center (NHC)	USA
	To predict cyclone track data for coming 24 h, based on past 12-h locations at six hourly intervals besides the present position about the latitude and longitude [34]	Artificial Neural Network (ANN)	32 Years tropical cyclone data on Indian ocean from joint typhoon warning center (JTWC), USA	India
	Detect storm surge using no linear model from data collected at coastal station [35]	Time series and chaos theory	Water level, surge, atmospheric pressure and wind speed/direction data from seven coastal stations along the Dutch coast are monitored and provided by the North Sea Directorate	Netherland
Behavioral	Study effect of the hurricane 'Hugo' on life events birth, death, divorce, etc. [36]	Statistical analysis	Life event data from all the counties of South Carolina	USA
Disaster management	Identify the concerns of people, stay duration of the 'concerns', conduct analysis by gender [37]	Sentiment analysis, normal text processing techniques	Twitter	USA
	Analysis of people's sentiment, after Hurricane Sandy and also to gather and decimate important information through social media [38]	Text processing techniques	Twitter	USA
	Analysis of public behaviors during and after a disaster through visualization and spatial temporal analysis [39]	Spatial, temporal techniques, visualization	Twitter	USA

Table 2g Tsunami data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Disaster management	Viability of use of twitters by government agencies, to inform public about natural disaster. It was compared against traditional sources and proved its value as a complementary source [40,41]	Text processing techniques	Twitter	Indonesia, USA
Prediction	Build an early warning system for Tsunami [42]	Flood filling algorithm, classification algorithms	Bathymetry data, Seismic data, sea wave conditions, web service and API to collect the data	Indonesia

Numerical weather prediction (NWP) models, which apply advanced mathematical modeling, have been used for short term forecasting for long time. These models are employed to solve a closed set of atmospheric equations. Most of the meteorological departments have adapted using this model. However, the actual events cannot be predicted from the NWP models directly. Some other statistical techniques are required for prediction. In the paper [47], the authors list down applicability of data mining models in many weather prediction tasks.

6. Twitter and social media as sources

Twitter as a data source has gained lot of prominence in recent years. It is ranked as one of the top 10 popular websites, having 400 million registered users and over 500 million tweets generated everyday [48]. Additionally, information about disasters can be extracted from news channels and blogs through APIs, RSS feed or web scraping. Sentiment analysis [49,50], stock market [51], public health [52], general public mood and finding political alignments [53,54] are some of the areas where twitter data have been used.

Some of the advantages of tweets are as follows:

- Although it is unstructured, it has some structure by the limitation of 140 characters.
- It can use hashtags, which give semantic annotations of the tweets.
- The tweets have geocodes, which can help us spatially map the sources of the tweets.

Table 2h General data, model and task summary.

Task	Detailed objective	Model techniques used	Data source, type	Country
Disaster management	Build a tool to extract important information from tweets for relief workers [43]	Text processing techniques	Twitter	USA
	Build a system, for disaster discovery and humanitarian relief based on tweets. The system consists of a stream reader, a data storage and a visualization module. [44]	SVM, LDA, topic clustering	Twitter	USA
Prediction	Build a geo hazard database for early prediction system, by using the Google news service. Geo tagging is done for geo referencing. The purpose is using this database extensively for disaster management [45]	Text processing	Google news service, RSS feed	Italy

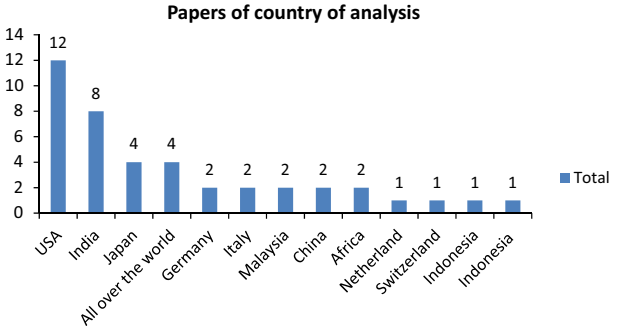

Figure 3 #Papers by country of analysis.

Table 3 Prediction tasks for each disaster type.

Disaster type	Prediction task
Earthquake	Predicting the time, place, magnitude of the earthquake
Cloudburst	Predicting cloudburst, predicting amount of rainfall
Storm	Predict the track and wind speed of the storm
Flood	Identify flood susceptible areas, predict flood, build flood routing model, build a model to assess damage to property etc. due to flood
Landslide	Predicting landslide, predicting landslide susceptible areas
Volcanic eruption	Predict eruptions; predict the trajectory of the ash cloud
Tsunami	Build early warning system with tsunami

Following are the fields that are available from twitter:

- *archive source:* API source of the tweet (twitter-search or twitter-stream),
- *text:* contents of the tweet itself, in 140 characters or less,
- *to_user_id:* numerical ID of the tweet recipient (for @replies) (*not always set, even for tweets containing @replies*),
- *from_user:* screen name of the tweet sender,
- *id:* numerical ID of the tweet itself,
- *from_user_id:* numerical ID of the tweet sender,
- *iso_language_code:* code (e.g. en, de, fr, ...) of the sender's default language (*not necessarily matching the language of the tweet itself*),
- *source:* name or URL of the tool used for tweeting (*e.g.*, tweet deck, ...),
- *profile_image_url:* URL of the tweet sender's profile picture,
- *geo_type:* form in which the sender's geographical coordinates are provided,
- *geo_coordinates_0:* first element of the geographical coordinates,
- *geo_coordinates_1:* second element of the geographical coordinates,
- *created_at:* tweet timestamp in human–readable format (*set by the tweeting client—inconsistent formatting*),
- *time:* tweet timestamp as a numerical unix timestamp.

In many countries, twitter has been used to effectively manage disasters; however, in Indian context we have not seen a lot of referential work. Twitter has been effectively deployed to:

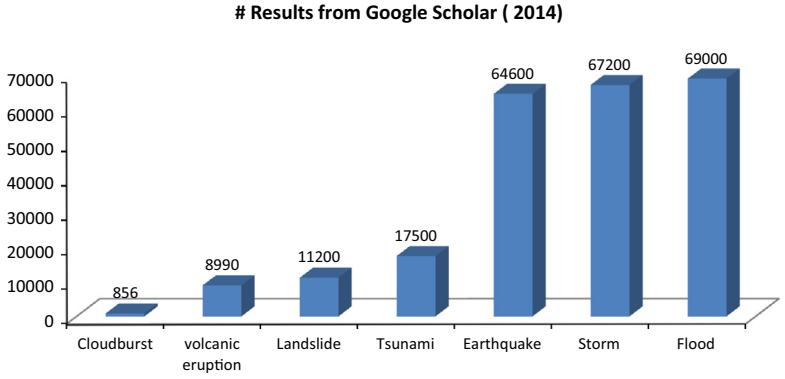

Figure 4 #Results from Google Scholar.

Figure 5 Word cloud of data mining techniques on disaster management.

Table 4 Data types and sources of data related to disasters.

Disaster types	Data category	Data format	Types of data	Agencies/bodies	Availability (paid/free)
Earthquakes, tsunami, etc.	Seismological data	Magnitude of Richter scale & microsoft excel sheet	Small scale (regional) seismic macro zonation at scales 1:5,000,000 to 1:50,000, and large scale (local) seismic micro zonation at scales of 1:50–25,000 to 1:10,000	ISC, NEIC, IMD, IRIS, NGDC, NWS, USGS, EM-DAT, JMA, NDMA	Paid
Tsunami, flood, drought, etc.	hydrological data	Google earth image data	Water level, pressure & density of water, etc.	INCOIS, NOAA, EM-DAT, JMA, NDMA	Paid
Landslide	Geological data	Different devices have different storage formats	Rainfall, moisture, pore pressure, tilt, vibrations, etc.	USGS, NOAA, NODC	Paid
All types of disasters	Remote sensing data	Mainly image file formats	Spatial, temporal, and thematic data (satellite data)	CRSSP, NASA, ISRO, NRSC	Open source
All types of disasters	Google Information System (GIS) & Google mapping data	Stored into microsoft excel spreadsheets or text file	Spatial, temporal, and thematic data (floods and pre-flood SAR images) are collected	NOAA, NIDM, NASA, ISRO	Non-paid (sometimes copyright assertions)
Flood, Tsunami, etc.	Wireless Sensor Network (WSN) data	Transmitting and receiving data through wireless transmitter and receiver	Form of analog warning signal. [needed an analog to digital converter]	GSI.	Mainly from Geological Survey of India (GSI) and paid
Volcanic eruption	Geological data & seismological data	Magnitude of Richter scale & MySQL and xml format (e.g. WOVOml)	Spatial, temporal, and thematic data (e.g. angle, EDM, GPS, inSAR image, gravity, magnetic field, etc.- WOVOdat1.1 document)	WOVOdat, JMA, DATAGOV, NGDC, USGS.	Open source – freely available

- Detect disasters [11,13] faster than observatories.
- Identify key entities in disaster management and relief organization [41,43,44,24,40].
- Temporally study needs and concerns after the disaster [38,37,55].

One of the limitations of using twitter data, is that the 'free access' (streaming API) only provides 1% of sample data, and on the other hand, the alternative way (fire hose) which provides full access is prohibitively expensive. In [48], the authors conducted a study, between both of these ways of extraction. The results obtained therein, though there is some agreement between them, to get a truer picture and the coverage of sampling in terms of parameters of steaming API, need to be varied.

7. Proposed system

We intend to build a database for natural disasters happening in India. In Phase 1 of the project we want to concentrate on sources such as (1) twitter, (2) news, and (3) other, social media and internet sources.

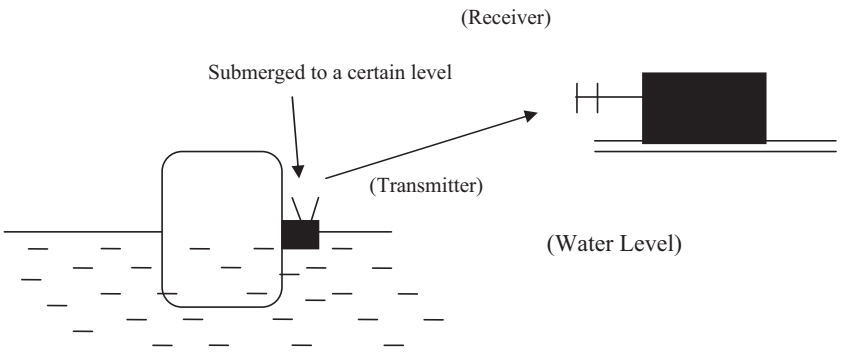

Figure 6 Flash floods with advanced WSN technology.

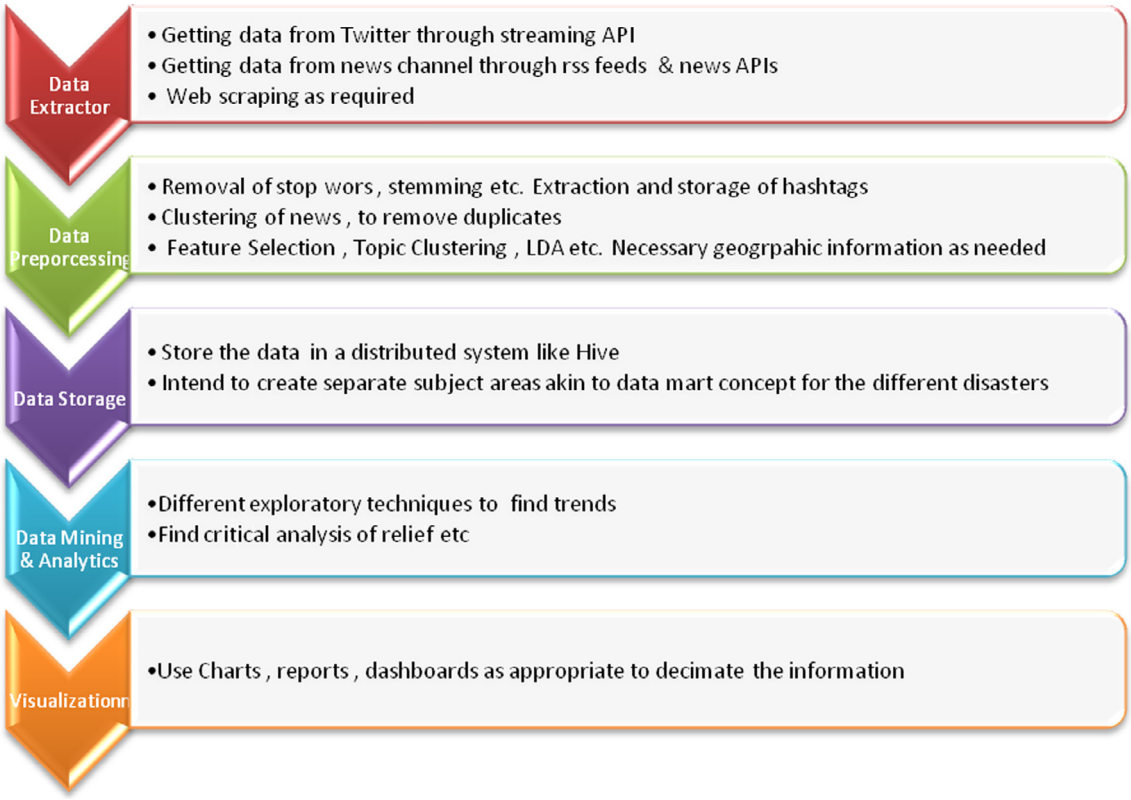

Figure 7 Process flow of proposed system.

Like any standard systems, we will need the following components as shown in Fig. 7.

Tweet tracker [44] and tweedr [43] are couple of systems built in USA to streamline relief and disaster response. In [45], authors have proposed a geo hazards inventory store, with focus on geo tagging and entity resolution from news service. They have built a detailed database of geographical feature (listing of mountains, rivers, etc.) for Italy.

Our proposal is different from the above approaches, in the following ways:

- We are targeting much wider sources, not only twitter.
- The focus would be, on understanding peoples need during the disaster and evaluate the social impact and changes due to natural hazards.

In extracting information we plan to implement methods as described in [56].

- In Phase 2, of our proposed system, we intend to use other sources of data apart from Internet based sources.

In Fig. 8, a Hadoop based open source system has been proposed.

Data sources: We are focusing on twitter and RSS feeds of news at this point. The data extraction scripts will be written in flume, scoop or R [57] as applicable.

Data storage: The data storage is envisaged to be on Hadoop [45]. After data processing using 'pig' [58] the data can reside in HDFS (Hadoop distributed file system) or hive [59], a NOSQL based database.

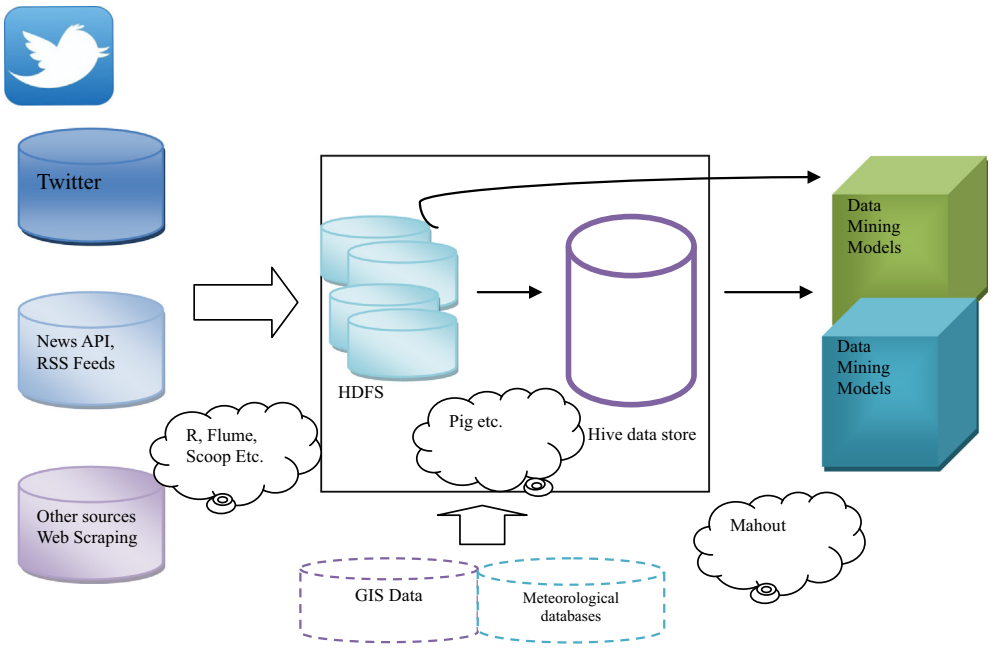

Figure 8 High level architecture of the proposed system.

Model building: The plan is to build our algorithms based on Mahout [60] which can leverage the HDFS and has extensive text processing capabilities and can be extended using Java.

Some portions of the proposed system are shown with dashed lines as we plan to integrate those sources in next phases.

8. Conclusion

Natural disasters in the form of earthquake, floods, landslides, and storms that claim numerous lives, cause significant damage to the properties. The effects have been much more severe in a developing country such as India compared to developed countries. There have been many efforts to predict the disasters based on various sources of data. In our literature survey, we explore the multidisciplinary nature of the task, where data mining models are being applied on various types of data, requiring deep subject matter expertise. Recently social media and internet have also emerged as an important source of information. These sources may not be used in prediction of the disasters, but they have contributed significantly to early detection and adoption for appropriate disaster response. We observe in our study that, there have not been enough works done in this area to tap the potential of these sources especially in the context of India. We propose to build a data store for natural disasters from these sources in Phase 1. In Phase II, we intend to integrate it with other sources of information.

Appendix A

Data agency	URL
ISC – International Seismological Center	http://www.isc.ac.uk/standards/datacollection/
NEIC – National Earthquake Information Center	http://earthquake.usgs.gov/regional/neic/
IMD – India Meteorological Department	http://www.imd.gov.in/
IRIS – Global Seismographic Network	http://www.iris.edu/
USGS – United State Geological Survey	http://www.usgs.gov/
SMA – Social Media Analytic	http://www.datalabs.com.au/
NODC – National Oceanographic Data Center	http://www.nodc.noaa.gov/
CRSSP – Commercial Remote Sensing Space Policy	http://crssp.usgs.gov/
NRSC – National Remote Sensing Center	http://www.nrsc.gov.in
EM-DAT (CRED)	http://www.emdat.be/additional-disaster-data-resources
WOVOdat – A database of Volcanic unrest	http://www.wovodat.org/
JMA – Japan Meteorological Agency	http://www.jma.go.jp/jma/en/Activities/earthquake.html
DATAGOV	https://catalog.data.gov/dataset/global
NDMA – National Disaster Management Authority	http://www.ndma.gov.in/

References

[1] Brooks Nick, Adger W Neil. Country level risk measures of climate-related natural disasters and implications for adaptation to climate change. Clim Res. Tyndall Centre for Climate Change; 2013. p. 1–30.

[2] Datar Ashlesha, Liu Jenny, Linnemayr Sebastian, Stecher Chad. The impact of natural disasters on child health and investments in rural India. Soc Sci Med 2013;76:83–91.

[3] Zheng Li, Shen Chao, Tang Liang, Zeng Chunqiu, Li Tao, Luis Steven, Chen S-C. Data mining meets the needs of disaster information management. IEEE Trans Human–Mach Syst 2013;43(5):451–64.

[4] Nimmagadda Shastri L, Dreher Heinz. Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along Eurasian–Australian continental plates. In: 5th IEEE international conference on industrial informatics; 2007. p. 597–602.

[5] Zhang Xiao Yu, Li Xiang, Lin Xiao. The data mining technology of particle swarm optimization algorithm in earthquake prediction. Adv Mater Res 2014;989:1570.

[6] Bhargava Neeti, Katiyar VK, Sharma ML, Pradhan P. Earthquake prediction through animal behavior: a review. Ind J Biomech 2009:159–65.

[7] Adeli Hojjat, Panakkat Ashif. A probabilistic neural network for earthquake magnitude prediction. Neural Netw 2009;22(7):1018–24.

[8] Nimmagadda Shastri L, Dreher Heinz. Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along Eurasian–Australian continental plates. In: 5th IEEE international conference on industrial informatics, vol. 1; 2007. p. 597–602.

[9] Aydin Ilhan, Karakose Mehmet, Akin Erhan. The prediction algorithm based on fuzzy logic using time series data mining method. World Acad Sci Eng Technol 2009;51(27):91–8.

[10] Yuen Dave A, Kadlec Benjamin J, Bollig Evan F, Dzwinel Witold, Garbow Zachary A, da Silva Cesar RS. Clustering and visualization of earthquake data in a grid environment. Vis Geosci 2005(10):1–12.

[11] Liu SB, Bouchard B, Bowden DC, Guy M, Earle P. USGS tweet earthquake dispatch (@USGSted): using twitter for earthquake detection and characterization. AGU fall meeting abstracts, vol. 1; 2012.

[12] Kradolfer Urs. SalanderMaps: a rapid overview about felt earthquakes through data mining of web-accesses. EGU general assembly conference abstracts, vol. 15; 2013.

[13] Sakaki Takeshi, Okazaki Makoto, Matsuo Yutaka. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web. ACM; 2010.

[14] Takako Hashimoto, Chakraborty Basabi, Kuboyama Tetsuji, Shirota Yukari. Temporal awareness of needs after east japan great earthquake using latent semantic analysis. Information modelling and knowledge bases XXV, vol. 260. IOS Press; 2014. p. 200–12.

[15] Doan Son, Vo Bao-Khanh Ho, Collier Nigel. An analysis of Twitter messages in the 2011 Tohoku earthquake. In: 4th International Conference on eHealth. Electronic Healthcare. Berlin Heidelberg: Springer; 2012. p. 58–66.

[16] Panigrahi Sangram, Verma Kesari, Tripathi Priyanka, Sharma Rika. Knowledge discovery from earth science data. In: Fourth international IEEE conference on communication systems and network technologies (CSNT); 2014. p. 398–403.

[17] Pabreja Kavita, Datta Rattan K. A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube. Int J Data Anal Tech Strat 2012;4(1):57–82.

[18] Srivastava Kuldeep, Bhardwaj Rashmi. Real-time now cast of a cloudburst and a thunderstorm event with assimilation of Doppler weather radar data. Nat Hazard 2014(70):1357–83.

[19] Merz B, Kreibich H, Lall U. Multi-variate flood damage assessment: a tree-based data-mining approach. Nat Hazard Earth Syst Sci 2013(13):53–64.

[20] Shafapour Mahyat, Pradhan Biswajeet, Jebur Mustafa Neamah. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 2013;504:69–79.

[21] Sahay Rajeev Ranjan, Srivastava Ayush. Predicting monsoon floods in rivers embedding wavelet transform, genetic algorithm and neural network. Water Resour Manage 2014(28):301–17.

[22] Thiemig V, Bisselink B, Pappenberger F, Thielen J. A pan-African flood forecasting system. Hydrol Earth Syst Sci Discuss 2014;11:5559–97.

[23] Karahan Halil, Gurarslan Gurhan, Geem Zong Woo. A new nonlinear Muskingum flood routing model incorporating lateral flow. Eng Opt 2014:1–13.

[24] Cheong France, Cheong Christopher. Social media data mining: a social network analysis of tweets during the 2010–2011 Australian floods. PACIS 2011.

[25] Wan S, Lei TC, Chou TY. A novel data mining technique of analysis and classification for landslide problems. Nat Hazard 2010(52):211–30.

[26] Venkatesan M, Thangavelu Arunkumar, Prabhavathy P. An improved Bayesian classification data mining method for early warning landslide susceptibility model using GIS. In: Proceedings of seventh international conference on bio-inspired computing theories and applications (BIC-TA 2012). Springer, India; 2013.

[27] Korup Oliver, Stolle Amelie. Landslide prediction from machine learning. Geol Today 2014(30):26–33.

[28] Sheth Anmol, Thekkath Chandramohan A, Mehta Prakshep, Tejaswi Kalyan, Parekh Chandresh, Singh Trilok N, Desai Uday B. Senslide: a distributed landslide prediction system. ACM SIGOPS Oper Syst Rev 2007(41):75–87.

[29] Ho Jui-Yi, Lee Kwan Tun, Chang Tung-Chiung, Wang Zhao-Yin, Liao Yu-Hsun. Influences of spatial distribution of soil thickness on shallow landslide prediction. Eng Geol 2012;124:38–46.

[30] Chen Huangqiong, Zeng Zhigang. Deformation prediction of landslide based on improved back-propagation neural network. Cogn Comput 2013(5):56–62.

[31] Di Salvo Roberto, Montalto Placido, Nunnari Giuseppe, Neri Marco, Puglisi Giuseppe. Multivariate time series clustering on geophysical data recorded at Mt. Etna from 1996 to 2003. J Volcanol Geoth Res 2013;251:65–74.

[32] Webley Peter, Mastin Larry. Improved prediction and tracking of volcanic ash clouds. J Volcano Geotherm Res 2009(186):1–9.

[33] Das Himangshu S, Jung Hoonshin. An efficient tool to assess risk of storm surges using data mining. Coast Hazard 2013(2):80–91.

[34] Ali MM, Kishtawal CM, Jain Sarika. Predicting cyclone tracks in the north Indian Ocean: an artificial neural network approach. Geophys Res Lett 2007(34).

[35] Siek M, Solomatine DP. Nonlinear chaotic model for predicting storm surges. Nonlinear Process Geophys 2010(17):405–20.

[36] Cohan Catherine L, Cole Steve W. Life course transitions and natural disaster: marriage, birth, and divorce following Hurricane Hugo. J Family Psychol 2002(16):14–25.

[37] Mandel Benjamin et al. A demographic analysis of online sentiment during hurricane Irene. In: Proceedings of the second workshop on language in social media. Association for Computational Linguistics; 2012.

[38] Dong Han, Halem Milton, Zhou Shujia. Social media data analytics applied to hurricane sandy. In: International IEEE conference on social computing (SocialCom); 2013. p. 963–6.

[39] Chae Junghoon, Thom Dennis, Jang Yun, Kim Sung Ye, Ertl Thomas, Ebert David S. Public behavior response analysis in

disaster events utilizing visual analytics of microblog data. Comput Graph 2014;38:51–60.
[40] Chatfield Akemi Takeoka, Brajawidagda Uuf. Twitter early tsunami warning system: a case study in Indonesia's natural disaster management. In: 46th Hawaii IEEE international conference on System Sciences (HICSS); 2013. p. 2050–60.
[41] Chatfield Akemi, Brajawidagda Uuf. Twitter tsunami early warning network: a social network analysis of Twitter information flows. In: 23rd Australasian conference on information system; 2012.
[42] Siahaan Daniel, Wenas Royke, Widodo Amien, Yuhana Umi. Web-based tsunami early warning system. IPTEK, J Technol Sci 2013(24).
[43] Ashktorab Zahra, Brown Christopher, Nandi Manojit, Culotta Aron. Tweedr: mining twitter to inform disaster response. In: Proceedings of the 11th international ISCRAM conference; 2014.
[44] Kumar Shamanth, Barbier Geoffrey, Abbasi Mohammad Ali, Liu Huan. Tweet tracker: an analysis tool for humanitarian and disaster relief. ICWSM; 2011.
[45] Battistini Alessandro, Segoni Samuele, Manzo Goffredo, Catani Filippo, Casagli Nicola. Web data mining for automatic inventory of geohazards at national scale. Appl Geogr 2013;43:147–58.
[46] Flash flood event with advanced GPS activity: http://www.nws.noaa.gov/om/brochures/flood and http://www.dailymail.co.uk/indiahome/indianews/article-2345372/Delhi-Noida-risk-flash-floods-rampant building-boom-destroys-forest-farmland-helped-absorb-rain.html.
[47] Pabreja Kavita. Clustering technique to interpret numerical weather prediction output products for forecast of cloudburst. Int J Comput Sci Inf Technol (IJCSIT) [ISSN: 0975-9646].
[48] Morstatter Fred, Pfeffer Jürgen, Liu Huan, Carley Kathleen M. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehouse. ICWSM; 2013.
[49] Jiang Long, Yu Mo , Zhou Ming, Liu Xiaohua, Zhao Tiejun. Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics, human language technologies, vol. 1; 2011. p. 151–60.
[50] Pak Alexander, Paroubek Patrick. Twitter as a corpus for sentiment analysis and opinion mining. LREC 2010:1320–6.
[51] Bollen Johan, Mao Huina, Zeng Xiaojun. Twitter mood predicts the stock market. J Comput Sci 2011;2(1):1–8.
[52] Paul Michael J, Dredze Mark. You are what you tweet: analyzing twitter for public health. Proceedings of the fifth international AAAI conference on weblogs and social media 2011:265–72.
[53] Bollen Johan, Mao Huina, Pepe Alberto. Modeling public mood and emotion: twitter sentiment and socio-economic phenomena. Proceedings of the fifth international AAAI conference on weblogs and social media 2011:450–3.
[54] Conover Michael D, Gonçalves Bruno, Ratkiewicz Jacob, Flammini Alessandro, Menczer Filippo. Predicting the political alignment of twitter users. In: 3rd IEEE international conference on privacy, security, risk and trust (Passat); 2011. p. 192–9.
[55] Kireyev Kirill, Palen Leysia, Anderson K. Applications of topics models to analysis of disaster-related twitter data. NIPS workshop on applications for topic models: text and beyond, vol. 1; 2009.
[56] Panem Sandeep, Gupta Manish, Varma Vasudeva. Structured information extraction from natural disaster events on twitter. In: Proceedings of the 5th international workshop on web-scale knowledge representation retrieval & reasoning. ACM; 2014.
[57] Core Team R. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013 [ISBN 3-900051-07-0, http://www.R-project.org/].
[58] Olston Christopher, Reed Benjamin, Srivastava Utkarsh, Kumar Ravi, Tomkins Andrew. Pig latin: a not-so-foreign language for data processing. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM; 2008. p. 1099–110.
[59] Thusoo Ashish, Sarma Joydeep Sen, Jain Namit, Shao Zheng, Chakka Prasad, Anthony Suresh, et al. Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endow 2009;2(2):1626–9.
[60] Mahout Apache. Scalable machine-learning and data-mining library. Available at mahout.apache.org.

Saptarsi Goswami: He is an Assistant professor at Institute of Engineering and Management in Computer Science Department, India, and a Research Scholar at A.K.Choudhury School of Information Technology, University of Calcutta. He has 10+ years of working experience in IT industry. His areas of interest are feature selection, outlier detection, mining unstructured data, etc.

Sanjay Chakraborty: He has completed his B-Tech from West Bengal University of Technology, India, on Information Technology in the year 2009. He has completed his Master of Technology from National Institute of Technology, Raipur, India, in the year of 2011. Now, He is working as an Assistant Professor at Department of Computer Science & Engineering in Institute of Engineering & Management, Kolkata. His areas of interests are Data Mining, Cryptography & Network Security, Cloud computing and Quantum Computing. He is a professional member of IAENG and UACEE.

Sanhita Ghosh: She is an Assistant professor at Institute of Engineering and Management in Computer Science Department, India. She has completed her graduation and masters from abroad.

Amlan Chakrabarti: He is at present an Associate Professor and HoD at the A.K. Choudhury School of Information Technology, University of Calcutta. He has done his Doctoral research on Quantum Computing and related VLSI design at Indian Statistical Institute, Kolkata, 2004–2008. He was a Post-Doctoral fellow at the School of Engineering, Princeton University, USA, during 2011–2012. He is the recipient of BOYSCAST fellowship award in the area of Engineering Science from the Department of Science and Technology Govt. of India in 2011. He has held Visiting Scientist position at the GSI research laboratory Germany and Department of Computer Science and Engineering at the New York State University at Buffalo, USA, during Sept–Oct, 2007. He has published around 50 research papers in referred journals and conferences. He is a Sr. Member of IEEE and life member of Computer Society of India. He has been the reviewer of IEEE Transactions on Computers, IET Computers & Digital Techniques, Elsevier Simulation

Modeling Practice and Theory, Springer Journal of Electronic Testing: Theory and Applications. His research interests are as follows: Quantum Computing, VLSI design, Embedded System Design, Video and Image Processing Algorithms and pattern recognition.

Basabi Chakraborty: She received B.Tech, M. Tech and Ph.D degrees in RadioPhysics and Electronics from Calcutta University, India. She worked in National Center for Knowledge based Computing Systems and Technology affiliated to Indian Statistical Institute, Calcutta, India, until 1990. From 1991 to 1993 she worked as a part-time researcher in Advanced Intelligent Communication Systems Laboratory in Sendai, Japan. She received another Ph.D in Information Science from Tohoku University, Sendai, in 1996. From 1996 to 1998, she worked as a post doctoral research fellow in Research Institute of Electrical Communication, Tohoku University, Japan (under Telecommunication Advancement Organization (TAO) fellowship for a period of 10 months). In 1998 she joined as a faculty in Software and Information Science department of Iwate Prefectural University, Iwate, Japan, and currently an Associate Professor in the same department. Her main research interests are in the area of Pattern Recognition, Image Processing, Soft Computing Techniques, Biometrics, Trust and Security in Computer Communication Network. She is a senior member of IEEE, member of ACM, Japanese Neural Network Society (JNNS) and Information Processing Society of Japan (IPSJ), executive committee member of IUPRAI (Indian Unit of Pattern Recognition and Artificial Intelligence), IEEE JC WIE (Women In Engineering) and ISAJ (Indian Scientists Association in Japan).