Summarization Approach From Microblog During Disaster Events

Pooja B. Kawade
Student of Masters(CSE)
Department of Computer Engineering
Maharashtra Institute of Technology, Pune

N. N.Pise, PhD
Associate Professor
Department of Computer Engineering
Maharashtra Institute of Technology, Pune

P. V. Kulkarni
Assistant professor
Department of Computer Engineering
Maharashtra Institute of Technology, Pune

ABSTRACT
During bulk convergence events such as natural disasters, microblogging platforms like Twitter are broadly used by affected people to post situational awareness messages. As soon as natural_disaster events happen, users are willing to know more about them. Twitter is a great source that can be exploited for obtaining such fine-grained arranged information for fresh natural disaster events. These crisis-related messages disperse among multiple categories like infrastructure damage, information about bomb blast, missing, injured, and dead people etc. The challenge here is to create summary from disaster related tweets and filter the short spam url containing tweets.

General Terms
URL Spaming Detection,Summarization Approach.

Keywords
Disaster events, Twitter, situational information, classification, summarization.

1. INTRODUCTION
Crisis situations such as disasters carried on by natural hazards present unique challenges to those who study them, creating conditions that call for particular research methods. In this paper, we survey approaches for studying disasters from the perspective of information processing and management.We know that information posted to social media platforms in time and safety critical situations can be of great value to those tasked with making decisions in these fraught situations.Microblogging sites like Twitter have become imperative sources of real time information during disaster events. A natural disaster is a main inimical event resulting from natural processes of the Earth; examples include floods, hurricanes,tornadoes, volcanic eruptions, earthquakes, tsunamis, and other geologic processes. A natural disaster can cause loss of life or properties damages, and typically leaves some economic damage in its wake, the severity of which depends on the affected populace's resilience, or ability to recover and also on the infrastructure available[16].In response to an event, a lot of short messages are posted on social media. Specifically, microblogging platforms such as twitter provide rapid access to situation-sensitive messages that people post during mass convergence events such as natural disasters.Studies show that these messages contain situational awareness and other useful information such as reports of urgent needs, missing or found people that, if processed timely, can be very effective for humanitarian organizations for their disaster response efforts. However, this information is immersed among hundreds of thousands of tweets, generally containing opinion of the masses, that are posted during such events. To effectively exploit microblogging sites during disaster events, it is essential to (i) Extract the situational information from among the large amounts of opinion, and (ii) Summarize the situational evidence.It helps to decision-making processes when time is critical and (iii) Spam URL detection to detect whether URL containing tweets are spam or notspam.Work on filter tweets gives the maximum accuracy.The American Red Cross (ARC), in a survey, reported the effectiveness of social media and mobile apps. ARC recently opened their Social Media Digital Operations Center for Humanitarian Relief. The aims of this center are to source additional information from affected areas during emergencies to better serve those who want help; spot trends and better anticipate the publics needs; and connect people with the resources they need, like foodstuff, water, shelter or even emotional support[9]. Typically, the first step in extracting situational awareness information from these tweets involves classifying them into different informational categories such as infrastructure damage, shelter needs or offers, relief supplies.

2. RELATED WORK
Aforementioned research has shown that information which contributes to situational awareness is reported via Twitter. The list of work on disaster events and the list of survey on methods of detection of URL spamming below In [3], Axel Brun-2013, the authors have proposed a solution for tracking hashtags. YourTwapperkeeper is the open source tool. Building on PHP and MySQL, it draws mainly on the Twitter streaming API to track a number of keywords nominated by its user, using the search API to fill any gaps which may exist in the data received from the streaming API. In [8], Chao Shen-2013, propose a participant-based event summarization method that zooms-in the Twitter event streams to the participant level TF-IDF approach to extract the representative sentences from a collection of tweets In this approach, each tweet was considered as a sentence. The sentences were ranked according to the average TF-IDF score of the consisting words; top weighted sentences were iteratively extracted, while excluding those that have high cosine similarity with the existing summary sentences In [4], Olariu-2014, the Olariu introduces us to TOWGS, a highly efficient algorithm capable of online abstractive microblog summarization. A word graph, along with optimization techniques such as decaying windows and pruning is introduced.In [1], Koustav Rudra-2015, a novel content-word based summarization approach (COWTS) to summarize the situational tweet stream by optimizing the coverage of important content words in the summary, using an Integer

Linear Programming (ILP) framework. The authors have recommended working with tweet fragments rather than entire tweets. Distinct lexical and syntactic features present in tweets can be used to separate out situational and non-situational tweets, which leads to significantly better summarization. In his work [12], Alan Ritter has introduced the first open domain event-extraction and categorization system for twitter, named TwiCal. A scalable and open-domain approach to extracting and categorizing events from status messages. In [2], Koustav Rudra 2016, an Integer-linear programming (ILP) based optimization technique and content word based abstractive summarization technique to produce the final summary. They have developed a complete system to generate summaries in real time from the incoming stream of tweets. In[18], Xianghan Zheng et. al. They have introduced a machine learning based spammer detection solution for social networks. The solution considers the user's content and behavior feature, and apply the min to SVM based algorithm for spammer classiffication. Through a multitude of analysis, experiment, evaluation and prototype implementation work, they have shown that proposed solution is feasible and is capable to reach much better classification result than the other existing approaches.In [19], Sangho Lee et. al. introduces us to The major goal of the WARNINGBIRD is to detect the suspicious URLs. Suspicious URLs are nothing but the doubtful URLs which contains malicious elements. Malicious elements like viruses, malwares, phishing etc. Conventional twitter suspicious URL detection system is based on correlated URL redirect chain methodology. It detects the suspicious URLs which were often shared. It will examine the correlated URL redirect chain and tweet context information to detect the suspicious URLs.In[21], Hailu Xu introduces us to a new perspective to distinguish between spam and legitimate contents in Twitter and Facebook. They collected two datasets through their APIs and analyzed their content information. They used several traditional classifiers such as Random Forest, Random Tree, J48, Logistic, and Nave Bayes to evaluate these two original datasets. Random Forest shows the best performance with a nearly 94.7percent accuracy and 66percent recall for Twitter Spam dataset, and 97.7percent accuracy and 84.4percent recall for Facebook Spam dataset.

3. OUR APPROACH

User tweets collected from dataset.This system perform all operation on this dataset and used detection of URL spamming to detect whether URL containing tweets are spam or not spam with the help of random forest algorithm. This type of system never used this type of filter before.This system used url spam detection filter to improve accuracy. Because work on fake tweets give us the fake and less accuracy result.Tweet analysis is important to analyze that our dataset have valid data or not for work. Analysis come up with the graph.Tweet classification is most important function in this system.This function have done using Naïve Bayes algorithm.In tweet classification done category flood,bombblast,earthquick.For better accuracy of summarization this system provided some names that can used by people when disaster happens. I. for flood-missing, shelterless, foodless, destroy. II.for bomblast-damage,dead,injured,collapse. III.for earthquick-die,sound,area,ruined.Tweet Summarization have done using Random generation abstractive summarization technique and Genetic algorithm.This new approach implemented by using Java.

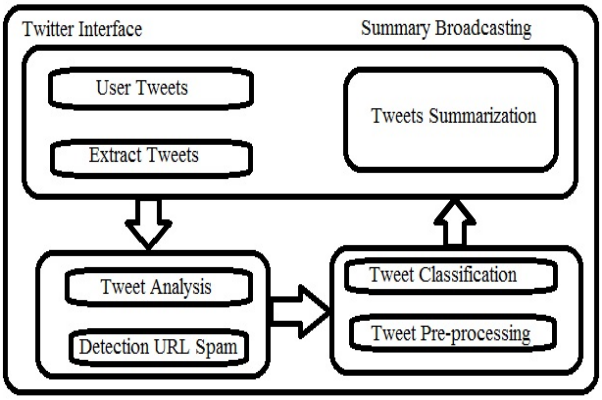

Fig1.System Architecture

Twitter Interface-
a.User tweets along with the all user details.
User Tweets-
a.Contains the user tweets with user details.
Extract Tweets-
a.Extract Tweets from dataset.
Tweet Analysis-
a.Analyzing Extracted tweets to normalize.
Detection of Spam URL-
a.Random Forest algorithm to detect URL containing tweets are spam or not spam..
Tweet Pre-Processing-
a.Stemming algorithm for tweet Pre-Processing.
Tweet classification-
a.Naïve Bayes algorithm for tweet Classification For example: Bomb blast, Floods.
Category-
I. for flood-missing, shelterless, foodless, destroy.
II.for bomblast-damage,dead,injured,collapse.
III.for earthquick-die,sound,area,ruined.
Tweet Summarization-
a.Random generation abstractive summarization technique and genetic algorithm used for tweet summarization.The use of genetic algorithm, was an idea to avoid problems with local search techniques. Local search may find a local maximum and declare it as the answer.In our problem, the goal is to find a summary with high readability, high cohesion, and high topic relation. One approach to find such a summary using Genetic Algorithm.
Summary Broadcasting-
a.Final Summary generated.

4. EXPERIMENTAL SETUP AND SYSTEM

In this section, we compare the performance of our proposed framework with state-of-the-art abstractive and disaster-specific summarization techniques. We first describe the COWABS technique as well as the experimental settings.It classified messages from three classes 1.Flood 2.Earthquake 3.Bomb blast. We perform this data-wise split from Adataset.This Dataset have tweets of crisis events.Adataset downloaded from www.crisislex.org .

Table 1.Comparison of ROUGE-1 recall (with classification, Twitter specific tags,hashtags, mentions, urls, removed and standard rouge stemming and stopwords for GAST(the Proposed methodology)

Rouge1 Recall score				
	Flood	**Earthquake**	**Bomb blast**	**Normal**
COWABS	0.97	0.026	0.0186	0.232
GAST	0.98	0.027	0.0186	0.356

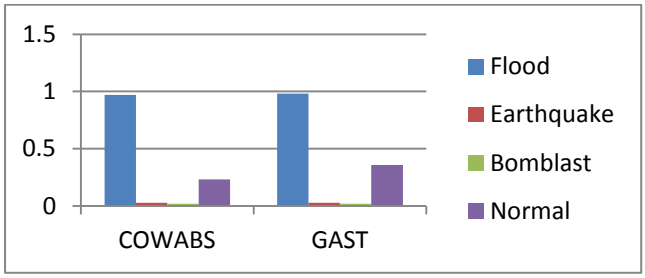

Fig 2. ROUGE-1 Recall Score Graph

Redundancy in summaries: Apart from ROUGE-1 score, we also measure redundancy score of the summaries as this can indicate if the summaries contain distinct or redundant information. We compute redundancy score for a summary as follows: For each sentence included in the summary (a sentence can be a tweet or a path), we assign it a sentence redundancy score as its maximum cosine similarity (excluding #,@,URLs,stopwords) with any other sentence in the summary. Finally, we take an average of the individual sentence redundancy scores to compute the redundancy value for the summary.Table shows redundancy values of different methods.

Evaluation using crowdsourcing: Next, we perform crowdsourced evaluation using the CrowdFlower crowdsourcing platform.

Quality of Information Summarized: Beyond the mere numbers proving our superiority, we also looked into the tweets and checked its quality with respect to (a). number of distinct places mentioned (b). number of event phrases used and (c). extent of numbers present in the summary. Details of which follow - Location coverage: During large scale disaster like earthquake, flood et al. several parts of a country are damaged and coverage of information from these different places are necessary. Location coverage corresponds to the information about different places a summary contains. For instance, a summary with diverse information from many locations is considered better in terms of location coverage. The problem is challenging in the sense that there is overwhelmingly more information about big cities/towns in the tweet. For example, during Nepal earthquake most of the information are available in Twitter from its capital city Kathmandu but there is a scarcity of information from local villages like Barpak, Lamjung etc. Our proposed GAST is able to capture information about more number of locations.

Table2.Redundancy score for different methods of summarization

Redundancy Score			
	Flood	**Earthquake**	**Bomb blast**
COWABS	0.1775	0.2099	0.1433
GAST	0.1675	0.1902	0.1432

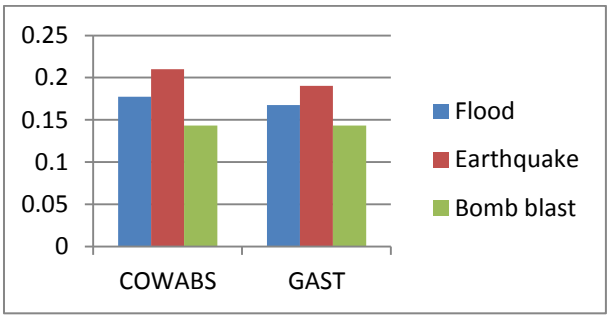

Fig 3. Redundancy Score Graph

Length of summary-Here,Two methods are compared COWABS and Genetic algorithm summarization text.From the COWABS method we got the length 369 words and from the GAST we got the 256 words.From this length of summary experiment we get less words summary in GAST.

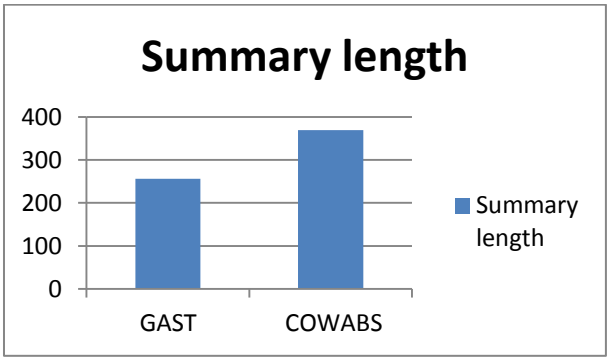

Fig 3. Summary Length Graph

Summary by COWABS:Times of india live blog earthquake in kathmandu , 25 04 2015. Chairs follow-up meeting to review situation following earthquake in decades.5 commercial flights have landed in kathmandu was painted in 1850 ad. Iaf's c-130j aircraft carrying 55 passengers , 30 people dead in Kathmandu.including four infants , lands at delhi's palam airport. Nepal quake photos show historic buildings reduced to rubble as survivor search continues.

Summary by GAST:

ayo alert at kathmandu 29 people dead commercial ights have landed in kathmanduInfrastructure ruined at kathmandualert Ayo bombast at Kathmandu 200 people injured kathmandu high building collapse.

5. CONCLUSION AND FUTURE SCOPE

This system used to collect different tweet during disaster for tweet summary creation. Additionally used URL spamming detection to know that URL containing tweets are spam or not.GAST gives the best summarization result. We have developed a complete system to generate summaries from incoming tweets.We have specifically taken the tweets generated during disasters events and generate comphrehensive abstractive summaries for three important class-Bomblast,Flood,Earthquake.We have performed an extensive evaluation of our algorithm by roping in disaster related experts in the loop-results show that our GAST perform significantly better than all existing approaches. It is important to filter spam URL in the system to give the best summarization result. Because this system is useful for the public or formal response organizations, that it has the prospective to save lives or property during an emergency.In future scope, If this system works in a real time way it will give more benefit. Because when we get alert from real time streaming tweets it will easy to serve help to needy people.

6. ACKNOWLEDGMENT

I express my deepest sense of gratitude towards my respected guide Dr.N.N.Pise and Prof.P.V.Kulkarni for technical support, valuable guidance, encouragement and consistent help without which it would have been difficult for me to complete this work.Special thanks to my family, friends and well wisher for their valuable support and those who are directly and indirectly involved in making of this work possible.

7. REFERENCES

[1] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Extracting Situational Information from Microblogs during Disaster Events: a Classification- Summarization Approach. CIKM, Melbourne, VIC, Australia, 2015.

[2] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Summarizing Situational Tweets in Crisis Scenario. HT , Halifax, NS, Canada,2016.

[3] Axel Bruns,Yxian Liang, Tools and Methods for capturing twitter data during natural disaster. In First Monday, Volume 17, Number 4-2April 2013.

[4] Andrei Olariu, Eficient Online Summarization of Microblogging Streams. In Pro-ceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 236-240,April 26-30 Gothenburg, Sweden, 2014.

[5] Sandeep Panem, Manish Gupta, Vasudeva Varma,Structured Information Extraction from Natural Disaster Events on Twitter. KDD , xian,china,2014.

[6] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, On Summarization and Timeline Generation for Evolutionary Tweet Streams. In IEEE Transactions on Knowledge and Data Engineering, DOI 10.1109/TKDE, 2013.

[7] Muhammad Imran, Fernando Diaz, Carlos Castillo, Processing Social Media Messages in Mass Emergency: A Survey ACM Computing Surveys, Vol. 47, No. 4,Article 67,2015.

[8] Chao Shen, Fei Liu, FuliangWeng,Tao Li, A Participant-based Approach for Event Summarization Using Twitter Streams,KDD,Xian,China,2014.

[9] Muhammad Imran, Carlos Castillo, Ji Lucas, AIDR: Artificial Intelligence for Disaster Response. In WWW Companion, Seoul, Korea,April 7,2014.

[10] Sarah Vieweg, Carlos Castillo, and Muhammad Imran, Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. In Springer LNCS 8851, pp. 444-461,2014.

[11] Miles Osborne, Elizabeth Cano, Craig Macdonald, Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media,2014.

[12] Alan Ritter, Mausam, Open Domain Event Extraction from Twitter. In KDD,Beijing, China,August 1216, 2012.

[13] Robert Power, Bella Robinson, John Colton, Emergency Situation Awareness:Twitter Case Studies. In Springer ISCRAM-med , pp. 218231, 2014.

[14] Pengyi Zhang, Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In CSCW , Hangzhou, China March 19-23, 2011.

[15] https://en.wikipedia.org/wiki/Natural_disaster

[16] Chao Chen, Jun Zhang, Wan lei Zhou, 6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection. IEEE ICCCommunication and Information Systems Security Symposium, 2015 .

[17] Shigang Liu, Jun Zhang, Yang Xiang, Statistical Detection of Online Drifting Twitter Spam. ASIA CCS, Xian, China,May 30June 3, 2016.

[18] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Detecting spammers on social networks. Neurocomputing159(27-34),2015.

[19] Sangho Lee, Jong Kim, WARNINGBIRD: A Near Real-time Detection System for Suspicious URLs in Twitter Stream. IEEE Transaction On Dependable And Secure Computing, Vol. 10, no. 2, January 2013.

[20] Sandeep Kumar Rawat, Saurabh Sharma,A Review on Spam Classification of Twitter Data Using Text Mining and Content Filtering. In International Journal of Advanced Research in Computer Science and Software Engineering Volume 5,Issue 6, pp. 485-488,June 2015.

[21] Hailu Xu, Weiqing Sun, Ahmad Javaid, Eficient Spam Detection across Online Social Networks.In Big Data Analysis (ICBDA), IEEE International Conference12-14, Hangzhou, China,March 2016 .

[22] Jyoti D. Halwar, Sandeep Kadam, Vrushali Desale,Detection of Suspicious URL in Social Networking Site Twitter: Survey Paper . In International Journal of Computer Applications Volume 110 No. 8, January 2015.

[23] Monika Verma,Sanjeev Sofat, Techniques to Detect Spammers in Twitter- ASurvey. In International Journal of Computer Applications Volume 85 No 10,January 2014.

[24] Guofei Gu, Chao Yang, Amit A. Amleshwaram, CATS: Characterizing Automation of Twitter Spammers.In Fifth International Conference,Bangalore, India,7-10 Jan 2013.

[25] Chao Yang, Robert Harkreader, and Guofei Gu, Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers.In IEEE Transactions on

Information Forensics and Security Volume: 8, Issue: 8, Aug. 2013.

[26] Nasim eshraqi, Mehrdad Jalali, Mohammad Hossein Moattar, Detecting Spam Tweets In Twitter Using a Data Stream Clustering Algorithm Second International Congress on Technology, Communication and Knowledge (ICTCK) November, Mashhad Branch, Islamic Azad University, Mashhad, Iran,11-12jan, 2015 .

[27] Cheng Cao and James Caverlee, Detecting Spam URLs in Social Media via Behavioral Analysis. In ECIR, LNCS 9022, pp. 703714, Springer International Publishing Switzerland,2015.

[28] Sangho Lee, Jong Kim, Early filtering of ephemeral malicious accounts on Twitter Pages 48-57, Volume 54, 1 December 2014.

[29] Nikitha.R, Anand.R, Detection of Suspicious URLs through Vision Techniques in Twitter Stream. In International Journal of Advancements in Research and Technology, Volume 3, Issue 5, May-2014.

[30] Nour El-Mawass and Saad Alaboodi, Hunting for Spammers: Detecting Evolved Spammers on Twitter. In arXiv:1512.02573v2 [cs.IR] 15 Dec 2015.