#ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters using Social Media

Bhuvaneswari Anbalagan
Research Scholar
Department of Computer Technology
MIT, Anna University, Chennai, India.
e-mail: bhuvana.cse14@gmail.com

Dr. Valliyammai. C
Asst. Professor (Sr. Grade)
Department of Computer Technology
MIT, Anna University, Chennai, India.
e-mail: cva@annauniv.edu

Abstract— **The recent emergence of ubiquitous smart communication devices accelerate people to post the current trending topics in real time as micro blogs, tweets, posts and multimedia content on social media sites along with geographical location tags (geo-tags). Specifically, during recent floods in Tamilnadu 2015, the early warnings about flooded areas emerged to get posted in popular social media with geo-parsed hash tags continuously. In the humanitarian view, the real-time crisis sparked great interest in designing an innovative methodology using big social media data analysis along with supervised machine learning techniques to actuate immediate disaster response and rescue efforts in near future. The proposed system performs disaster tweet collection based on trending disaster hash tags. Our system performs Naive-Bayesian (multinomial) and SSVM classification on collected tweets to identify the severity of the disaster. Based on location-to-interpolation cluster proximity, disaster geographic map is generated for the affected area. Our approach detects the tweets fitted into correct classifier label, and generate an output with detection rate of 79% to 91% of the time. The predicted disaster mapping results are highly accurate up to 89% for real time geo-parsed tweets that matched with actual location at-risk during the flood.**

Keywords - Emergency Informatics; Social media; Geo-parsing; Geo-coding; Hash tags; Crowd sourcing; Naive-Bayesian; SSVM Classifier; Disaster Management;

I. Introduction

Social media sites such as Twitter, Facebook and Tumblr which provide micro-blogging service, allows people to get connected for ease way of sharing the ideas and updates of activity as posts, tweets, uploading photos, videos in their profile, pages, groups, events and live streaming. The users have gained much more attention recently as they rely closely on this social media platform to communicate during and after disasters like earthquake, floods, flash floods, tsunamis, hurricanes, tropical storms, landslides and debris flow etc. During times of crisis, people come together on Social Networks [1] to support one another through effective communication. The hash tags carries the geographic location during natural disasters help the disaster management and rescue team to identify and classify the early disaster warnings, damage prediction during and after disaster.

The recent emergency crisis [2],[3] particularly earthquake struck in Nepal, Paris Attack, acute rainfall in Tamilnadu [4],[5] where people turned to Facebook, Twitter to find out if their loved ones are safe. Popularly known Facebook introduces the feature called Disaster Message Board later termed as "Safety Check". This feature facilitates the people at the time of natural or man-made disasters to quickly and promptly determine whether people are safe in the affected geographical area. In November 2015, Facebook deployed the feature during Paris attacks and activated the feature button for a violent attack for the first non-natural disaster.

Many features that are available in social media facilitate the people to ensure their current situation, location exposed to their friends to whom they are connected. Google Person Finder (GPF) is an open source feature that provides a registry detail and message board for family, and close ones affected by a natural disaster to send and look for information about each other's current status and whereabouts situation. GPF was created during 2010 Haiti earthquake [6] by volunteer Google engineers. The statistics [7] can give clarity on the usage of social media and sites at the time of crisis to find the whereabouts and how about of people under disaster to their connections in real life. In this paper, disaster event is interpreted and analyzed using big social media data extracted from Twitter. The tweets, geo-parsed hash tags are collected and geo-coded maps are generated to identify the precision of location-level tweet. The proposed system can visually alert public, local volunteers and civil guard authorities along with geographical map assistance for effective recovery in action.

The rest of the paper is organized as follows. Section II contains the disaster related model on Twitter hash tags using various implemented frameworks is studied. Section III presents the detail study on emergency informatics related to Chennai Flood 2015, rain gauge geographical view created by expert groups during floods. The Naïve Bayes and SVM machine learning models involved in the proposed system is discussed in Section IV. The proposed machine learning predictive analysis model for disaster response and recovery is explained in Section V. Section VI contains implementation details of datasets and crowd sourcing metrics for computing accuracy of tweet datasets. Additionally, Natural Language Processing (NLP) on flood tweets are explained. It shows the process of extracting the locations using geo-parsing and geo-coding of tweets and

978-1-5090-5773-3/16 $31.00 © 2016 IEEE
DOI 10.1109/HiPCW.2016.17

mapping. The experimental results of the proposed model is discussed and its performance is evaluated in Section VII. Finally, Section VIII discusses the conclusion and future work.

II. Related Work

Many researchers have proposed event monitoring and detection systems for disasters related incidents, such as flood, earthquakes and hurricanes, by physical sensor networks and real-time web monitoring. An effective SOA-based framework [8] for disaster service management for very large-scale observing systems was developed. It can be used as the dominant means of study for a large range of natural phenomena including natural disasters. A web-based system primarily focused on data collection and visualization, but more featured analytics on disaster response was studied with Twitcident [9]. Twitter Earthquake Detector (TED) precisely examines data from social networks [10] and delivers risky information to the social community based on the quantity of significance in a particular earthquake.

An efficient algorithm to observe tweets [11] and sense earthquake events by taking into account of each Twitter user as a sensor developed platform using client tools called Emergency Situation Awareness-Automated Web Text Mining system. The system can identify tweets relevant to emergency incidents. Stream Web [12] was proposed, which is a real-time Web monitoring system on top of a stream computing system called System S. It was developed by IBM Research, which provides a platform for developers to examine steaming data such as Twitter streaming. Artificial Intelligence for Disaster Response (AIDR) [13] platform can be used to classify and order Twitter messages for community health monitoring. It is a web-based platform designed to allow analysts to collect tweets and classify micro blog posts based on search keywords. More specially, AIDR classify tweets into a set of user-requested categories of information. The platform continuously stream data, processes it using machine learning classification technique, and combine human intelligence with the help of volunteers.

More than 300 million tweets that are posted before and after the Great East Japan Earthquake [14] analyzed to reveal how people share the disaster information on Twitter. The users of social media changed their behavior and reasons to use social media after serious Seattle-Tacoma crisis events. A parallel visualization model using stream graph and relational graph as a spring model was proposed. They viewed the flow of associated topic words which change their phenomenon in the relational graph. The information of temporal tweets includes mood of various user topic and group interests. During Haiti Earthquake, tweets from social media are analyzed to classify the tweets and evaluated the accuracy. During the earthquake and Tsunami people cooperated to change their Twitter mode from communication to an information sharing tool to report their status after the earthquake. Of course, no certainties exist that twitter will be used during subsequent disasters. Their findings include social media as useful tools for information sharing and communication, because of the changes in user behaviors during serious situations.

III. Emergency Informatics: A Study on Chennai Flood Tweets 2015

The Chennai city is situated in Tamilnadu (India). Its suburb areas unexpectedly recorded manifold heavy rainfall measures during November-December 2015. The rain gauge geographical areas are highly affected by Chennai floods are taken as reference for our work. The flood submerged the coastal districts of Chennai, Kancheepuram, and Tiruvallur, and blown up more than 4 million people with financially viable reimbursement, damages and death tolls of about 250 lives. These rainfall record matches with a rate of Twitter tweets about the flood on the dates mentioned as follows: For the period of the 24 hours on December 2, 2015, "Extremely Intense Rainfall" is recorded in Chennai. For three days of intense rainfall (Nov 30, Dec 1, Dec 2) and Adyar river-basin flooded, where the basic amenities like food, shelter, drinking water, electricity, mobile networks all failed. But the Online Social Networks (OSN) and social media are used to connect the flood-affected people with rescue teams, worried family, relatives, and friends to get information about their loved ones.

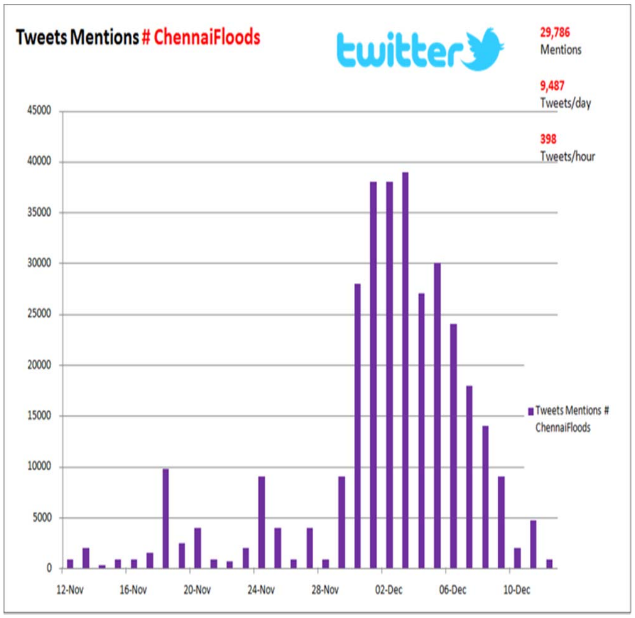


Figure 1. Tweet Mention analysis on Chennai Flood from 12-Nov to 12-Dec 2015.

According to our analysis on Twitter, Fig.1 shows during the normal and peak rainfall people show up in social sites where 2,342 Blogs Mentions, 1,453 Forums Mentions, 298,123 Twitter Mentions, 19,342 News , Mentions in the period of 30 days during 12-Nov to 12-Dec 2015. The leading Indian Express reports #ChennnaiFloods hash tag for tracking disaster on Twitter. Once the rainfall shutters the whole city with flood where people are paralyzed to evacuate and needs help, they rely on social media sites to report the status of rainfall, water logged level in their location through micro blogs with official and trending hash tags. The twitter tweets recorded frequent rates of tweets from the places where streets, roads, reservoirs got flooded. Twitter traffic records total mentions of tweets during the Chennai flood due to which Government of India declared the Tamilnadu State flood as National Disaster.

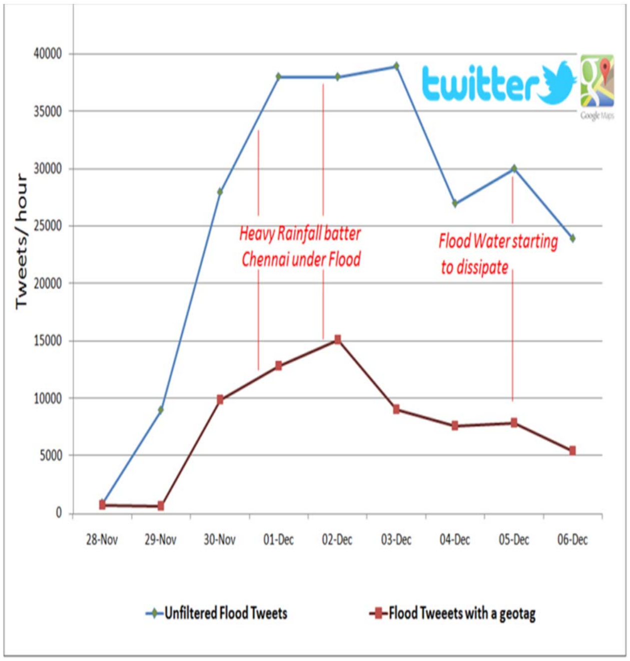


Figure 2. Twitter traffic recorded using Chennai Flood keywords over period of November–December 2015 resulted in heavy rainfall at Tamil Nadu. Peak tweet traffic noted 38,000 tweets per day, with 12% of tweets using the TamilNadu time zone UTC +05:30 containing a geo-tag

The top trending hash tags views and responses to those pictures across Twitter in India from Tamilnadu are the first clues to get the attention of people around the world so come to know that Chennai city is submerging under flood water. The early warning signs of photos and videos about flooded areas, submerged houses, cars, bridges, railway tracks emerged to get posted in Twitter with famous hash tags #ChennaiFloods, #TNflood, #ChennaiRains and #PrayForChennai continuously. The records ingest tweets with a geographical hash tag (geo-tag) is 12% of overall unfiltered flood tweets is shown in Fig 2. Twitter analysis on overall Tweets, @reply, Retweets posted by users to spread the emergency situation in Chennai is shown in Fig 3. Further it reports majority 71% of the flood tweets during the crisis are Re-tweets. Its shows people need the tweets about the disaster should be spread to everyone. In that case, any respondent in that location could help by seeing the tweet. The recent emergence of disaster management and recovery deals with how people could be reached and saved through easy and fast way with the help of Social Networks. The tweets and posts in OSN are mainly used extensively to relay information about flooded areas, rescue agencies, food and relief centers. The popular hash tags and accounts identified are #Chennairains, #ChennaiVoluntees, @Chennairescue, #Chennairainhelp, @ChennaiRainsOrg, #ChennaiMicro.

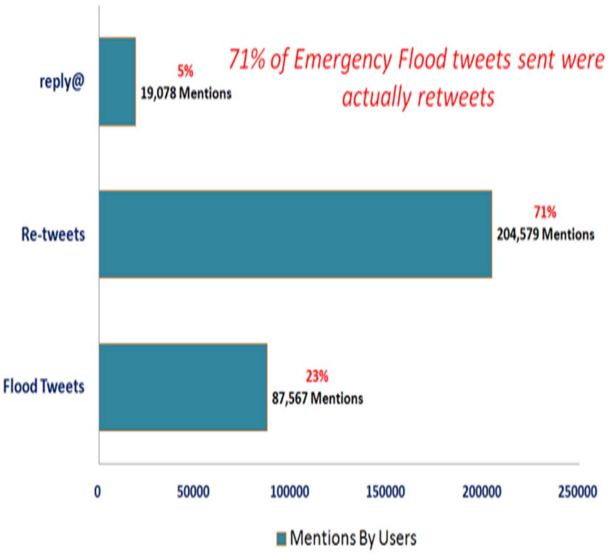


Figure 3. Total Mentions on Flood Tweets, Retweets, @reply

IV. Machine Learning Methods

A. Naïve Bayes Classifier

The tweets are trained using Naïve Bayes, which is a classifier based on Bayes hypothesis theorem. It performs probabilistic prediction based on statistical values. The classifier works by expecting that; value of the attribute is conditionally independent. For Naïve Bayes classification, the following condition is satisfied:

$$P(B_i|X) = \frac{P(X|B_i)\,P(B_i)}{P(X)} \qquad (1)$$

Equation (1) above, the simple Naïve Bayesian classifier, work as follows.

i. Let S be the set of tuples for training and their related class labels. The tuple is identified by an ‘n’ dimensional attribute vector, $X = (X_1, X_2, .., X_n)$, describing n dimensions prepared on the tuple from n attributes, correspondingly as $A_1, A_2, .., A_n$

ii. Considering there are m classes $B_1, B_2, .., B_m$. For the given tuple X, the classifier will predict that X is belonging to the class having the maximum posterior probability with conditioned on the tuple X.
It is the obvious that the Naïve Bayesian classifier predicts that tuple X belongs to the class B_i if and only if as following condition (2) is satisfied.

$$P(B_i|X) > P\left(B_j\middle|X\right) \text{ for } 1 \le \text{j} \le \text{m; j} \ne 1 \qquad (2)$$

Thus the system have maximum $(B_i|X)$. The class B_i for which $P(B_i|X)$ is maximized in (3) is called the maximum a posteriori (MAP) hypothesis.

$$\hat{z} = \underset{M\in\{1,..,m\}}{argmax}\; p(B_m)\prod_{i=1}^{n} p(x_i|B_m) \qquad (3)$$

iii. From (1), only $P(X|B_i)\,P(B_i)$ need be maximized, as P(X) remains constant for all classes. Then the classifier predicts the data item X belonging to class B_i if and only if it holds the highest probability when compared to other class label.

Specifically, multinomial Naïve Bayes model is a classifier, which is used to obtain the frequencies with which certain events are generated using multinomial values as $p_1, p_2, .., p_n$ where p_a is the probability that event 'a' occurs. In a particular instance, a feature vector is transformed to a histogram, along with counting the number of times event 'a' is observed.

$$P(x|B_k) = \frac{(\Sigma_{\,i}\, x_i)\,!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i} \qquad (4)$$

The probabilistic event model can used for text based classification, in which events representing the existence of a term in the given input text document. Using (4), the system calculates the maximum likelihood of observing a frequency of a term x. While expressing in log-space the multinomial naive Bayes classifier becomes a linear classifier.

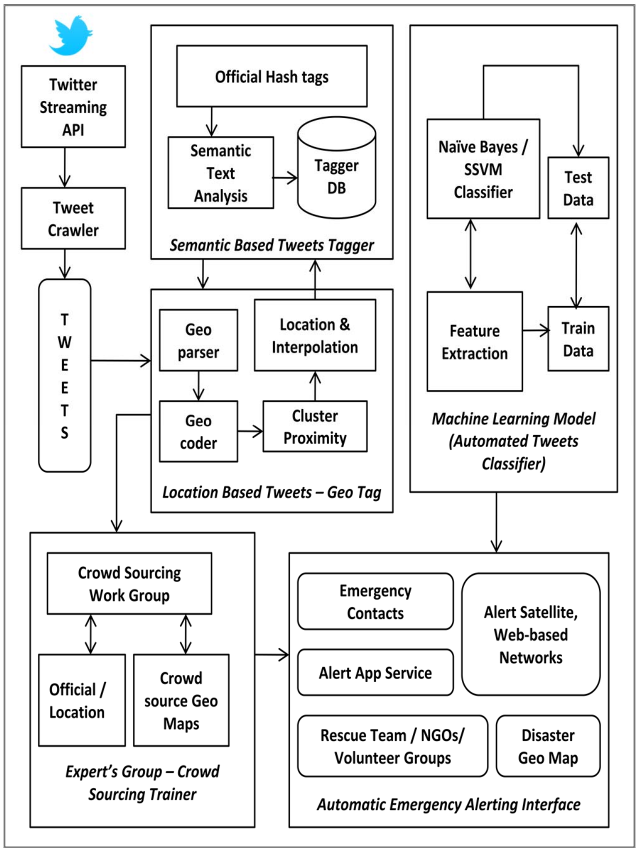
B. Structured Support Vector Machines

Structured SVM is a Support Vector Machine (SVM) learning algorithm to predict multivariate structured output labels. The prediction is accomplished on complex objects including Natural Language Parsing, Parts Of Speech tagging using Markov models which performs multi-class SVMs under simple linear kernel function. The learning algorithm will perform supervised learning by approximating a mapping $\mathrm{h}{:}\, \mathrm{X} \rightarrow \mathrm{Y}$ using labeled training samples $(\mathrm{x}_1, \mathrm{y}_1), \dots .(\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})$. The initial learning model parameters are set and the pattern-label pairs are taken as input with kernel functions. The user defined special constraints are then set and the learning model is initialized.

Subsequently, a cache is created combining with feature vectors and then the learning process begins. For each training tweet data, the label related with most violated constraint for the pattern is found. After learning a model is created for classification. Then, the feature vector describing the relationship among the pattern and the label is calculated and the loss is also computed with loss function. For the testing phase, the learned model is taken as reference and the testing pattern-label example pairs are given as input. Then, it repeats over all the testing tweets, classifies each example, writes the label and then may evaluate the prediction and accumulate statistics. Then the system define a loss function measuring how well the prediction matches the truth label.

V. Disaster Response And Recovery Model

Disaster management operations require high reliable communication and connectivity to estimate the level of damages, public security, safety measures, mitigation services and hazardous area of rescue. The emergency support function falls under the categories referred by experts such as mass care, transportation, tactical team deployment, volunteer and missing-persons. The proposed model for emergency management is shown in Fig 4. Social media contribute enhanced mortality and morbidity support during many natural disasters and man-made disasters.



Twitter Streaming API
Tweet Crawler
T W E E T S
Official Hash tags
Semantic Text Analysis
Tagger DB
Semantic Based Tweets Tagger
Geo parser
Location & Interpolation
Geo coder
Cluster Proximity
Location Based Tweets – Geo Tag
Naïve Bayes / SSVM Classifier
Test Data
Feature Extraction
Train Data
Machine Learning Model (Automated Tweets Classifier)
Crowd Sourcing Work Group
Official / Location
Crowd source Geo Maps
Expert's Group – Crowd Sourcing Trainer
Emergency Contacts
Alert App Service
Alert Satellite, Web-based Networks
Rescue Team / NGOs/ Volunteer Groups
Disaster Geo Map
Automatic Emergency Alerting Interface

Figure 4. System Model for Disaster Management

A. Disaster Identification

Climatologically natural hazards have the high risk of damage to human life and infrastructure. A disaster happens, micro-bloggers start to post tweets with hash tags

#disaster, #flood, #storm, #tsunami and so on. The proposed model can identify the type of disaster by analyzing crowded source of tweets with same hash tag in a particular sliding time window size (say 1 hour).

B. *Social Media Crisis Tweets Collector*

The disaster is handled instantaneously so that the people who got affected by the disaster obtain the immediate response to the incident. The proposed model uses the Streaming API service to collect tweets of trending disaster hash tags. The API crawls all tweets using keywords of location, road, area, and landmark of disaster to collect the most precise and accurately tweets. The collector contains the tweet id, tweet name, actual tweet content, time, total reply to that tweet, retweet information, verified tweet or not, official or not. These are the features extracted for our classifiers using tagger.

C. *Semantic Based Tweets Tagger*

The main process of a tagger is to compute the feature extraction which classifies individual tweet to get a tag or category. Tweets are micro-text with unigrams and bigrams semantic annotations. It sends to a classifier by the previously extracted key tag phrases. The extraction system will identify the tweet belongs to caution and warning, damage of infrastructure, casualties, relief donations offer, medical care, person missing and seen, immediate need for money, shelter, services, water and so on. These tags are labeled so that the model can learn during the tagging phase. The significance of official tweets and verified tweets are propagated to retweets through official tagging. Example #ChennaiMicro, #ChennaiFloods leads to the credibility of the disaster happened. The adaptive- filtering and LDA algorithm is used in our system to analyze semantically and tag the tweets. The tagger performs official tweet tags are verified and stored in database.

D. *Location based Geo-Tagger for Crisis Mapping*

The tagging of tweets cannot simply serve as a responding force of immediate action. It needs in-depth information of the geographical location details to fasten the rescue operations. In the case of medical emergency, the tweet should identify the location where the emergency causalities can get help from nearest clinic, hospital or move to the nearest doctor available. The geo tagger will map tweet text to extract location by finding the geographical names and if it match exactly with true positive rates, then it is mapped to that particular location under healthcare emergency category.

E. *Expert Group Crowd Sourcing Trainer and Decision Support*

An external man- assisted crowd sourcing is required to ensure the quality, reliability, credibility of the tweets are reported for immediate response. In order to precisely improve and fasten the disaster recovery, the trainer model should be well trained to ensure high quality in such a way it maximizes marginal quality gains per human label so that decision making is possible with True positive rates are calculated. This will label event classification during the learning phase. Thus the tweets are trained by the human of expertise group in disaster respondent assistance.

F. *Automated Tweets Classifier Accuracy towards Mapping*

The streaming tweet data source is taken as input for the test datasets. Already tweets are trained with the help of crowd sourcing and learning classifier. When a new tweet arrives with the help of hash tags, geo tags, location maps geometry, it gets classify automatically to a category based on previously supervised learning method while training data. The bench mark datasets collected during Chennai floods is used. The metrics of precision for true-positive rate, recall for false-positive rate is calculated to visualize the quality of classification under ROC (Receiver Operating Characteristic) curve and AUC (Area under Curve) coefficient values. The tweets of chennai flood mapped using the geo tagger and automatic tweet classifier shows the high percent of matching accuracy with the actual post-flood assessment map with less recall threshold.

G. *Automatic Emergency Alerting Interface*

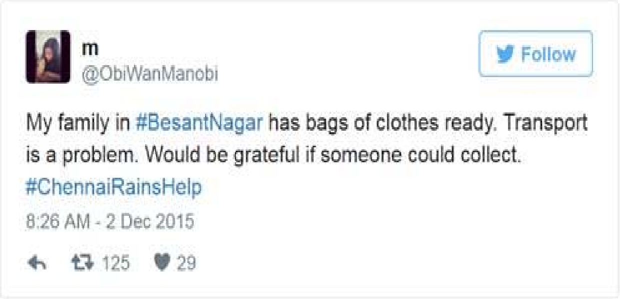
The emergency alerting interface requires registered emergency contacts, crowd sourcing task generator to alert the causalities connections and identify category of operation required from disaster mitigation. The interface includes notification through “Safety Check” enabling in Facebook, notifications to rescue team, NGOs, Volunteer groups, Mobile apps service, and offline apps services.

VI. Implementation Details

A. *Dataset*

Twitter has an official API called OAuth, a token-based authentication system that indexes tweets in order to match a given search string and writes the output to a file. The tweet extraction service is free and convenient up to certain limit to perform a quick and efficient extraction of tweets, it has a crucial limitation: it can retrieve tweets only from the previous week of 7 days. Hence, the tweets are searched and extracted for the search term is the hash tag #ChennaiFloods during the beginning period of December 2015 month where the flood rate is high and people affected lot during that tenure.

The platform independent Java language is used to perform the extraction; the extracted tweets are written to a comma-separated value file. Various hash tags for the same flood topics are observable. This would make it more challenging in order to separate all tweets on the particular subject. The geo tags indicate to the place Chennai Airport is logged with water. The geo parser will extract the text to get mapped into the geographical coordinate of Chennai airport in training map. For instance, here is a tweet in Fig 5 discharged in Twitter about flood in Chennai.



m
@ObiWanManobi

My family in #BesantNagar has bags of clothes ready. Transport is a problem. Would be grateful if someone could collect. #ChennaiRainsHelp

8:26 AM - 2 Dec 2015

125 29

Figure 5. Tweet Sample

B. *Flood Tweets – Natural Language Processing (NLP)*

The flood tweets are collected and basically parsed into a corpus for effective text analysis using NLP [15]. The following steps are executed to clean the corpus and make it ready and prepared for further analysis. Only the text portion of the tweet (the actual message) is considered. The NLP steps is shown in Fig 6.

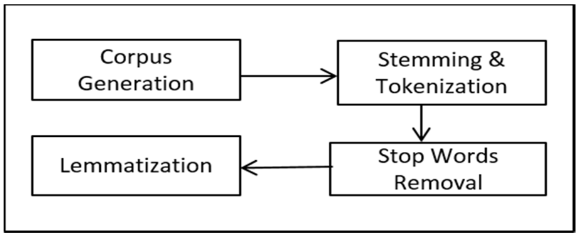


Figure 6. NLP Steps

Various steps involved in NLP is shown as follows.

1) Evacuating Stop Words: Tweet IDs are number generated by Twitter to identify each tweet. Numbers as such don’t serve any purpose for text analysis and hence they are discarded. Stop words are words that are regularly utilized as a part of each sentence, yet have no systematic essentialness. These words are expelled by coordinating the corpus with the stop words list in the text mining (tm) package of R. Swearwords are additionally expelled.

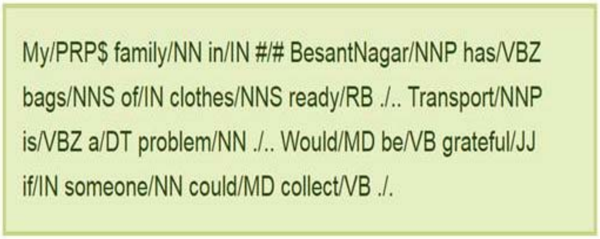
2) Stemming words : In content investigation, stemming is 'the way toward lessening curved (or infrequently inferred) words to their assertion stem, base or root structure. Stemming is done to diminish inflectional structures and ‘now’ and ‘then’ derivationally related-types of a word to a typical base structure.

3) Suffix-dropping algorithm: The last parts of all the words in the tweet get truncated initially. For example, words like ‘producing’, ‘producer’, ’produced’, can all be stemmed to the root ‘produce’. On the other hand, ‘rescued’, ‘rescuing’, ‘rescue’ are stemmed to form ‘rescue’, which is not a word or a root.

4) Lemmatization: Every tweet word is the determination of the lemma for a word in the corpus. This is finished with the comprehension of the connection, grammatical form and the vocabulary.

5) N-gram corpus analysis: Each tweet word is broken into a piece of its entire by "n" characters. For instance, for n=1 (unigram), the letters 'f', 'l', 'o', 'o', "d" are exclusively parsed from 'surge'. For a higher n (say n=5), "surge" is held from 'flooding'.

The sample response of the collected tweets refer to the location, water level, relief camps, diseases and precautions, food crops damaged, house, car, factory, air flights flooded into water, volunteers in the area of needy, preparedness of the first aid kits, water bottles, cloths, foods is the sample response for the tweets. Some handles belong to prominent celebrities. The tweet is posted with the geographical tag that is given a sample response. Using the geo tags, the accessibility of the transport could be estimated and predicted. It ensures which area is safe to relocate and to travel at the time of disaster. The tagger block represents the way of geo coding which is done after identifying the tweet location using geo parser. Once the location is identified, the tweet is coded onto the map co-ordinates.



```
My/PRP$ family/NN in/IN #/# BesantNagar/NNP has/VBZ
bags/NNS of/IN clothes/NNS ready/RB ./.. Transport/NNP
is/VBZ a/DT problem/NN ./.. Would/MD be/VB grateful/JJ
if/IN someone/NN could/MD collect/VB ./.
```

Figure 7. Tagging, Chunking and Named Entity Recognition with NLTK

C. *Tokenization and Part of Speech (POS) tagging*

The crude content of the dataset is part into sentences utilizing sentence segmentation, and every sentence is further subdivided into words utilizing a tokenizer. Next, every sentence is labeled with grammatical feature labels, which will demonstrate exceptionally accommodating in the following stride, named substance identification. In this progression, we hunt down notice of conceivably fascinating substances in every sentence. The output is the set of tags, chunking with Noun, Verbs, and Adjectives as shown in Fig 7.

D. *Noun Phrase Chunking and Relation Detection*

The assignment of expression lumping, NP-piecing, where tweets is compared to individual phrases. The content with NP-lumps stamped utilizing sections is analyzed. For the above input the regular expression parser will evaluate the sentence in input document.

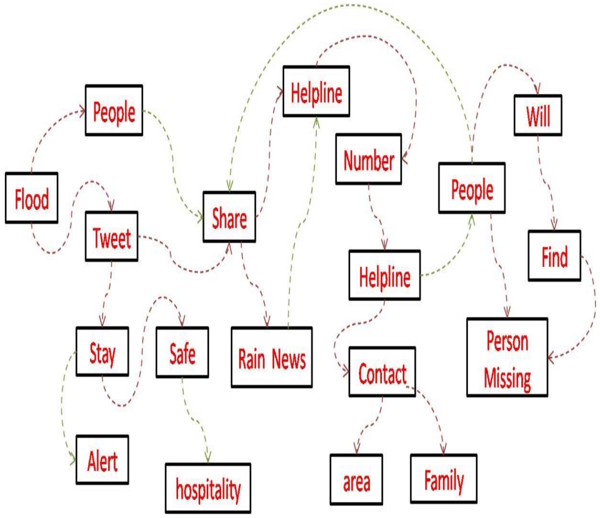


Figure 8. Word Semantics - Relation Detection on Flood Tweet

E. *Semantic Based Tweet Tagger*

The semantic ontology is constructed to illustrate the tweet type to which the word belongs is shown in Fig 8. Tweets that discuss about general data about influenced people, news channels and information about the emergency. Tweets that depict the climate along with gauges of downpour needs further improvements. Tweets provide alert to individuals on unsafe territories and offer data on help endeavors.

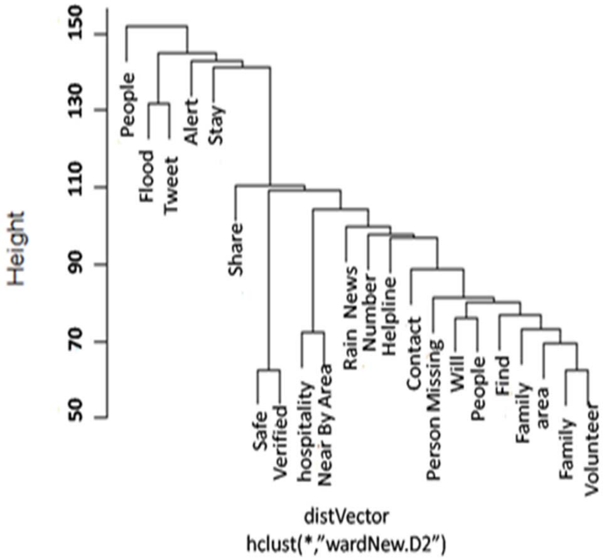


Figure 9. Cluster Dendrogram

For example, the hash tag popularity (defines a total number of tweets containing the hash tag divided by the total number of tweets of the same period of time) of #ChennaiFloods jumped to 0.14 percent from 0.67 percent. The accompanying unmistakable bunches of tweets are detectable from the dendrogram is shown in Fig 9.

F. *Crisis - Location Mapping*

During Chennai Floods, tweets contain area mentioned in micro text format that simply contain misspelled local language (Tamil) that are mistaken, abbreviated or highly localized. The geo-hash tags frequently replicate the increase and drop off events and trends obsessed by user attention concerning the area of flood and its location. This increase in popularity corresponded with the flood affected area which needs the immediate response. In our experiment, by analyzing the tweet, the system come to know people requested for help that needs the attention of the volunteers using the geo tags (#Chennai).

1) Geo-Parsing Phase: Geo parsing is the technique to convert the free text of place descriptors to the exact mapping of the identifiers in the geography called as geo coordinates with mentions on latitude and longitude interpolation. It is an explicit process of identifying area or region or exact location, the street named within the tweet text by identifying hash tag popularity. The process of converting free-text descriptions of places (such as "Chennai") into definite geographic identifiers such as latitude (Lat.) and longitude (Lon.) coordinates. For example, the latitude 13.000120 and longitude 80.2565 for Adyar River is obtained using geo parser is listed in Table 1. The confidence is measured between 0 and 1. Further, geo-parsing goes beyond geo-coding in that, rather than analyzing structured location references like limited address and numerical coordinates, by handling uncertain place names in unstructured text.

TABLE 1. OUTPUT FROM GEOPARSER LIBRARY

Place / Location	Con.	Geo. Type	Co-ordinates	
			Lon.	Lat.
India / India	1	independent Country entity	79	22
Adyar/ Chennai	1	Place / River	13	80

For our experiment, the open source R package is used for parsing text (tm-geo parser) into geographical locations. By declaring the DOM element that will contain map and as the result Map shown is executed and loaded for the datasets.

2) Geo-Coding Phase: It is the method of converting address into geographic coordinates like latitude 13.000120 and longitude 80.2565, to place markers on a map and co-ordinates position the map is shown in Fig 10. The system implements geo-coding and geo-parsing using Maps of India which is provided as the open source API to map the coordinates. The coordinates are searched for a provided location, city, address, place or point of interest (POI) and returns all available details along with its coordinates on the map using location-to-interpolation cluster proximity method. Point locations on map are displayed using markers.

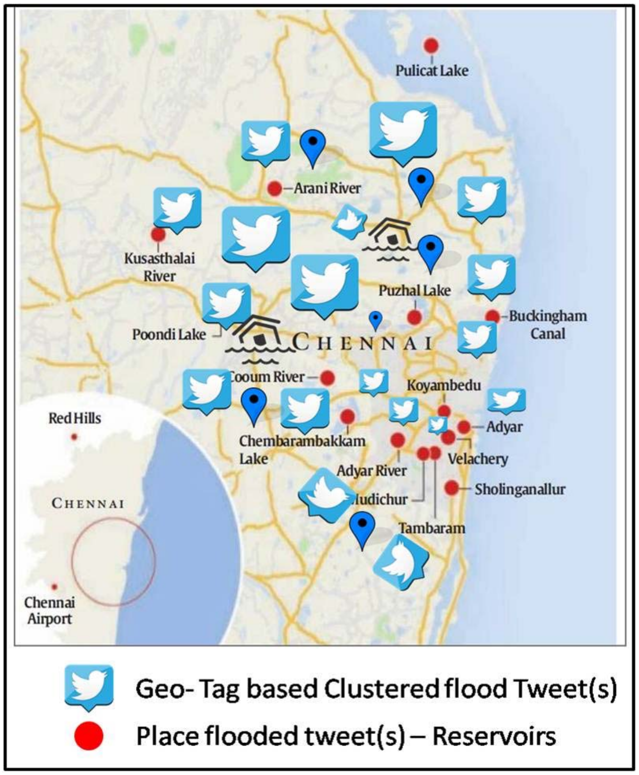


Figure 10. Geo coding points on Map – Location Based Plots

VII. Experimental Results

During the crowd sourcing task, context is extracted to showcase each tweet and the type (and sub-type, if accessible) determined during the classification phase. The data is labeled using the same procedure as for other datasets. In this task, approximately 7500 unique tweets are collected, in which after cleaning, preprocessing the system identified 5000 valid tweets related to Chennai floods. During our implementation, the work groups are requested to label an individual tweet is listed in Table 2 is used to separate informative tweets from personal and for an informative tweet specify what information it conveys.

There are two aspects of measurement that are related to the sensitivity of the system namely hit and detection rate. The Naïve-Bayesian and SSVM classifier is applied on the data set to determine the detection rate and hit rate. Detection rate measures the fraction of tweets in which humans found a relevant part of information and our system also found somewhat, even if that some tweets classification is incorrect. Hit rate measures the fraction of tweets for which our system found the classification success and that could be measured accurate or correct by humans.

TABLE 2. Tweet Classifier by Crowdsourcing Task

Type of Tweet	Key	Tweet Tagger
Requests for help	HLP	#NeedDrinkingWater, #NeedFood
Sympathy	SMP	#SaveChennai, #PrayForChennai, #ChennaiRainsHelp
Information on Specific areas	INF	#Chennai, #airport, #Chromepet, #Mudichuir, #Velachery
Cautionary Messages / Public Warnings	PWC	#ChennaiDrowning, #verified, #FloodAlert, #PorurLake, #ChennaiFloodRelief
Information on further weather	IW	#chennaiweather, #porurLake, #WeatherChanges, #AdyarRiver
Volunteers	VOL	#ICanAccomodate, #FoodSupply, #MobileRecharge, #Volunteers
Infrastructure Damages	IFD	#HouseDrowned, #CarFlooded, #VehicleDamaged, #FloorSinked
Forecasts	FRC	#OverNightsRains, #BadWeather

The feature extraction is carried out through which 8 unique features are identified: tweet ID, tweet text, username, location, geographical coordinates, hashtags, time, date. Table 3 shows the results of classification on the disaster dataset. The first two rows shown in Table 3 are scenarios where a particular model is created, and the remaining corresponds to various class-specific models.

TABLE 3. Classification Vs Crowd source Results

Train on 65% of Flood Data, Test on 35% Flood Data						
Train Data	*Train Count*	*Test Data*	*Test Count*	*Detected Correct*	*Detection Rate*	*Hit Rate*
All	3750	All	1250	887	71%	79%
All	3750	HLP, INF, SMP	486	413	84%	90%
All	3750	PWC	156	68	44%	39%
All	3750	IW, IFD	368	228	62%	55%
All	3750	VOL	541	362	67%	66%

The classification can be improved if the dataset is tested with tweets datasets in which more than one type of information is present in a given tweet, as shown in the second and third rows (Table 3).The system is computed by comparing its output with the responses provided by humans. The proposed system is trained on 65% data of the (crowd sourced) human-provided labels, and tested on the 35% of data. For the experiment, Naive Bayesian and SSVM classifiers are used to run on the test data sets.

TABLE 4. CONFUSION MATRIX

Predicted Actual	**Yes**	**No**	**Total**
Yes	**FP** = 259	**TN**=1390	1649
No	**TP** = 3210	**FN**=141	3351
N=5000	3469	1531	5000

The system is trained over the entire training set and tested on the test set, it is observed with relatively high detection rate using SSVM classifier when compared with NB classifier. The system considers the correct output if it matches in at least one word with the given human label. The confusion matrix is constructed using actual and predicted classification as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The evaluation is carried out based on these formula (SSVM classifier) with confusion matrix as described in Table 4. The system identified significant improvements in hit ratio for all tweets and cumulative class-specific models using SSVM.

TABLE 5. NAÏVE BAYES VS SSVM CLASSIFIER

Performance Measures	**Classifier Method**	
	Naïve Bayes	***SSVM***
Accuracy	78.51%	91.01%
Misclassification (Error Rate)	21.49%	8.99%
Recall	0.81	0.95
Precision / Confidence	0.77	0.91
Specificity	0.31	0.41
ROC	0.78	0.92

Performance evaluations are measured using Naïve Bayes and SSVM algorithms based upon Accuracy, Misclassification (Error Rate), Recall, Precision, Specificity and ROC curve is shown in Table 5 using MATLAB software. Accuracy of a classifier is the proportion of test set tuples that are properly classified by the classifier on a given test set. Recall/Sensitivity/True positive rate it is the ratio of overall positive cases that are exactly identified. Precision/Confidence represents the ratio of predicted positive cases that are correctly real positives. Specificity is the ratio of true negative cases found that are actually false. Misclassification Rate also known as "Error Rate" is the ratio of sum of false cases to total number of cases. Receiver Operator Characteristic (ROC) curve is a graphical way for representing the tradeoff between true positive rate and false positive rate of a classifier. The ROC graph of Naïve Bayes and SSVM algorithms is shown in Fig.11. The ROC has got number of properties depending on the value of its area under the curve. Furthermore our experimental results shows that SSVM found to be the best in terms of Accuracy (91.01%), Precision (0.91), Recall (0.95) and ROC (0.92) when compared to Naïve Bayes results in terms of Accuracy (78.51%), Precision (0.77), Recall (0.81) and ROC (0.78).

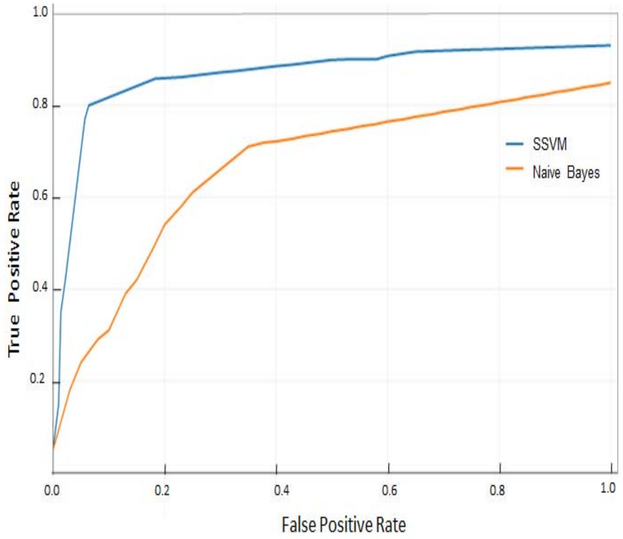


Figure 11. ROC graph – Naïve Bayes Vs SSVM

VIII. CONCLUSION AND FUTURE WORK

Social media platform combines human efforts and machine computation to process highly accurate tags and labels for subset of micro tweets in Twitter. It coordinates the role of humans and smart-technology to work together and improve disaster response efforts. From the experiments conducted using machine learning algorithms on tweets containing approximately 7500 tweets, our approach can detect the tweets containing this type of information, and generate an output that is correct 78% to 91% of the time. By considering the scalability of algorithm used in our work, we are extending this work with core vector machines. The geo coder map is generated which matches with almost 89% accuracy in actual flood affected areas. In future, streaming multi- window model can be deployed in distributed ecosystem with no single point of failure for real time decision support model during disasters. Moreover, random forests (supervised machine learning algorithm will be applied to train multi-level and multi-class classifiers under different disaster training data. The training can also be done semantically which is more similar to classifiers using the past category of disasters.

REFERENCES

[1] Dennis Thom, Robert Krcuger And Thomas Ertl, "Can Twitter Save Lives? A Broad-Scale Study On Visual Social Media Analytics For Public Safety", IEEE Transactions On Visualization And Computer Graphics, Vol. 22, No. 7, July 2016.

[2] Stuart E. Middleton, Lee Middleton, and Stefano Modafferi, "Real-Time Crisis Mapping of Natural Disasters Using Social Media", Social Intelligence and Technology, IEEE Computer Society, March/April 2014.

[3] Muhammad Imran and Carlos Castillo, "Towards a Data-driven Approach to Identify Crisis-Related Topics in Social Media Streams", Social Web for Disaster Management (SWDM) at WWW2015, Florence, Italy, 2015.

[4] Balaji Narasimhan, S. Murty Bhallamudi, Arpita Mondal, Subimal Ghosh, Pradeep Mujumdar, "Chennai Floods 2015 - A Rapid Assessment", Interdisciplinary Centre for Water Research, Indian Institute of Science, Bangalore ,May 2016.

[5] Amalorpavanathan, J., M Ramakumar, S Sivasubramanian, "Preparedness in Disaster Situations Lessons from Chennai Floods 2015", Economic & Political Weekly, vol L1, No. 8, pp. 30 – 34 2016.

[6] P.S. Earle, D.C. Bowden, and M. Guy, "Twitter Earthquake Detection: Earthquake Monitoring in a Social World", Annals Geophysics, vol. 54, no. 6, 2011.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors", Proc. Int. Conf. World wide web, pp. 851–860, 2010.

[8] C. Cotofana, L. Ding, P. Shin, S. Tilak, T. Fountain, J. Eakins, and F. Vernon, "An soa-based framework for instrument management for large-scale observing systems," in Proc. Int. Conf. Web Services, pp. 815–822, 2006.

[9] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics + filtering + search= twitcident. exploring information in social web streams," in Proc. 23rd ACM Conf. Hypertext Social Media, pp. 285–294, 2012.

[10] M. Guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," in Proc. 9th Int. Conf. Advances Intelligent Data Analysis, pp. 42–53, 2010.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proc. WWW, pp. 851–860, 2010.

[12] T. Suzumura and T. Oiki, "StreamWeb: Real-time web monitoring with stream computing," in Proc. Int. Conf. Web Services, pp. 620–627, 2011.

[13] Muhammad Imran, and Carlos Castillo, "Volunteer-powered Automatic Classification of Social Media Messages for Public Health in AIDR", Workshop in Public Health in the Digital Age Seoul, Korea, 2014.

[14] Fujio Toriumi , Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, Itsuki Noda, "Information Sharing on Twitter During the 2011Catastrophic Earthquake", Proceedings of the 22nd International Conference on World Wide Web, pp. 1025-1028, 2013.

[15] A. Bird, E. Klein, and E. Loper. 2009. Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit, O'Reilly Media (2009).