EXTRACTING AND SUMMARIZING INFORMATION FROM MICROBLOGS DURING DISASTERS

Koustav Rudra

EXTRACTING AND SUMMARIZING INFORMATION FROM MICROBLOGS DURING DISASTERS

Thesis submitted to the
Indian Institute of Technology Kharagpur
For award of the degree

of

Doctor of Philosophy

by

Koustav Rudra

Under the supervision of

Prof. Niloy Ganguly

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

April 2018

APPROVAL OF THE VIVA-VOCE BOARD

Date:

Certified that the thesis entitled **“Extracting and Summarizing Information from Microblogs during Disasters”** submitted by Koustav Rudra to the Indian Institute of Technology, Kharagpur, for the award of the degree of Doctor of Philosophy has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

Prof. Sourangshu Bhattacharya (Member of DSC)	Prof. Bivas Mitra (Member of DSC)	Prof. Anirban Mukherjee (Member of DSC)
Prof. Niloy Ganguly (Supervisor)	Prof. Utpal Garain (External Examiner)	Prof. Sudeshna Sarkar (Chairman/HoD CSE)

CERTIFICATE

This is to certify that the thesis entitled **"Extracting and Summarizing Information from Microblogs during Disasters"**, *submitted by* **Koustav Rudra** *to the Indian Institute of Technology Kharagpur, for the award of the degree of Doctor of Philosophy, is a record of bona fide research work carried out by him under our supervision and guidance. The thesis, in our opinion, is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute. To the best of our knowledge, the results embodied in this thesis have not been submitted to any other University or Institute for the award of any other Degree or Diploma.*

Niloy Ganguly
Professor
Computer Science and Engineering
IIT Kharagpur

Date:

DECLARATION

I certify that

a. The work contained in this thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Koustav Rudra

ACKNOWLEDGMENTS

A PhD is a long and onerous journey, not completed without the help and support of many. I would like to take this opportunity express my gratitude to everybody who have inspired me, provided me mental support, and contributed directly or indirectly in realizing this thesis. My parents have been the major source of inspiration and strength throughout my life and without their support and sacrifice, this thesis would never have taken shape.

I would like to express my gratitude to my supervisor, Prof. Niloy Ganguly for his guidance, support and encouragement. The discussions with him not only helped me in formulating the problems of this thesis, but also helped me to realize the real motivation behind doing the research. I would like to thank Prof. Pawan Goyal (IIT Kharagpur), Prof. Saptarshi Ghosh (IIT Kharagpur), Dr. Muhammad Imran (QCRI, Qatar), Dr. Prasenjit Mitra (Penn State, USA), Dr. Monojit Choudhury (Microsoft Research, Bangalore), and Dr. Kalika Bali (Microsoft Research, Bangalore). I would also like to thank the members of my DSC Prof. S. Bhattacharya, Prof. B. Mitra, and Prof. A. Mukherjee who have enriched this thesis with their insights and suggestions.

I consider myself very fortunate to have been part of Complex Network Research Group (CNeRG). I want to extend my gratitude towards all my CNeRG lab mates Abhijnan Chakraborty, Abhik Jana, Amrith Krishna, Abir De, Ankan Mullick, Madhumita Mallick, Mayank Singh, Parantapa Bhattacharya, Rijula Kar, Rohit Verma, Soumajit Pramanik, Sandipan Sikdar, Soumya Sarkar, Satadal Sengupta, Suman Kalyan Maity, Subhendu Khatuya, Sankarshan Mridha, and Surjya Ghosh.

I would also like to thank my student co-authors Ashish Sharma, Shruti Rijhwani, and Siddhartha Banerjee.

Finally, I am grateful to God who provide me mental strength to carry out this long and arduous journey.

Koustav Rudra
Kharagpur, India

Author's Biography

Koustav Rudra received his B. E. (Hons.) degree in Computer Science and Engineering from Bengal Engineering and Science University Shibpur in 2011 and M. Tech. degree in Computer Science and Engineering from Indian Institute of Technology Kharagpur in 2013. He was awarded the Tata Consultancy Services Ph.D. Fellowship in 2014.

Publications made out of this thesis (at the time of defence)

1. *K. Rudra*, N. Ganguly, P. Goyal, and S. Ghosh. Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters. *ACM Transactions on the Web (TWEB)*, 2018.

2. *K. Rudra*, A. Sharma, N. Ganguly, and S. Ghosh. Characterizing and Countering Communal Microblogs during Disaster Events. *IEEE Transactions on Computational Social Systems*, DOI: 10.1109/TCSS.2018.2802942, 2018.

3. *K. Rudra*, A. Sharma, N. Ganguly, and M. Imran. Classifying and Summarizing Information from Microblogs during Epidemics. *Special Issue on "Exploitation of Social Media for Emergency Relief and Preparedness" in the journal Information Systems Frontiers (Springer)*, 2018.

4. *K. Rudra*, A. Sharma, N. Ganguly, and M. Imran. Classifying Information from Microblogs during Epidemics. In *Proceedings of the 7th ACM Digital Health Conference (DH 2017)*, London, July 2-5, 2017.

5. *K. Rudra*, A. Chakraborty, N. Ganguly, and S. Ghosh. Understanding the Usage of Idioms in Twitter Social Network. *Pattern Recognition and Big Data, World Scientific*, Pages 767–788, February 2017, (Editors: Amita Pal, Sankar K Pal), (DOI: https://doi.org/10.1142/9789813144552_0024), (ISBN: 978-981-3144-54-5).

6. *K. Rudra*, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra. Summarizing situational tweets in crisis scenario. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, Pages 137–147, 2016.

7. *K. Rudra*, A. Sharma, N. Ganguly, and S. Ghosh. Characterizing Communal Microblogs during Disaster Events. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Pages 96–99, 2016.

8. *K. Rudra*, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra. Summarizing situational tweets in crisis scenario. Accepted in *the 4th International Workshop on Social Web for Disaster Management (SWDM'16) co-located with CIKM 2016*, (arXiv preprint: https://arxiv.org/abs/1610.015610).

9. *K. Rudra*, S. Rijhwani, R. Begum, K. Bali, M. Choudhury, and N. Ganguly. Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pages 1131–1141, 2016.

10. *K. Rudra*, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, Pages 583–592, 2015.

11. *K. Rudra*, A. Chakraborty, M. Sethi, S. Das, N. Ganguly, and S. Ghosh. #FewThingsAboutIdioms: Understanding Idioms and Its Users in the Twitter Online Social Network. In *Proceedings of the 2015 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Pages 108–121, 2015.

ABSTRACT

In this thesis, we focus on extracting and summarizing information from Twitter during disaster and explore specific traits of situational and non-situational tweets to develop methods which is able to extract relevant information and assist government, rescue agencies in their work.

Prior researches on Twitter have shown that microblogging sites like Twitter have become important sources of real-time information during disaster events. A significant amount of valuable *situational information* (information provide updates about current situation) is available in these sites; however, this information is immersed among hundreds of thousands of tweets, mostly containing sentiments and opinion of the masses, that are posted during such events. To effectively utilize microblogging sites during disaster events, it is necessary to (i) extract the situational information from among the large amounts of sentiment and opinion, and (ii) summarize the situational information in real-time, to help decision-making processes when time is critical. In this thesis, we propose a low level lexical feature based situational tweet classifier which classifies situational tweets from non-situational ones. After separating situational tweets, we observe that some specific words like nouns, numerals, locations, verbs provide key information about the present situation. We call these words *content words* and propose an integer linear programming based summarization framework which tries to maximize the coverage of content words. Side by side, certain numerical information, such as the number of casualties, vary rapidly with time. We also devise a scheme where we utilize the direct objects of disaster-specific verbs (e.g., ‘kill’ or ‘injure’) to continuously update important, time-varying actionable items such as the number of casualties. We observe that apart from English, people also post situational updates in their local languages (predominantly Hindi in India). In this thesis, we also extend our classification-summarization framework to Hindi tweets.

These large volume of situational tweet streams are scattered across various humanitarian categories like ‘infrastructure damage’, ‘missing or found people’ etc. We also observe that each of these humanitarian categories contain information about various small scale sub-events like ‘airport shut’, ‘building collapse’ etc. We develop a noun-verb pair based method to detect sub-events which are more explainable compared to random collection of words. It is observed that different stakeholders are looking for different kinds of summaries like overall high level

summary, humanitarian category based summary etc during disaster. To satisfy their needs, we develop an ILP-based generic summarization technique which combines information about sub-events, content words, and humanitarian categories to generate summaries from various perspectives.

Further, we observe that lots of situational tweets posted during disaster contain similar information with slight variations. Combining information from multiple related tweets help to cover more situational information in a summary within a given word limit. In this thesis, we develop an abstractive summarization method which creates word graph from tweets, generates path from the word graph, and combines path importance and content words into an ILP framework to produce final summary.

It is observed that taking advantage of panic situation, some people post offensive content targeting specific religious communities during disaster. Such communal posts deteriorate law and order situation. In this thesis, we have developed method to detect such communal tweets and characterize their users. Non-situational tweets are mostly used for expressing opinion and sentiment of masses. We observe that users mostly prefer vernacular languages such as Hindi over English to post communal tweets, negative sentiments, and slangs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online social networks (OSNs) and Microblogging sites like Twitter and Facebook are currently important sources of information on the web [88]. Now a days they are extensively used for communication purpose by a diverse set of users. For example, common people use these platforms to keep in touch with their friends and families, celebrities use these mediums to communicate with their fans, promote their upcoming movies, business organizations stay connected with their employers, clients etc.

Apart from communication, in recent years, social media sites are also helpful in gathering information on various topics (sports, politics, movies etc.) and current events. Especially, Twitter is increasingly being used to gather real-time information on events happening 'now', including disasters, emergency situations, political / social movements, sports updates, and so on [102, 132, 143].

Recent researches [20, 82, 111, 129] have shown the importance of microblogging sites in enhancing situational awareness [113], i.e., information which helps the concerned authorities (e.g., Governmental and non-governmental agencies) to gain a high-level understanding of the situation during disasters. In fact, recent research shows that Twitter reports the same events as any news media sites (e.g., Newswire) do, and even captures many minor events which are ignored by news providers [98]. In some cases, updates about disaster are available in Twitter much earlier than news media.

In a disaster situation, various types of information, including situational updates, sentiment (e.g., sympathy for those affected by the disaster) and personal opinion (e.g., on the adequacy of relief operations) are posted by users in *huge volume and at rapid rates*. Broadly, two categories of tweets are posted during disaster — (i). Situational tweets which provide situational information, and (ii). Non-situational tweets which consist of opinion of masses. Different types of information have different utilities, *situational information* (including the actionable items [129] such as the number of affected people) is critical for the authorities to understand the situation and plan relief efforts accordingly. On the other hand, common masses express their grievances, thoughts via non-situational tweets. Often, taking advantage of such disaster situation, hatred, communal venom and misinformation are also propagated in the affected zone, which may result in serious deterioration of law and order situation. Social media acts as a fertile ground in spreading hatred and specially Twitter is increasingly used as a powerful tool [17].

Most of the existing studies focus on events in only a few countries such as the USA, European countries and Japan [43, 44, 111, 132] where Twitter is used by large fractions of the population. Only a few studies have focused on emergencies in countries such as India [42]. There are additional challenges while dealing with disaster situations in countries such as India where usage of online social networks is not so prevalent, including scarcity of data [2], lack of updates by authoritative users [42], and so on. Side by side, lots of information are posted in local languages like Hindi. Hence mechanisms to utilize information posted in other languages (Hindi) during emergency situations in India need to be developed.

1.1 Motivation

Recent researches [20, 111] established that social media can work as a sentinel during emergency situations. However, there exist several challenges in utilizing information from social media. Some common challenges in utilization of Twitter during emergencies are described below.

1.1.1 Non-situational tweets and irrelevant content

As stated earlier, during disaster lots of opinions are also posted along with situational tweets. However, such opinions do not carry any meaningful information which can be used for decision making purpose. It is necessary to filter out situational tweets from non-situational ones because unfiltered content creates problems for users in terms of information processing [40]. We observe that on an average 60% tweets are situational. As huge amount of tweets are posted within small interval of time, it is not feasible for human beings to properly separate them into two classes. Hence it is important to develop automated methods to *extract microblogs / tweets which contribute to situational information* [56, 57, 130].

1.1.2 Information overload and summarizing content

In response to an event, a lot of 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 [131]. Enabling rapid crisis response requires processing of these messages as soon as they arrive. However, even after the automatic classification step, it still contains thousands of important situational messages—also increasing each passing minute. Hence, this requires a coherent situational awareness summary preparation for disaster managers to understand the situation. To get a quick overview of the event and what Twitter is saying about it, a summary of these tweets is very valuable [25, 34, 86, 87, 115]. Sometimes simple extractive summaries where sentences / tweets are selected based on importance are not enough; combining information from related tweets, i.e., producing abstractive summaries are useful in reducing information overload [34, 87].

1.1.3 Communal content

The huge amount of tweets posted during a disaster event include information about the present situation as well as the emotions / opinions of the masses (non-situational tweets). Recent researches [14, 16, 17] showed that a large amount of communal tweets, i.e., abusive posts targeting specific religious / racial groups are posted during disaster scenario. A disaster generally affects the morale of the masses making them vulnerable. Posting of such communal contents during disaster may deteriorate the law and order situation and affect the communal harmony. It is necessary to filter out such posts as soon as they are tweeted and also counter their adverse effects.

1.1.4 Information extraction from non-English (Hindi) tweets

As stated earlier, most of the prior works [43,44,111,132] tried to retrieve situational information from English tweets. However, in developing countries like India, usage of social media is not popular (Out of 1.34 billion users around 250M users use social media in India[1]). Further, only 10% users in India speak English[2]. After disaster, most of the users post tweets in local languages mainly in Hindi. Hence, it is useful to extract information from both English and Hindi tweets specially in Indian context.

1.2 Objectives of the thesis

The primary goal of this thesis is to develop methodologies to assist government, NGOs in their decision making process during disaster. During disaster, different stakeholders are looking for different kind of information. For example, some volunteers are looking for infrastructure related damages, some NGOs are looking for

[1] https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/

[2] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

updates about missing and injured persons, and so on. Governmental organizations try to get an overall high level view of the current situation. On the other hand, non-situational communal tweets pose a threat to the communal harmony and law order situation. Hence, detection and filtering of such messages is necessary under such situation. Considering the scarcity of the data and use of Hindi in India, gathering situational information from Hindi tweets and analyzing opinions to understand language preference is also necessary.

In this thesis, we are trying to extract and summarize information from Twitter microblog. Twitter is a popular microblogging platform consists of 328 million monthly active users. Around 500 million tweets are posted daily on Twitter. In Twitter, most of the data streams are public and Twitter also provides public APIs to collect tweets in a structured format [127]. Hence, collection of tweets for information extraction and summarization purpose is relatively easy compared to other social medias like Facebook.

To achieve our overall goal to assist government, different NGOs, and stakeholders during disaster scenario and analyze the tweets posted during disaster, we set forth the following four objectives:

1.2.1 Objective 1: To identify situational tweets and summarize information during crises

Large volume of information is posted on social media during disasters. For instance, the largest observed peak was during the Sandy hurricane in which around 16 thousand messages per minute were posted using hashtag #Sandy. However, all of these information are not important from situational point of view; non-situational tweets carry personal opinion and sentiment of masses whereas situational tweets carry important information about infrastructure damages (collapse of roads, bridges), missing or trapped or injured people, shelter and volunteer related information and so on. It is very time consuming for humans to manually segregate situational tweets from non-situational ones and process them accordingly. Therefore, an automated methodology is required for this classification task. Standard bag-of-words / vocabulary based classifiers [130] depend on the words used during a particular event and vocabulary used in various disaster events like flood, earthquake, typhoon, shooting, bomb blast etc. are quite different. Hence, standard bag-of-words based classifier trained on the vocabularies used in a particular event do not perform well if applied directly for future disaster events due to vocabulary mismatch. Our objective is to develop a situational tweet classifier which is independent of the vocabularies used during an event and can be directly applied over future disaster events.

Even after classification, the number of messages is quite high. A related, yet different, challenge is to deal with the rapid rate at which microblogs are posted during such events — this calls for summarization of the situational information. Further, some of the situational information, such as the number of casualties or injured / stranded persons, changes rapidly with time, asking for special treatment. Since time is critical in a disaster situation, these classification and summarization tasks have to be performed in near real-time, so that the processed information is readily available to the authorities. Side by side, the tweets posted in a disaster scenario have certain specific traits, which can be exploited for the above tasks. Our objective is to develop a real-time summarizer which can exploit specific peculiarities of the situational tweets posted during disaster.

During disasters in countries like India, we observe that a large number of tweets are posted in local resource-poor languages like Hindi. Some of these information are not available in English tweets, whereas sometimes we get such information much earlier from the Hindi tweets, as compared to what we get from the English tweets. Hence, our objective is to develop classification-summarization framework to handle both English and Hindi tweets.

1.2.2 Objective 2: To identify sub-events and summarize information at different levels of granularity

Microblogging platforms like Twitter provide real-time information about disasters. Tweets posted during a particular event contain diverse set of topics or sub-events [1, 99]. Identifying these sub-events help end users to get a quick overview of the situation. For example, in case of disaster, information about collapsing of bridges, blockage of roads etc. helps government, responding agencies, and users in their planning and rescue mission. On the other hand, going through a set of tweets is quite tedious and time consuming. Traditional sub-event detection methods identify sub-events as clusters of related tweets and select high frequency words from each cluster to describe a sub-event. We observe that words selected to represent a cluster of tweets may not be semantically related and they are difficult to grasp for end users. Our objective is to represent sub-events through semantically related words which can describe a topic in a more meaningful way and different stakeholders can get a clear view of the present situation from those words.

When a disaster happens, responders to disasters use such information obtained from microblogging sites to plan and respond to the needs of people located in disaster areas. Volunteers and other support personnel generate summaries and reports based on short messages such as tweets posted via Twitter that are then used by the responders [55]. Different stakeholders and responders need information at varying levels of granularity. Some stakeholders may want to obtain overall situational updates for a given day as a short summary or report (overall high-level information need) or specific updates for a particular class or category such as ‘infrastructure damage’, ‘shelter’ etc. (humanitarian category-based need). They may also want to get updates at a much finer granularity with very specific focus on events, persons and locations connected with the disaster. For example, one may not only be interested in ‘missing people’, but, more specifically, they may be interested in finding out about the Australian mountain climbers who were at the foothills of Mt. Everest when the earthquake hits Nepal.

Our objective is to filter out relevant tweets from the huge set, classify the rapidly streaming tweets posted after a disaster into appropriate categories, extract small scale sub-events from tweets, and finally create a summary of these tweets generated from multiple perspectives tailored towards different needs of the different stakeholders.

1.2.3 Objective 3: To generate abstractive summaries during disaster

Tweets posted during disasters contain similar information with slight variations. For example, if we consider the following two tweets — (i). *7 people died, 20 injured in bomb blast* and (ii). *7 died, 20 injured in Hyderabad blast*, is posted during Hyderabad blast, we get very similar information from both the tweets except that the second tweet also provides information about the location of the blast. In case of extractive summarization, we have to select both the tweets to cover complete information. Considering the word limit in final summary, it is better to combine information from multiple related tweets, i.e., to produce abstractive summaries. Abstractive summaries are helpful to capture information from different dimensions within a specific word limit compared to simple extractive summarization. Hence, it is useful in terms of information coverage (cover various situational information), diversity (situational information from different dimension like ‘road blockage’, ‘building collapse’ etc.), and redundancies (less repetitive information). Our objective is to produce class specific (‘infrastructure damages’, ‘missing or trapped person’, ‘injured and dead person’, ‘shelter and service’ etc.) abstractive summaries from the situational tweet streams in real-time.

1.2.4 Objective 4: To analyze non-situational tweets

Non-situational tweets posted during disaster scenario contain various kinds of tweets like sentiments / opinions of masses, analysis about the disaster event, communal tweets targeting specific religious communities and so on. Out of these classes, communal tweets are potentially dangerous during disaster because their

Table 1.1: Examples of communal tweets posted during disasters.

Communal Tweets
F**k these Missionaries who are scavenging frm whatever's left after the #NepalEarthquake Hav some shame & humanity
RT @polly: #HillaryClinton's reply when asked if war on terror is a war on "radical Islam" #DemDebate

spread may lead to deterioration of the law and order situation, hamper communal harmony etc. During man-made disasters like bomb blast or shooting, common people target specific religious communities to which attackers belong [16, 17]. However, we observe that such kind of communal tweets also exist in case of natural disasters like earthquake and flood [110]. Table 1.1 shows examples of communal tweets. These communal tweets are posted not only by common people but also by popular politicians, media houses etc. who are followed by thousands of users in Twitter. Hence, such communal tweets receive high exposure.

Along with many users posting communal tweets, there are some users who post anti-communal content, e.g., asking people to stop spreading communal posts. However, we observe that the users who post anti-communal content are, in general, far less popular (have far fewer followers) than the users who post communal content.

Considering the detrimental effect of communal tweets, our first objective is to automatically identify such tweets, analyze users who posted these tweets, identify anti-communal posts, and use them to counter the effect of communal tweets.

1.3 Contributions of the thesis

The thesis fulfills the four fold objective laid out in the previous section. In the first part of the thesis, we develop a vocabulary independent situational tweet classifier. After classifying situational tweet stream, we propose an integer linear programming (ILP) based summarization approach which tries to maximize the coverage of *content words* (nouns, verbs, numerals, locations). We also devise a scheme where we utilize the direct objects of disaster-specific verbs (e.g., ‘kill’ or ‘injure’) to continuously update important, time-varying actionable items such as the number of casualties. Apart from English tweets, we also extend this classification-summarization framework to Hindi tweets. In the second part, we develop a noun-verb pair based sub-event detection method and *content words*, *sub-events* based summarization approach to produce summaries at various granularities. In the third chapter, we propose content words and language model based abstractive summarization approach. Finally in the last chapter, we analyze the non-situational tweets posted during disaster. We develop a communal tweet classifier, characterize users who posted such contents, and propose a method to identify anti-communal tweets.

1.3.1 Contribution 1: Classifying and summarizing English and Hindi situational tweet streams from Twitter

As stated in Section 1.2.1, our objective is to develop vocabulary independent situational tweet classifier which can be directly used over future disaster events without any further training. We observe that during disaster, non-situational tweets follow some specific patterns like these tweets contain exclamation and question marks, intensifiers (‘too’,‘so’), modal verbs (‘could’,‘should’), strong subjective terms etc. Earlier, Verma *et al.* [130] reported that non-situational tweets are written in a subjective way. These patterns remain more or less uniform across various disaster events. We explore such vocabulary and event independent features to develop our situational tweet classifier.

Side by side, the tweets posted in a disaster scenario have certain specific traits, which can be exploited for the above tasks. For instance, we observe that most of the important information is centered around a limited set of specific words, which we call content words (verbs, nouns, numerals). It is beneficial to focus on these content words while summarizing the situational tweets. Furthermore, a significant fraction of tweets posted during disasters have a mixture of situational and non-situational information within the same tweet (e.g., *ayyo! not again! :(Blasts in Hyderabad, 7 Killed: tv reports*). Separating out the different fragments of such tweets is vital for achieving good classification and summarization. We develop an ILP-based[3] summarization method which optimizes the coverage of important content words to produce final summary. Further, information about missing or injured persons, victims vary rapidly over time. In order to handle such fast changing numerical information, we propose a dependency parser [68] based summarization method which utilizes direct object of disaster specific key verbs like 'injure', 'strand', 'die' etc. to provide continuous update about the number of casualties.

Finally, we observe that lots of situational information available in Hindi tweets are not available in English ones. Hence, extraction of information from other languages is also necessary especially in countries like India where important event updates are available in the long tail [2]. To satisfy information needs, we extend our classification-summarization framework to Hindi tweets.

1.3.2 Contribution 2: Identifying sub-events and summarizing situational information at different granularity levels

As mentioned in Section 1.2.2, single high level situational summary is not able to satisfy needs of multiple stakeholders. In previous part (Section 1.3.1), we classify the tweets into situational and non-situational category and then summarize whole set of situational tweets. However, situational information is distributed across various informative classes/ humanitarian categories like 'infrastructure damage', 'missing or trapped people', 'shelter and services', 'volunteer services' and so on [58] and tweets present in these informative classes contain information about various small scale sub-events [1, 100]. Hence, we develop a generic summarization technique to produce summaries at various granularity levels by combining information from different information classes.

We develop a noun-verb pair based sub-event detection approach using Twitter dependency parser [68]. These noun-verb pairs like 'airport shut', 'communication

[3] Henceforth we represent integer linear programming approach as ILP-based approach.

cut', 'building collapse' etc. represent sub-events in a more semantic way compared to traditional clustering based sub-event detection techniques [1, 100]. After identifying sub-events, we develop an ILP-based general summarization technique by optimizing two parameters — content words (nouns, verbs, numerals) and sub-events. Our summarization approach can be tuned to produce different summaries like high level summaries, humanitarian category specific summaries (e.g. 'infrastructure and damage', 'missing or trapped people', 'shelter and services') etc. to satisfy the needs of different stakeholders. Along with tweets, our summarization method also highlights sub-events present in a tweet and class-specific information. Highlighting the phrase 'airport shut' in the tweet 'Kathmandu **airport shut**, flights from India canceled' helps users to comprehend the summary better.

1.3.3 Contribution 3: Generating abstractive summaries of situational tweet stream

As reported in Section 1.2.3, extractive summaries cannot combine information from related tweet streams. However, combining information from related situational tweets, i.e., abstractive summaries is helpful for producing more informative, diverse and less redundant summaries [22]. Abstractive techniques produce more compact summaries compared to extractive methods but they are more time consuming. If such methods [7, 33] are applied directly over large volume of situational tweets posted during disaster, then real-time summarization is difficult to achieve. Hence, we develop a two stage summarization framework. In the first step, we extract a set of important tweets from the whole situational tweet stream using content word based extractive summarization technique as proposed in contribution 1.3.1. In the second step, we propose a word graph and content word based abstractive summarization technique to produce the final summary.

1.3.4 Contribution 4: Analyzing non-situational microblogs during disaster

Within non-situational tweets, it is observed that people post communal tweets targeting specific religious communities during disaster. We notice that such communal tweets are mostly posted in three different forms — (i) contain religious slang terms like 'Christianity shiz', 'Muhammad pigs' etc. (ii) contain strongly negative term or slang term in the vicinity of neutral religious terms like 'Muslim' or 'Christian', or (iii) contain wh-words or intensifiers with neutral religious terms. Based on above patterns, we develop a rule based classifier to classify communal tweets which performs significantly better than n-grams and hate term based communal tweet detection methods proposed by Burnap *et al* [17].

After identifying communal tweets, we analyze the users who posted such tweets to understand the reason and impact of such communal tweets. We observe that along with common masses, popular politicians, and TV reporters are also involved in initiating and propagating such tweets. A very small fraction of users have a regular habit of posting communal content. On the other hand, most of the users suddenly get outraged as a reaction to certain event.

Along with communal tweet some people also post anti-communal content which appeal to people to refrain themselves from attacking any religious community. We also propose a rule based classifier to automatically identify anti-communal tweets. However, these tweets receive very low exposure compared to communal ones. These tweets can be utilized to counter the adverse effect of communal tweets.

1.4 Organization of the thesis

The rest of the thesis is organized as follows:

In Chapter 2, we provide a comprehensive review about the recent researches related to the topics discussed in this thesis.

In Chapter 3, we present a novel classification-summarization framework for English and Hindi tweets posted during six recent disaster events. Our classification framework depends on low-level lexical features which are independent of any specific event. In the summarization scheme, we propose an ILP-based content words (noun, verb, numeral) optimization technique to provide real-time summaries. Side by side, we also propose a dependency parser based method to handle fast changing numerical information.

In Chapter 4, we present a noun-verb pair based sub-event detection approach and ILP-based summarization method which optimizes content words and sub-events to produce summaries at various granularity level like high level or humanitarian category specific summaries.

In Chapter 5, we present a word graph and content word based abstractive summarization method to generate real-time summaries which cover diverse set of situational information.

In Chapter 6, we analyze non-situational tweets posted during disaster. First, we identify the communal tweets and characterize their users during five disaster events which include both natural (earthquake, flood) and man-made (terrorist attack) disasters. We also propose a set of rules to identify anti-communal content which asks people to refrain themselves from posting communal content.

Finally, in Chapter 7, we conclude this thesis, along with future directions for taking the research further. We have uploaded relevant codes of this thesis in the github repository (`https://github.com/krudra/koustav_phdthesis_2018`). We make the tweet-ids of the tweets related to disaster events publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterSummarizer/dataset.html`,`http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`.

Chapter 2

Related work

In this chapter we provide an overview of the recent researches related to the work present in this thesis. The main objective of this thesis is to develop a framework which can be used to extract and analyze information from microblogging sites during disaster. In order to develop such a framework, detailed understanding of various kinds of information (situational, non-situational etc.) posted during disaster is necessary. Exploring specific traits of tweets posted during disaster is essential to develop tools which can satisfy the requirement of different stakeholders like government, rescue agencies etc.

Keeping the above in consideration, we first provide a small description about Twitter platform and show how recent researches use Twitter as a sentinel during disaster. Next, we provide a review of research works that have tried to classify tweets into situational and non-situational classes and further classify situational tweets into various humanitarian categories like ‘infrastructure damage’, ‘missing or trapped people’ etc. After that, we have provided a detailed overview of different kinds (extractive and abstractive) of document and tweet summarization techniques over English and Hindi. This is followed by a review of research works which have tried to detect sub-events during crises. Finally, in the last section, we provide a detailed research overview of the detection of hate speeches and communal tweets.

2.1 The Twitter microblogging platform

The Twitter microblogging network was founded in 2006. Now a days, it has become a popular microblogging platform consisting of 328 million monthly active users as of the first quarter of 2017 [126]. Around 500 million tweets are posted per day [125]. Most of the information is public in Twitter [74] and such kind of openness, data availability, structured API [127] makes Twitter a popular resource for social media researchers. In this thesis, all the experiments are conducted over Twitter data.

2.2 Twitter as a sentinel during disaster

Microblogging sites are serving as useful sources of situational information during disaster events [20, 82, 102, 111, 129, 132, 143]. When a disaster happens, responders to disasters use such information obtained from microblogging sites to plan and respond to the needs of people located in disaster areas. Volunteers and other support personnel generate summaries and reports based on short messages such as tweets posted via Twitter that are then used by the responders [55]. Real-time information posted by affected people and other observers on Twitter helps in improving disaster relief operations [35, 55]. Recently, Imran *et al.* [55] provides a survey of research works that have used social media during emergencies.

2.3 Classifying situational tweets during disaster events

Large volume of tweets are posted during disaster. However, for practical utility, such situational information has to be extracted from among a lot of conversational and sentimental information. This section discusses some recent studies on classification of tweets.

2.3.1 Classifying tweet streams into situational and non-situational categories

Several studies have attempted to extract situational information during disaster events [129, 130]. Specifically, Verma *et al.* [130] observed that situational tweets are written in a more formal, objective, and impersonal linguistic style as compared to non-situational tweets, and used bag-of-words classifier models to classify tweets based on these features. However, as reported by Verma *et al.* themselves [130], this approach is heavily dependent on the vocabulary of a specific event, and does not work well in the practical cross-domain scenario where the classifier is trained on tweets of some past events and is used to classify tweets of a new disaster event. To overcome the limitations of bag-of-words model, in this thesis we use lower-level lexical and syntactic features of tweets to build an event-independent classifier for situational and non-situational tweets which outperforms the bag-of-words model.

2.3.2 Classifying tweets into humanitarian categories

Two level classification does not provide a detailed overview about various kind of tweets because situational tweets contain information from various informative classes like ‘infrastructure’, ‘missing’, ‘shelter’, and so on [58]. Imran *et al.* [56, 57] proposed unigram, bigram based classifier to classify tweets into various informative classes. They also identified non-relevant tweets which should be discarded from further processing.

2.4 Summarizing information

Situational tweets need to be summarized in real-time. We discuss some recent studies on document and tweet summarization in this section. We also discuss some of the prior works on processing non-English text, especially text in Hindi (Devanagari script).

2.4.1 Document summarization

Document summarization can be of two types : extractive and abstractive. In case of extractive summarization, sentences are selected from the document, whereas in the case of abstractive one, new sentences may be generated by combining related sentences from the document.

Extractive summarization: Researchers adopted graph based summarization approaches in different ways like TextRank [80], LexRank [32] to summarize a set of documents. First they construct a graph where sentences are nodes and weight between two nodes is calculated as similarity between corresponding sentences. Finally, PageRank [90] based iterative updates are applied over this graph to rank the sentences. Li *et al.* [70] proposed bigram based integer linear programming (ILP) technique for summarization. Earlier, Parveen and Strube [93] combined sentence importance and non-redundancy to generate extractive summaries. Later, they combined sentence importance, non-redundancy and coherence into an ILP framework to generate extractive summaries for medical documents [94]. Neural network based models are also proposed to summarize documents. Kageback *et al.* [59] used continuous vector space model to compute semantic similarity between sentences and finally produce extractive summaries. Cao *et al.* [21] applied recursive neural networks (R2N2) to rank sentences for multi document summarization. Recently, Cheng and Lapata [28] proposed a data-driven approach based on neural networks and continuous sentence features to summarize single documents. Lots of existing approaches tried to compute the importance of sentences using standard techniques like tf-idf score, eigen-vector centrality etc [8, 119, 144].

Abstractive summarization: Ganesan *et al.* [34] proposed an abstractive summarization of product reviews using word-graphs. Gerani *et al.* [37] proposed abstractive summarization of product reviews using discourse structure. Liu *et al.* [73] proposed semantic representation based abstractive summarization. In the first step, they parsed the source text into a set of Abstract Meaning Representation (AMR) graphs. After that, they converted AMR graphs to summary graphs and generated final summary. Semantic role labeling based multi document summarization technique was proposed by Khan *et al.* [64]. Li [71] extracted semantic information from multiple documents to construct semantic link network and finally produced abstractive summaries from that network. Bing *et al.* [10] presented multi-document abstractive summarization via phrase selection and merging. Recently, Banerjee *et al.* [7] proposed a graph-based abstractive summarization method on news articles. Several new sentences are generated using the graph and an optimization problem is formulated that selects the best sentences from the new sentences to optimize the overall quality of the summary. The optimization problem ensures that redundant information is not conveyed in the final generated summary. However, the graph construction and path generation is computationally expensive in real-time.

However, such extractive and abstractive document summarization techniques do not consider temporal evolution of data streams, noises, ungrammatical construction of tweets, and real-time requirements. Hence, they are not directly applicable to tweet streams. In recent times, researchers put separate effort to summarize tweet streams. We describe these approaches in the next section.

2.4.2 Tweet summarization

Most of the prior research on tweet summarization focused on summarizing a set of tweets, e.g., tweets posted during the course of a sports event [25, 65, 120]. O'Connor proposed a system *TweetMotif* which searches tweet, groups near duplicate tweets, and summarizes topic [85]. Chakrabarti and Punera [25] used Hidden Markov Model to learn the underlying hidden states present in set of tweets and summarize an event. Khan *et al.* [65] extracted three kinds of words (nouns, verbs, adjectives) from the tweet corpus, constructed a graph among those words, and applied PageRank based algorithm to learn the importance score of individual words. Next, weight of a tweet is calculated by adding the weights of the words present in that tweet. Finally, a summary is created by taking top ranking tweets and discarding similar ones based on Simpson score [117]. Xu *et al.* [141] extracted events from tweets and developed an event graph. Finally, PageRank based algorithm is used to rank the events and summarize the top ranking tweets. Nichols *et al.* [84] presented methods for summarizing sporting events. They used spikes in the volume of status updates to identify important moments during a match and summarize tweets based on sentence ranking method.

All of the above methods consider whole tweet set as a document and did not capture their time evolving nature. However, what is necessary during a disaster event is online / real-time summarization of continuous *tweet streams*, so that the government authorities can monitor the situation in real-time. A few approaches for online summarization of tweet streams have recently been proposed [87,115,137,145]. For instance, Shou *et al.* [115] proposed a scheme based on first clustering similar tweets and then selecting few representative tweets from each cluster, finally ranking these according to importance via a graph-based approach (LexRank) [32]. Osborne *et al.* [89] proposed a real event tracking system using greedy summarization. Recently, a lightweight and scalable tweet summarization approach was developed by Suwaileh *et al.* [118].

Most of the summarization algorithms developed for Twitter are extractive in nature. Due to noise, ungrammatical constructs, and incomplete senses, high quality abstractive summarization is very difficult for tweets. Olariu [86] first proposed a hierarchical clustering based abstractive summarization technique for tweets. Later, Olariu [87] proposed a graph-based abstractive summarization scheme where bigrams extracted from the tweets are considered as the graph-nodes.

2.4.3 Tweet summarization during disaster

Along with standard summarization approaches, a few recent studies [62, 63, 83] have also focused specifically on summarization of news articles and tweets posted during disasters. Kedzie *et al.* [63] used news headlines, tf-idf score of words to learn the salient score of news articles. Finally, they apply affinity clustering and salient score based technique to summarize news articles related to disasters. In their recent work [62], they used features like position of a sentence in a document, an event-type-specific language model built from Wikipedia articles related to the event-type domain, general newswire language model etc., which are difficult to use in applications over noisy, short, and informal texts. Nguyen *et al.* [83] proposed a disaster specific summarization method TSum4act. They prepared clusters of situational tweets using latent dirichlet allocation (LDA), extracted numerals, geo-location information, and events from tweets using the Twitter NER tool [107], constructed a weighted graph among the tweets using cosine similarity as the edge weights, applied weighted PageRank [90], and finally selected tweets based on Simpson similarity measure from each cluster.

Though there have been *separate* prior works on extracting situational information during disasters and on summarization of tweets (as discussed above), to our knowledge, no prior work has attempted to combine the two classical tasks. In this thesis, we show that summarization of tweets during disaster events can be better accomplished if different types of information (e.g., situational and non-situational) are first separated out, and then summarized separately. Additionally, the method proposed in this thesis separately identifies and summarizes time-varying actionable information such as the number of casualties, which constitute some of the most important information during disaster events, but has not been considered in any prior work.

2.4.4 Processing Devanagari documents

During a disaster in a developing region such as the Indian subcontinent, situational information is sparse, and hence it is important to fully utilize whatever information is being posted. We observe that a significant amount of information is posted in local languages (e.g., Hindi), which is not available in the English tweets. This motivated us to process and extract information from Hindi tweets along with English tweets, to produce an informative summary even for disasters in developing regions. There have been prior attempts to summarize Devanagari documents [124], and to develop basic natural language processing tools such as parts-of-speech (POS) taggers [106] and subjectivity lexicons [6] for Devanagari. Gupta *et al.* [45] developed a summarization method for Hindi and Punjabi text. However, it is known that classification / summarization techniques developed for longer and more formal text do *not* perform well for tweets which are very short and mostly written informally [49]. As such, research on processing of tweets written in regional languages such as Devanagari is still in its infancy, and in this thesis, we try to develop a systematic approach to extract situational information from Devanagari tweets.

2.5 Sub-event detection during crises

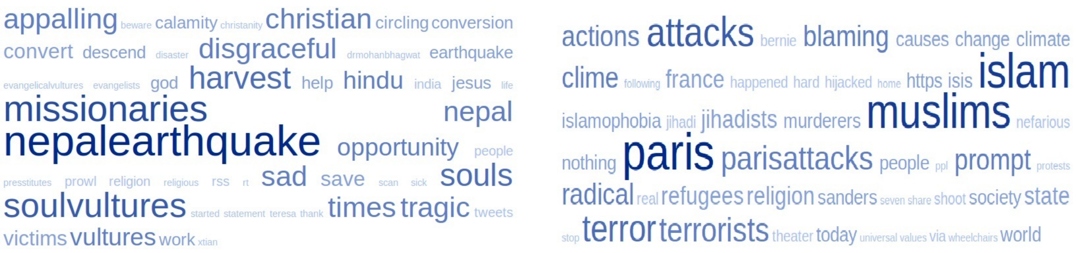
Several studies have tried to extract sub-events from tweets during disasters [1, 99, 100, 121]. Recent approaches attempted to identify topics from evolving tweet streams [79, 115, 137]. However, most existing topic/sub-event detection approaches clustered tweets using different approaches (self organizing map, latent dirichlet allocation (LDA), biterm topic modeling, nearest neighbour, etc.) and represented each cluster or some top frequency words from each cluster as topics or sub-events. Dhekar and Toshniwal [1] collected data from various social medias like Flickr, Twitter, Facebook and developed clusters from different features like text, latitude, longitude, and date. Finally, they applied single pass clustering algorithm to determine final set of clusters. End users find it difficult to understand a bag-of-words representing a sub-event. For example, traditional biterm topic models represent a topic using most probable terms like ‘to’, ‘relief’, ‘Nepal’, ‘material’, ‘NDRF’, etc. However, if we can identify sub-events like ‘relief sent’, ‘NDRF rush’, ‘material carry’, then it will be easier for end users to take decisions regarding relief and rescue operations. Information nuggets consisting of a noun (relief) and a verb (sent) represent more meaningful sub-events compared to a simple collection of words. We try to go beyond traditional clustering and bag-of-words based sub-event detection methods, and provide a more comprehensive and meaningful noun and verb pair based sub-event detection scheme, which is useful in disaster scenarios.

2.6 Identifying and analyzing hate speeches and communal tweets

Microblogs, online forums are increasingly being used by the masses to post offensive content and hate speeches. In recent times, researchers put a lot of effort for automatic identification of such offensive content [17,26,39,69,116]. This section briefly discusses such studies, and points out how the present study is different from the prior works.

Several studies have attempted to identify online content which are potentially hate speeches or offensive in nature. For instance, Greevy *et al.* [41] proposed a supervised bag-of-words (BOW) model to classify racist content in webpages. Along with words, context features are also incorporated to improve the classification accuracy in a later version [97]. Chen *et al.* [27] identified offensive content in Youtube comments using obscenities, profanities, and pejorative terms as features with appropriate weightage. Similarly, *Cyberbullying* is identified by Dinakar *et al.* [30], using features like parts-of-speech tags, profane words, words with negative connotations, and so on. Mahmud *et al.* [76] identified online flaming behaviour using relationship between terms, insulting syntactic constructs. More recently, Burnap *et al.* [14,16,17] proposed hate term and dependency feature based model to identify hate speech posted during a disaster event (the Woolwich attack). Alsaedi *et al.* [3] proposed classification and clustering based technique to predict disruptive events like riot. Burnap *et al.* [15] proposed model to detect cyber hate on Twitter across multiple protected characteristics like race, disability, sex etc.

The present attempt to identify and characterize communal content in Twitter is motivated by the following two perspectives. First, hate speech can come under various categories where people target specific characteristics of users like gender, race, sex, nationality, religion, ethnicity, and so on. Prior studies [116] showed that most prevalent hate speech is targeted towards certain races, while religion-induced hate speech is very sparse. Hence, a general purpose hate speech identifier may fail to capture all the nuances of a rare category (say religion-based hate speech), especially, when for a short period of time such category of tweets are tweeted in huge number. We actually demonstrate in this thesis that the classifier proposed by [116] can hardly


appalling beware calamity christanity christian circling conversion
convert descend disaster disgraceful drmohanbhagwat earthquake
evangelicalvultures evangelists god harvest help hindu india jesus life
missionaries nepal
nepalearthquake opportunity people
presstitutes prowl religion religious rss rt sad save scan sick souls
soulvultures started statement teresa thank times tragic tweets
victims vultures work xtian
actions attacks bernie blaming causes change climate
clime following france happened hard hijacked home https isis islam
islamophobia jihadi jihadists murderers muslims nefarious
nothing paris parisattacks people ppl prompt protests
radical real refugees religion sanders seven share shoot society state
stop terror terrorists theater today universal values via wheelchairs world

(a) Nepal earthquake **(b) Paris Attack**

Figure 2.1: Word cloud of tweets posted in two events.

capture communal tweets. Consequently, in recent times, researchers focus on more granular levels of hate speech detection in Twitter. For example, Chaudhry [26] tried to track racism in Twitter and Burnap *et al.* [17] detected religious hate speeches posted during Woolwich attack.

Second, most of the prior studies on hate speech have focused on content posted in blogs or webpages [31, 39]. On the contrary, in this thesis, we focus on Twitter, and it has been widely demonstrated that standard Natural Language Processing-based methodologies, that have been developed for formally-written text, do not work well for short, informal tweets [38]. Hence new methodologies are necessary to deal with noisy content posted on Twitter.

Burnap *et al.* [14, 16, 17] detected hate speech (religious, racial) posted during the Woolwich attack using a bag-of-words model, where n-grams containing specific hate terms and some dependencies like ‘det’ (determiner) and ‘amod’ (adjectival modifier) are considered as features. However, the bag-of-words model has a known limitation – classifiers based on this model are heavily dependent on event-specific n-grams extracted from the training data, which might not be suitable for applying the classifier to different types of events. For instance, Figure 2.1 shows the tag clouds of communal tweets posted during two different events – the Nepal earthquake (April 2015) and Paris terrorist attack (November 2015). It is clear from the figure that the different religious community being targeted, and hence the vocabularies are significantly different for these two events. As a result, a bag-of-words based classifier is unlikely to perform well if trained on one of these events and used on the other. Recently Magdy *et al.* [75] used post-event tweets to learn users stances towards

Muslims and exploited pre-event interactions, posted tweets to build a classifier to predict post-event stances. However, it is observed that overlap among the users who post communal tweets during multiple events is very low (Chapter 6). Hence, such user specific classifier has very low chance to perform well on future events. On the other hand, using low-level lexical and content features (instead of specific terms) can make the classifier's performance largely independent of specific disaster events considered for training as demonstrated in situational-non-situational tweet classifier. This finding motivated us to propose an *event-independent classifier* for identifying communal tweets.

The focus of almost all the prior works is on identifying offensive hate speech contents. However, very little efforts were made to characterize the users who post such contents. Recently, Silva *et al.* [116] tried to detect the sources and targets of such hate speeches. However, detailed characterization of users who post offensive contents is necessary. In this thesis, we take the first step in this direction by characterizing the users posting communal tweets based on their popularity, interests, and social interactions.

Chapter 3

Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters

Microblogging sites like Twitter have become important sources of real-time information during disaster events. A large amount of valuable *situational information* is posted in these sites during disasters; however, the information is dispersed among hundreds of thousands of tweets containing sentiments and opinions of the masses. To effectively utilize microblogging sites during disaster events, it is necessary to not only *extract* the situational information from the large amounts of sentiment and opinion, but also *summarize* the large amounts of situational information posted in real-time. During disasters in countries like India, a sizeable number of tweets are posted in *local resource-poor languages* besides the normal English language tweets. For instance, in the Indian subcontinent, a large number of tweets are posted in Hindi (the national language of India), and some of the information contained in such non-English tweets are not available (or available at a later point of time) through English tweets.

In this work, we develop a novel classification-summarization framework which handles tweets in both English and Hindi – we first extract tweets containing situational information, and then summarize this information. Our proposed methodology is developed based on the understanding of how several concepts evolve in Twitter during disaster. This understanding helps us achieve superior performance compared to the state-of-the-art tweet classifiers and summarization approaches on English tweets. Additionally, to our knowledge, this is the first attempt to extract situational information from the tweets posted in Indian languages such as Hindi.

3.1 Introduction

Microblogging sites such as Twitter and Weibo have become important sources of information in today's Web. These sites are used by millions of users to exchange information on various events in real-time, i.e., as the event is happening. Especially, several recent studies have shown that microblogging sites play a key role in obtaining situational information during *disaster events* [20,82,102,111,129,132,143].

During a disaster event, various types of information, including situational updates, personal opinions and sentiments are posted by users. Out of this, *situational information* helps the concerned authorities (e.g., Governmental and non-governmental agencies) to gain a high-level understanding of the situation [113]. Hence it is important to develop automated methods to *extract microblogs / tweets which contribute to situational information* [83,130][1] and *summarize those situational updates.* Since time is critical in a disaster situation, these tasks have to be performed in *near real-time*, so that the processed information is readily available to the authorities.

Several recent studies have attempted to address the challenges of extracting situational information from microblogs [130] and summarizing such information [63, 83]. However, these prior works have certain limitations, as detailed in Chapter 2. For instance, most of the classifiers developed to distinguish

[1] Tweets which provide situational information are henceforth referred to as *situational* tweets, while the ones which do not are referred to as *non-situational* tweets.

between situational and non-situational tweets rely on the vocabulary of particular events, and hence do not generalize to various types of disaster events. Again, most of the summarization methodologies do not consider the salient features of tweets posted during disaster events. Most importantly, all the prior studies focus only on English tweets, in order to extend it to a resource-poor Indian language (say, Hindi[2]), several modifications need to be made. This is particularly important from Indian context where a portion of information is present only in Hindi tweets and is not available via English ones (details in Section 3.2).

In this chapter, we propose a novel framework for extracting and summarizing situational information from microblog streams posted during disaster scenarios. Our major contributions are listed below.

(i) Analyzing tweets posted during several recent disaster events (detailed description of dataset is in Section 3.2), we observe that a significant fraction of tweets posted during disasters have a mixture of situational and non-situational information within the same tweet (e.g., *'ayyo! not again! :(Blasts in Hyderabad, 7 Killed: tv reports'*). Again, many tweets contain partially overlapping information (e.g. an earlier tweet 'seven people died', followed by a later tweet 'seven died. high alert declared'). We show that separating out the different fragments of such tweets is vital for achieving good classification and summarization accuracy.

(ii) We develop a classifier using low-level lexical and syntactic features to distinguish between situational and non-situational information (Section 3.3). Incorporating vocabulary independent features enables our classifier to function accurately in cross-domain scenarios, e.g., when the classifier is trained over tweets posted during earlier disaster events and then deployed on tweets posted during a later disaster event. Experiments conducted over tweet streams related to several diverse disaster events show that the proposed classification model outperforms a vocabulary based approach [130] for cross-domain settings.

(iii) We observe that most of the important information posted during disasters is

[2] In India only about 10% of the population speaks English, according to `https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population`.

centered around a limited set of specific words, which we call *content words* (verbs, nouns, numerals). It is beneficial to focus on these content words while summarizing the situational tweets. We propose 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 (Section 3.4). The proposed approach surpasses various state-of-the-art tweet summarization approaches [65, 83, 115] in terms of ROUGE-1 recall and F-score (Section 3.5). We also devise a scheme where we utilize the direct objects of disaster-specific verbs (e.g., 'kill' or 'injure') to continuously update important, time-varying actionable items such as the number of casualties (Section 3.4.4). We try to provide global as well as local location specific updates about victims who were killed, stranded, trapped, died etc.

(iv) For Hindi tweets we cannot directly use classification-summarization framework designed for English tweets due to the following reasons – (i) Most of the lexicons (subjective, question framing words, slangs etc.) used in classification phase is not available (sometimes not enriched) in a consolidated manner in Devanagari. (ii) POS taggers specific to Hindi tweets are not available. We have to apply standard Hindi POS tagger [52]. However, Hindi tweets are mixed with several English words, Twitter tags etc. Hence, a preprocessing phase is necessary for Hindi tweets. To solve the first problem, we gather necessary lexicons for Hindi from different online sources. Preprocessing of Hindi tweet is presented in Section 3.3.5. To the best of our knowledge, this is the first attempt to summarize tweets in Indian languages such as Hindi. Experiments show that the proposed scheme performs significantly better than several state-of-the-art summarization approaches.

3.2 Dataset

This section describes the datasets of tweets that are used to evaluate our classification–summarization approach.

3.2.1 Disaster events

We considered tweets posted during the following disaster events – (i) **HDBlast** – two bomb blasts in the city of Hyderabad, India [54], (ii) **SHShoot** – an assailant killed 20 children and 6 adults at the Sandy Hook elementary school in Connecticut, USA [112], (iii) **UFlood** – devastating floods and landslides in the Uttaranchal state of India [128], (iv) **Hagupit** – a strong cyclone code-named Typhoon Hagupit hit Philippines [48], (v) **NEQuake** – a devastating earthquake in Nepal [81], and (vi) **HDerail** – two passenger trains got derailed near Harda in India [50]. Note that the selected events are widely varied, including both man-made and natural disasters occurring in various regions of the world. Hence, the vocabulary / linguistic style in the tweets can be expected to be diverse as well.

We collected relevant tweets posted during each event through the Twitter API [127] using keyword-based matching. For example, the keywords 'Hyderabad and bomb', 'Hyderabad and bomb and blast', '#Hyderabadblast' were used to identify tweets related to the HDBlast event, while the keywords 'Sandyhook and shoot', and '#sandyhookshooting' were used to collect tweets related to the SHShoot event.

Among the events listed above, we used the first four to develop and evaluate our classification-summarization framework. For each of these four events, we selected the around first 5,000 English tweets in chronological order. We then used the two more recent events NEQuake and HDerail to demonstrate (i) the utility of the framework on large-scale data collected during future events, and (ii) the generalizability of the framework to tweets posted in other languages, by adapting it to Hindi tweets. For these two events, we collected both English and Hindi tweets using the Twitter API, by searching for tweets containing the hashtags #NepalEarthquake and #Harda[3]. For these two events, we gathered 19,970 and 4,171 English tweets, and 6,349 and 1,083 Hindi tweets respectively.

[3] Note that even tweets in other languages use English hashtags for greater visibility.

Table 3.1: Examples of Hindi tweets that contain information that is not available in the English tweets on the same event (case (i)).

Event	Tweet
HDerail	मुंबई से वाराणसी जा रही कामायनी एक्सप्रेस रात करीब 11:45 बजे हरदा के पास पटरी से उतर गई । (Kamyani Express moving from Mumbai to Varanasi gets derailed near Harda at 11:45)
	अंधेरे के कारण राहत बचाव कार्य में मुश्किलें 25 यात्रियों को बचाया गया। (Rescue operation was affected due to darkness; 25 people were rescued)
NEQuake	लखनऊ भूकंप के झटको के चलते लखनऊ के मॉल्स कराए गए खाली शहर के दर्जन भर से अधिक मॉल खाली कराए गए । (Due to earthquake aftershocks, malls in Lucknow were evacuated; more than a dozen malls in the city got evacuated)
	बिहार में अब तक 48 लोगों की मौत । (So far 48 people died in Bihar)

3.2.2 Utility of Hindi tweets

Hindi tweets can be useful in two ways – (i) if we are able to gather new situational information from the Hindi tweets, i.e., information which is present in Hindi tweets but not available in English tweets, and (ii) if we can extract situational information from Hindi tweets *earlier than* what we can from the English tweets. We observed several examples of both the above cases in our datasets. Table 3.1 shows some sample Hindi tweets containing information that is not available in the English tweets for the same event. Similarly, Table 3.2 shows sample tweets where the information present in the Hindi tweets is covered by some English tweets, but the Hindi tweet was posted earlier (timestamp wise) compared to the English ones. Note that, in Table 3.2, the Hindi tweets provide important information such as the exact time of the HDerail event, the effect of NEQuake event in diverse places like Bihar, Uttarpradesh etc., and that this information is obtained from the Hindi tweets earlier than when they are available from the English tweets.

To quantify the utility of including Hindi tweets, we derived the above two statistics over the tweets collected during the two events - NEQuake and HDerail. For this analysis, we take Hindi and English tweets posted during same time span. First,

Table 3.2: Examples of Hindi tweets which contain the same information as some English tweets, but are posted earlier than all such English tweets (case (ii)).

Event	Language	Timestamp	Tweet
Harda	Hindi	2015-08-04 23:19:25	खराब मौसम के कारण राहत और बाचव कार्यों में बाधा आ रही है : सुरेश प्रभु ।
	English	2015-08-05 00:25:18	@jayprakashindia – Raining restarted at ground zero at Harda ....due to this Search and Rescue opration is beeing affected
	Hindi	2015-08-05 01:04:17	ट्रेन हादसाः हरदा एसपी प्रेमबाबू शर्मा का बयान अब तक 12 शव निकाले गए ।
	English	2015-08-05 01:14:30	Madhya Pradesh train accidents: 12 bodies recovered so far, says Prem Babu Sharma, Superintendent of Police, Harda
NEQuake	Hindi	2015-04-25 08:55:16	में भूकंप से 100 लोगों के मरने की आशंका ।
	English	2015-04-25 09:05:04	more than 100 people died by indian news #NepalEarthquake
	Hindi	2015-04-25 08:57:27	नेपाल के जनकपुर में जानकी मंदिर को नुकसान ।
	English	2015-04-25 08:58:40	Mother Sita's palace "Janaki temple" also damaged in #NepalEarthquake

we remove duplicate tweets from the Hindi dataset. After this step, we get 230 and 128 Hindi tweets for HDerail and NEQuake events respectively[4]. Three human annotators individually analyzed the tweets. First, duplicates were removed from both Hindi and English tweet sets. After that, the annotators went through the whole set of deduplicated English tweets to get an overview of the information content of tweets. Then they went through the Hindi tweets one by one, and for each of the tweets they checked the following two scenarios —

1. Whether the same information is missing in English tweets, i.e. the information is exclusively available in Hindi.

2. The same information is also present in English tweets but we can extract

[4] Only situational tweets were considered for this analysis, as identified by the approach described later in this chapter.

that information from Hindi tweets earlier than what we can from the English tweets (based on the timestamps of the tweets).

In order to check whether any English tweet contains similar information corresponding to a Hindi tweet, the annotators particularly relied on the content words present in both the tweets. We got a very high Fleiss Kappa agreement score of 0.92 in this annotation process. For the rest of the cases, there were some disagreements in deciding whether the same information appeared in the English tweets; these disagreements were resolved through discussions among the annotators.

We found that 15.45% & 21.43% of Hindi tweets contain *new information* which is not available from the English tweets, for the HDerail and NEQuake events respectively. Additionally, in 8.13% & 14.29% cases, the information was obtained earlier in Hindi tweets than from the English tweets. These observations establish the need to process tweets in regional Indian languages like Hindi.

We make the tweet-ids of the collected tweets publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterSummarizer/dataset.html`.

3.2.3 Types of tweets

As stated earlier, tweets posted during a disaster event include both tweets contributing to situational awareness, and non-situational tweets. Earlier studies [102, 130] showed that situational tweets contain information about the current situation, whereas non-situational tweets mostly consist of opinion, sentiments, abbreviations, and so on. Recently, Imran *et al.* [58] showed that situational tweets can be of various types, such as victims looking for help, humanitarian organizations providing relief, and so on. Also, the types of situational tweets are not the same for different kinds of disasters. On the other hand, the non-situational tweets mention about the event but do not contain any factual information. Some ambiguity exists in case of tweets related to donation or charities, as to whether they should be considered situational or otherwise. Prior works such as

Table 3.3: Examples of various types of situational tweets (which contribute to situational awareness) and non-situational tweets.

Type	Event	Tweet text
Situational tweets (which contribute to situational awareness)		
Situational updates	Hagupit	typhoon now making landfall in eastern samar, with winds of 175 to 210 kph, and rainfall up to 30mm per hour
	SHShoot	state police are responding to a report of a shooting at an elementary school in newtown [url]
	UFlood	call bsnl toll-free numbers 1503, 09412024365 to find out last active location of bsnl mobiles of missing persons in uttarakhand
	HDBlast	blood banks near dilsuknagar, slms 040-64579998 kamineni 39879999 hima bindu 9246373536 balaji
	Hagupit	#Oxfam have raced hygiene kits with soap, toothpaste, toothbrushes, sleeping mats, blankets and underwear to areas hit by Typhoon #Hagupit
	SHShoot	If you want to donate blood, call 1-800-RED CROSS. @CTRedCross @redcrossbloodct
Non-situational tweets		
Sentiment / opinion	SHShoot	There was a shooting at an elementary school. I'm losing all faith in humanity.
	Hagupit	thoughts/prayers for everyone in the path of #typhoon hope lessons from #haiyan will save lives.
Event analysis	UFlood	#Deforestation in #Uttarakhand aggravated #flood impacts. Map showing how much forestland diverted [url]
	HDBlast	#HyderabadBlasts: Police suspect one of the bombs may have been kept on a motorcycle; the other in a tiffin box.
Charities	SHShoot	r.i.p to all of the connecticut shooting victims. for every rt this gets, we will donate $2 to the school and victims
	Hagupit	1$ usd for a cause-super-typhoon hagupit, i'm raising money for eye care global fund, click to donate, [url]

Qu *et al.* [102] considered donation or charity related tweets as non-situational tweets. In this work, we are following the same protocol and categorize donation related tweets as non-situational. Some example tweets of each category are shown in Table 3.3.

Situational awareness tweets: Tweets in this category contain diverse information like infrastructure damage, information about missing, trapped or injured people, number of casualties, shelter and volunteer and relief information, and so on [58]. Relief information includes information about helping organizations, necessary requirements of affected victims, phone numbers of nearby hospitals, etc. Such information can immediately help in relief operations.

Table 3.4: Examples of mixed tweets containing multiple fragments, some of which convey situational information while the other fragments are conversational in nature.

ayyo! not again! :(Blasts in Hyderabad, 7 Killed: TV REPORTS
oh no !! unconfirmed reports that the incident in #newtown #ct may be a school shooting. police on the way
58 dead, over 58,000 trapped as rain batters Uttarakhand, UP.....may god save d rest....NO RAIN is a problem....RAIN is a bigger problem
”@IvanCabreraTV: #Hagupit is forecast to be @ Super Typhoon strength as it nears Philippines. [url]” Oh no! Not again!

Non-situational tweets: Non-situational tweets (which do not contribute to situational awareness) are generally of the following types: (i) *Sentiment / opinion* – sympathizing with the victims, or praising / criticizing the relief operations, opinion on how similar tragedies can be prevented in future, (ii) *Event analysis* – post-analysis of how and why the disaster occurred, findings from police investigation in case of man-made emergencies, and (iii) *Charities* – tweets related to charities being organized to help the victims.

The next two sections discuss our proposed methodology of first separating the situational and non-situational tweet streams (Section 3.3), and then summarizing the situational information (Section 3.4).

3.3 Classification of tweets

In this section, we focus on separating the situational and non-situational tweets by developing a supervised classifier. Since training such a classifier requires gold standard annotation for a set of tweets, we used human annotators to obtain this gold standard (details below). During annotation, it was observed that *a significant number of tweets posted during disaster events contained a mixture of situational and non-situational information.* Table 3.4 shows some examples of such tweets. Note that none of the prior attempts to distinguish between situational and non-situational tweets reported this phenomenon of the same tweet containing both types of information. The presence of such tweets motivated us to identify different fragments of a tweet and process them separately for classification and summarization steps. This preprocessing stage is described next.

3.3.1 Preprocessing and fragmentation of tweets

To effectively deal with tweets containing a mixture of situational and non-situational information, we perform the following preprocessing steps.

(i) We use a Twitter-specific part-of-speech (POS) tagger [38] to identify POS tags for each word in the tweet. Along with normal POS tags (nouns, verbs, etc.), this tagger also labels Twitter-specific keywords such as emoticons, retweets, URLs, and so on. We ignore the Twitter-specific words that are assigned tag 'U', 'E', '@', '#', 'G' by the POS tagger [38] because they represent URLs, emoticons, mentions, hashtags, abbreviations, foreign words, and symbols which do not contribute to meaningful information.

(ii) We apply standard preprocessing steps like case-folding and lemmatization. Additionally, it is observed that many phonetic variations are created in case of modal verbs contained in the tweets, primarily because of the strict limitation on the length of tweets (140 characters). For example, 'should' is represented as 'shld', 'shud', while 'could' is often represented as 'cud', 'cld'. In our work, we attempt to unify such variations of modal verbs, which helps in the classification phase (Section 3.3). First we collect standard modal verbs for English. Next, we manually collect different phonetic (out-of-vocabulary) variations of such modal verbs from a list of out-of-vocabulary words commonly used in social media [77]. Table 3.5 shows examples of some modal verbs and their variations.

We also attempt to maintain uniformity across different representations of numeric information (e.g. '7' and 'seven'). Specifically, we use the *num2words* Python module (`https://pypi.python.org/pypi/num2words`) for this purpose. This step primarily helps in summarization (Section 3.4).

(iii) Subsequently, we focus on particular end-markers (e.g., '!', '.', '?') to split a

Table 3.5: Different out-of-vocabulary variations of modal verbs.

Modal verb	Out-of-vocabulary variations
Should	‘shud’, ‘shld’, ‘sud’
Could	‘cud’ ,‘cld’, ‘culd’
Would	‘wud’, ‘wuld’, ‘wld’
Would not	‘wont’, ‘wouldnt’, ‘wouldn’t’, ‘wudnt’, ‘wudn’t’

Table 3.6: Number of tweets and fragments present in each dataset.

	HDBlast	SHShoot	UFlood	Hagupit	NEQuake	HDerail
#Tweets	4,930	4,998	4,982	4,996	19,970	4,171
#Fragments	5,249	5,790	6,236	5,444	19,102	4,361

tweet into multiple fragments. We use Twitter parts-of-speech tagger to identify the three sentence boundaries: (‘!’,‘?’,‘.’). Finally, we keep only those fragments satisfying minimum length constraint of five.

As a result of these preprocessing steps, each tweet is decomposed into multiple fragments, and all the subsequent steps are carried out on these fragments. Table 3.6 shows the total number of tweets and the total number of fragments obtained from these tweets, for each of the datasets (as described in Section 3.2).

3.3.2 Establishing gold standard

For training the classifier, we considered 1000 randomly selected tweet fragments related to each of the first four events described in Section 3.2. Three human volunteers independently observed the tweet fragments. All the volunteers are regular users of Twitter, have a good knowledge of English and Hindi. Before the annotation task, the volunteers were acquainted with some examples of situational and non-situational tweets identified in prior works [130, 132].

Each volunteer was asked to decide whether a certain tweet fragment contributes to situational awareness. We obtained unanimous agreement (i.e., all three volunteers labeled a fragment similarly) for 82% of the fragments, and majority opinion was considered for the rest of the fragments.

After this human annotation process, we obtained 416, 427, 432 and 453 tweet-fragments that were judged as situational, for the HDBlast, UFlood, SHshoot and Hagupit events respectively. From each of these four datasets, we selected an equal number of tweet-fragments that were judged non-situational, in order to construct balanced training sets for the classifier.

Apart from classifying the tweet-fragments, we also develop a classifier for the raw tweets. We follow same annotation process also for raw tweets. As identified earlier, some raw tweets may contain both situational and non-situational information. In the annotation phase, a tweet is marked as situational if it contains some situational information. For all the four events, we randomly sampled 1000 tweets and these tweets were annotated as situational or non-situational by the same volunteers as mentioned above. Finally, we obtained 376, 427, 439, and 401 tweets that were judged as situational for the HDBlast, UFlood, SHShoot and Hagupit events respectively. From each of these four datasets, we selected an equal number of tweets that were judged non-situational in order to develop balanced training set for raw tweets.

3.3.3 Classification features and performance

Prior research [130] has shown that the situational tweets are written in a more formal and less subjective style, and from a more impersonal viewpoint, as compared to the non-situational tweets. We consider a set of eleven low-level lexical and syntactic features, as listed in Table 3.7, to identify the more complex notions of subjectivity and formality of tweets. Briefly, situational tweets / tweet-fragments are expected to have more numerical information, while non-situational tweets are expected to have more of those words which are used in sentimental or conversational content, such as subjective words, modal verbs, queries and intensifiers.

We compare our classifier with a standard bag-of-words (BOW) model similar to that in [130], where the classifier is used considering as features – the frequency of every distinct unigram and bigram (Twitter specific tags are removed using POS tagger),

Table 3.7: Lexical features used to classify between situational and non-situational tweets.

Feature	Explanation
Count of subjective words	Number of words listed as strongly subjective in a subjectivity lexicon for tweets [133]. Expected to be higher in non-situational tweets.
Presence of personal pronouns	Presence of commonly used personal pronouns in first-person (e.g., *I, me, myself, we*) and second-person (e.g., *you, yours*). Expected to be higher in non-situational tweets.
Count of numerals	Expected to be higher in situational tweets which contain information such as the number of casualties, emergency contact numbers.
Presence of exclamations	Expected to be higher in non-situational tweets containing sentiment and exclamatory phrases (e.g., 'Oh My God!', 'Not Again!').
Count of question marks	Expected to be higher in non-situational tweets containing queries / grievances to the authorities (e.g., 'Can't they spend some of the #Coalgate cash for relief?').
Presence of modal verbs	Expected to be higher in non-situational tweets containing opinion of people and event analysis e.g., ('should', 'could', 'would', 'cud', 'shud').
Presence of wh-words	Number of words such as 'why', 'when', etc. Expected to be higher in non-situational tweets containing queries of people, e.g., 'Why don't you submit your coalgate scam money to disaster'.
Presence of intensifiers	Existence of frequently used intensifiers [103], more used in non-situational tweets to boost sentiment, e.g., 'My heart is *too* sad', 'Hyderabad blasts are *so* saddening'.
Presence of non-situational words	We identify a set of words (96 words) which only appear in non-situational tweets across all events, such as 'pray', 'God', 'donate', 'condemn'. Then we find the presence of such event-independent non-situational keywords.
Presence of religious words	Religious words are used to target specific religious communities and they are usually present in non-situational tweets [110], e.g., 'Is it cong supporter right wing *hindutva* extremists behind bomb blasts in the indian city of hyderabad'.
Presence of slangs	Slang words are present mostly in non-situational tweets, e.g., 'But some *f***ing* bastards use religion as cover'.

Table 3.8: Statistics of distinct unigrams, bigrams, feature space, and training data size for fragmented tweets across four different disaster events in BOW model.

Event	#Unigrams	#Bigrams	#Feature space	#Training data (#Tweets)
HDBlast	2,029	4,451	6,502	832
UFlood	2,517	5,290	7,829	854
SHShoot	1,212	3,410	4,644	864
Hagupit	2,211	5,033	7,266	906

POS tags, count of strong subjective words, and presence of personal pronouns. In case of BOW model, for each of the events, total feature space consists of number of distinct unigrams, bigrams, POS tags, strong subjective word count, and personal pronouns. Table 3.8 shows number of distinct unigrams, bigrams, total feature space, training data size for each of the four events.

We compare the performance of the two feature-sets (using the same classifier) under two scenarios — (i) *in-domain classification*, where the classifier is trained and tested with the tweets related to the *same event* using a 10-fold cross validation, and (ii) *cross-domain classification*, where the classifier is trained with tweets of one event, and tested on another event. In this case, all the annotated tweets of a particular event are used to train / develop the model and then it is tested over all the tweets of rest of the events.

Performance of a classifier is heavily dependent on the appropriate model selection. In the next part, we try to select most appropriate model for both sets of features based on some specific criteria.

Model selection: We consider four state-of-the-art classifier models — a. SVM with default rbf kernel b. SVM with linear kernel, c. Logistic regression, and d. Naive Bayes for both the feature sets. For each of these models we use Scikit-learn [96] package. To judge the performance of all these models on above mentioned feature sets we set following evaluation criteria.

1. **Average in-domain accuracy**: Average accuracy of the classifier across the

four events in in-domain scenario.

2. **Average cross-domain accuracy**: Average accuracy of the classifier in different cross-domain scenario among the four events. In this case, we have twelve different cross-domain settings.

3. **Average precision for situational tweets**: Detection of situational tweets with high precision is a necessary requirement for the classifier. Hence we consider average precision across the four datasets.

4. **Average recall for situational tweets**: The classifier should ideally capture all the situational posts, i.e., have high recall. Hence we consider the recall averaged over the four datasets.

5. **Average F-score for situational tweets**: F-score of the classifier indicates the balance between coverage / recall and accuracy / precision.

Table 3.9: Score of different evaluating parameters for four different classification models, using (i) bag-of-words features (BOW), (ii) proposed lexical features (PRO).

Model	**Evaluation parameters**									
	In-domain accuracy		Cross-domain accuracy		Precision		Recall		F-score	
	BOW	PRO	BOW	PRO	BOW	PRO	BOW	PRO	BOW	PRO
SVM (rbf)	0.70	0.85	0.54	0.82	0.56	0.83	0.22	0.83	0.25	0.82
SVM (linear)	0.82	0.85	0.66	0.81	0.76	0.79	0.46	0.86	0.53	0.82
Logistic regression	0.85	0.85	0.69	0.82	0.80	0.80	0.54	0.86	0.62	0.82
Naive Bayes	0.88	0.85	0.78	0.82	0.90	0.83	0.63	0.82	0.73	0.82

We report the performance of different classification models on two different sets of features in Table 3.9. From Table 3.9, it is clear that bag-of-words features are very much dependent on the model and only Naive Bayes model shows promising performance compared to other three models. Naive Bayes model shows superior performance compared to others because it considers each of the features (unigrams, bigrams) independently. Our proposed set of features more or less show equal performance for different classification models. This clearly reveals the benefit of working with event independent features. Finally, we select SVM with default RBF kernel for our proposed set of features and Naive Bayes model for bag-of-words features. All the subsequent results are produced using these two models.

Table 3.10: Classification accuracies of Naive Bayes and Support Vector Machine (SVM) classifier on tweet fragments, using (i) bag-of-words features (BOW), (ii) proposed lexical features (PRO) respectively. Diagonal entries are for in-domain classification, while the non-diagonal entries are for cross-domain classification.

Train set	**Test set**							
	HDBlast		UFlood		SHShoot		Hagupit	
	BOW	PRO	BOW	PRO	BOW	PRO	BOW	PRO
HDBlast	**86.890%**	84.260%	73.653%	**78.220%**	86.689%	**89.583%**	77.262%	**82.339%**
UFlood	81.850%	**82.451%**	**84.988%**	79.609%	85.532%	**89.814%**	81.333%	**81.456%**
SHShoot	82.211%	**83.052%**	75.058%	**79.859%**	**93.179%**	90.042%	78.366%	**80.242%**
Hagupit	73.557%	**77.283%**	65.573%	**75.644%**	71.875%	**86.458%**	**87.737%**	85.862%

In-domain classification: The BOW model performs well in the case of in-domain classification (diagonal entries in Table 3.10) due to the uniform vocabulary used during a particular event. However, the performance of our proposed features is at par with the baseline BOW model even without considering the event-specific words. The result is specially significant since it shows that similar accuracy can be achieved even without considering the event-specific words.

Cross-domain classification: The non-diagonal entries of Table 3.10 represent the accuracies, where the event stated on the left-hand side of the table represents the training event, and the event stated at the top represents the test event. The proposed model performs much better than the BOW model in such scenarios, since it is independent of the vocabulary of specific events. We also report recall and F-score of our proposed classification model over the situational tweets, in Table 3.11.

Table 3.11: Classification scores (recall(F-score)) of Naive Bayes and Support Vector Machine (SVM) classifier on situational tweet fragments, using (i) bag-of-words features (BOW), (ii) proposed lexical features (PRO) respectively. Non-diagonal entries are for cross-domain classification.

Train set	**Test set**							
	HDBlast		UFlood		SHShoot		Hagupit	
	BOW	PRO	BOW	PRO	BOW	PRO	BOW	PRO
HDBlast	**0.88(0.86)**	0.85(0.84)	0.56(0.68)	**0.75(0.77)**	0.74(0.85)	**0.87(0.89)**	0.60(0.72)	**0.77(0.81)**
UFlood	0.83(0.82)	**0.85(0.83)**	**0.85(0.85)**	0.81(0.80)	0.72(0.83)	**0.88(0.89)**	**0.80(0.83)**	0.78(0.81)
SHShoot	0.80(0.82)	**0.85(0.83)**	0.64(0.72)	**0.77(0.79)**	**0.91(0.93)**	0.87(0.89)	0.65(0.75)	**0.75(0.79)**
Hagupit	0.48(0.64)	**0.89(0.79)**	0.36(0.51)	**0.86(0.78)**	0.44(0.61)	**0.88(0.86)**	0.83(0.87)	**0.94(0.87)**

Recall and F-score of classification models: It is observed that bag-of-words model achieves more or less reasonable accuracy in cross-domain settings despite its dependency on the vocabularies used during a particular event. We find that non-situational tweets across different events contain similar kind of words like 'god', 'pray', 'condemn', 'condolence', 'heart' etc. However, during disaster, situational tweets carry more importance and detection of such tweets with high precision and recall is the prime requirement. Unfulfillment of such requirement will in turn create a two-way bottleneck in disaster handling — (i) low precision value indicates lots of non-situational tweets are misclassified as situational tweet and such tweets hamper the summarization phase, (ii) on the other hand, low recall means many situational tweets are missing which indicates loss of information. Our objective is to make a balance between precision and recall scores so that both classification and summarization phases can be optimized. Recall and F-score values of our proposed and baseline classifier are shown in Table 3.11 during in-domain and cross-domain scenario. Recall and F-score of the classifier in cross-domain is more important because our primary intention is to deploy the classifier directly over future events without any further training. Our proposed model achieves 36%, 14% improvement over baseline model in terms of recall and F-score in cross-domain settings. Further, we notice if vocabulary significantly differs between two events then recall and F-score values of situational tweets fall drastically. For example, in case of Hagupit, vocabularies used in the situational tweets are significantly different compared to other three events which results in a low recall and F-score in cross domain settings.

Benefit of fragmentation and preprocessing before classification: As described earlier, our methodology consists of preprocessing and fragmenting the tweets before classification. A natural question that arises is whether the preprocessing and fragmentation steps help to improve the classification performance. To answer this question, we apply the same classifier as stated above on the *raw tweets*; the classification accuracies are reported in Table 3.12. Comparing the classification accuracies in Table 3.10 (on preprocessed and fragmented tweets) and Table 3.12 (on raw tweets), we can verify that the initial fragmentation and preprocessing steps help to improve the performance of both the BOW model as well as the proposed model. We shall also show later (in Section 3.5) that the preprocessing phase in turn

Table 3.12: Classification accuracies of Naive Bayes and SVM on *raw tweets*, using (i) bag-of-words features (BOW), (ii) proposed features (PRO) respectively. Diagonal entries are for in-domain classification, while the non-diagonal entries are for cross-domain classification.

Train set	**Test set**							
	HDBlast		UFlood		SHShoot		Hagupit	
	BOW	PRO	BOW	PRO	BOW	PRO	BOW	PRO
HDBlast	**82.866%**	81.899%	73.292%	**76.112%**	83.599%	**81.890%**	77.177%	**77.431 %**
UFlood	78.191%	**80.984%**	**81.838%**	77.062%	82.004%	**82.118%**	75.561%	**80.548%**
SHShoot	80.186%	**78.723%**	72.131%	**75.058%**	**87.922%**	**84.738%**	78.301%	**76.059%**
Hagupit	73.234%	**78.590%**	65.140%	**75.409%**	71.095%	**79.612%**	**83.027%**	79.667%

helps in information coverage during the summarization process.

Table 3.13: Feature ablation experiments for the situational tweet classifiers for both in-domain and cross-domain scenarios. NONE represents the case when all the features were used.

Ablated Feature(s)	**In-domain accuracy**	**Cross-domain accuracy**
NONE	0.8494	0.8220
subjective word	0.8249	0.8094
religion	0.8465	0.8217
slang	0.8451	0.8217
non-situational word	0.8161	0.7857
pronoun	0.8346	0.8074
wh-word	0.8451	0.8073
intensifier	0.8444	0.8216
modal verb	0.8404	0.8209
question mark	0.8471	0.8215
exclamation	0.8393	0.8112
numeral	0.8243	0.8110

Feature ablation: In this part, we try to judge the importance of individual features in the classification, through feature ablation experiments. Table 3.13 reports the in-domain and cross-domain accuracies of the situational tweet classifier for feature ablation experiments, averaged over all the datasets. Presence of numerals, pronouns, exclamation mark, subjective words, and non-situational words appear to be most determining factors. However, all the features help in increasing the accuracy of the situational tweet classifier.

Thus the proposed classification scheme based on low-level lexical and syntactic features performs significantly better than word-based classifiers [130] under various experimental settings. However, since the best achieved classification accuracy is still around 80%, a question naturally arises as to whether the 20% mis-classification would substantially impact the subsequent summarization step. We shall discuss the effect of mis-classification on summarization in Section 3.5.

3.3.4 Applying classifier on future disaster events

The good cross-domain performance of the proposed classification scheme (as stated above) implies that the selected low-level lexical and syntactic features can robustly distinguish between situational and non-situational tweets *irrespective of* the specific type of event under consideration, or the vocabulary / linguistic style related to specific events. Additionally, since we train our classifier using low-level features, we expect that the accuracy of the classifier will not vary significantly based on the size and diversity of training set (e.g., if multiple past disasters of various types are used to train the classifier).

To demonstrate this, we perform another set of experiments taking Hagupit (the most recent of the four events under consideration) as the test event, and instead of training the classification model with only one event, we combine the remaining two / three events for training. The classifier achieves accuracy values of 81.89%, 82.23%, 81.23% and 82.34% respectively when trained on (HDBlast and UFlood), (HDBlast and SHShoot), (UFlood and SHShoot), and all three events taken together. These accuracy values show that as the classifier is trained on more patterns expressing situational and non-situational information related to various types of disasters, the classifier's accuracy with cross-domain information becomes almost equal to that when it is trained with in-domain information. Thus, we conclude that the proposed classification framework can be trained over tweets related to past disaster events, and then deployed to classify tweets posted during future events.

Later in Section 3.5.3, we actually deploy the classifier trained on earlier events over tweets related to the two later events, NEQuake and HDerail.

3.3.5 Classifying Hindi tweets

For classifying Hindi tweets, we need to extend our classification framework to the Hindi language. We now describe the challenges in extending the methodology to Hindi tweets, and how we address those challenges.

Challenges in Hindi tweet classification: From our datasets, we observe that Hindi tweets are often *not* written in proper Devanagari script; rather Devanagari script is frequently mixed with many English terms, and Twitter-specific elements such as mentions, hashtags, and URLs. To our knowledge, there does not exist any Twitter-specific part-of-speech tagger for Hindi. Hence, we have to apply Hindi POS tagger [52] which is designed for *formal* Devanagari text. Hence, we apply the following pre-processing techniques to remove English terms and Twitter specific symbols from Hindi tweets, before applying the parts-of-speech tagger.

1. English terms and Twitter specific symbols (‘mentions’,’ ‘hashtags’, ‘urls’, ‘emoticons’) are removed from tweets based on regular expressions. After this step, tweets contain only numerals and Devanagari terms.
2. Finally, tweets are fragmented based on endmarkers ‘!’, ‘?’, ‘।’

The lexical and syntactic features that are listed in Table 3.7 (for classification of English tweets) are based on the presence or absence of some specific types of words in the tweets, such as personal pronouns, modal verbs, wh-words, intensifiers, and so on. To extend this methodology to tweets in a non-English language, it is necessary to develop lexicons of these types of words in that language. For identifying subjective words, we use a subjectivity lexicon for Hindi developed as part of an earlier study [6]. All the other lexicons like pronouns, intensifier, wh-words etc. are collected from Wikipedia and online sources. All these lexicons also contain many morphological variations of a particular word (e.g., अपना, अपनी, अपने).

We apply the same methodology as described in Section 3.3 — tweets are partitioned into fragments, the features listed in Table 3.7 are computed for each fragment, and the fragments are then classified into situational or non-situational. In case of Hindi, we check the performance of four different classification models as used in English. We observe similar kinds of trend as in English tweets. Unigram, bigram features are heavily dependent on model and show good performance for Naive Bayes model. On the other hand, SVM model with RBF kernel performs best for our proposed set of features. Hence, we select these two models for Hindi tweet classification. Finally, we compare the performance of SVM model over proposed set of features with Naive Bayes model and bag-of-words features.

Evaluating the performance of the classifier on Hindi tweets: As in the case of English tweets, we use human volunteers to obtain a gold standard annotation of the fragments of the Hindi tweets. Three human volunteers – each having a good knowledge of the Hindi language — independently observed the tweet fragments (after removing duplicate fragments), deciding whether they contribute to situational awareness. We obtained unanimous agreement for 87% of the fragments (i.e., all three volunteers labeled these similarly), and majority opinion was considered for the rest. After this human annotation process, we obtained 281 and 120 tweet fragments that were judged as situational, for the NEQuake and HDerail events respectively. From each of these two datasets, we next selected an equal number of tweet fragments that were judged non-situational, and constructed balanced training sets for the classifier.

Table 3.14: Classification accuracies of SVM on fragmented Hindi tweets, using (i) bag-of-words features (BOW), (ii) proposed lexical and syntactic features (PRO).

Train set	**Test set**			
	NEQuake		HDerail	
	BOW	PRO	BOW	PRO
NEQuake	**85.412%**	81.305%	69.191%	**72.222%**
HDerail	74.377%	**77.935%**	**74.833%**	74.222%

Table 3.14 shows the in-domain (diagonal entries) and cross-domain accuracies (non-diagonal elements) for Hindi tweet classifier. It clearly shows that the proposed features lead to better classification of Hindi tweets than the

bag-of-words model in the cross-domain scenarios. Further, we notice that in case of Hindi tweet classification, we achieve low accuracy compared to English tweets (Table 3.10) due to unavailability of resources. However, we are able to achieve comparable accuracy for Hindi tweets under such resource constraints.

Finally, for summarizing the situational tweets (discussed in the next section), we only considered those tweets which were classified with a certain confidence level. For this, we tested our proposed English and Hindi situational tweet classifiers on manually annotated datasets. We checked various confidence scores – $(0.6, 0.7, 0.8, 0.9)$. At 0.9 confidence level, the recall score drops drastically to 0.10. For rest of the three cases the precision, recall and F-scores are comparable (F-score is around 0.84). For both English and Hindi tweets, we decided to set the confidence level to 0.8, i.e., we selected only those SVM classified situational messages for which the classification confidence was ≥ 0.80.

3.4 Summarization of tweets

After separating out situational tweets using the classifier described in the previous section, we attempt to summarize the situational tweet stream in real-time. For the summarization, we focus on some specific types of terms which give important information in disaster scenario – (i) numerals, (e.g., number of casualties or affected people, or emergency contact numbers), (ii) nouns (e.g., names of places, important context words like people, hospital etc.), and (iii) main verbs (e.g., 'killed', 'injured', 'stranded'). We refer to these terms as *content words*. This section describes our proposed method, which we call COWTS (COntent Word-based Tweet Summarization).

3.4.1 Need for disaster-specific summarization approach

We observe a specific trend in case of situational tweets posted during disaster events, which is very different from tweet streams posted during other types of events. As

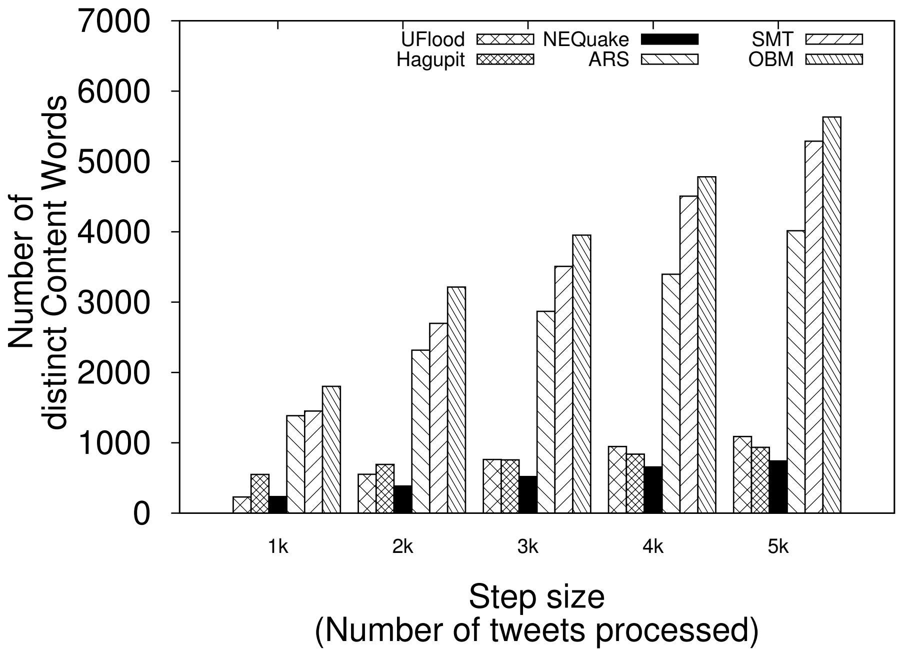


Figure 3.1: Variation in the number of distinct content words with the number of tweets in chronological order, shown for disaster events (three left bars in each group), and other events (three right bars in each group (Arsenal, Smartphone, Obama)).

tweets are seen in chronological order, the number of *distinct content words* increases very slowly with the number of tweets, in case of disaster events.

To demonstrate this, we compare tweet streams posted during disaster events with those posted during the three political, sports, and technology-related events; these streams were made publicly available by a previous study [115]. Figure 3.1 plots the variation in the number of distinct content words seen across the first 5,000 tweets in these three tweet streams, as well as the situational tweet streams posted during three disaster events. It is evident that the number of distinct content words increases very slowly in case of the disaster events. We find that this is primarily due to (i) presence of huge number of retweets or near-duplicates of few important tweets, and (ii) presence of large number of tweets giving latest updates on some specific contexts, such as the number of people killed or stranded. This leads to heavy usage of some specific content-words (primarily, verbs) – such as ‘killed’, ‘injured’ and ‘stranded’ – and rapidly changing numerical information in the context of these content-words.

The above observations indicate that summarizing situational information in disaster scenarios requires a different approach, as compared to approaches developed for other types of events. Hence, we (i) remove duplicate and near-duplicate tweets using the techniques developed in [123], (ii) focus on the content words during summarization (as described in Section 3.4.2), and (iii) adopt specific strategies for the heavily-repeated content words associated with frequently changing numerical information (described in Section 3.4.4).

3.4.2 Content word based summarization

The summarization framework we consider is as follows. Tweets relevant to the disaster event under consideration are continuously collected (e.g., via keyword matching), and situational tweets are extracted using the classifier. At any given point of time, the user may want a summary of the situational tweet stream, by specifying (i) the starting and ending timestamps of the part of the stream that is to be summarized, and (ii) a desired length L which is the number of words to be included in the summary.

Considering that the important information in a disaster situation is often centered around content words, an effective way to attain good coverage of important information in the summary is by optimizing the coverage of *important content words* in the tweets included in the summary. The importance $Score(j)$ of a particular content word j is computed using the *tf-idf* score with sub-linear *tf* scaling considering the set of tweets containing it:

$$Score(j) = (1 + log(|T_j|)) * log(n/|T_j|) \tag{3.1}$$

We use an Integer Linear Programming (ILP)-based technique [93] to optimize the coverage of the content words. Table 3.15 states the notations used. The summarization is achieved by optimizing the following ILP objective function:

$$max(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} Score(j).y_j) \tag{3.2}$$

Table 3.15: Notations used in the summarization technique.

Notation	**Meaning**
L	Desired summary length (number of words)
n	Number of tweets considered for summarization (in the time window specified by user)
m	Number of distinct content words included in the n tweets
i	index for tweets
j	index for content words
x_i	indicator variable for tweet i (1 if tweet i should be included in summary, 0 otherwise)
y_j	indicator variable for content word j
$Length(i)$	number of words present in tweet i
$Score(j)$	tf-idf score of content word j
T_j	set of tweets where content word j is present
C_i	set of content words present in tweet i

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \tag{3.3}$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m] \tag{3.4}$$

$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n] \tag{3.5}$$

where the symbols are as explained in Table 3.15. The objective function considers both the number of tweets included in the summary (through the x_i variables) as well as the number of important content-words (through the y_j variables) included. The constraint in Eqn. 3.3 ensures that the total number of words contained in the tweets that get included in the summary is at most the desired length L (user-specified) while the constraint in Eqn. 3.4 ensures that if the content word j is selected to be included in the summary, i.e., if $y_j = 1$, then at least one tweet in which this content word is present is selected. Similarly, the constraint in Eqn. 3.5 ensures that if a particular tweet i is selected to be included in the summary, i.e., if $x_i = 1$, then the content words in that tweet are also selected.

We use GUROBI Optimizer [46] to solve the ILP. After solving this ILP, the set of tweets i such that $x_i = 1$ represents the summary at the current time.

3.4.3 Summarizing Hindi tweets

We now describe how the summarization scheme is extended to summarize Hindi tweets, and the challenges therein. As mentioned in previous section, performance of our proposed summarization algorithm depends on extraction of content words. Usability of the various summarization algorithms on Hindi tweets is limited by the unavailability of natural language processing tools for Hindi tweets.

Extraction of content words: To our knowledge, there does not exist any Twitter-specific part-of-speech tagger for Hindi. Hence we apply a standard Hindi POS tagger [52] to identify nouns and verbs. For English tweets, we use the standard Twitter-specific POS tagger [38] having accuracy $\geq 90\%$. Hence, for English tweets we can detect content words with $\geq 90\%$ accuracy. In order to check how accurately we are able to detect such important words for Hindi tweets, we take five random samples of Hindi tweets, each sample containing 100 tweets. Content words (numeral, noun, and verb) were extracted from these tweets as marked by Hindi POS tagger [52]. Three annotators manually checked these content words and identified what fraction of these content words are correct. Overall, a mean accuracy close to 85% was achieved in detecting such content words. The accuracy of detecting content words is lower for Hindi than for English, because many general words also got annotated as content words. Hence, the limitation of POS tagging affects the performance of summarization of non-English tweets. This limitation hampers the diversity of information in the final summary generated.

3.4.4 Summarizing frequently changing information

As stated earlier, a special feature of the tweet streams posted during disaster events is that some of the numerical information, such as the reported number of victims or injured persons, changes rapidly with time. For instance, Table 3.16 shows how, during the HDBlast event, the reported number of victims / injured persons changed during a period of only seven minutes. Since such information is important and time-varying, we attempt to process such actionable information separately from summarizing the rest of the information. Additionally, disasters like hurricanes, floods and earthquakes often affect large geographical regions, spanning different locations. In such cases, numerical information usually varies across locations, such as *'19 People Killed In* **Bihar** *, 28 in* **India***, and 500+ killed in* **Nepal** *. #NepalEarthquake'*. To our knowledge, none of the prior works on processing tweet streams during disaster events have attempted to deal with such location-specific rapidly changing (or even conflicting) information[5].

Specifically, we consider particular disaster-specific key verbs like 'kill', 'die', 'injure', 'strand', and report the different numerical values attached to them, coupled with the number of tweets reporting that number. For instance, considering the tweets in Table 3.16, the information forwarded would be: 'seven people killed' is supported by two tweets, while 'ten killed' and 'fifteen killed' is supported by one tweet each.

Table 3.16: Variation in casualty information within a short time-span (less than 7 minutes), on the day of the Hyderabad blast (Feb 21, 2013).

Timestamp	**Extract from tweet**
14:13:55	seven killed in hyderabad blast [url]
14:16:18	at least 15 feared dead in hyderabad blast, follow live updates, [url]
14:19:01	10 killed in hyderabad blast more photos, [url]
14:20:56	hyderabad blast, 7 people are feared dead and 67 others are missing following a blast

Assigning numeral values to keywords: It is often non-trivial to map numeral values to the context of a verb in a tweet. For instance, the number 'two' in the tweet *'PM visits blasts sites in hyderabad, three days after two powerful bombs killed'* is *not* related with the verb *'killed'*, as opposed to the number 'seven' in the tweet

[5] Note that we only attempt to report all versions of such information; verifying which version is correct is beyond the scope of the current work.

'seven people were killed'. Therefore, whenever the numeral is not directly associated with the main verb, we extract the direct object of the main verb and check whether (i) the numeral modifies the direct object, and (ii) the direct object is a living entity. For example, in case of the tweet *'7 people killed in Hyderabad blast'*, the dependency tree returns the following five relations — (7, people), (people, killed), (in, killed), (blast, in), (Hyderabad, blast). In this tweet, 'people' is the direct object which is associated with the main verb 'killed' and the numeral 7 modifies the direct object 'people' which is a living entity. We use the POS tagger and dependency parser for tweets [68] to capture this information. If a numeral is directly associated with a main verb (i.e., if an edge exists between numeral and the verb in the dependency tree), we associate that numeral with the verb (e.g., 'seven' with 'killed' in *'seven killed in hyderabad blast'*). The list of living-entity objects for disaster specific verbs was pruned manually from the exhaustive list obtained from Google syntactic *n*-grams[6].

Assigning locational information to keyverbs: Next, we attempt to associate such key verbs to specific locations (as tagged by the named entity recognizer). Note that it is often non-trivial to map locations to the context of a verb in a tweet. For instance, the number '17' in the tweet *'More than 450 killed in a massive 7.9 earthquake in Nepal and 17 killed in India , #NepalEarthquake.'* is *not* related with the location 'Nepal', rather it is related with the location 'India'. Therefore, whenever the numeral is associated with a main verb directly or through some living entity, we check whether any location is associated with that verb (verb and location are connected within a 2-hop distance in dependency parse tree). If there is no specific location information, as in the tweet *'More than 150 people died in Earthquake'*, we associate the global location name to that value. For example, in our case, we associate this information to *Nepal.* Hence, our methodology is able to simultaneously provide **global** updates as well as more granular location-specific **local** updates. The performance of our methodology is discussed in the next section.

[6] Available at `http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html`

3.5 Experimental results

This section compares the performance of the proposed framework (COWTS) with that of four state-of-the-art summarization techniques (baselines). We first briefly describe the baseline techniques and the experimental settings, and then compare the performances.

3.5.1 Experimental settings: baselines and metrics

We considered the first four disaster events described in Section 3.2 for the experiments. For each dataset, we considered the first 5000 tweet fragments in chronological order, extracted situational tweet-fragments using our classifier, and passed the situational tweets to the summarization modules. We considered two breakpoints at 2K, and 5K tweets, i.e., the summaries were demanded at the corresponding time-instants.

Establishing gold standard summaries: At each of the breakpoints, three human volunteers (same as those involved in the classification stage) individually prepared summaries of length 250 words from the situational tweets. In this step, volunteers were allowed to combine information from multiple related tweets but new words are not included as it may hamper overall computation. For example, if we have two tweets in hand — (i) *7 people died, 20 injured in bomb blast*, and (ii) *7 died, 20 injured in Hyderabad blast*, the annotators were allowed to form a tweet like *7 people died, 20 injured in Hyderabad bomb blast.* To prepare the final gold standard summary at a certain breakpoint, we first chose those tweet fragments which were included in the individual summaries of all the volunteers, followed by those which were included by the majority of the volunteers. In this final step also, we combine information from multiple related tweets. Thus, we create a single gold-standard summary containing 250 words for each breakpoint, for each dataset.

Baseline approaches: We compare the performance of our proposed summarization scheme with that of four prior approaches, which consist of recent disaster-specific extractive summarization techniques and real-time extractive tweet summarization methods. Note that the selected baselines include both generic tweet summarization approaches and disaster-specific approaches.

(i) **NAVTS:** since COWTS considers nouns, numerals and main verbs as content words, a question arises as to whether the choice of content words is prudent. To verify this, we devise a competing baseline where noun, verbs and adjectives are taken as content words; these parts of speech were found to be important for tweet summarization (not online) in a prior study by Khan *et al.* [65].

(ii) **Sumblr:** the online tweet summarization approach by Shou *et al.* [115], with a simplifying assumption – whereas the original approach considers the popularity of the users posting specific tweets (based on certain complex functions), we give equal weightage to all the users.

(iii) **APSAL:** is an affinity clustering based summarization technique proposed by Kedzie *et al.* [63]. It mainly considers news articles and focuses on human-generated information nuggets to assign salience score to those news articles while generating summaries. In our case, we apply it over tweets after removing Twitter specific tags like URLs, hashtags, mentions, emoticons etc using the POS tagger [38].

(iv) **TSum4act:** the methodology proposed by Nguyen *et al.* [83]. They prepare clusters of situational tweets using LDA [11], extract numerals, geo-location information, and events from tweets using the Twitter NER tool [107], construct a weighted graph among the tweets using cosine similarity as the edge weights, apply weighted PageRank [90], and finally select tweets based on Simpson similarity measure from each cluster.

We apply COWTS and all the above baseline methods on the same situational tweet stream (obtained after classification), and retrieve summaries of the same length, i.e., the number of words present in the gold standard summary for a certain breakpoint (described earlier). *To maintain fairness, the same situational tweet stream (after classification) was given as input to all the summarization approaches.* Note that while computing the length of the summaries, we do not consider the following seven tags as marked by the CMU POS tagger [38] – #(hashtags), @(mentions), (Twitter-specific tags), U(urls), E(emoticons), G(garbage), and punctuations. We maintain this scheme uniformly for the gold standard summaries, and the summaries generated by our method as well as all the baseline methods.

Evaluation metrics: We use the standard ROUGE [72] metric for evaluating the quality of the summaries generated. Due to the informal nature of tweets, we actually consider the *recall and F-score* of the ROUGE-1 variant. Formally, ROUGE-1 recall is unigram recall between a candidate / system summary and a reference summary, i.e., how many unigrams of reference summary is present in candidate summary normalized by the count of unigrams present in reference summary. Similarly, ROUGE-1 precision is unigram precision between a candidate summary and a reference summary, i.e., how many unigrams of reference summary is present in candidate / system summary normalized by the count of unigrams present in the candidate summary. Finally the F-score is computed as harmonic mean of recall and precision.

3.5.2 Performance comparison

Table 3.17: Comparison of ROUGE-1 F-scores (with classification, Twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard ROUGE stemming(-m) and stopwords(-s) option) for COWTS (the proposed methodology) and the four baseline methods (NAVTS, Sumblr, APSAL, and TSum4act) on the same situational tweet stream, at breakpoints 2K, and 5K tweets.

Step size	**ROUGE-1 F-score**									
	HDBlast					UFlood				
	COWTS	NAVTS	Sumblr	APSAL	TSum4act	COWTS	NAVTS	Sumblr	APSAL	TSum4act
0–2000	**0.6326**	0.6030	0.5374	0.5765	0.5167	**0.4817**	0.3686	0.2663	0.3980	0.4152
0–5000	**0.5893**	0.5441	0.4313	0.5606	0.3951	**0.4028**	0.2954	0.2638	0.3894	0.3825

Step size	**ROUGE-1 F-score**									
	SHShoot					Hagupit				
	COWTS	NAVTS	Sumblr	APSAL	TSum4act	COWTS	NAVTS	Sumblr	APSAL	TSum4act
0–2000	**0.6353**	0.6060	0.5669	0.6060	0.4329	**0.4736**	0.4084	0.3425	0.3141	0.3701
0–5000	**0.5705**	0.5705	0.5425	0.4672	0.4187	**0.4000**	0.3394	0.2486	0.2141	0.2321

Table 3.17 gives the ROUGE-1 F-scores for the five algorithms for the four datasets, at breakpoints 2K, and 5K respectively. It is evident that COWTS performs significantly better than all the baseline approaches. For instance, mean scores indicate an average improvement of more than 38% in terms of F-score over

Table 3.18: Summary of 100 words, generated at 5K breakpoint of the UFlood dataset by (i) COWTS (proposed methodology), (ii) TSum4act, another disaster-specific summarization methodology.

Summary by COWTS	Summary by TSum4act
Google launches Person Finder to help people. WATCH Uttarakhand , 100 houses collapse , 10 dead , 50 missing as rain batters Uttarakhand. Uttarakhand helplines , For Pauri , Haridwar , Nainital , 999779124 , 9451901023. Uttarakhand , Almora , Bageshwar , Pithoragarh helpline numbers are 9456755206 , 9634535758. Call 011-24362892 and 9968383478. Uttarakhand tragedy continues , death toll touches 200 , Hindustan Times. Monsoon fury , Toll rises to 131 , Kedarnath temple in mud. Landslides destroyed roads to towns. 50,000 stranded , 5000 stranded in Badrinath. Uttarakhand Floods relief nos , Uttarkashi , 01374-226126 , Chamoli , 01372-251437 Tehri , 01376-233433 , Rudraprayag 01732-1077. Uttarakhand , Chopper deployed for rescue operations crashes. 1000 Uttarakhand pilgrims sighted , work to identify bodies begins ht.	DAY-4 , RSS Swayamsevaks actively involved in relief activities at Uttarakhand , RSS appeals for Help , Uttarakha. Uttarakhand flood , Death toll crosses 550 , says CM , 50,000 still stranded , The Economic Times. Thousand of people still stranded in Uttarakhand. In Uttarkashi , Uttarakhand , flash floods triggered by heavy rains wash away houses along the river. 10 Crore for flood-affected-people in Uttarakhand. Narendra Modi lands in Uttarakhand , flies out with 15,000 Gujaratis. Uttarakhand flood helpline numbers , 0135-2710334 , 0135-2710335 , 0135-2710233. Uttarakhand flood , Stranded Karnataka pilgrims begin their journey back. Uttarakhand floods , These people are missing. Uttarakhand CM is going to Switzerland. Sources , Uttarakhand Govt rejected 24 choppers offered by Gujarat Govt for rescue work in the flood affected areas.

Sumblr [115] which is a general-purpose (i.e., not disaster-specific) summarization scheme. The proposed methodology also performs better than the disaster-specific summarization techniques TSum4act [83], and APSAL in all cases – on an average, we obtain improvement of 34%, and 24% for F-score over TSum4act and APSAL respectively. However, in some cases performance of TSum4act and APSAL is at par with our proposed method COWTS. Further, the higher F-scores for COWTS than those for NAVTS indicate that our selected content words lead to better summarization. We also see that the better performance of COWTS remains consistent even if we increase the number of tweets for summarization.

To give an idea of the nature of the summaries generated by the methods, Table 3.18 shows summaries of length 100 words, generated by COWTS and TSum4act (both disaster-specific methodologies) from the same tweet stream — at the 5K breakpoint during the UFlood event. The two summaries are quite distinct, with most of the tweets being different. We find that the summary returned by COWTS is more informative, and contains crucial information about hotline numbers, rescued and stranded victims, critical areas and infrastructure damages. On the other hand, the summary returned by TSum4act mostly contains similar types of information (about the relief efforts and evacuated people) expressed in various ways.

Time taken for summarization: Since time is critical during disaster events, it is important that the summaries are generated in real-time. Hence, we analyze the execution times of the various techniques. At the breakpoints of 2K and 5K tweets, the COWTS takes 7.759, and 9.562 seconds on average (over the four datasets) respectively to generate summaries. The time taken increases sub-linearly with the number of tweets. TSum4act requires huge amount of time due to cluster validation phase, computation of large similarity graphs, and execution of PageRank algorithm [83]. In our case, it takes 2350 seconds and 5500.52 seconds at 2K and 5K breakpoints. APSAL needs high time due to similarity matrix computation and affinity clustering method. In case of APSAL, running time increases exponentially with number of tweets. APSAL takes 61.51 seconds and 293.38 seconds at 2K and 5K breakpoints.

Benefit of classification before summarization: We verify that separating out situational tweets from non-situational ones significantly improves the quality of summaries. Considering all the four events together, the mean ROUGE F-score at breakpoint 2000 for COWTS was 0.4916 *without* prior classification (i.e., when all tweets were input to the summarizer) as compared to 0.5558 after classification. Table 3.19 gives the F-score of COWTS on classified and unclassified tweets, for all four events at two breakpoints. As time progresses, fraction of non-situational tweets is also increased in number which affects the summarization step to a great extent. As is evident from Table 3.19, F-score at 5K is significantly low when we consider whole set of tweets.

Table 3.19: ROUGE-1 F-score of COWTS on classified and unclassified tweets, over all four events at breakpoints 2K and 5K.

Events	**ROUGE-1 F-score**			
	Breakpoint-2k		**Breakpoint-5k**	
	Classified	**Unclassified**	**Classified**	**Unclassified**
HDBlast	0.6326	0.5478	0.5893	0.4200
UFlood	0.4817	0.4630	0.4028	0.3532
SHShoot	0.6353	0.5095	0.5705	0.4244
Hagupit	0.4736	0.4462	0.4000	0.3573

Effect of misclassification on summary recall: As stated in Section 3.3, the proposed classifier achieves around 80% accuracy and 0.81 recall in classifying between situational and non-situational tweets. We now investigate how the 20% error in classification affects the subsequent summarization of situational information. It is evident that 20% situational tweets are misclassified as non-situational tweets which is more critical during disaster.

We further check what fraction of content-words are really missed out due to misclassification. Across all the four datasets, more than 85.41% of the content-words present in the *mis-classified tweets* are also covered by the correctly classified situational tweets. Correctly classified situational tweets cover 84.48%, 83.55%, 87.57%, 86.07% content words present in misclassified tweets. This implies that only a small fraction of the content-words are missed in the stream sent for summarization.

Effect of choice of content words: Choosing what type of words to focus on is important for achieving a good summarization of tweet streams, as also observed in [65]. As stated in Section 3.4, we consider three types of content words – numerals, nouns, and verbs. From the comparison between COWTS and NAVTS, it has already been established that our choice of content words achieves better summarization for tweets posted during *disaster events*, than the information words proposed in [65].

We now analyze whether all the three chosen types of content words are effective for summarization, by comparing the quality of the summaries generated in the *absence* of one of these types of content words. Figure 3.2 compares the F-scores

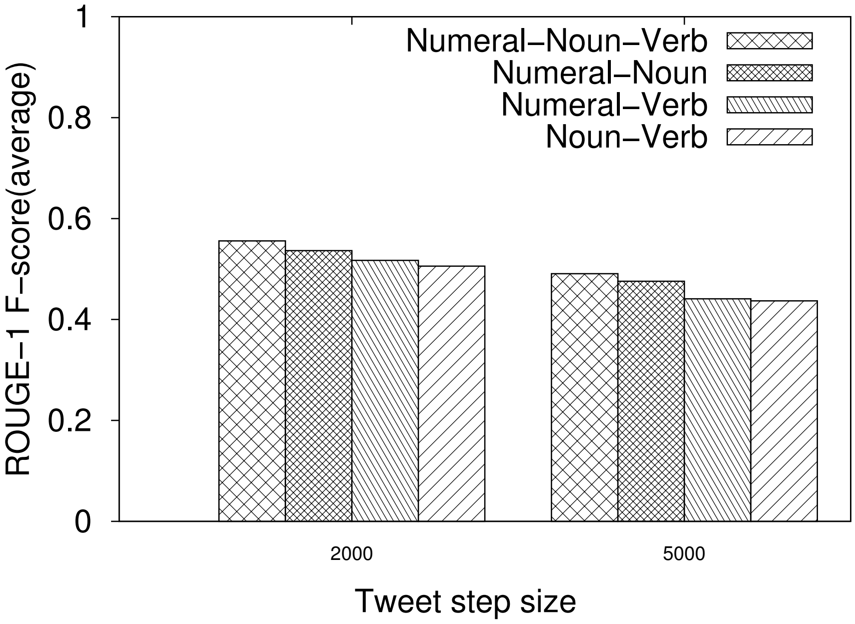


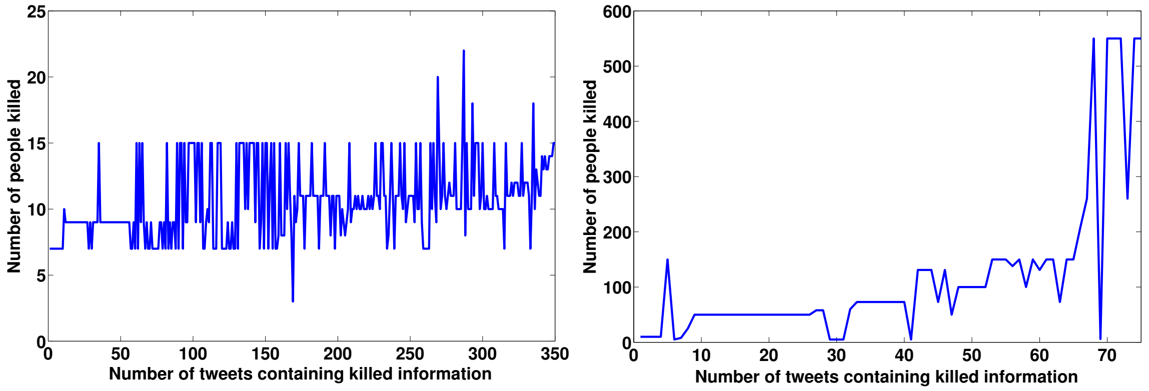
Figure 3.2: Effect of individual types of content words on the summary.

(averaged over all four datasets) considering all three types of content words, with those obtained by considering any two types of content words. It is clear that all three types of content words are important for summarization, numerals and nouns being the most important (since the numeral-noun combination outperforms the other 2-combinations).

Note that most of the earlier summarization frameworks *discarded* numerals contained in the tweets, whereas we show that numerals play a key role in tweets posted during disaster events, in not only identifying situational updates but also in summarizing frequently changing information (which we evaluate next).

Handling frequently changing numerals: Figure 3.3 shows how the numerical value associated with the key verb 'kill' changes with time (or sequence of tweets, as shown on the x-axis) during two different disaster events, HDBlast and UFlood. Clearly, there is a lot of variation in the reported number of casualties, which shows the complexity in interpreting such numerical information.

We now evaluate the performance of our algorithm in relating such numerical information with the corresponding key verb (as detailed in Section 3.4.4).



(a) HDBlast

(b) UFlood

Figure 3.3: Variation in the reported number of people killed, during two disaster events. The x-axis represents the sequence of tweets which contain such information.

Specifically, we check what fraction of such numerical information could be correctly associated with the corresponding key verb. We compare the accuracy of our algorithm with a simple baseline algorithm where numerals occurring within a window of 3 words on either side of the verb are selected as being related to the verb. Considering all the four datasets together, the *baseline algorithm has a precision of 0.63, whereas our algorithm has a much higher precision of 0.95*. Also, we achieve 100% accuracy for location tagging. These statistics show the effectiveness of our strategy in extracting frequently changing numerical information.

3.5.3 Application of the summarizer on future events

We envisage that the proposed classification-summarization framework will be trained over tweets related to past disaster events, and then deployed to extract and summarize situational information from tweet streams posted during future events. In this section, we demonstrate the utility of the framework by training it on the earlier four disaster events mentioned in Section 3.2, and then deploying it on tweets posted during the two most recent disaster events – NEQuake (the earthquake in Nepal in April 2015) [81], and HDerail (train derailment at Harda, India in June 2015) [50].

Table 3.20: Comparison of ROUGE-1 F-scores (with classification, Twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard ROUGE stemming(-m) and stopwords(-s) option) for COWTS (the proposed methodology) and the four baseline methods (NAVTS, Sumblr, APSAL, and TSum4act) on the same situational tweet stream, at two breakpoints (B1 = 10000, 2000 and B2 = 19102, 4361 for NEQuake and HDerail respectively).

Step size	**ROUGE-1 F-score**									
	NEQuake					HDerail				
	COWTS	NAVTS	Sumblr	APSAL	TSum4act	COWTS	NAVTS	Sumblr	APSAL	TSum4act
B1	**0.3650**	0.3224	0.2160	0.3444	0.3563	**0.4870**	0.4554	0.4312	0.3740	0.4489
B2	**0.3500**	0.2666	0.2000	0.3467	0.3368	**0.4864**	0.4653	0.4472	0.4673	0.4134

Summarization of English tweets: We directly use COWTS for summarizing the English tweets. We compute summaries at two breakpoints – in the middle of the stream B1 (at 10,000 and 2,000 tweets for the NEQuake and HDerail events respectively) and at the end of the stream. Three human volunteers were used to prepare gold standard summaries at these breakpoints following the approach used in Section 3.5. Table 3.20 states the ROUGE-1 F-scores for COWTS and the four baseline strategies NAVTS, Sumblr, APSAL, and TSum4act. It is evident that COWTS *has the highest F-score*. Further, COWTS takes 21.92 and 38.57 seconds respectively to summarize the tweets related to the NEQuake event, at the 10,000 and 19,102 breakpoints respectively, which is comparable or less than the time taken by the baseline approaches.

Summarization of Hindi tweets: We apply COWTS to situational Hindi tweets, and compare its performance with that of three baseline techniques NAVTS, Sumblr, and APSAL. We could not apply the TSum4act method [83] due to unavailability of named entity recognizers for tweets posted in Hindi. Similar to earlier evaluation frameworks, three human volunteers were used to prepare gold standard summaries. We compute recall, precision, and F-scores for Hindi tweets based on unigrams and after removing stopwords, doing lemmatization etc. as per the ROUGE-1 F-score [72][7]. The results are stated in Table 3.21 – it is evident that COWTS outperforms the baseline approaches in terms of coverage and quality of

[7] We could not use the standard ROUGE toolkit for Hindi tweets because it depends on English stopwords and lemmatization.

Table 3.21: Comparison of unigram F-scores for COWTS (the proposed method) and two baseline methods (NAVTS, Sumblr, and APSAL) on the same situational Hindi tweet stream.

Event	ROUGE-1 F-score			
	COWTS	NAVTS	Sumblr	APSAL
NEQuake	**0.5694**	0.4700	0.3588	0.5539
HDerail	**0.6833**	0.6146	0.5495	0.6717

the summaries.

The experiments in this section show that (i) COWTS is able to extract and summarize situational information from tweet streams posted during new disaster events satisfactorily, and in near real-time, and (ii) COWTS is extendable to any other language for which basic NLP tools are available, such as POS taggers.

3.5.4 Discussion on performance

A deeper look at various baseline techniques helps us to understand their shortcomings and the reasons behind the superior performance of COWTS. Again, the inferior performance of NAVTS, which is a variation of COWTS with different types of content words, brings out the importance of choosing proper content words for summarization.

Among the other baseline techniques, Sumblr [115] does not discriminate among different types of parts-of-speech, which potentially reduces the focus on important words. Additionally, Sumblr maintains clusters of related information and finally chooses one top scoring tweet from each cluster. Tweets within a cluster are ranked based on LexRank method; however, clusters are not ranked. In Sumblr, it is assumed that each cluster is of equal importance which may not be true because some clusters may contain more informative situational tweets compared to other clusters. Determining importance of clusters is also necessary for preparing the final summary. Similar types of tweet selection problems also arise in case of TSum4act [83]. TSum4act [83] captures disaster specific terms like numerals, events, noun-phrases, locations but it has two limitations — (i). determining importance of different clusters (same as Sumblr), (ii). determining appropriate number of clusters, and (iii). PageRank based iterative update takes long time for large datasets which creates a bottleneck for real-time summarization. APSAL uses salience scores of tweets to determine importance of different clusters. However, it is specifically designed for news articles which are formal and less noisy in nature. This might affect the output of APSAL. Both TSum4act and APSAL take large time to produce summaries; hence, they are not suitable for real-time updates of tweet streams. To resolve both the issues — information coverage and summarization in real-time, focusing on particular POS-tags, and ILP-based technique (as used in COWTS) prove to be very handy.

To be fair to other methods, most of them are *not* specifically designed to summarize tweet streams posted during disaster-specific events, which have their own peculiarities. We observe that across all types of disaster events, numerals, nouns, and key verbs provide salient situational updates during disasters. Hence, we set our summarization objective to maximize the coverage of these parts of speech in the final summary, by using an ILP-based technique. The strong points in favor of COWTS is that it is completely unsupervised and can be applied to any type of disaster events.

In case of Hindi tweets, we have less (varied) tweets compared to English. Hence, capturing important content words is relatively easy in case of Hindi tweets. Due to this, our system obtained high ROUGE-1 F-scores (Table 3.21) for Hindi tweets inspite of such resource constraints.

3.6 Conclusion

In this chapter, we present a novel classification-summarization framework for disaster-specific situational information on Twitter. We derive several key insights – (i) it is beneficial to work with tweet fragments rather than an entire tweet, (ii) low-level lexical and syntactic features present in tweets can be used to separate out situational and non-situational tweets, which leads to significantly better summarization, (iii) content words are especially significant for summarization of disaster-specific tweet streams, and (iv) special arrangements need to be made to deal with a small set of actionable keywords which have numerical qualifiers.

We had several realizations during the course of this work. For instance, whereas some disasters are instantaneous (such as bomb blast, or shooting incidents) and span short time durations, other events such as floods and hurricanes span much longer time periods. Such long ranging disasters consist of information related to various humanitarian categories ('infrastructure damage', 'missing people', $\cdots$) and small scale sub-events. Side by side, different stakeholders are looking for different kind of situational updates like overall update (proposed in this chapter), category specific updates etc. In the next chapter, we develop an efficient method to extract those sub-events and propose a general summarization method to satisfy the needs of different stakeholders.

Chapter 4

Identifying Sub-events and Summarizing Information during Disasters

Humanitarian organizations looking for information on social media microblogging platforms have the unenviable task of finding actionable information from among the vast amount of information posted during a disaster. Their operations can be streamlined if they can be presented with the information they need in a concise summary to help address their situational awareness needs.

In the previous chapter, we introduce the idea of low level lexical features and content words to classify and summarize situational tweets during disasters. However, it is observed that different stakeholders have different information needs at varying levels of granularities, such as, (a) high level situational update, (b) individual class level (infrastructure and damage, missing or trapped etc.) information or (c) a very special focus summary like detail information about people missing.

In this chapter, we first identify small scale sub-events from tweets and use those *sub-events*, *content words*, *class/ humanitarian category information* to summarize tweets posted across various humanitarian categories (infrastructure and damage, missing or trapped etc.) during disaster. Interestingly, the importance of the content words, sub-events, and class of interest can be tuned to generate high-level, class-level, and subject-focused summaries. Moreover, we propose an innovative scheme to represent the summary whereby underlying sub-events are highlighted and we provide the fraction of the tweets that are from a particular class (e.g., "infrastructure") in the summary. These sub-events and class information help the end-user get a grasp of the situation quickly.

4.1 Introduction

Microblogging platforms such as Twitter provide rapid access to situation-sensitive information that people post during mass convergence events such as natural or man-made disasters. Volunteers and other support personnel generate summaries and reports based on short messages such as tweets posted via Twitter that are then used by the responders [55].

Different stakeholders and responders need information at varying levels of granularities. Some stakeholders may want to obtain overall situational updates for a given day as a short summary or report (**high-level information need**) or specific updates for a particular class or category[1] such as 'infrastructure damage', 'shelter' etc. (**humanitarian category-based need**). They may also want to get updates at a much finer granularity with very **specific focus** on events, persons and locations connected with the disaster. For example, one may not only be interested in 'missing people', but, more specifically, they may be interested in finding out about the Australian mountain climbers who were at the foothills of Mt. Everest when the earthquake hits Nepal. Furthermore, these needs change over time after a disaster. For example, right after a disaster a responder may be interested in 'missing people', but after several days, the focus may be more on 'health and sanitation issues' or on 'infrastructure repair'.

In this chapter, we address all the above mentioned challenges. Instead of a

[1] The term category and class is used interchangeably throughout this thesis.

one-size-fits-all summary, as proposed in Chapter 3, we propose a perspective-based tweet summarization. Our perspective-based tweet summarization allows the end-user to generate summaries on any humanitarian class with varying granularity. To satisfy these various needs from different information-seekers, we propose a system to not only generate summaries of tweets based on topic classes, e.g., 'infrastructure damage' in disaster scenarios, but also to generate a summary that provides a high-level overview by combining information from different classes. First, we classify tweets into humanitarian classes. Information from a disaster can be classified into various categories and contains discrete sub-events. Our system identifies sub-events using noun-verb pairs in the tweets belonging to these classes. Apart from the noun-verb tuples, the system also identifies *content words*, i.e., nouns, numerals, and verbs that help improve information coverage in the generated summaries (Section 4.5). Crowdsource evaluation reveals that providing sub-event information along with the summary is helpful for comprehension (Section 4.5). For example, highlighting the phrase 'airport shut' in the tweet 'Kathmandu **airport shut**, flights from India canceled' helps users to understand the summary better.

The major contributions of this chapter are as follows:

- To provide a rapid, yet fine-grained overview of a crisis event, we propose a simple yet powerful noun-verb pair based sub-event detection approach that outperforms state-of-the-art sub-event detection approaches. Experimental results in Section 4.3 confirm that our extracted sub-events outperform traditional LDA, and, biterm topic modeling based methods [1, 99, 142].
- To address the multi-dimensional needs of different stakeholders, we propose a perspective-based tweet summarization technique using an integer linear programming (ILP) framework. The framework provides flexibility to add constraints that capture the information needs of end-users (Section 4.4).

We evaluate our proposed methods on 1.87M, 0.49M, and 0.24M tweets collected using the AIDR platform [56] corresponding to the 2015 Nepal earthquake, the 2014 Typhoon Hagupit, and the 2014 Pakistan flood respectively using both traditional IR metrics and crowdsourcing. Experiments conducted (reported in Section 4.5) over tweet streams show that the proposed tweet summarization method performs 6-30%

better in terms of ROUGE-1 score than existing methods. We use crowdsourcing to evaluate the quality of summaries and show that our method generates summaries that are significantly more useful compared to prior approaches [63, 83] in terms of information coverage, diversity, and readability (see Section 4.5). Almost all crowd-sourced responders opined that our key phrase highlighting feature quickly helps them to grasp the situation summarized.

4.2 Dataset and supervised classification

In this work, we are interested in long ranging natural disasters like flood, earthquake, typhoon etc. We collected crisis-related messages using the AIDR platform [56] from Twitter posted during three major natural disaster events:

1. **Nepal Earthquake (NEQuake):** This dataset consists of 1.87 million messages posted between April 25th and April 27th, 2015 fetched from Twitter using different keywords (e.g., Nepal Earthquake, NepalQuake, NepalQuakeRelief etc.).
2. **Typhoon Hagupit/Ruby (Hagupit):** This dataset consists of 0.49 million messages posted between December 6 and December 8, 2014 obtained using different keywords (e.g., TyphoonHagupit, TyphoonRuby, Hagupit, etc.).
3. **Pakistan Flood (PFlood):** This dataset consists of 0.24M messages posted on September 7th and 8th, 2014 obtained using different keywords (e.g., pakistanflood, PakistanFlood, Pakistanflood, etc.).

Preprocessing: We discard URLs, mentions, hashtag signs, emoticons, punctuation, and other Twitter specific tags and special characters from tweets. To identify such tags, we have used a Twitter POS tagger [38].

The dataset is then classified into (a) broad humanitarian categories (using AIDR [56]) and (b) individual sub-events. We develop an unsupervised algorithm for the detection of sub-events, which we discuss below.

4.2.1 Supervised classification into broad humanitarian categories

We train classifiers to categorize Twitter messages into categories that are useful for humanitarian operations. We discard tweets that are classified as "not-related" or "irrelevant" and keep the rest. This is an indispensable step to remove noise before performing event identification and summarization. We train a Logistic Regression classifier to classify messages using the labeled data taken from CrisisNLP [58]. The labeled datasets consist of around 2K labeled messages in each dataset and are annotated by humans into several humanitarian categories. These categories vary across different kinds of disasters like earthquake, flood etc. Some of the common categories are listed below.

1. **Injured or dead people:** Casualties due to the crisis
2. **Missing, trapped, or found people:** Questions and/or reports about missing or found people
3. **Displaced people:** People who have been relocated due to the crisis, even for a short time (includes evacuations)
4. **Infrastructure and utilities:** Buildings or roads damaged or operational; utilities/services interrupted or restored
5. **Shelter and supplies:** Needs or donations of shelter and/or supplies such as food, water, clothing, medical supplies or blood
6. **Money:** Money requested, donated or spent
7. **Volunteer or professional services:** Services needed or offered by volunteers or professionals
8. **Animal management:** Pets and animals, living, missing, displaced, or injured/dead
9. **Caution and advice:** Warnings issued or lifted, guidance and tips

Table 4.1: Description of dataset across three different events. NA indicates the absence / less number of tweets of a particular category for an event (i.e., no labeled data).

Category	NEQuake	Hagupit	PFlood
Missing, trapped, or found people	10,751	NA	2797
Infrastructure and utilities	16,842	3517	998
Donation or volunteering services	1,530	4504	27,556
Shelter and supplies	19,006	NA	NA
Caution and advice	NA	25,838	NA
Displaced people and evacuations	NA	18,726	NA

10. **Personal updates:** Status updates about individuals or loved ones

11. **Sympathy and emotional support:** Thoughts and prayers

12. **Other relevant information:** Other useful information that helps one understand the situation

13. **Not related or irrelevant:** Unrelated to the situation or irrelevant

In this chapter, we have taken some of the categories from above mentioned list. The categories are shown in Table 4.1. We used unigrams, bigrams, and POS tags as our features.

We use 10-fold cross-validation for the evaluation of the trained models. The classifiers produce AUC scores of 0.81, 0.72, and 0.70 for the Nepal, Hagupit, and Pakistan datasets respectively. We use the trained models to predict the labels for the rest of the unlabeled messages in the datasets. Messages with machine prediction confidence ≥ 0.80 are then selected for the next stages, i.e., sub-event identification and summarization. Table 4.1 shows the selected categories and proportion of messages for each of the datasets. Tweet-ids of the tweets of above mentioned datasets are available at `http://crisisnlp.qcri.org/lrec2016/lrec2016.html`.

While supervised classification works for determining the broad categories into which the tweets can be classified, for the purposes of sub-event detection, such a supervised approach cannot be used because we do not know the sub-event classes a priori.

4.3 Unsupervised identification of small-scale sub-events

A major disaster results in a number of small-scale sub-events, such as ‘power outage’, ‘bridge closure’, etc. Identification of these sub-events is crucial for a thorough understanding of the situation. Each category (e.g., infrastructure damage) is further divided into sub-events like ‘airport shut’, ‘building collapse’, etc. Latent Dirichlet Allocation (LDA) is a popular topic-detection algorithm [11] which can be used to generate topics and sub-topics. It outputs the most probable words belonging to each topic. However, domain experts at the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) found that LDA-based topics are too general to act upon [131].

Upon analyzing a few hundred tweets from each category and events’ time-lines from web sources[2], we find that messages which report the most important sub-events after a major disaster, consist of two nuggets: 1) entity/noun (e.g., person, place, organization, etc.), i.e., the entity that the event is about, and 2) an action-part/verb (e.g., destroyed, closure, etc.), i.e., the part that specifies the type of incident that happened to the reported entity.

Table 4.2 provides examples of some sub-events from various categories. These sub-events show important yet very specific information after the Nepal earthquake disaster. We seek to generate these automatically.

Forming noun-verb pairs: We extract nouns and verbs present in each message by using Twitter POS tagger [38]. However, detecting correct associations between nouns and verbs is a non-trivial task. For example, in the tweet: `#China media says buildings toppled in #Tibet _URL_`, both the words, ‘says’ and ‘toppled’ are identified as verbs. The noun ‘building’ is related to the term ‘toppled’ but it is not

[2] `goo.gl/mKfdiy`

Table 4.2: Popular sub-events learned from the first day of the Nepal earthquake (Apr 25, 2015).

Category	Sub-events
Infrastructure	'service affect', 'airport shut', 'road crack', 'building collapse', 'tower topple'
Missing	'family stuck', 'tourist strand', 'rescue location', 'database track', 'contact number'
Shelter	'field clean', 'medicine carry', 'emergency declare', 'deploy transport', 'aircraft deploy'

related to the verb 'says'. Hence, ('building','toppled') forms a valid sub-event whereas ('building','says') does not. Note that sometimes such nouns may not always appear prior or adjacent to the verbs in a tweet. For example, in the tweet: `India sent 4 Ton relief material, Team of doctors to Nepal`, ('material','sent') is a valid sub-event but the noun 'ton' appears closer to the verb 'sent' than the noun 'material'. Earlier, Cai *et al.* [18] showed dependency grammar based subject verb evaluation in formal sentences. Following their approach, we associate a noun to a verb accurately using the dependency edge information as obtained from the Twitter dependency parser [68].

Ranking sub-events: Since a sub-event is represented by a noun-verb pair (e.g. ('airport','shut')), we postulate that an event is important if the constituent words in the pair have not (rarely) occurred separately in the document. Accordingly, we compute the Szymkiewicz-Simpson overlap score of a sub-event S (N, V) using Equation 4.1:

$$Score(S) = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{4.1}$$

where X indicates the set of tweets containing N and Y indicates the set of tweets containing V. A higher value would mean that the particular noun (verb) is solely used for signifying the sub-event - highlighting its importance - this phenomenon is observed in the dataset.

However, Equation 4.1 does not discriminate between frequent and infrequent sub-events. For example, suppose we have two sub-events ($N1$,$V1$) and ($N2$,$V2$), where N_1, V_1 represent sets of tweets containing $N1$, and $V1$ respectively (similarly for $N2$, and $V2$). Let both of these sub-events satisfy following conditions — (i). $min(|N_1|,\ |V_1|) = 100$, (ii). $min(|N_2|,\ |V_2|) = 1$, (iii). $|N_1 \cap V_1| = 100$, and (iv). $|N_2 \cap V_2| = 1$. Equation 4.1 provides equal score to both the sub-events. To overcome this problem, we apply a discounting factor δ proposed by Pantel and Lin [91] to Equation 4.1. The discounting factor reduces the score of infrequent events.

$$\delta(S) = \frac{|X \cap Y|}{1 + |X \cap Y|} * \frac{min(|X|, |Y|)}{1 + min(|X|, |Y|)} \tag{4.2}$$

The weight of a sub-event S is computed as follows:

$$Weight(S) = Score(S) * \delta(S) \tag{4.3}$$

Finally, our system ranks the sub-events based on their weights. We term our DEPendency parser based SUB-event detection approach as **DEPSUB**. We evaluate the performance of our proposed sub-event detection approach in Section 4.5.

4.4 Sub-event based extractive summarization

Different humanitarian organizations look at information from social media using different perspectives. Some NGOs are interested in specific information like missing persons or volunteer services (class-specific update), whereas others need an overall view of the current situation (high-level update) so that they can make high-level decisions. Some information may be time-critical, e.g., a person trapped under some building or missing from some location whereas other information does not need immediate action, e.g., Dharara Tower needs to be rebuilt, but that will possibly be over years. To capture information from different information classes ('infrastructure', 'missing' etc.) at various granularity levels ('class specific update', 'high level update' etc.), we propose a generalized summarization framework, which can be customized to fulfill specific requirements of different stakeholders based on their needs at run-time. First, we explain our general summarization framework and then explore its application in three different scenarios[3].

4.4.1 Disaster-specific summarization

In Chapter 3, we show that in a disaster scenario, content words are important and an effective summary can be generated by maximizing the number of content words in the summary. In this chapter, we use the observation that small scale sub-events are also useful in determining the importance of tweets because such sub-events more or less capture real world events. Beyond sub-events, the importance of humanitarian categories needs to be considered while generating summaries - depending upon the view one seeks. Hence to sum-up, **content words**, **sub-events** and tweets pertaining to certain **humanitarian categories** are taken into consideration while producing a summary. This is put together in an ILP framework to summarize a set of tweets as discussed below.

The importance of a content word is computed using the tf-idf score (with sub-linear tf scaling). Similarly, weight of each sub-event is computed using Eqn. 4.3. The weight of sub-events are in the [0,1] scale. Hence, all the weights of content words are also normalized in [0,1] scale. Our summarization framework tries to optimize the parameters: content words and sub-events. First, it tries to maximize the coverage of important content-words to capture situational awareness. The framework uses the weight of sub-events present in that tweet to determine its importance. It also tries to cover more informative tweets within a fixed word limit. In Table 4.3, we provide examples of some sub-events and their corresponding short summaries[4] (shortened due to space limitations). Further, we take into consideration the category of the tweet (described next) while including them in the summary. We term our summarization framework as SUB-event based COntent Words Summarization (SUBCOWTS).

ILP Formulation: The summarization of L words is achieved by optimizing the following ILP objective function, whereby the highest scoring *tweets* are returned as

[3] Note that this is not an exhaustive list.

[4] We have removed the following tags from tweets [E,U,@,#,G,~] based on the POS tagger [38]

Table 4.3: Sub-events and short extracts from the tweets posted on the first day of the Nepal Earthquake (25th April).

Sub-event	**Sub-event summary**
communication cut	China's Tibet severely affected by Earthquake; houses collapsed, communications cut off
flight cancel	Flights to Kathmandu hit: Flight services to Kathmandu were today cancelled or put on ho; Kathmandu airport closed Saturday after a strong earthquake struck the country. All flights canceled.
medicine send	4 Tonne relief materials carrying food & medicines Earth excavation equipments have been sent to Nepal from India; India Sends Medicines, Blanket & Other Relief Materials To Nepal

the output of summarization. We use the GUROBI Optimizer [46] to solve the ILP. After solving this ILP, the set of *tweets* i such that $x_i = 1$, represents the summary. The symbols used in the following equations are as explained in Table 4.4.

$$\begin{aligned} max((1-\lambda_1-\lambda_2).\sum_{i=1}^{n} x_i.ICL(CL(i)) + \\ \lambda_1.\sum_{j=1}^{m} Con\_Score(j).y_j.max_{i\in TC_j}(ICL(CL(i))) + \\ \lambda_2.\sum_{k=1}^{p} Sub\_Score(k).z_k.max_{i\in TS_k}(ICL(CL(i)))) \end{aligned} \tag{4.4}$$

In Eqn. 4.4, the scores of each of the content words and sub-events are multiplied by the weight of the highest informative category in which this content word or sub-event is present. For example, 'airport' belongs to both infrastructure and shelter categories and its weight also varies based on the number of tweets present in a category. In our ILP framework, the weight of 'airport' is multiplied by the informative score of either 'infrastructure' or 'shelter' class depending on which one is higher. The importance of tweets, content-words, and sub-events is regulated by the parameters λ_1, λ_2.

The equation is, however, subject to the following constraints which are explained

Table 4.4: Notations used in the summarization technique.

Notation	**Meaning**
L	Desired summary length (number of words)
n	Number of *tweets* considered for summarization (in the time window specified by user)
$m\,,p$	Number of distinct content words and sub-events included in the n *tweets* respectively
q	Number of *categories* considered for summarization (each of the tweets belong to some category)
$i\,,j\,,k\,,a$	index for *tweets*, *content words*, *sub-events*, *categories*
x_i	indicator variable for *tweet i* (1 if *tweet i* should be included in summary, 0 otherwise)
y_j	indicator variable for content word j
z_k	indicator variable for sub-event k
$Length(i)$	number of words present in *tweet i*
$Con\_Score(j)$	tf-idf score of content word j
$Sub\_Score(k)$	tf-idf score of sub-event k
$ICL(a)$	importance of class a
TC_j	set of *tweets* where content word j is present
$C_i\,,S_i$	set of content words and sub-events present in *tweet i*
TS_k	set of *tweets* where sub-event k is present
$CL(i)$	category of tweet i
TCL_a	set of *tweets* belonging to *category a*
λ_1, λ_2	tuning parameter – relative weight for content word, and sub-event score

below.

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \tag{4.5}$$

Eqn. 4.5 ensures that the total number of words contained in the *tweets* that get included in the summary is at most the desired length L (user-specified).

$$\sum_{i \in TC_j} x_i \geq y_j, j = [1 \cdots m] \tag{4.6}$$

$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n] \tag{4.7}$$

Eqn. 4.6 ensures that if the content word j is selected to be included in the summary, i.e., if $y_j = 1$, then at least one *tweet* in which this content word is present is selected. Eqn. 4.7 ensures that if a particular *tweet* is selected to be included in the summary, then the content words in that *tweet* are also included in the summary.

$$\sum_{i \in TS_k} x_i \geq z_k, k = [1 \cdots p] \tag{4.8}$$

$$\sum_{j \in S_i} z_k \geq |S_i| \times x_i, i = [1 \cdots n] \tag{4.9}$$

Eqn. 4.8 ensures that if sub-event k is selected in the final summary, i.e., if $z_k = 1$, then at least one *tweet* which covers that sub-event is selected. Eqn. 4.9 ensures that if a particular *tweet* is selected to be included in the summary, then the sub-events in that *tweet* are also considered in the summary to compute the optimal value.

$$\sum_{a=1}^{q} ICL(a) = 1, a = [1 \cdots q] \tag{4.10}$$

Eqn. 4.10 ensures that sum of weight/ importance of all the classes is 1.

$$\sum_{i \in TCL_a} x_i \geq \delta, a = [1 \cdots q] \text{ if } ICL_a > 0 \tag{4.11}$$

Eqn. 4.11 ensures that at least δ tweets from each class whose importance is greater than 0 will be included in the final summary.

4.4.2 Scenario specific summarization

Humanitarian organizations analyze an event from different perspectives and accordingly their information needs vary. We take into account these varying information needs as different settings for our framework under which summaries are generated. Specifically, we generate: (i) **High level summarization**, (ii) **Humanitarian category specific summarization** and (iii) **Missing person summarization**. We explain how the constraints set in the generalized equation is customized for each case.

Scenario 1: High level summarization

Government agencies, and, NGOs look for a high-level overview of the current situation. In such cases, we have to generate an informative summary by taking important information from all the categories. Some simple customizations of Equations. 3.2 - 4.11 help us achieve that - (a) Given q classes, the importance of each class (ICL) is set to $\frac{1}{q}$. (b) The parameter δ in Eqn. 4.11 is set to 2, which represents the minimum number of items (tweets) that must be included in the final summary.

Scenario 2: Humanitarian category specific summarization

End users demand an overview for each of the higher level informative categories like ‘infrastructure’, ‘missing’ etc. Hence, we summarize information for each of the categories separately. To achieve this, we perform a simple customization of Eqns. 4.4 - 4.11 - given q categories, importance of all categories except the category for which summarization needs to be generated is set to 0. It is set to 1 for the desired class.

Scenario 3: Missing person summarization

Ground-level rescue workers need specific details about missing persons like their name, last location, contact number, age etc. to launch search operations. Note

Table 4.5: Examples of missing person information posted during Nepal earthquake.

Missing my friend Azhar. 23 years age. Last location Sindhupalchok. Pls help.
Last seen at Birjung. Family members trying 2locate Krija (mother)n Piu(child) pl rt @tajinderbagga

that this is an example of specialized summary – several of which may be required at various stages of disaster management. Here the customization of the general framework is done at various levels: (a) Since for now the rescue workers are only interested in the "missing class", the importance of all other classes is set to 0 and $ICL(missing)$ is set to 1. (b) Such tweets do not contain any sub-events and important information is centered around 'name', and 'relation' of missing persons. Table 4.5 shows examples of such tweets. Hence, λ_2 in Eqn. 4.4 is set to 0. (c) The definition of "content word" is changed to fit the requirement. We consider the following parameters as content words for this summarization task: (i) **Name:** name of the missing person[5], (ii) **Relation:** personal relations like 'brother', 'wife', 'son', 'friend' mentioned in the tweet.

The performance of our proposed summarization techniques is discussed in the next section.

4.5 Experimental setup and results

We analyze the performance of our proposed sub-event detection and summarization framework over three recent disaster events.

4.5.1 Evaluation of sub-events

We perform a thorough evaluation of automatically identified sub-events to check their coverage with real-world events, their accuracy, and quality.

[5] We have used the Stanford named-entity-tagger [67] for name detection

4.5.1.1 Experiment settings: baselines and metrics

Given the sub-events identified by various methods from our datasets: NEQuake, Hagupit, and PFlood, we perform both qualitative and quantitative analysis of sub-event detection approaches. In this part we describe different evaluation metrics and baseline techniques. Performance of different methods is evaluated in the next part.

Baseline approaches: We use following four state-of-the-art disaster-specific sub-event detection approaches as our baselines:

1. **COS-clustering:** Clustering based sub-event detection approach proposed by Dhekar *et al.* [1]. We discard URLs, mentions, hashtag signs, emoti- cons, punctuation, and other Twitter specific tags using Twitter POS tagger. Finally, in the labelling phase each sub-event cluster is represented by top four words having the highest term frequency among all the words belong to that cluster.
2. **LDA-clustering:** Clustering based sub-event detection approach similar to the approach proposed by Dhekar *et al.* [1]. However, instead of identifying cosine similarity between tweets, LDA based topic modeling is used.
3. **BTM-clustering:** Similar approach like LDA based topic modeling. However biterm topic modeling [142] is designed for short and informal texts like tweets.
4. **SOM-clustering:** Self organizing map based automatic sub-event detection proposed by Pohl *et al.* [99].

None of the baseline techniques were designed to extract sub-events as noun-verb pairs. Each of the baseline techniques represents sub-events as a cluster and labels each cluster with top four words which can describe that sub-event cluster. Hence, for this experiment, each sub-event cluster is represented by four words having the highest probability of belonging to that cluster.

Evaluation metrics: One of the objectives of this chapter is to identify small scale sub-events in a way which can be useful for crisis responders in their decision making process. Hence we propose a new method to represent / identify sub-events instead of clustering based approaches. Apart from quantitative measures some qualitative subjective analysis is also required to validate the utility of the proposed approach. We perform two different kinds of evaluation —

1. **Evaluation using crowdsourcing:** To evaluate the importance and utility of our identified sub-events, we perform a user study using CrowdFlower. Over each day, we extract top fifteen sub-events based on our proposed method for each of the humanitarian categories. In a similar way, we identify fifteen sub-events using the above mentioned four baseline approaches.
2. **Evaluation using gold standard sub-events:** Sub-event identification is mainly a detection task. We use standard metrics *precision*, *recall*, and *F-score*. Because our sub-event detection approach (noun-verb pair) is different from other baseline approaches, we compute values of these metrics only for our proposed approach. For this, we consider all the sub-events identified by our proposed method and gold standard sub-events (described next) from each class and each day.

Establishing gold standard sub-events: Three human volunteers individually prepare sub-events as pair of nouns and verbs for each information class, each day, and each dataset. To prepare the final gold standard sub-events for a particular class, first, we choose those sub-events that are included by all the volunteers, followed by those sub-events that are selected by the majority of the volunteers. Thus, we create a gold-standard list of sub-events for each class, each day, and each dataset.

4.5.1.2 Performance evaluation of sub-event detection methods

Evaluation using crowdsourcing: We use crowdsourcing to judge the utility of our sub-event detection approach over four clustering based techniques. We ask three questions to the workers on CrowdFlower as follows:

- (Q1) Which of the five methods identifies least number of irrelevant sub-events?

A sub-event is irrelevant if it is a random selection of words / terms from tweets. For example (airport closed) is more relevant compared to (airport man).

- (Q2) Which of the five methods identifies the most useful sub-events for crisis responders to understand the situation in the disaster region? From a sub-event like (building collapse) rescue workers come to know they have to reach the place to save trapped people.
- (Q3) Which of the five methods is able to provide a clear situational overview (through the identified sub-events) of the disaster situation stated above?

A *crowdsourcing task*, in this case, consists of a task description, which asks the crowd-worker to read a list of fifteen sub-events generated by the five methods, and answer the above-mentioned three questions. In total we have 12 tasks for the NEQuake and Hagupit and 6 tasks for the PFlood. Each task is performed by 15 crowd-workers. For each class, each date and for each dataset, the option that gets the most votes from the 15 respondents is chosen as the winner.

Table 4.6 shows that sub-events identified by our proposed approach are able to provide a clear situational overview of the disaster situation (Q3). Finally, as mentioned in Section 4.1, the main objective of our sub-event identification approach is to come up with a meaningful set of sub-events rather than a random collection of words for each topic / sub-event. Results show that the sub-events identified by our method is not a random collection of words (Q1) and that they are also useful for crisis responders to understand the situation in a disaster region (Q2). It is interesting to note that our method represents sub-events as a pair of words whereas baseline methods represent sub-events as a collection of four words; still the sub-events identified by DEPSUB are easier to understand. These results justify that meaningful collection of words is more important compared to the collection of large number of words to represent a sub-event.

Evaluation using gold standard sub-events: Table 4.7 shows the precision, recall, and F-scores for our proposed approach for the three datasets, over various days and classes respectively. In case of disaster, any prespecified list of sub-events

Table 4.6: Results of the crowdsourcing based evaluation of sub-events for DEPSUB (our proposed method) and the four baseline techniques (COS, LDA, BTM, SOM).

Datasets	Method	Q1	Q2	Q3
NEQuake	DEPSUB	**0.84**	**0.67**	**0.92**
	COS	0.08	0	0
	LDA	0.08	0.33	0.08
	BTM	0	0	0
	SOM	0	0	0
Hagupit	DEPSUB	**1**	**0.67**	**1**
	COS	0	0.16	0
	LDA	0	0.17	0
	BTM	0	0	0
	SOM	0	0	0
PFlood	DEPSUB	**1**	**0.67**	**0.67**
	COS	0	0	0
	LDA	0	0.33	0.33
	BTM	0	0	0
	SOM	0	0	0

(like yellow card, half-time in sports matches) are not available and it will vary across different disasters and humanitarian categories. Hence, we propose an unsupervised sub-event detection approach in this chapter. From Table 4.7, we can see that our proposed approach is able to achieve precision around 65%—70% and recall around 85%—90%. Finally, average F-scores for NEQuake, Hagupit, and PFlood are 0.76, 0.75, and 0.78 respectively. Our proposed approach is able to cover most of the small scale sub-events (recall is around 90%). However, we observe that some of the rarely / infrequently occurring noun-verb pairs (‘afternoon fly’, ‘terminal flee’) do not make any sense. Actually, for this noun-verb association, we rely on Twitter dependency parser [68] which has its own limitations due to noisy and informal nature of tweets. These shortcomings slightly hamper the precision of the system.

Table 4.7: Precision (P), recall (R), F-scores (F) over different datasets for our proposed approach (DEPSUB).

Event	Date	Infrastructure			Missing			Shelter			Volunteer		
		P	R	F	P	R	F	P	R	F	P	R	F
NEQuake	25/04/2015	0.57	0.91	0.70	0.72	0.79	0.75	0.85	0.96	0.90	0.41	0.71	0.52
	26/04/2015	0.52	0.88	0.66	0.83	0.88	0.85	0.78	0.96	0.87	0.71	0.84	0.77
	27/04/2015	0.46	0.83	0.60	0.81	0.95	0.87	0.69	0.98	0.81	0.77	0.83	0.80

Event	Date	Infrastructure			Caution			Displaced			Volunteer		
		P	R	F	P	R	F	P	R	F	P	R	F
Hagupit	06/12/2014	0.80	0.85	0.83	0.67	0.83	0.74	0.57	0.83	0.68	0.83	0.85	0.84
	07/12/2014	0.81	0.92	0.86	0.65	0.88	0.75	0.54	0.85	0.66	0.68	0.86	0.76
	08/12/2014	0.65	0.67	0.66	0.68	0.91	0.77	0.63	0.85	0.72	0.76	0.86	0.81

Event	Date	Infrastructure			Missing			Volunteer		
		P	R	F	P	R	F	P	R	F
PFlood	07/09/2014	0.60	0.76	0.67	0.71	0.93	0.80	0.68	0.95	0.82
	08/09/2014	0.78	0.74	0.76	0.75	0.90	0.82	0.67	0.93	0.80

4.5.2 Evaluation of summarization method

We analyze the performance of our proposed summarization framework with recent disaster-specific and real-time summarization approaches. We discuss the baseline techniques and the experimental settings briefly, and then compare the performance of the techniques.

4.5.2.1 Experiment settings: baselines and metrics

Given the machine-classified messages from our datasets: NEQuake, Hagupit, and PFlood, we split the tweets by date: 25th April to 27th April, 2015 for NEQuake, 6th December to 8th December, 2014 for Hagupit, and 7th September to 8th September, 2014 for the PFlood.

Establishing gold standard summaries: For each information class over each day, three human volunteers individually prepared summaries of length 200 words from the tweets. To prepare the final gold standard summary for a particular class, first, we chose those tweets that were included in the individual summaries written by all the volunteers, followed by those tweets that were included by the majority of the volunteers. Thus, we create a gold-standard summary containing 200 words for each class. Finally, we have prepared a single gold-standard summary containing 200 words for each day and for each dataset by combining information from all the classes.

Baseline approaches: In Chapter 3, we have shown the importance of our chosen *content words* and efficiency of our proposed approach COWTS over general summarization techniques like Sumblr. In this chapter, we use the following three state-of-the-art disaster specific summarization approaches as our baselines:

1. **COWTS:** is an extractive summarization approach specifically designed for generating summaries from disaster-related tweets (Chapter 3).
2. **APSAL:** is an affinity-clustering based extractive summarization technique proposed by Kedzie *et al.* [63].
3. **TSum4act:** is a disaster-specific summarization approach proposed by Nguyen *et al.* [83].

Evaluation metrics: We perform two types of evaluations. First, we use the standard ROUGE [72] metric for evaluating the quality of summaries generated using the proposed as well as the baseline methods. In this case, due to the informal nature of tweets, we consider the recall and F-score of the ROUGE-1 variant only. Second, we perform user studies using paid crowdsourcing (described below). In both the cases we have generated a system summary of 200 words for SUBCOWTS and each of the baselines over each dataset classes. For the SUBCOWTS method, we determine the best values for λ_1, and λ_2 as 0.5, and 0.5 for NEQuake and 0.5, and 0.3 for Hagupit and PFlood respectively.

4.5.2.2 Performance evaluation of category/class based summarization

The format in which the summary is produced is highlighted in Table 4.10. We produce extractive summary by choosing the tweets which have been selected through the ILP framework. Besides that, if there is a sub-event responsible for selection of a tweet, it is highlighted. This helps in making the summary more explainable to the

Table 4.8: Comparison of ROUGE-1 F-scores (with classification, Twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard ROUGE stemming(-m) and stopwords(-s) option) for SUBCOWTS (the proposed methodology) and the three baseline methods (COWTS, APSAL, and TSum4act) on the same situational tweet stream for each class, for each day, and for each dataset.

Step size	**ROUGE-1 F-score (NEQuake)**															
	Infrastructure				Missing				Shelter				Volunteer			
	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act
25/04/2015	**0.4966**	0.4842	0.3691	0.3758	**0.5407**	0.5353	0.3162	0.1901	**0.5503**	0.5165	0.4513	0.4742	**0.4417**	0.4127	0.4405	0.3174
26/04/2015	**0.3719**	0.3496	0.3071	0.2387	**0.3848**	0.3066	0.3496	0.3694	**0.3689**	0.3674	0.3275	0.3610	**0.5704**	0.5524	0.4982	0.3426
27/04/2015	**0.4971**	0.3631	0.3657	0.3765	**0.3574**	0.3494	0.3478	0.2825	**0.4573**	0.4340	0.3238	0.3631	**0.7069**	**0.7069**	0.6941	0.6934

Step size	**ROUGE-1 F-score (Hagupit)**															
	Infrastructure				Caution				Displaced				Volunteer			
	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act
06/12/2014	**0.6200**	0.6190	0.4946	0.5655	**0.4658**	0.4498	0.2922	0.3566	**0.3989**	0.3955	0.2881	0.2558	**0.4966**	**0.4966**	0.4814	0.4444
07/12/2014	**0.6177**	0.6173	0.4339	0.4852	**0.3363**	0.3303	0.3202	0.3281	**0.3718**	0.3585	0.2500	0.2307	**0.4972**	0.4782	0.4294	0.2902
08/12/2014	**0.4857**	**0.4857**	0.3891	0.4413	**0.4175**	0.4169	0.3803	0.4125	**0.4277**	**0.4277**	0.3376	0.3812	**0.3829**	0.3701	0.3823	0.3816

Step size	**ROUGE-1 F-score (PFlood)**											
	Infrastructure				Missing				Volunteer			
	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act	SUBCOWTS	COWTS	APSAL	TSum4act
07/09/2014	**0.7306**	0.7232	0.6894	0.7191	**0.6039**	**0.6039**	0.5787	0.5769	**0.3651**	0.3378	0.2646	0.2092
08/09/2014	**0.7235**	0.7206	0.6781	0.6315	**0.4758**	**0.4758**	0.4705	0.4498	**0.3844**	0.2865	0.2105	0.2631

viewers as it is highlighted in the results given below.

Evaluation using gold-summaries: Table 4.8 shows the ROUGE-1 F-scores for the four algorithms for the three datasets, over various days and classes, respectively. For Nepal earthquake (NEQuake), we have shown results for first three classes, infrastructure, missing, and shelter. In case of volunteer class, we obtain similar results. It is evident that SUBCOWTS performs significantly better than all the baseline approaches. For instance, mean scores indicate an average improvement of more than 6%, 24%, and 30% in terms of ROUGE-1 F-score over COWTS, APSAL [63], and TSum4act [83] (disaster specific extractive summarization scheme) considering all the datasets.

Evaluation using crowdsourcing: We use the CrowdFlower[6] crowdsourcing platform. We take summaries generated from each class for each day using our proposed method and all three baselines. In total we have 12 instances (hence 48 summaries) for the NEQuake and Hagupit and 6 instances (hence 24 summaries) for the PFlood. A crowdsourcing task, in this case, consists of four summaries (i.e., one proposed and three from baseline methods) and the four evaluation criteria with their descriptions (as described below). Each task requires at least ten different

[6] `http://www.crowdflower.com/`

workers' agreement on an answer before we finalize it. The exact description of the crowdsourcing task is as follows:

"The purpose of this task is to evaluate machine-generated summaries using tweets collected during the Nepal Earthquake of 2015, the Typhoon Hagupit which happened in 2014, and the flood in Pakistan in 2014. We aim to built an automatic method to generate such summaries/reports useful for situational awareness (information that helps understand the situation on the ground after an incident) for crisis responders. For this purpose, we have used four different methods and we want to compare which one is better."

Given the summaries and their topic, we asked four questions to the workers on CrowdFlower as follows:

1. (Q1) Overall, which method in your opinion has the best information coverage?
2. (Q2) Overall, which method provides the most diverse information?
3. (Q3) Overall, which summary helps you quickly understand and comprehend the situation?
4. (Q4) Overall, do you prefer summaries with highlighted topics or without them?

Q1. Information coverage corresponds to the richness of information a summary contains. For instance, a summary with more informative sentences (i.e., crisis-related information) is considered better in terms of information coverage. From Table 4.9, we can see that our proposed method is able to capture more informative summary compared to other baseline approaches in around 67% cases.

Q2. Diversity corresponds to the novelty of tweets in a summary. A good summary should contain diverse informative tweets. While we do not use any explicit parameter to control diversity, the ILP framework relies on importance score of sub-events and different content words, which helps in capturing information from various dimensions. It is quite clear from Table 4.9 that the proposed summaries are found diverse in around 64% cases.

Table 4.9: Results of the crowdsourcing based evaluation of class based system summaries for SUBCOWTS (our proposed methodology) and the three baseline techniques (COWTS, APSAL, TSum4act). Values in the table indicate percentage(%) of times a method is preferred for a particular question (NA indicates question is not valid for a method).

Datasets	Method	Q1	Q2	Q3	Q4
NEQuake	SUBCOWTS	59	34	75	83
	COWTS	33	33	25	NA
	APSAL	8	25	0	NA
	TSum4act	0	8	0	NA
Hagupit	SUBCOWTS	75	75	75	92
	COWTS	8	25	17	NA
	APSAL	17	0	8	NA
	TSum4act	0	0	0	NA
PFlood	SUBCOWTS	67	83	83	83
	COWTS	33	17	17	NA
	APSAL	0	0	0	NA
	TSum4act	0	0	0	NA

Q3. Summary understanding attempts to measure how easy it is to comprehend the summary. This is where we ask the workers whether they get a mental picture of the situation and can think of some action after reading the summary. From Table 4.9, we see that overwhelming number of respondents (78% cases) found that SUBCOWTS facilitates quick understanding of the situation.

Q4. Necessity of sub-event highlight tries to measure whether users prefer such highlighting and whether that helps in improving comprehension. In Table 4.9, we can see that almost all respondents (Nepal - 83%, Hagupit - 92%, Pakistan - 83%) found highlighting helpful and provide more vivid picture compared to the flat versions.

Table 4.10 shows summaries generated by SUBCOWTS and TSum4act (both disaster-specific methodologies) from the same set of messages (i.e., tweets from infrastructure class posted on 26th April). The two summaries are quite distinct. On manual inspection, we find that summary returned by SUBCOWTS is more informative and diverse in nature compared to TSum4act. For instance, we can see

Table 4.10: Summary of length 50 words (excluding #,@,RT,URLs), generated from the situational tweets of the infrastructure class (26th April) by (i) SUBCOWTS (proposed methodology), (ii) TSum4act.

Summary by SUBCOWTS	Summary by TSum4act
All flights canceled as **airport closes** down after quake. Reporter kathmandu airport closed **following** 6.7 **aftershock** no planes allowed to land. Metropolitan police department **rescue team** at airport to nepal. Kathmandu **airport reopened**. Nepal quake photos show historic **buildings reduced** to rubble as survivor search continues. Death **toll** in the earthquake in nepal **exceeded** 2 thousand people	RT @MEAIndia: #NepalEarthquake update A fourth aircraft with 160 is expected to leave Kathmandu by about 1 am tonight. Reports state so far 9K death tolls.Heartbreaking.Deep condolences.Be strong Nepal. #NepalEarthquake #PrayersForNepal #IndiaStandsWithNepal. RT @Online_Salman: SRK Is Such A Badluck He Went Nepal And #Earthquake In Nepal. RT @ANI_news: Baba Ramdev had a narrow escape after stage he was addressing from in Kathmandu collapsed after #earthquake struck Nepal.

the SUBCOWTS summary contains information about flights, damages of buildings, airport close and reopen.

Table 4.11: Runtime (seconds) of different algorithms for each of the classes averaging over three days.

Datasets	Class	SUBCOWTS	COWTS	APSAL	TSum4act
NEQuake	infrastructure	130.17	12.88	1719.79	16.79K
	missing	103.96	7.20	646.18	7.97K
	shelter	226.70	16.78	2685.67	21.45K
	volunteer	22.58	1.98	10.35	0.84K
Hagupit	infrastructure	63.92	3.02	57.50	2.01K
	caution	205.97	19.91	3846.34	33.30K
	displaced	152.10	17.06	2144.39	22.22K
	volunteer	38.86	4.07	103.67	2.70K
PFlood	infrastructure	13.62	1.82	11.37	0.78K
	missing	42.32	3.61	100.13	2.55K
	volunteer	390.68	56.02	11542.43	75.69K

Time taken for summarization: As stated earlier, one of our primary objectives is to generate the summaries in real-time. Hence, we analyze the execution times of the various techniques. Table 4.11 provides detailed information about runtime of our proposed SUBCOWTS method and three other baselines. The time taken by our method is slightly higher to that of other real-time summarization approaches such as COWTS. TSum4act is disaster specific method but it is not suitable for real-time

Table 4.12: Effect of sub-events and content words on summarization.

Datasets	Class	SUBCOWTS	SUBCOWTS - subevents	SUBCOWTS - content words
NEQuake	infrastructure	**0.4552**	0.3989	0.4193
	missing	**0.4276**	0.3977	0.3404
	shelter	**0.4588**	0.4393	0.3958
	volunteer	**0.5730**	0.5578	0.5338
Hagupit	infrastructure	**0.5744**	0.5740	0.4385
	caution	**0.4065**	0.3990	0.3226
	displaced	**0.3994**	0.3946	0.3107
	volunteer	**0.4589**	0.4483	0.3917
PFlood	infrastructure	**0.7270**	0.7195	0.6112
	missing	**0.5260**	0.5250	0.5363
	volunteer	**0.3747**	0.3263	0.3742

summarization. APSAL is suitable for small datasets but running time increases exponentially with the number of tweets making it infeasible for real-time updates.

Discussion on performance: We have already shown the individual importance of content words and sub-events in the summarization phase. TSum4act [83] maintains a set of clusters and selects one top ranking tweet from each cluster. However, it assumes that each cluster is of equal importance which is not true in real life scenario. Some clusters may contain many useful information whereas some may be noises. APSAL [63] maintains clusters of related information and finally chooses an exemplar tweet from each cluster. Finally, it selects the exemplar tweets from each cluster based on its salience score (resolves the uniform selection issue of TSum4act). Accuracy of this method heavily depends on cluster formation and correct exemplar tweet selection. However, this method is designed for formal news articles; hence, many of its features are missing for tweets which are noisy and informal in nature which affects the performance of APSAL. SUBCOWTS performs better than COWTS because it works with both content words, and sub-events, and we have already shown in Section 4.3 that such sub-events are related to real-world events. We also perform extensive experiments to understand the individual importance of content words and sub-events in summarization. Table 4.12 compares the F-scores

Table 4.13: Comparison of ROUGE-1 F-scores for SUBCOWTS (the proposed methodology) and the three baseline methods (COWTS, APSAL, TSum4act) on the same tweet stream for each dataset, for each day.

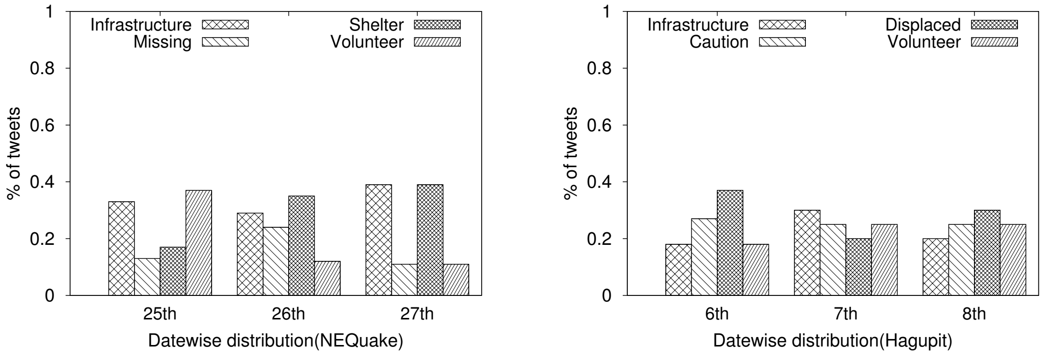
Datasets	Day	SUBCOWTS	COWTS	APSAL	TSum4act
NEQuake	25/04/2015	**0.4117**	0.3662	0.2215	0.3241
	26/04/2015	**0.3055**	0.2896	0.3055	0.2666
	27/04/2015	**0.3853**	0.3726	0.2866	0.3087
Hagupit	06/12/2014	**0.3223**	0.3008	0.1943	0.2460
	07/12/2014	**0.4124**	0.3569	0.2314	0.2492
	08/12/2014	**0.3475**	0.3002	0.2128	0.2359
PFlood	07/09/2014	**0.4524**	0.4141	0.2016	0.3014
	08/09/2014	**0.4145**	0.3085	0.1823	0.2030

(averaged over three days) obtained considering both sub-events and content words, with those obtained considering any one of these parameters. The results show that both content words and sub-events contribute to the quality of the summary, and removing either decreases the overall performance in all the cases.

4.5.2.3 Performance evaluation of higher level summarization

We have proposed a method that combines important information from all the classes for a particular day and generates an overall summary. The summary presentation has the following components:(a) tweets selected by the ILP framework, (b) highlights of sub-events and a mention of the class from which the tweet has been selected, and, (c) classwise distribution of tweets in the summary. For example, on 25th April in NEQuake event, the overall summary contains 33% tweets from infrastructure, 13% from missing, 17% from shelter and 37% from volunteer classes respectively.

Table 4.13 gives the ROUGE-1 F-scores for the four algorithms for the three datasets over different days. As noted above, the SUBCOWTS method generates additional class distribution information that the other baseline methods do not generate. To be fair, we exclude the additional class distribution information generated by



(a) NEQuake **(b) Hagupit**

Figure 4.1: Variation in the categories distribution during two disaster events across days.

SUBCOWTS while computing the ROUGE-1 F-scores. SUBCOWTS performs significantly better than the other baselines in terms of information coverage.

Role of humanitarian categories: We also check what happens if we consider each of the classes separately and compute the summaries of equal number of words from each of these classes using SUBCOWTS, COWTS, APSAL, and TSum4act and then combine them to form the overall summary. In many cases, the performance of these methods deteriorates in comparison to the present summarization scheme where tweets from all the classes are considered as a whole to produce final summary of a desired word length. This is because the importance of individual classes varies over the day and a summary needs to capture that.

Figure 4.1 shows the distribution of different classes across various days as captured by SUBCOWTS. Note that the trend is different for different disaster events. For the Nepal earthquake, the importance of the class ‘shelter’ increases and that of the class ‘volunteer’ decreases over days. In case of the Hagupit, information about ‘caution’ class decreases and ‘volunteer’ class increases. Keeping this high level view in mind, one can look into the summaries of specific categories for more details.

Evaluation using crowdsourcing: For general summarization, we asked same questions (Q1—Q4) to the crowd workers. We take summaries for each of the events over different days. Table 4.14 shows results for our crowdsource evalution.

Table 4.14: Results of the crowdsourcing based evaluation of high level system summaries for SUBCOWTS (our proposed methodology) and the three baseline techniques (COWTS, APSAL, TSum4act). Values in the table indicate percentage(%) of times a method is preferred for a particular question (NA indicates question is not valid for a method).

Datasets	**Method**	**Q1**	**Q2**	**Q3**	**Q4**
NEQuake	SUBCOWTS	100	100	100	100
	COWTS	0	0	0	NA
	APSAL	0	0	0	NA
	TSum4act	0	0	0	NA
Hagupit	SUBCOWTS	67	67	100	74
	COWTS	33	0	0	NA
	APSAL	0	33	0	NA
	TSum4act	0	0	0	NA
PFlood	SUBCOWTS	100	100	100	90
	COWTS	0	0	0	NA
	APSAL	0	0	0	NA
	TSum4act	0	0	0	NA

Our proposed approach performs the best for 88% of the cases with respect to information coverage and diversity. For all the cases, SUBCOWTS performs the best in terms of summary understanding and comprehension. The workers found that providing **humanitarian category, sub-event** information as well as the **distribution of class** is preferable and helpful for comprehension. We have shown these summaries to several disaster management experts. They particularly appreciated the meta-information provided, and said that the very information about temporal shift in the importance of constituent classes in the overall summary can be used to explore some classes with a finer granularity.

4.5.2.4 Performance of missing person information

Since other methods do not provide such specialized summarization, we concentrate on finding its coverage vis-a-vis the produced ground truth.

Establishing gold standard summaries: The ground-truth generation is a bit different than the previous cases because the required kind of information is very sparse. Hence we do not put any restriction on the number of words while generating a gold standard summary; the tweets which pass unanimous judgement from all the (three) volunteers are considered. For three days (25th, 26th, and 27th April), we have created summaries of 30, 305, 130 words respectively, for NEQuake event reflecting the availability. Similarly for PFlood event, we have created summaries of 110 and 80 words for 7th and 8th September. Our system also generates summaries of the same length as the ground truth.

Evaluation: Since we are primarily interested in coverage/recall score, we consider the recall of the ROUGE-1 variant only. We have obtained 100%, 82%, 87% score over three days (25th, 26th, 27th) respectively. For 26th and 27th, our proposed method fails to cover some information about missing persons. In case of PFlood, we have obtained 81%, 83% recall score for 7th and 8th September respectively. The mistakes specially occur where instead of name - only relationship information is present (25%) - e.g., *"My brother is missing"*. Also there are spelling mistakes and short-hand expressions (doughter, bro etc.) which our system fails to capture.

4.6 Discussion and conclusion

After interacting with several responders to disasters, we realize that categorization and summarization of information in the tweets along with finer granularity of sub-event detection is a pressing problem in the real world. Thus, in this chapter, we propose a generalized disaster-specific summarization approach, which can generate summaries across various scenarios. We consider three scenarios: (i) generating a general overview of the situation, (ii) generating summaries specific to various humanitarian classes / categories, and, (iii) generating summaries about missing people. The importance of the different humanitarian classes / categories (infrastructure, missing, shelter etc.) varies over days. Though in the case studied, we weight each category uniformly, but category weight / importance can be set accordingly in Eqn. 4.4 as per expert knowledge. Our system allows changes to the definition of content words based on users' requirements. We also highlight the utility and need for a comprehensive multi-faceted summarization approach.

In this chapter, we present a noun-verb pair based approach to represent sub-events. We need parse tree information about tweets to detect sub-events (forming noun-verb pairs). Next, we propose an ILP-based summarization method which tries to maximize the coverage of sub-events and content words (noun, verb, numeral). During long ranging natural disasters, it is observed that number of tweets increases over days. It may be difficult to produce summaries in near real-time because the summarization method is dependent on sub-events and detection of sub-events require parsing information. In the next chapter, instead of parsing (to make the summarization method independent of sub-events), we try to combine information from multiple related tweets (abstractive summarization) to produce the final situational summary.

Chapter 5

Abstractive Summarization of Information during Disasters

In Chapter 3 and 4, we develop extractive summarization methods to retrieve concise situational updates during disaster. However, it is observed that combining information from related tweets can help in better information coverage. In this chapter, we propose a two stage summarization framework which first extracts a set of important tweets from the whole set of information through an integer-linear programming (ILP) based optimization technique and then follows a word graph and content word based abstractive summarization technique to produce the final summary.

5.1 Introduction

In Chapter 3 and 4 we observe that microblogging platforms such as Twitter provide rapid access to situation-sensitive messages that people post during mass convergence events such as natural disasters. To get a quick overview of the event, the first step involves classifying them into different humanitarian categories such as infrastructure damage, shelter needs or offers etc (using AIDR [56]) and then summarizing information present in various humanitarian categories.

To this end, a straightforward and fast way would be to pick the messages that maximize the coverage of the content words (extractive summarization proposed in Chapter 3). In Chapter 4, we show that combining sub-events with content words helps in extracting more informative messages. In this chapter, we propose a summarization method which combines related information from several messages (abstractive summarization) to maximize the coverage of information within the specified word limit and produces the output in near real-time.

For example, consider the following tweets from Nepal earthquake that occurred in 2015 — 1. `Dharara Tower built in 1832 collapses in Kathmandu during earthquake`, 2. `Historic Dharara Tower Collapses in Kathmandu After 7.9 Earthquake`. Both tweets provide information about the collapsing of the Dharahara tower. Our objective is to combine important information from both of these tweets and generate a single meaningful situational tweet that contains all the relevant information like, `Dharara tower built in 1832 collapses in Kathmandu after 7.9 earthquake`.

Despite progress in natural language generation, generating abstractive summaries remains a hard problem. The algorithms, in general, are time-consuming. Hence if the abstractive approach is allowed to run over the entire incoming set of tweets, it may not be possible to produce the results in real-time (which is one of the important requirements during disaster).

In order to circumvent this problem, in this chapter, we propose a two stage summarization framework. Broadly, following steps are executed to generate the final summary —

1. First, tweets are classified into different humanitarian classes (‘infrastructure’, ‘missing’ etc.) using the AIDR platform [56].
2. Next, we try to generate summary for each of the classes. However, due to real-time constraint, we can’t apply abstractive approach over the entire tweet set. Hence, first we extract a set of important tweets from the whole set of a

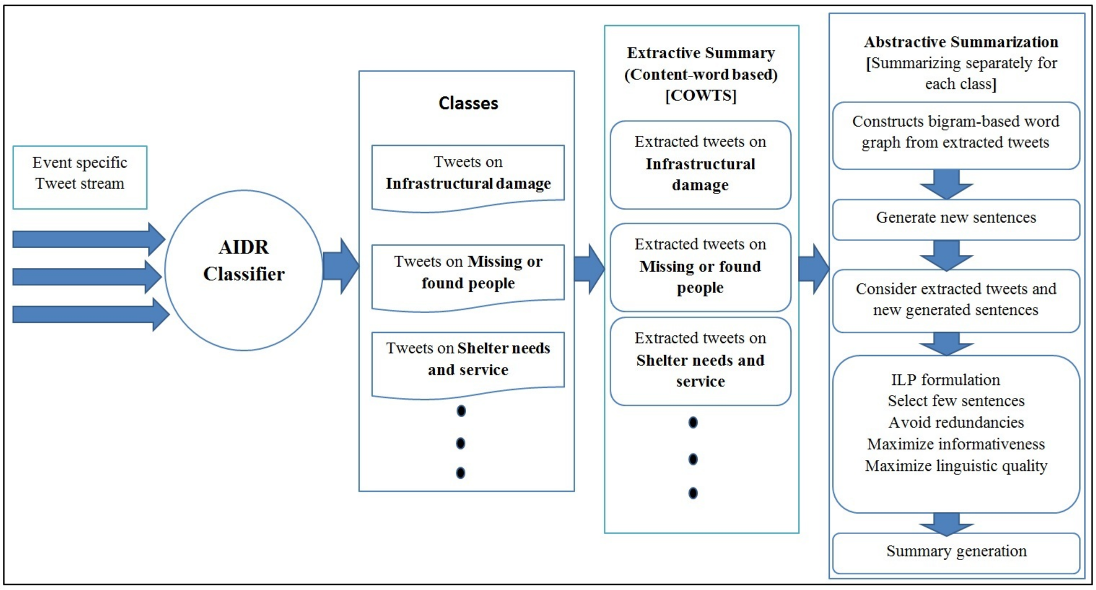

Event specific Tweet stream
AIDR Classifier
Classes
Tweets on Infrastructural damage
Tweets on Missing or found people
Tweets on Shelter needs and service
Extractive Summary (Content-word based) [COWTS]
Extracted tweets on Infrastructural damage
Extracted tweets on Missing or found people
Extracted tweets on Shelter needs and service
Abstractive Summarization [Summarizing separately for each class]
Constructs bigram-based word graph from extracted tweets
Generate new sentences
Consider extracted tweets and new generated sentences
ILP formulation
Select few sentences
Avoid redundancies
Maximize informativeness
Maximize linguistic quality
Summary generation

Figure 5.1: Our proposed framework for abstractive summarization of disaster-specific tweets.

particular humanitarian class using a fast but effective extractive summarization method COWTS (Chapter 3).

3. After that, we propose a word graph based technique which combines information from semantically similar tweets and generate new paths / sentences (*abstractive phase*).

4. Finally, we consider tweets and paths for a particular class and apply ILP-based content word coverage method to generate the final summary for each of the classes respectively.

Figure 5.1 provides an overview of our proposed approach. We test our proposed summarization approach on 1.87M, 0.49M, and 0.24M tweets collected using the AIDR platform [56] corresponding to the 2015 Nepal earthquake, the 2014 Typhoon Hagupit, and the 2014 Pakistan flood respectively. Our proposed *extractive-abstractive* summarization strategy performs better than other state-of-the-art disaster specific summarization methods.

Table 5.1: Description of dataset across three different events. NA indicates the absence / less number of tweets of a particular category for an event (i.e., no labeled data).

Category	NEQuake	Hagupit	PFlood
Missing, trapped, or found people	10,751	NA	2797
Infrastructure and utilities	16,842	3517	998
Donation or volunteering services	1,530	4504	27,556
Shelter and supplies	19,006	NA	NA
Caution and advice	NA	25,838	NA
Displaced people and evacuations	NA	18,726	NA

5.2 Dataset and classification of messages

In this chapter, we use the same three datasets — (i). Nepal earthquake (25th April - 27th April, 2015), (ii). Typhoon Hagupit (6th December - 8th December, 2014), (iii.) Pakistan flood (7th September - 8th September, 2014), as mentioned in Chapter 4. The categories taken from different datasets are shown in Table 5.1.

5.3 Automatic summarization

Given the categorized messages by AIDR for which the machine-confidence score is ≥ 0.80 (as described in Section 5.2), in this section we present our two step automatic summarization approach to generate summaries from each class. We consider the following key characteristics/objectives while developing an automatic summarization approach:

1. A summary should be able to capture the most important situational updates from the underlying data. That is, the summary should be rich in terms of information coverage.
2. As most of the messages on Twitter contain duplicate information, we aim to

produce summaries with less redundancy while keeping important updates of a story.

3. Twitter messages are often noisy, informal, and full of grammatical mistakes. We aim to produce readable summaries.

4. The system should be able to generate the summary in real-time, i.e., the system should not be heavily overloaded with computations such that by the time the summary is produced, the utility of that information is only marginal.

The first three objectives can be achieved through abstractive summarization and near-duplicate detection, however, it is very difficult to achieve that in real-time (hence violating the fourth constraint). In order to fulfill these objectives, we follow an extractive-abstractive framework to generate summaries. In the first phase (extractive phase), we use the summarization approach COWTS proposed in Chapter 3 and select a subset of tweets that cover most of the information produced and then in the second phase, apply an abstractive summarization method to obtain the final summary.

5.3.1 Extractive summarization approach

Disaster related tweets have distinct features that we use to construct our extractive summaries.

Content words: As identified in Chapter 3, in crisis scenarios some specific type of words play a key role by capturing important events and snapshots. Such useful words which we term as *content words* are — (i). numerals, (ii). nouns, (iii). location, (iv). main verbs.

Duplicates: Moreover, a large proportion of messages on Twitter contain redundant information. For instance, in the following five tweets, the same information related to the closure of Kathmandu airport and flights cancellation is conveyed in different ways:

1. `Nepal quake , Kathmandu airport shut, flights from India cancelled via @timesofindia`

2. `Flights to Kathmandu put on hold following powerful earthquake Read more here`
3. `Kathmandu airport shut, flights from India cancelled`
4. `K'mandu airport shut, flights from India cancelled via @timesofindia`
5. `After massive 7.9 earthquake, commercial flights to Kathmandu put on hold`

To handle duplicate or near duplicate information in the messages and to find disaster specific content words, we follow two schemes — (i) we remove duplicate and near-duplicate tweets (using the similar technique developed by Tao *et al.* [123]), and (ii) we focus on the content words during summarization.

We consider each class (infrastructure and utilities, missing, trapped or found people, shelter and supplies, $\cdots$) separately and try to extract concise summaries for each of these classes. Specifically, we take day-wise snapshots of each class, i.e., the system produces a summary of the desired length (number of words) over each day for each of the classes using COWTS. First we extract a set of content words, i.e., words with numeral, noun or verb pos-tags from the messages and try to maximize the coverage of these set of content words. In this phase, our main objective is to capture all the content words within a small number of tweet set such that the next phase of abstractive summarization can generate paths from these tweets and also rank those paths in near real-time. We observe that within 1,000 words limit, around 80% of the content words (present in the entire tweet set) can be covered within the chosen limited set of tweets. We illustrate the rationale behind the 1,000 word limit as follows.

Content-word coverage vis-a-vis length: In Figure 5.2, we show how the coverage of content words varies with the number of tweets extracted from the whole dataset for different classes of tweets posted during Nepal earthquake on 25th April, 2015. We observe a similar pattern for the other days. Within 1,000 word limit we are able to capture around 80% content words and the number of extracted tweets are manageable such that abstractive phase (described next) is able to construct paths from these tweet set in real-time. An informative set of 1,000 words turns out to be sufficient for the next stage of summarization; hence, we extract a set of tweets with 1,000 word limit constraint in our initial extractive phase of summarization.

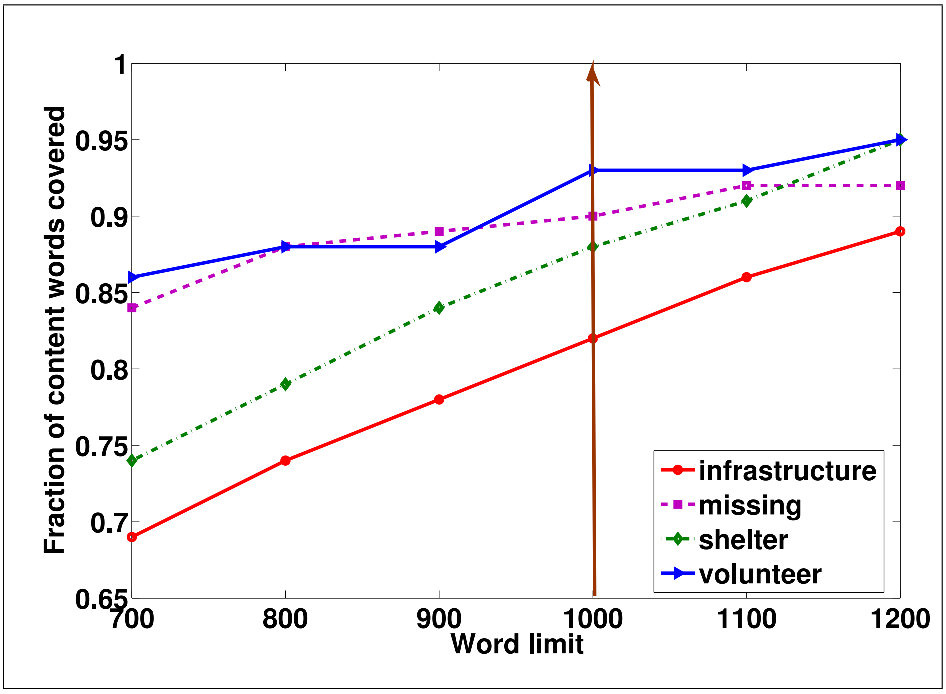


Figure 5.2: Variation in the coverage of content words with the number of extracted tweets for NEQuake event.

After extracting a set of important and informative tweets, we try to prepare a more concise and comprehensive summary through a COntent Words based ABStractive Summarization (**COWABS**) approach using these tweets (described next).

5.3.2 Abstractive summarization

The goal of this step is to generate an abstractive summary by combining information from multiple tweets. The generated summary must be comprehensive in the sense that it contains more information than extractive summaries of the same length (in words). Our abstractive summarization method is aimed at maximizing the informativeness of tweets while also avoiding redundancy of information. We follow an over-generate and select [134] strategy where we combine multiple tweets to generate a new sentence. Our method tries to select the best sentences from the set of generated sentences and creates a summary by optimizing three factors: **Informativeness, Redundancy** and **Readability**. Informativeness and readability have to be maximized, while redundancy is required to be minimized. *Informativeness* is defined as the amount of information in the summary, measured using a centroid-based ranking score. *Redundancy* is minimized such that we do not convey same or similar information in multiple sentences in the summary. We use a trigram-based log-likelihood score using a language model as a dummy representative of the *Readability* of the generated content. We adapt the ILP-based method for summarization proposed by Banerjee *et al.* [7] for news summarization; however, we make several modifications to make it usable for tweet summarization. Instead of a unigram-based sentence generation technique, we employ a bigram-based method. This adaptation improves the grammaticality of the resulting summaries. We also introduce a content-word based parameter in the ILP to tackle informativeness and redundancy.

Sentence generation process: In order to generate sentences, we build up a **word-graph** [33] with the entire tweet set where each tweet is iteratively added to the graph with the bigrams (adjacent words along with their parts-of-speech (POS) tags[1]) representing the nodes. An edge in the graph represents consecutive words in a sentence. When a new tweet is added to the graph and it contains a bigram that already exists in the graph, the nodes of the new tweet are merged with the existing nodes. We merge the nodes if the words in the bigrams have the same lexical form as well as the same POS tag. POS tags help maintain grammaticality and avoid potentially spurious fusions.

An example of our bigram-based word-graph construction is shown in Figure 5.3. Each node has been labeled with the form $w1 \parallel w2$, where $w1$ and $w2$ refer to the first and the second word in every bigram, respectively. We mark two nodes as the start and the end nodes that indicate the beginning and end of the tweets. The graph is generated considering the following two tweets, which were tweeted on a particular day and were assigned to the infrastructure class by the AIDR system — (i) `dharara tower built in 1832 collapses in kathmandu during`

[1] We employ a Twitter specific POS tagger [38]. In addition to the regular parts-of-speech tags, it also tags hashtags, retweet mentions, URLs separately. We ignore such words that have these specific hashtags because they are not important in the context of summarization as it might affect readability.

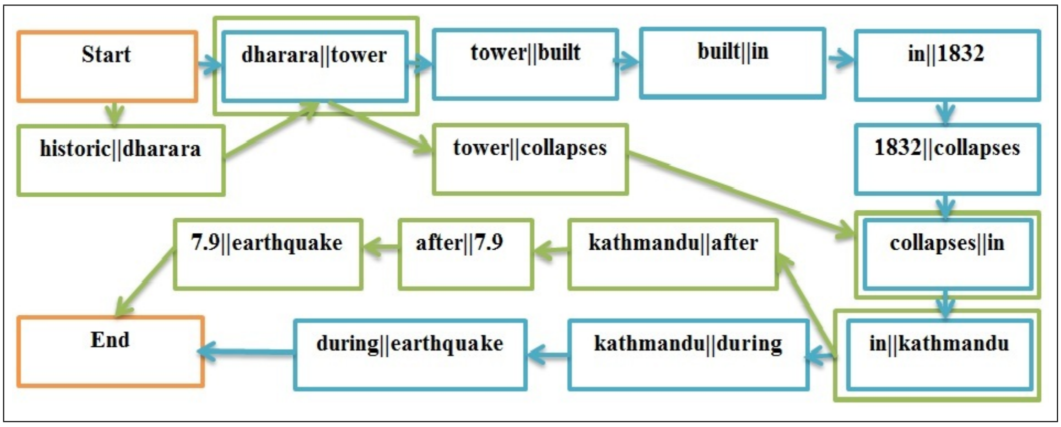


Figure 5.3: Bigram word graph generated using above two tweets (we do not show POS tags in the figure to maintain clarity). Nodes from different tweets are represented by different colours. Common nodes contain both the colours. Start and End are special marker nodes.

`earthquake`, and (ii) `historic dharara tower collapses in kathmandu after 7.9 earthquake`. We lower-case all words during the graph construction.

Once the graph is formed, sentences, which we term as *tweet-paths* are generated by **traversing paths in the graph** between the dummy *Start* and the *End* nodes. For example, from the graph in Figure 5.3, we can easily generate a *tweet-path* such as *dharara tower built in 1832 collapses in kathmandu after 7.9 earthquake*. Several such sentences might hold more information than the original tweets, yet containing the same or similar number of words. We set a minimum (10 words) and maximum (16 words) length for a sentence to be generated. We apply such constraints to avoid very long sentences that might be grammatically ill-formed and very short sentences that are often incomplete. In a real-scenario, the number of generated *tweet-paths* can be of several thousands, because there can be multiple points of merging across several tweets.

After this step, we have a set of tweets (extracted in first step) and tweet-paths (generated from the extracted tweets using graph traversal) in our hand and our goal is to select the best tweets or tweet-paths with the objective of generating a readable and informative summary. We formulate an ILP problem to select the final tweets, tweet-paths and construct the summary.

ILP Formulation

The ILP-based technique optimizes based upon three factors - (i) weight of content words (this is similar to that adopted during the extractive phase): The formulation tries to maximize the number of important content words in the final summary. Importance of a content word is captured through its weight. (ii) Informativeness of a tweet or tweet-path, and (iii) *Linguistic Quality Score* that captures the readability of a tweet-path using a trigram confidence score.

Weight of content words $(Score(j))$**:** tf-idf scores of content words are computed in the first step (extractive phase) of summarization as proposed in Chapter 3. These weights are used as a proxy to determine the importance of content words. **Informativeness** $(I(i)))$**:** We use a centroid based ranking as a proxy of sentence importance as one of the system configurations in our experiments. Centroid-based ranking [104] implies selection of sentences that are more central to the topic of the document. Each sentence is represented as a tf-idf vector. The centroid is the mean of the tf-idf vectors of all the sentences. Cosine similarity values between the sentences and the centroid is computed and used as the informativeness component in the ILP formulation. Importance of a tweet-path is normalized in [0,1] scale. For original raw tweets, we use machine predicted confidence scores as their informativeness score.

Linguistic quality score $(LQ(i))$ **:** The linguistic quality score is computed using a language model. A language model assigns probabilities to the occurrences of words. We use a Trigram language model [51] to compute a score with the goal of assigning higher scores to more probable sequences of words. For raw tweets which are extracted in the first step, linguistic score is taken as 1.

$$LQ(s_i) = \frac{1}{(1 - ll(w_1, w_2, w_3, \cdots, w_q))} \tag{5.1}$$

where $ll(w_1, w_2, w_3, \cdots, w_q)$ is computed as:

$$ll(w_1, w_2, w_3, \cdots, w_q) = \frac{1}{L} log_2 \prod_{t=3}^{q} P(w_t | w_{t-1} w_{t-2}) \tag{5.2}$$

Assuming the sentence consists of the words $w_1, w_2, w_3, \cdots w_q$, the value of $LQ(i)$ is computed using the above two equations (Eqns. 5.1 and 5.2).

Table 5.2: Notations used in the summarization technique.

Notation	**Meaning**
L	Desired summary length (number of words)
n	Number of *tweets and tweet-paths* considered for summarization (in the time window specified by user)
m	Number of distinct content words included in the n *tweets and tweet-paths*
i	index for *tweets and tweet-paths*
j	index for content words
x_i	indicator variable for *tweets and tweet-paths* i (1 if *tweet or tweet-path* i should be included in summary, 0 otherwise)
y_j	indicator variable for content word j
$Length(i)$	number of words present in *tweet or tweet-path* i
$Score(j)$	tf-idf score of content word j
$I(i)$	Informativeness score of the *tweet or tweet-path* i
$LQ(i)$	Linguistic quality score of a *tweet or tweet-path* i
T_j	set of *tweets and tweet-paths* where content word j is present
C_i	set of content words present in *tweets or tweet-paths* i

The summarization of L words is achieved by optimizing the following ILP objective function, whereby the highest scoring *tweets and tweet-paths* are returned as output of summarization, The equations are as follows:

$$max(\sum_{i=1}^{n} I(i).LQ(i).x_i + \sum_{j=1}^{m} Score(j).y_j) \tag{5.3}$$

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \tag{5.4}$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m] \tag{5.5}$$

$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n] \tag{5.6}$$

where the symbols are as explained in Table 5.2. The objective function considers both the number of *tweets and tweet-paths* included in the summary (through the x_i variables) as well as the number of important content-words (through the y_j variables) included. The constraint in Eqn. 5.4 ensures that the total number of words contained in the *tweets and tweet-paths* that get included in the summary is at most the desired length L (user-specified) while the constraint in Eqn. 5.5 ensures that if the content word j is selected to be included in the summary, i.e., if $y_j = 1$, then at least one *tweet or tweet-path* in which this content word is present is selected. Similarly, the constraint in Eqn. 5.6 ensures that if a particular *tweet or tweet-path* is selected to be included in the summary, then the content words in that *tweet or tweet-path* are also selected.

We use the GUROBI Optimizer [46] to solve the ILP. After solving this ILP, the set of *tweets and tweet-paths* i such that $x_i = 1$, represents the summary at the current time.

5.4 Experimental setup and results

In this section, we compare the performance of our proposed framework with state-of-the-art disaster-specific summarization techniques. We first describe the baseline techniques as well as the experimental settings.

5.4.1 Experimental settings

Given the machine-classified messages from our datasets: NEQuake, Hagupit, and PFlood, we split the tweets by date: 25th April to 27th April, 2015 for NEQuake, 6th December to 8th December, 2014 for Hagupit, and 7th September to 8th September, 2014 for the PFlood. In this work, we are using the same gold standard data as reported in Chapter 4.

Baseline approaches: We use three state-of-the-art summarization approaches as our baseline that are described below:

1. **COWTS:** is an extractive summarization approach specifically designed for

Table 5.3: Comparison of ROUGE-1 F-scores (with classification, Twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard ROUGE stemming(-m) and stopwords(-s) option) for COWABS (the proposed methodology) and the three baseline methods (COWTS, TSum4act, and APSAL) on the same situational tweet stream for each class, for each day, and for each dataset.

Step size	**ROUGE-1 F-score (NEQuake)**															
	Infrastructure				Missing				Shelter				Volunteer			
	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act
25/04/2015	**0.4947**	0.4842	0.3691	0.3758	**0.5407**	0.5353	0.3162	0.1901	**0.5165**	**0.5165**	0.4513	0.4742	0.4127	0.4127	**0.4405**	0.3174
26/04/2015	**0.3642**	0.3496	0.3071	0.2387	0.3066	0.3066	0.3496	**0.3694**	**0.3674**	**0.3674**	0.3275	0.3610	**0.5644**	0.5524	0.4982	0.3426
27/04/2015	0.3631	0.3631	0.3657	**0.3765**	**0.3494**	**0.3494**	0.3478	0.2825	**0.4340**	**0.4340**	0.3238	0.3631	**0.7069**	**0.7069**	0.6941	0.6934

Step size	**ROUGE-1 F-score (Hagupit)**															
	Infrastructure				Caution				Displaced				Volunteer			
	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act
06/12/2014	**0.6190**	**0.6190**	0.4946	0.5655	**0.4498**	**0.4498**	0.2922	0.3566	**0.3955**	**0.3955**	0.2881	0.2558	**0.4966**	**0.4966**	0.4814	0.4444
07/12/2014	**0.6173**	**0.6173**	0.4339	0.4852	**0.3303**	**0.3303**	0.3202	0.3281	**0.3585**	**0.3585**	0.2500	0.2307	**0.4782**	**0.4782**	0.4294	0.2902
08/12/2014	**0.4876**	0.4857	0.3891	0.4413	**0.4183**	0.4169	0.3803	0.4125	**0.4277**	**0.4277**	0.3376	0.3812	0.3701	0.3701	**0.3823**	0.3816

Step size	**ROUGE-1 F-score (PFlood)**											
	Infrastructure				Missing				Volunteer			
	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act	COWABS	COWTS	APSAL	TSum4act
07/09/2014	**0.7399**	0.7232	0.6894	0.7191	**0.6039**	**0.6039**	0.5787	0.5769	**0.3743**	0.3378	0.2646	0.2092
08/09/2014	**0.7206**	**0.7206**	0.6781	0.6315	**0.4758**	**0.4758**	0.4705	0.4498	**0.3227**	0.2865	0.2105	0.2631

generating summaries from disaster-related tweets (Chapter 3).

2. **APSAL:** is an affinity clustering based summarization technique proposed by Kedzie *et al.* [63].

3. **TSum4act:** is an extractive summarization method proposed by Nguyen *et al.* [83]. It is specifically designed for generating summaries from disaster-related tweets.

Evaluations: We use the standard ROUGE [72] metric for evaluating the quality of summaries generated using the proposed as well as the baselines methods. In this case, due to the informal nature of tweets, we consider the recall and F-score of the ROUGE-1 variant only.

5.4.2 Performance comparison

Table 5.3 depicts the ROUGE-1 F-scores for the four algorithms for each class and day. We can see that COWABS performs better compared to other three baselines. COWABS performs better compared to TSum4act and APSAL in 90% and 87%

Table 5.4: Summary of length 50 words(excluding #,@,RT,URLs), generated from the situational tweets of the infrastructure class (26th April) by (i) COWABS (proposed methodology), (ii) COWTS.

Summary by COWABS	**Summary by COWTS**
RT @cnnbrk: Nepal quake photos show historic buildings reduced to rubble as survivor search continues http://t.co/idVakR2QOT. Reporter: Kathmandu Airport closed following 6.7 aftershock; no planes allowed to land - @NepalQuake https://t.co/Vvbs2V9XTX. #NepalEarthquake update: Flight operation starts from Tribhuvan International Airport, Kathmandu. Pakistan Army Rescue Team comprising doctors, engineers & rescue workers shortly after arrival at #Kathmandu Airport http://t.co/6Cf8bgeort	#PM chairs follow-up meeting to review situation following #earthquake in #Nepal @PMOlndia #nepalquake. @SushmaSwaraj @MEAcontrolroom Plz open help desk at kathmandu airport. @Suvasit thanks for airport update. #NepalQuake. Pakistan Army Rescue Team comprising doctors, engineers & rescue workers shortly after arrival at #Kathmandu Airport http://t.co/6Cf8bgeort. RT @cnnbrk: Nepal quake photos show historic buildings reduced to rubble as survivor search continues. http://t.co/idVakR2QOT http://t.co/Z.

cases. Performance of COWABS on the remaining cases (where TSum4act or APSAL perform better) is more or less comparable with these two baselines. Combining tweet-paths with raw tweets helps in increasing the coverage over COWTS by 1% to 2%. However, our objective is to generate summaries in real-time. TSum4act and APSAL do not fulfill this criteria (Table 4.11). On an average COWABS takes 20.49 seconds, 21.83 seconds, and 29.73 seconds for NEQuake, Hagupit, and PFlood events respectively which is at par with other real-time methods like COWTS.

To give a flavor of the kind of summaries produced by the proposed summarization approach, Table 5.4 shows summaries generated by COWABS and COWTS (both disaster-specific methodologies) from the same set of messages (i.e., tweets form infrastructure class posted on 26th April). The two summaries are quite distinct. We find that summary returned by COWABS is more informative and diverse in nature compared to COWTS. For instance, we can see the COWABS summary contains information about flights, airport updates, damages of buildings, and information sources.

Reason behind better performance: We try to dissect the three baseline algorithms and identify their limitations and thus understand the reason behind superior performance of COWABS. APSAL and TSum4act suffer from the same clustering problems as reported in Chapter 4. COWTS although extractive, performs the best among all the baselines according to the ROUGE-1 scores, perhaps due to its simplicity. However, COWTS suffers from the fundamental problem of extractive summarization, namely, redundancy. Same or similar information might exist in two different tweets, yet they can be the part of the summary. COWABS provides more or less similar information coverage like COWTS but achieves 1-2% improvement.

5.5 Conclusion

A large number of tweets are posted during disaster scenarios and a concise, categorical representation of those tweets is necessary. In this chapter, we develop a summarization method to generate summaries in real-time from the incoming stream of tweets. We specifically take the tweets generated during the three disaster events and generate comprehensive abstractive summaries for some important classes like infrastructure, missing, shelter etc. Results show that combining information from related tweets helps in better information coverage which satisfies the objective of this chapter. So far, we have dealt with classification and summarization of situational tweet streams. In the next chapter, we explore non-situational tweets and show their impact in disaster situation.

Chapter 6

Analyzing Non-situational Information during Disasters

In Chapter 3, 4, and 5, we explore several traits of situational tweet streams posted during disaster and provide different real-time summarization approaches which satisfy the needs of various kinds of end users. However, along with situational tweets, huge amount of non-situational tweets are also posted during a disaster event which contains the emotions / opinions of the masses. While looking through these tweets we realized that a large amount of communal tweets i.e., abusive posts targeting specific religious / racial groups are posted even during natural disasters - this work focuses on such category of tweets. Considering the potentially adverse effects of communal tweets during disasters, in this work, we develop a classifier to distinguish communal tweets from non-communal ones, which performs significantly better than existing approaches. We also characterize the communal tweets posted during five recent disaster events, and the users who posted such tweets. Interestingly, we find that a large proportion of communal tweets are posted by popular users (having tens of thousands of followers), most of whom are related to the media and politics. Further, users posting communal tweets form strong connected groups in the social network. As a result, the reach of communal tweets are much higher than non-communal tweets. We also propose an event-independent classifier to automatically identify anti-communal tweets and also indicate a way to counter communal tweets, by utilizing such anti-communal tweets posted by some users during disaster events.

6.1 Introduction

A disaster generally affects the morale of the masses making them vulnerable. Often, taking advantage of such situation, hatred and misinformation are propagated in the affected zone, which may result in serious deterioration of law and order situation. Social media acts as a fertile ground in spreading hatred and specially Twitter is increasingly used as a powerful tool [17]. In this work, we primarily analyze offensive content in details posted during disaster.

There have been lots of research in recent years for automatic identification of offensive content, trolls, and hate speeches [17, 26, 39, 69, 116]. However, hate speech or trolling can come under various categories where people target religion, gender, sex, ethnicity, nationality etc. Out of this, especially harmful and potentially dangerous are the *communal tweets*, which are directed towards certain religious or racial communities. In this chapter, we provide a detailed analysis of communal tweets like automatic identification of such tweets, analyzing users posting such tweets, and provide a way to counter these tweets.

Earlier it has been observed that such offensive tweets are often posted during man-made disasters like terrorist attacks. For instance, Burnap *et al.* [17] showed that UK masses targeted a certain religious community during Woolwich attack to which the attackers are affiliated. However, it is quite surprising that in certain geographical regions such as Indian subcontinent, communal tweets are also posted even during natural disasters like floods and earthquakes. Some examples of communal tweets are shown in Table 6.1. Such kind of communal tweets result in developing hatred and agnosticism among common masses which subsequently deteriorates communal harmony, law and order situation. In the midst of disaster, this kind of situation is really difficult for government to handle.

In this chapter, we try to identify communal tweets, characterize users initiating or

Table 6.1: Examples of communal tweets posted during disaster events.

F**k these *Missionaries* who are scavenging frm whatever's left after the #NepalEarthquake Hav some shame & humanity.
Dear #kashmirFloods take away all rapist *muhammad's piglets* out of kashmir with you, who forced out kashmiri *Hindus* from their motherland!!
Radical Muslims want to behead u, moderate Muslims want radical Muslims to behead you n liberals want to save thm. result. #GurdaspurAttack
RT @polly: #HillaryClinton's reply when asked if war on terror is a war on *"radical Islam"* #DemDebate
Jesus *F***ing Christ* ... Active shooter reported in San Bernardino, California @CNN

promoting such contents, and counter such communal tweets with anti-communal posts which ask users not to spread such communal venom. Our major contributions are listed below.

(i) We develop a simple *rule based classifier* using low-level lexical and content features to automatically separate out communal tweets from non-communal ones (Section 6.3). Keeping in mind the limitations of previous works [17, 31], we develop an *event-independent* communal tweet classifier which can be directly used to filter out communal tweets during future events. Experiments conducted over tweet streams related to several disaster events with diverse characteristics show that the proposed classification model outperforms vocabulary based approaches [17, 75].

(ii) After identifying communal tweets, we study the nature of communal tweets, and the users who post them (Section 6.4). Broadly, we have observed two categories of users — (a). **Initiators:** who initiate communal tweets, (b). **Propagators:** who retweet communal tweets posted by initiators or copy the content of some initiator and post their own tweet with minor changes. We observe that, alarmingly, a significant section of communal tweets is posted by some very popular users who belong to media houses or are in politics. Such communal tweets are retweeted more heavily compared to other kinds of tweets. These communal users are connected via a strong social bond among themselves.

(iii) Apart from communal tweets, in this work, we observe that the tweets posted during disaster events follow certain specific traits, which can be exploited to counter adverse effects of communal tweets. After first level classification, we obtain communal and non-communal tweets. Further analysis of this non-communal tweet set reveals that a small number of users post anti-communal tweets which try to dissuade people from posting communal content. However, it is observed that such anti-communal posts are less retweeted and receive less exposure compared to communal tweets. Hence, a convincing way to counter the communal venom during disaster is to promote such anti-communal content.

In the second step, we develop a classifier (Section 6.5) for automatically separating out anti-communal tweets from non-communal tweets (identified in first level). In this case also we rely on some low-level lexical features to make this classifier event-independent. This is the first study, to our knowledge, that looks at anti-communal tweets as a practical way of countering adverse effects of communal tweets.

6.2 Dataset

This section describes the datasets used for the study, and various types of tweets observed in the datasets.

We considered tweets posted during the following recent disaster events – (i) **NEQuake** - a destructive earthquake in Nepal [81], (ii) **KFlood** - floods in the state of Kashmir in India [60], (iii) **GShoot** - three gunmen dressed in army uniforms attacked the Dina Nagar police station in Gurudaspur district of Punjab, India [47], (iv) **PAttack** - coordinated terrorist attacks in Paris [92], (v) **CShoot** - a terrorist attack consisting of a mass shooting at the Inland Regional Center in San Bernardino, California [19].

Note that the first two events are natural disasters, and the last three events are man-made disasters. Additionally, we have considered events occurring in different geographical regions so that this study would not get influenced by any kind of

Table 6.2: Statistics of data collected.

Event	# Tweets	# Distinct users
NEQuake	5,05,077	3,26,536
KFlood	14,922	8,367
GShoot	53,807	29,293
PAttack	6,48,800	5,77,888
CShoot	2,93,483	1,64,276

demographics. Tweet-ids of these tweets are made publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`.

We applied keyword based matching to retrieve relevant tweets through the Twitter API [127] during each event. For example, to identify the tweets related to the NEQuake event, we search tweets with the keywords like '#NepalEarthquake', 'Nepal' and 'earthquake' etc. For each keyword, we collected *all* the tweets returned by the Twitter Search API. Further, we consider only English tweets based on the language identified by Twitter. For each event, we report the number of tweets collected and the number of distinct users who posted them in Table 6.2. We describe our communal tweet identification step in the next section.

6.3 Identifying communal tweets

This section focuses on extracting communal tweets from rest of the tweets, by developing a rule based classifier.

6.3.1 Establishing gold standard

To understand the pattern, specific traits of communal tweets and evaluate the proposed classifier, we require gold standard annotation for a set of tweets. For each of the events stated in the previous section, we randomly sampled 4,000 tweets (after

Table 6.3: Gold standard – number of tweets in different disaster events.

NEQuake	KFlood	GShoot	PAttack	CShoot
247	112	203	201	152

removing duplicates). These tweets were independently observed by three human volunteers, all of whom have a good knowledge of English. The volunteers were asked to identify whether a tweet is communal or not.

There was an unanimous agreement for 81% tweets, while we consider the majority decision for the rest. By this process, a total of 915 tweets were identified as communal. Table 6.3 shows the number of tweets in gold standard across five disaster events. From the rest of the tweets, we randomly sampled the same number of non-communal tweets to build gold standard dataset.

6.3.2 Features for classification

As stated earlier, we want our classifier to be event-independent, i.e., the classifier should be such that it can be directly used over tweets posted over later events. Hence, we take the approach of using a set of lexical and content features for the classification task, which is known to make the classifier's performance largely independent of specific events considered for training (as shown in Chapter 3).

We use the first three datasets, i.e., NEQuake, KFlood, GShoot as *training set*. In other words, the tweets from these three datasets are used to identify discriminating features and develop our classifier. The other two datasets, i.e., PAttack and CShoot, are used as *test set*, to check the performance of our proposed classifier over future disaster events. Next, we describe the features used for the classification.

(1) Presence of communal slang phrases: In order to develop the classifier, we need a lexicon of religious terms and antagonistic hate terms about religion and related nationality. For this, we consider the terms in a standard lexicon of religious terms `http://www.translationdirectory.com/glossaries/`. However, all these terms are *not* hate terms; rather, the lexicon contains many general religion related terms as well. Hence, we employed three human annotators (the same who judged the tweets) to mark the terms in the lexicon as hate-terms or normal religious term. We obtain an unanimous agreement for 84% of the terms, and for the rest, we follow majority verdict. Similarly, we collect all the hate terms related to religion and nationality from a repository of terms frequently used in hate speeches – www.hatebase.org.

(2) Presence of religious/racial negated or hate terms: We detect the presence of any strongly negative term or slang term in the vicinity of neutral religious terms like ‘Muslim’ or ‘Christian’. We use a subjectivity lexicon developed in [133] to identify strongly negative terms, and we obtain a standard list of slang terms from www.noswearing.com. Then, we check whether such terms appear within a left and right word window of size two each with respect to a religious term. Thus, presence of phrases like ‘bastard missionaries’, ‘islamic scoundrels’, ‘jesus f***tards’ are identified.

(3) Presence of communal hashtags: We observe that some specific hashtags are explicitly used across various events to curse certain religious communities, such as, ‘#SoulVultures’, ‘#evangelicalvultures’, ‘#WeAreThanklessMuslims’, ‘#TweetlikeSecularJamat’. Such hashtags are mostly present in communal tweets. We ourselves developed a lexicon of such communal hashtags. These lexicons can be downloaded and used for research purposes[1]. Note that these hashtags were identified by the annotators only from the training set, i.e., the NEQuake, KFlood, and GShoot datasets.

(4) Presence of religious terms with wh-words/ intensifiers: Sometimes wh-words / intensifiers with neutral religious terms like ‘Muslim’ or ‘Christian’ are used to target certain religious communities sarcastically specially in disaster scenario (e.g., “Why do all the Muslim guys barking endian endian?? If u dnt knw hw to write english jst dnt write.. #GurdaspurAttack”). Sometimes we also observe that a tweet which appears to be a normal tweet in the general scenario can actually become communal in the context of a disaster (e.g., ‘Why do Christians pray’, which is a sarcastic comment on the religious habits of a religious group). We use a list of intensifiers (so, too, really, $\cdots$) collected from Wikipedia[2].

[1] http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html

[2] https://en.wikipedia.org/wiki/Intensifier

6.3.3 Evaluating classification performance

We compare the performance of our proposed set of features under two scenarios — (i) *in-domain classification*, where the classifier is trained and tested with the tweets related to the *same event* using a 10-fold cross validation, and (ii) *cross-domain classification*, where the classifier is trained with tweets of one event, and tested on another event. In this case, all the annotated tweets of a particular event are used to train / develop the model and then it is tested over all the tweets of rest of the events.

Selection of classification model: Performance of a classifier is heavily dependent on the appropriate model selection. We now attempt to select the most appropriate model for our proposed set of features based on some specific criteria. We consider seven state-of-the-art classifier models for the above set of features — (1) SVM with default RBF kernel and $\gamma = 0.5$ (SVMG), (2) SVM with RBF kernel (SVM), (3) Random Forest (RF), (4) SVM with linear kernel (LSVC), (5) Logistic regression (LR), (6) Naive Bayes (NB), and (7) Rule based classifier (RL) – here we follow a simple approach – if any of the above mentioned features is present in a tweet, we mark that tweet as communal; otherwise non-communal.

For each of these models (except rule based) we use Scikit-learn [96] package. To judge the performance of these models on above mentioned feature sets, we set the following evaluation criteria. Each criterion is computed and averaged over the three training datasets.

(i). Average in-domain accuracy: Average accuracy of the classifier across the three events in the training set, in in-domain scenario. **(ii). Average cross-domain accuracy**: Average accuracy of the classifier in different cross-domain scenario among the three events in the training set. In this case, we have six different cross-domain settings. **(iii). Average precision for communal tweets**: Detection of communal tweets with high precision is a necessary requirement for the classifier. Hence we consider average precision across the three training datasets. **(iv). Average recall for communal tweets**: The classifier should ideally capture all the communal posts, i.e., have high recall. Hence we consider the recall averaged over the three training datasets. **(v). Average F-score for communal tweets**: F-score of the

Table 6.4: Score of different evaluating parameters for seven different classification models using proposed features.

Classifier	In-domain accuracy	Cross-domain accuracy	Precision	Recall	F-score
SVMG	0.9295	0.9308	0.9504	0.9106	0.9284
SVM	0.9267	0.9308	0.9502	0.9087	0.9274
RF	0.9304	0.9308	0.9404	0.9117	0.9258
LSVC	0.9295	0.9308	0.9504	0.9106	0.9284
LR	0.9254	0.8919	0.9513	0.8530	0.8941
NB	0.9267	0.9117	0.9509	0.8817	0.9112
RL	0.9308	0.9308	0.9494	0.9117	0.9291

classifier indicate the balance between coverage / recall and accuracy / precision.

We report the performance of different classification models on the proposed set of features in Table 6.4. From Table 6.4, it is clear that rule based classification model shows promising performance compared to other models. The superior performance of the simple rule based classifier is probably because number of features are less. These clearly reveals the benefit of working with event independent features. All the subsequent results are produced using the rule based model.

6.3.4 Comparison of proposed approach with baselines

We use the following state-of-the-art communal tweet detection approaches as our baselines:

BUR: religious and racial hate speech detection approach proposed by Burnap *et al.* [17] using n-grams(1-5), hateful terms (`http://www.rsdb.org/`) and Stanford typed dependencies like 'determiner', 'adjectival modifier'.

USR: Recently, Magdy *et al.* [75] has shown that past tweet history of users can be used to detect communal tweets. This method used pre-event interactions (mentions, replies), contents / tweets (unigrams, hashtags) posted by users to predict post-event stances of these users.

Table 6.5: Classification accuracies (AC), recall (R), and F-scores (F) for communal tweets, using baseline models (BUR, USR). Diagonal entries represent in-domain classification, while the non-diagonal entries represent cross-domain classification.

Train set	**Test set**																	
	NEQuake						KFlood						GShoot					
	BUR			USR			BUR			USR			BUR			USR		
	AC	R	F	AC	R	F	AC	R	F	AC	R	F	AC	R	F	AC	R	F
NEQuake	0.8659	0.84	0.8609	0.7013	0.8402	0.7321	0.6043	0.4086	0.5081	0.5654	0.7289	0.6265	0.55	0.3850	0.4638	0.55	0.6666	0.5970
KFlood	0.5260	0.5812	0.5631	0.5647	0.6352	0.5934	0.8488	0.7916	0.8362	0.5922	0.7654	0.6409	0.7075	0.4950	0.6285	0.5388	0.5888	0.5608
GShoot	0.5140	0.5307	0.4989	0.5235	0.5999	0.5573	0.6826	0.3826	0.5465	0.5654	0.6355	0.5938	0.7950	0.7750	0.7908	0.5222	0.6888	0.5796

Note that both the baseline methods are supervised, hence they require training. However, our proposed method is rule based (unsupervised) and can be used directly over future events. For training and testing of baseline methods, we have used the SVM classifier – specifically, the Scikit-learn package [96] with the linear kernel.

Performance of baseline classifiers: Table 6.5 shows the performance of the baseline classifiers when trained and tested on the NEQuake, KFlood and GShoot events.

(i) In-domain classification: Here, tweets from same event are used to train and test the baseline classifiers and accuracy is measured using 10-fold cross validation. The results are shown in the diagonal entries in Table 6.5. BUR method performs quite well in case of in-domain scenario and achieves around 83% accuracy averaging over all the three events. Given that USR method does not perform well, it is evident that users past history is not helpful in predicting future stances.

(ii) Cross-domain classification: In this case also, tweets of one event are used to train the baseline classifier and then it is tested over tweets of another event. Results are shown in the *non-diagonal* entries in Table 6.5, where the left-hand side event is used as the training event, and the event stated at the top represents the test event. In this case, the performance of the baseline models is often as low as that by random chance (accuracy 50%). Only in some cases where same community was targeted in both training and test event, BUR model achieves around 69% accuracy.

Performance of proposed classifiers: Table 6.6 shows the performance of the proposed rule based classifier on the same three events. Our proposed rule based

Table 6.6: Classification scores (precision, recall, F-score) for communal tweets and overall accuracies using rule based classifier with proposed features, for the events in the training set.

Event	Precision	Recall	F-score	Accuracy
NEQuake	0.9698	0.9000	0.9336	0.9360
KFlood	0.9173	0.9652	0.9406	0.9391
GShoot	0.9613	0.8700	0.9133	0.9175

Table 6.7: Classification scores (precision, recall, F-score) for communal tweets and overall accuracies using rule based classifier with proposed features, for future events.

Event	Precision	Recall	F-score	Accuracy
PAttack	0.9336	0.9849	0.9586	0.9575
CShoot	0.9006	0.9666	0.9324	0.9300

classifier achieves 94% precision and 91% recall on average (over three datasets) in communal tweet detection. It is clear that our proposed method performs significantly better compared to the baseline techniques. This improvement is 17% over method proposed by Burnap(BUR). Note that, since we define a set of rules which are independent of the vocabularies used in a particular event, no separate training is required for the proposed classifier.

6.3.5 Further analysis of proposed classifier

Application over future events: As stated in Section 6.5, our objective is to make the communal tweet classifier independent of the vocabularies used during a specific disaster. We use NEQuake, KFlood, and GShoot events to learn the patterns of communal tweets. In this part, we apply the classifier over other two events (PAttack and CShoot). Table 6.7 reports precision, recall, F-score and accuracy of the classifier for these two events. The proposed classifier achieves very high performance over these two future events as well. Hence, we see that people follow more or less similar patterns in targeting different religious communities during various disaster scenarios.

Table 6.8: Misclassified communal tweets posted during disasters.

"Allah ho Akbar" battle cry was raised by pigs killed in #GurdaspurAttack #presstitutes media will not show,because they r funded by Saudi's
Huh, its a Muslim behind California attack
Threat frm a kashmiri muslim not a terrorist. Every1 shd keep this as proof

Analyzing misclassified tweets: For our proposed method, we have also analyzed different types of errors i.e. how many times a communal tweet is marked as non-communal tweet or vice-versa. Table 6.6 reflects that we achieve precision of 0.94 over three training datasets which indicates around 6% non-communal tweets are marked as communal tweets. On the other hand, average recall score is 0.91. 9% of communal tweets are misclassified as non-communal tweets. Marking a communal tweet as non-communal is more serious problem compared to classifying a non-communal tweet as communal one.

Table 6.8 shows some examples of misclassified communal tweets. Almost in every case, tweets are posted in a sarcastic way i.e. particular communities are targeted in round-about fashion. In present work, we have tried to capture some part of sarcasm by checking the presence of wh-words, intensifiers along with religious terms. However, in future, we will try to capture more sarcastic patterns present in communal tweets considering event / vocabulary independent model [105].

Feature ablation: Finally, we attempt to judge the importance of individual features in the classification, through feature ablation experiments. One feature is dropped at a time, and the degradation of the classifier performance (as compared to performance using all the features) gives an idea of the importance of the dropped feature. Table 6.9 reports the accuracy, recall, and F-score of the communal tweet classifier for feature ablation experiments, averaged over all the datasets. Presence of communal slangs and religious / racial negated terms appear to be most determining factors. However, all the features help in increasing the accuracy of the communal tweet classifier.

The above results indicate that communal and non-communal tweets can be effectively classified based on low-level content-based features.

Table 6.9: Feature ablation experiments for the proposed classifier. NONE represents the case when all features were used.

Ablated Feature(s)	Accuracy	Recall	F-score
NONE	0.9360	0.9373	0.9357
Religious negated terms	0.8687	0.7744	0.8665
Communal slangs	0.7595	0.5518	0.6852
Communal hashtags	0.9112	0.8852	0.9048
Religious terms with wh-words/intensifiers	0.9101	0.8846	0.9061

6.4 Characterizing communal tweets and their users

In this section, we try to understand and characterize the communal tweets and the users who post them. We apply our proposed classifier described in the previous section, over the datasets; we refer to the tweets which were categorized as communal by our classifier as *communal tweets* (60K), and the users who posted them as *communal users* (48K). Specifically, we compare the set of communal tweets and communal users during a particular event with an equal number of randomly sampled non-communal tweets (as judged by our classifier) and the users who posted them (referred to as *non-communal users*) during the same event.

6.4.1 Characterizing communal tweets

Which communities are targeted? It is observed that during disaster scenario, people post communal tweets targeting specific religious communities. Examples of some communal tweets and communities targeted via those tweets is given in Table 6.10. We observe that these targeted communities do not remain same across different disasters. During man-made disasters, like terrorist attacks, common masses mainly target that community to which attackers are affiliated. Along with that, some other communities are also targeted. For example, during Paris attack, Islamic people were mostly targeted but Christians were also targeted side by side.

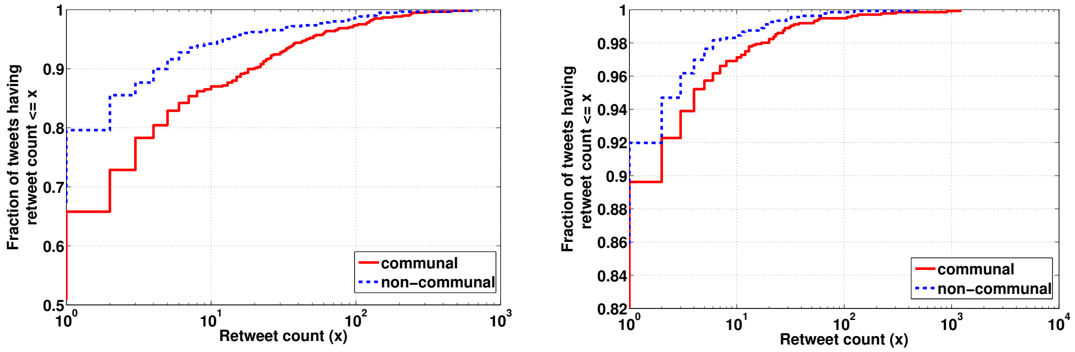
Table 6.10: Communities targeted during disaster events.

Event	**Communities Targeted**	**Sample communal tweets**
NEQuake	Christian	Meanwhile cheap *Christians* r busy spreading Christianity, Idiot morons #NepalEarthquake #earthquake
	Muslim	Y shd Allah waste his time killing *muslims* in #NepalEarthquake when de demselves r killing each others #SoulVultures [url]
KFlood	Muslim	I wish equal no of *muslims* perish & equal no are forced 2 leave their homes like KPs.. Then only they'll understand pain.. #kashmirFloods
GShoot	Muslim	#GurdaspurAttack is jihad on heart of Punjab ,brave sikhs will never forgive these *muslim* pigs for their coward act #Gurdaspur
PAttack	Persecuted Christian	RT @USER: If u think bringing any "persecuted Christians" into America from Syria and no terrorists will slip through, you're a fu
	Muslim	Slaughter. Like shooting fish in barrel. Now tell me what should be done with radical *Muslims* ?
CShoot	Muslim	https://t.co/xQ2Xo7WbPa via @USER- Take care of the radical *Islam* terrorists first Sir

It is interesting to note that users post communal tweets targeting specific religious communities like Christian missionaries, Muslims even during natural disasters like NEQuake or KFlood. During natural disasters, most of the people target core communities of the affected place which have been causing harm to the sentiments of other communities. For example, during Kashmir Floods, Muslims were targeted as some of the Muslim residents of Kashmir had maligned a temple of lord Shiva (Hindu mythological figure) before the disaster occurred. However, in some of the cases (e.g., Nepal Earthquake), people have specific reasons for targeting a community due to the behaviour and exertion of certain people of that community during post disaster scenario.

Popularity of communal tweets: In this part, we check whether communal tweets receive large attention from people. To measure the popularity of a tweet, we consider retweet-count of a tweet which is a standard metric to determine its exposure[3]. We show the distribution of retweet-counts for communal and non-communal

[3] All the tweets are re-crawled after several months from the date of the events, and hence such tweets contain more or less final retweet-count



(a) NEQuake

(b) CShoot

Figure 6.1: Cdf of retweet-count of communal and non-communal tweets. Communal tweets are retweeted more.

tweets, for the two events NEQuake, and CShoot in Figure 6.1. For this study, we discard retweets and only consider original tweets. From Figure 6.1, we can see that communal tweets become more popular compared to non-communal ones. We observe a similar pattern for other events.

Language of communal tweets: In Chapter 3, we observe that a sizeable number of tweets are posted in regional languages. Out of these regional languages, Hindi is used by majority of people. Earlier, we observe a significant amount of new information can be extracted from such Hindi tweets. In this chapter, we perform an initial study over NEQuake and GShoot events to understand the content of non-situational Hindi tweets. It is observed that users prefer Hindi over English tweets to post communal content and slang terms (1.75 and 1.25 times respectively). Side by side, we also observe that Hindi is mostly used to express negative sentiments in case of catastrophic as well as general events like sports, politics etc. [109].

6.4.2 Characterization of communal users

We next analyze the users who post communal tweets during the disaster events. For this, communal users are divided into following two categories –

1. **Initiators:** users who initiate communal tweets
2. **Propagators:** users who retweet the communal tweets posted by initiators or some other propagators or they copy the content of some initiator and post their own tweet with minor changes.

We next describe the construction procedure of initiator and propagator set and study the properties of initiators and propagators separately.

Construction of initiator and propagator set: For dividing users into initiators and propagators, we need to find the set of retweets Y of a particular tweet x. The users in set Y would then be classified as propagators while the users who posted x would be classified as initiator. As per the prototype, tweet x of user u is said to be propagated by tweet y of user v if y is formed by copying x, preceding it with RT and addressing u with @. However, due to the 140-character limitation on twitter and user's personal formatting preferences, a significant number of retweets do not follow this prototype [12]. Users like to add their own comments and sometimes even skip acknowledging the original users. As a consequence, some of the retweets lack distinguishable markers and patters which makes their identification difficult [5]. Thus, in order to get a near-accurate classification of users into initiators and propagators, there is a need to incorporate the inconsistent syntax a significant number of users follow while retweeting . We attempt to minimize error in this classification and try to find *true initiators and propagators.* We first compute normalized *Phrasal Overlap Measure* [101] between all pair of tweets in our corpus. This measure is based on the Zipfian relationship between the length of phrases and their frequencies in a text collection and is defined as follows:

$$phrasal\_overlap\_norm(t_1, t_2) = tanh\left(\frac{\sum_{i=1}^{n} m(i) * i^2}{|t_1| + |t_2|}\right) \tag{6.1}$$

where $m(i)$ is the number of i-gram phrases which match in tweets t_1 and t_2, n represents highest n-gram considered for computing phrasal overlap, $|t_1|$ is the length of tweet t_1. In Eqn. 6.1, higher n-grams get more weight which also helps in capturing the context rather than comparing unigrams. We then cluster together the tweets t_1 and t_2 having $phrasal\_overlap\_norm(t_1, t_2) \geq similarity\_threshold$ using Hierarchical Clustering Algorithm. We define the representative of each cluster as the tweet which was posted first on twitter among all the tweets of the cluster (i.e. tweet having smallest timestamp). Phrasal overlap between two clusters is defined as the overlap between the representative tweets of those clusters[4]. For a cluster of size k, one tweet is representative tweet and rest of the k-1 tweets are retweets of that tweet. The users corresponding to tweets become initiators and those corresponding to retweets become propagators. For our purposes, we take the value of n as 3 and similarity_threshold as 0.8[5].

Popularity of initiators and propagators: We next investigate popularity of users who posted communal tweets during disaster. Popularity of a user works as a major driving force in determining the popularity of tweets [23]. We observe a uniform phenomena across all the five disaster events — both common masses (27% having less than 100 followers) and popular users (10% having more than 10,000 followers) involve themselves in initiating and propagating communal content. Especially, some popular communal users belonging to media houses and politics have several tens or hundreds of thousands of followers. We provide examples of some such popular communal initiators and propagators in Table 6.11.

Do initiators also work as propagators? Next, we try to figure out whether during a disaster event communal tweet initiators also play the role of propagators during the same event. For this, during each event, the *Szymkiewicz-Simpson similarity* score [117] between initiator set and propagator set is computed. Table 6.12 shows the overlap score obtained across five disaster events for both communal and non-communal tweets. For communal tweets, we obtain a low similarity score of 0.15 averaging over all the events. Thus, communal tweet initiators hardly involve themselves in retweeting others contents; rather they are interested in posting their own views. Interestingly, this overlap score for natural disasters

[4] Please note that we remove “RT @user” from the tweet, if present, before finding the overlap similarity

[5] We have tried different values but this setting provides best result

Table 6.11: Sample popular users posted communal tweets.

Role	Screen_name	Follower count	Bio of the user
Initiator	abhijitmajumder	69.5k	Journalist. Managing editor, Mail Today. Views are personal, retweets are not necessarily endorsements
	HinduRajyam	11.9k	Om Namo Venkateshaya Namaha.Establishing Hindu Rashtra shud be the immediate goal of every hindu! Follow @noconversion
Propagator	SanghParivarOrg	133k	http://t.co/cF4rB7S56v is an independent initiative by Swayamsevaks. @RSSOrg is official Twitter Handle for RSS
	mediacrooks	98k	changing the way we consume news.... rts do not imply endorsements..

Table 6.12: Overlap score between initiators and propagators for communal and non-communal tweets across different events.

Tweet type	NEQuake	KFlood	GShoot	PAttack	CShoot
Communal	0.21	0.20	0.14	0.11	0.11
Non-communal	0.32	0.32	0.29	0.23	0.30

(NEQuake, KFlood) is twice the score of man-made disasters (GShoot, PAttack, CShoot). Generally, in case of man-made disasters, common masses get angry and they raise their voice. Hence, initiators hardly involve themselves in propagating such tweets. In case of natural disasters, communal sentiment among the common masses is not instinctive. Thus, initiators also play the role of propagators in order to activate communal belief among the people.

However, the overlap between communal initiators and propagators is less than that of non-communal initiators and propagators. For this overlap, we do not observe any significant difference between natural and man-made disasters.

User overlap across different events: We investigate whether a common set of users involved themselves in initiating/ propagating communal tweets during multiple events. For this, we consider events, which occurred in the same geographical region (e.g., NEQuake, KFlood, GShoot, all of which occurred in the Indian subcontinent). We find a small set of common users who posted tweets across all the three events. For instance, communal tweets are posted during all these three events by initiators like 'simbamara', 'RamraoKP_' and propagators like 'IndiaAnalyst', 'HinduRajyam'. In general, overlap among the communal users of three events is low (about 5%). This overlap score is three times higher (about 15%) in case of non-communal tweets. We define such common set of communal users as **core communal users**.

It is observed that only 22% of such core communal users are initiators and rest of the users help in propagating communal content during disaster. We also analyze the influence of such core users. Around 10% followers of these core users are popular(having more than 10,000 followers) and such users can help in getting wide exposure of communal content posted by core-users. Again 5% of these core communal users are popular, i.e., these users have more than 10,000 followers. If such users post communal content then they have high probability of getting large number of retweets and exposure.

Topical interests of communal users: In this section, we try to infer topical interests of communal users. Specifically, we attempt to match the interests of communal users to one of seven broad topics: (i) Media & Journalism (News), (ii) Politics, (iii) Movies & Entertainment, (iv) Writers / Authors, (v) Sports, (vi) Religion, and (vii) Business. We collected specific keywords from online sources[6] which help in characterizing above mentioned broad topics. Users whose topics of interest do not fit into any of the above mentioned categories are marked as others.

To perform this analysis, users are divided into two categories — (i) common users, having $< 5{,}000$ followers, and (ii) popular users, having $\geq 10{,}000$ followers. Twitter account bio is used to infer the topical interest of communal users. We check whether the keywords corresponding to any of the broad topics stated above are present in their bio. For popular users we not only rely on their bio but also use our prior method [114] which can infer topical interest of popular users. Finally, we match the topical characteristics with the keywords corresponding to any of the broad topics.

[6] `goo.gl/p4CPyX,goo.gl/Iqxo9T`

Table 6.13: Distribution of topics of interest of common and popular initiators of communal tweets.

User	Broad topic of interest							
	Media	**Politics**	**Sports**	**Religion**	**Writing**	**Entertainment**	**Business**	**Others**
Popular users	50%	33%	2%	5%	2%	4%	1%	3%
Common users	21%	25%	12%	19%	6%	8%	2%	7%

Table 6.14: Comparing the profile bio and tweets posted by users who posted communal tweets, and other users.

Most frequent words in bio	
communal	religion, india, hindu, life, endorsement
non-communal	fan, indian, music, lover, life

Most frequent words in tweets	
communal	hindu, religion, congress, media, muslim
non-communal	govt, india, life, people, movie

We show the distribution of topical interests of popular and common initiators in Table 6.13. We notice a similar phenomena for propagators. Most of the popular initiators belong to news media and politics. Interest of common masses is distributed across multiple topics like news, sports, politics, religion etc.

For active users, their profile and past history can also be useful in characterizing them. Thus, for further analysis, we process[7] the posted tweets and account bio of communal and non-communal users to infer their interest and behaviour. For each category of users, we show top five words which appear in their account bio and posted tweets in Table 6.14. As expected, we find presence of religion and politics related words in the bio and tweet of communal users. However, we do not find any topic specific alignment with the most occurring words in the bio and tweet of non-communal users. Such words are either normal chat words or they represent positive sentiment.

Are common communal users provoking popular users? Mentioning popular users to improve visibility of tweets is a common phenomenon on Twitter. Traditional communication theory states that a minority of users, called the *influentials*, excel in

[7] Case-folding, stopwords removal etc.

Table 6.15: % of times common users mention popular users in communal and non-communal tweets.

Event	Communal Tweets	Non-communal Tweets
Natural	63.74%	**69.06%**
Man-made	**74.07%**	68.89%

persuading others [108]. Thus, mentioning these influentials in the network helps in achieving a large-scale chain-reaction of influence driven by word-of-mouth [24, 61]. Popular users, i.e., users having a large number of followers on Twitter, are influential, and a retweet by popular users can help improve the visibility of a tweet [135]. Thus, common users, i.e., users with small number of followers on Twitter, often mention popular users in their tweets to increase the reachability and effectiveness of tweets. Table 6.15 shows the percentage of times a common user mentioned a popular user out of the total mentioning instances in communal and non-communal tweets respectively in case of natural (NEQuake, KFlood), and man-made (GShoot, PAttack , CShoot) disasters.

We find that the percentage of cases in which a common user (< 5000 followers) mentions a popular user (≥ 10000 followers) is larger for communal tweets than non-communal tweets in the case of man-made disasters and smaller for natural disasters. While computing these results, we have used the number of followers of a user as his/her measure of influence and popularity. It is clear that in case of man-made disasters when people are already angry towards some community, people try to provoke popular users by mentioning them in their communal posts. On the other hand, such trend is less in case of natural disasters.

6.4.3 Interactions among the users

In this part, we check the interaction pattern of non-communal and communal users among themselves. In Twitter, user u can interact with user v mostly in following two ways — (i) v can be mentioned (@mention) by user u in her tweet (ii) u can subscribe to the content posted by v by following v.

Table 6.16: Reciprocity and density of the mention and follow networks between two different groups of users.

Event	User group	Mention Network		Follow Network	
		Reciprocity	Density	Reciprocity	Density
NEQuake	communal	4.20%	0.0037	25.05%	0.0099
	non communal	3.31%	0.0002	16.88%	0.0007
GShoot	communal	4.74%	0.0047	26.14%	0.0133
	non communal	3.78%	0.0012	16.92%	0.0038

Two types of interaction networks are constructed among users — (i) **Mention network:** if user u has mentioned v we add a link $u \rightarrow v$, (ii) **Follow network:** if user u follows the content posted by v, we add a direct link $u \rightarrow v$. To quantify the level of interaction among the users, two structural properties of above mentioned networks are measured — (i) density, fraction of number of links present in a network and all possible links that can be present in a network (ii) reciprocity, what fraction of directed links are reciprocated, i.e. $v \rightarrow u$ and $u \rightarrow v$ both present in the network. Mutual friends generally have a high probability to share reciprocal links.

We report reciprocity and density values for mention and follow networks between two different groups of users in Table 6.16. A similar trend is observed across all the disaster events. Here we report the result for two disaster events — NEQuake and GShoot. From Table 6.16, we can see that communal users form a more dense network among themselves compared to non-communal users. Apart from density, we also observe that reciprocity of both the networks is higher for communal users. It indicates that a large fraction of communal users are mutual friends. Thus, there is a significant interaction among communal users and strongly-tied communities formed by them in social network.

6.4.4 Are the users getting outraged suddenly?

Previous studies argue that a significant rise is observed in communal hate online following 'trigger' events like disaster [4, 17, 138]. According to them, these trigger events work as activators to wake up the old feelings of hatred and negative sentiments towards suspected perpetrators and related groups. In this section, we check if such a sudden rise exist in the case of disasters and attempt to quantify it. We are also interested in finding out whether there exist users who have a general tendency to post communal tweets irrespective of the event and situation. In order to perform this analysis, we study the nature of tweets posted by the communal users for a particular time period surrounding the disaster which encompasses general as well as event-specific behavior of the communal users. Let a user u in our dataset first posted a communal tweet on day d. We define $TimeWindow(u, d)$, corresponding to a communal user u as a list of 31 days, comprising of 15 days before d and 15 days after d. For each communal user u, we scrapped all the tweets posted by her on $\forall\ d \in TimeWindow(u, d)$. We use Twitter Advanced Search[8] utility which can retrieve tweets posted by a user, given her screen name and a particular $TimeWindow(u, d)$. Our communal tweet detection algorithm is applied on these tweets which marked the retrieved tweets as communal and non-communal. Based on the classification, we define a vector v for each user u as following:

$$v[i] = \begin{cases} 1 & \text{if user u posted a communal tweet on d+i} \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

where $i \in [-15, 15]$.

Next, for each user u, we find her **regularity score**, $r_u = \sum_{i=-15}^{15} v_u[i]$, where r_u defines the number of days user u posted a communal tweet in her $TimeWindow(u, d)$. Figure 6.2 shows the cdf of regularity score for NEQuake, GShoot, and CShoot. From Figure 6.2, we observe two interesting phenomena,

1. Most of the users (80-90%) have regularity score < 5.
2. There are a small fraction of users (10-20%) having large values of regularity score ($\geq$5).

[8] `https://twitter.com/search-advanced`

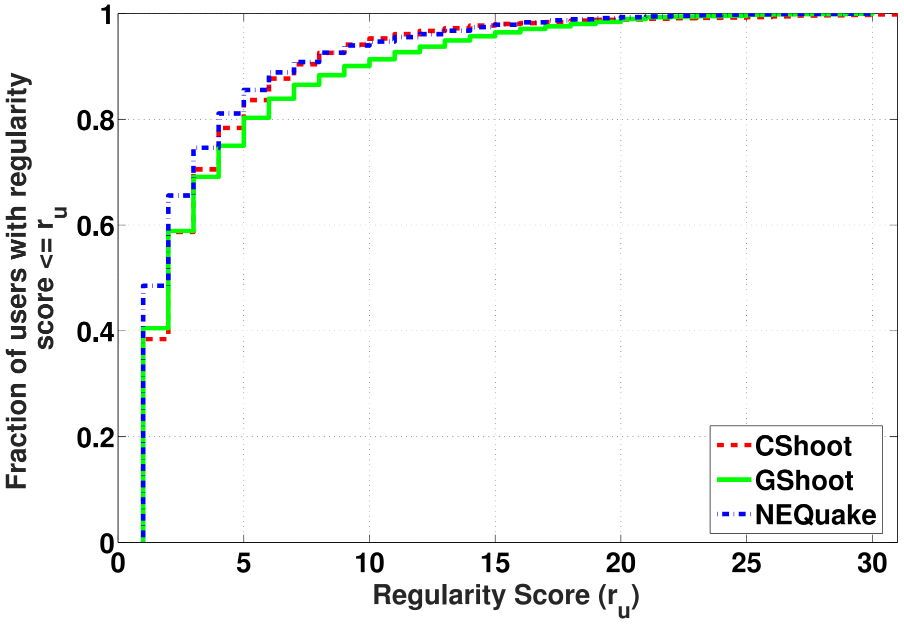


Figure 6.2: Cdf of regularity score of communal users.

Thus, a large fraction of users only get outraged at the time of disaster and do not express their hatred towards people of a particular religion or race otherwise. However, there are a few users who repeatedly post communal tweets irrespective of any trigger event. We define them as **regular communal users**. This phenomena also agrees with what prior works found [138].

Overlap between core communal users and regular communal users: We next find the overlap between core communal users (Section 6.4.2) and regular communal users using *Szymkiewicz-Simpson Similarity* [117]. For Regular communal users with $r_u >= 5$, we find an overlap score of 0.44 and for $r_u >= 10$, we obtain an overlap score of 0.22. These regular communal users play the role of core communal users in posting communal tweets across multiple events.

Table 6.17: Examples of anti-communal tweets posted during disasters.

Event	Tweet text
NEQuake	Sad commentary of our times that people bring religion even into the devastating
GShoot	A terrorist has no religion. No need 2 specifically mention d religion of a terrorist anywhere
PAttack	Tears & blood know no religion. All they know is pain. It's not just in Paris. It's everywhere. We are the killers & we are the victims.
CShoot	#California So sorry to hear abt the shooting & Killings of innocent people. There is no religion which allows that.

6.5 Countering communal tweets during disaster scenario

During a disaster event, when the masses are anxious, communal tweets may propagate venom among different religious communities and thus complicate the relief operations. Since online social media like Twitter work as important sentinels during disasters, shutting down online media during disasters is not a reasonable solution. On the other hand, if communal content is allowed to circulate freely and get large exposure, anti-government agencies can use such communal content for propaganda, causing certain religious communities to panic[9]. Hence, communal tweets posted during disasters need to be countered, so as to minimize their potential adverse effects. In this section, we discuss a potential way of countering such communal tweets.

Utilizing anti-communal tweets: During disasters, most of the people post communal tweets. However, it is observed that some users also post *anti-communal* content, asking people not to spread communal venom among society. Some examples of anti-communal tweets posted during different disaster events considered in this work are shown in Table 6.17. We also find that just as some communal hashtags

[9] For instance, after the mass shooting incident in California in November 2015, the American Muslims had to live in fear of demonization of Islam, according to the report by Reuters – https://t.co/GzMonqK9Js

Table 6.18: Examples of communal and anti-communal hashtags, which are used to attack or support certain religious communities during disasters.

Event	Anti-communal hashtags	Communal hashtags
NEQuake	#RespectAllReligion, #Intolerance, #stopit	#SoulVultures, #EvangelicalVultures, #EvanJihadis
PAttack	#MuslimsAreNotTerrorist, #ThisisNotIslam, #NothingToDoWithIslam	#KillAllMuslims, #IslamAttacksParis, #RadicalIslam

are introduced to target certain religious communities, certain other hashtags are introduced to *support* those religious communities. Table 6.18 shows some examples of hashtags of both types.

Thus, a potential way of countering communal content would be to utilize such anti-communal content. For this, first question arises about automatic identification of such anti-communal tweets.

6.5.1 Identifying anti-communal tweets

In Section 6.3, we have proposed a rule based classifier to detect communal tweets from large set of tweets. After separating out communal tweets, we try to capture anti-communal tweets from rest of the tweets.

Establishing gold standard: To understand the pattern of anti-communal tweets and define the rules for its detection, we require gold standard annotation for a set of tweets. For each event, first, we use the communal tweet classifier (proposed in Section 6.3) to identify communal tweets. Then, from rest of the tweets, we randomly sampled 2,000 tweets (after removing duplicates). These tweets were independently observed by three human volunteers, all of whom are regular users of Twitter, have a good knowledge of English. The volunteers were asked to identify whether a tweet is anti-communal or not.

There was an unanimous agreement for 78% tweets, while we consider the majority decision for the rest. By this process, a total of 196 tweets were identified as anti-communal. We can observe that much less number of anti-communal tweets are posted during such events. In fact, we were able to identify anti-communal tweets only for three events – NEQuake, GShoot, PAttack. For the other two events, no example of anti-communal tweet was found. Some examples of anti-communal tweets are shown in Table 6.17. From the rest of the tweets, we randomly sampled the same number of non-anti-communal tweets to build our training dataset.

Features for classification: As mentioned earlier, our main objective is to make our classifier independent of any specific event, i.e., the classifier should be such that it can be directly used over tweets posted over later events without further training. Following communal tweet classifier approach, in this section also, we rely on using a set of lexical and content features for the classification task. We describe the features next.

(1) Presence of anti-communal hashtags: While observing the three datasets, the annotators found that some specific hashtags are explicitly used across various events to post anti-communal tweets and ask users not to post communal contents, such as, "#RespectAllReligion", "#MuslimsAreNotTerrorist", "#ThisisNotIslam", "#NothingToDoWithIslam", "#stopit".

(2) Presence of collocations: Some collocations are frequently used in anti-communal tweets across all the three datasets, such as 'nature doesn't discriminate', 'has no religion', 'terrorism defies religion', etc.

(3) Mentioning multiple religious terms: The aim of anti-communal tweets is to ask people to treat all religions equally. Thus, either they do not mention religious terms explicitly or they mention multiple religions so as to create a sense of unity like 'WTF people are trying to save their life & this MORONs Tweeting *Hindu christian muslim* #earthquake #NepalEarthquake'.

We make above mentioned lexicons publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`. In future, we will try to enrich this lexicon set based on co-occurrence with current

Table 6.19: Classification accuracies (AC), recall (R), and F-scores (F) for anti-communal tweets, using bag-of-words model (BOW)

Train set	Test set								
	NEQuake			GShoot			PAttack		
	AC	R	F	AC	R	F	AC	R	F
NEQuake	0.8083	0.8166	0.7990	0.6875	0.8958	0.7413	0.7180	0.7067	0.7148
GShoot	0.56	0.76	0.6333	0.6875	0.9199	0.7382	0.5751	0.8947	0.6780
PAttack	0.6999	0.80	0.7272	0.6041	0.9166	0.6984	0.7596	0.8499	0.7828

Table 6.20: Classification scores (precision, recall, F-score) for anti-communal tweets and overall accuracies using rule based classifier with proposed features.

Event	Precision	Recall	F-score	Accuracy
NEQuake	0.8461	0.88	0.8627	0.86
GShoot	0.6351	0.9791	0.7704	0.7083
PAttack	0.8012	1	0.8896	0.8759

lexicons. We follow simple rule based classification approach to classify the tweets into two classes based on the features described above. If any of the above mentioned features is present in a tweet, we mark that tweet as anti-communal; otherwise non-anti-communal.

Evaluating classification performance: We compare our proposed features with the bag-of-words (BOW) model where we take unigrams as classification features and Naive-Bayes as classifier. Prior researches [130] showed that Naive Bayes model performs better compared to others when unigrams and bigrams are chosen as features. BOW is a supervised model; hence required training. Our proposed method is rule based and can be applied directly to any future event. Table 6.19 shows the performance of the classifier using the BOW model and Table 6.20 shows precision, recall, F-scores of anti-communal tweets and overall accuracies of our proposed rule based classifier. We compare the performance of two feature-sets with different classification models (rule based and Naive Bayes based). BOW model achieves 75% in-domain accuracy (training and testing events are same) but does not perform well in cross-domain setting (training and testing events are different). Our proposed

Table 6.21: Misclassified anti-communal tweets posted during disasters.

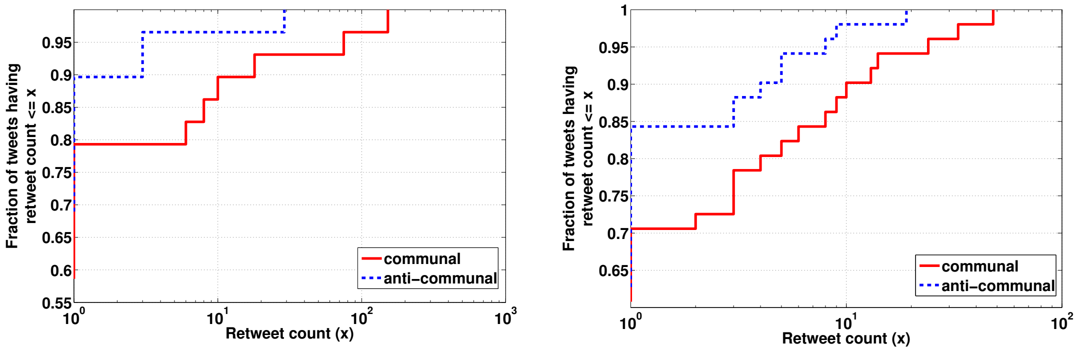
Could someone on the ground please ask about #gods involvement concerning the #NepalEarthquake ? Just curious.
earthquakes happen because of tectonic plates, they are not a result of lack of jesus. Christians and science, smh
Islam has nothing to do with #GurdaspurAttack. Stop spreading hatred among society

method performs better compared to vocabulary dependent model.

Analyzing misclassified tweets: For our proposed method, we have also analyzed different types of errors i.e. how many times an anti-communal tweet is marked as non-anti-communal tweet or vice-versa. We achieve precision of 0.76 over three datasets which indicates around 24% non-anti-communal tweets are marked as anti-communal tweets. On the other hand, Table 6.20 reflects that average recall score is 0.95. 5% of anti-communal tweets are misclassified as non-anti-communal tweets. It is observed that during disaster anti-communal tweets are posted in very low volume compared to other tweets. Hence, objective of the classifier is high recall so that we can utilize such tweets in maintaining communal harmony during emergency. Table 6.21 shows some example of misclassified anti-communal tweets. In most of the cases explicit signal for anti-communal tweets are missing. In future, we will try to capture such implicit senses and also try to enhance our feature sets.

6.5.2 Characterizing anti-communal tweets and its users

In this section, we study the anti-communal tweets and the users who post them. We apply the classifier described in the previous section, over the datasets; tweets which are identified as anti-communal by our classifier are referred as anti-communal tweets and the users who posted them as anti-communal users. Specifically, we compare the set of anti-communal tweets and anti-communal users during a particular event with an equal number of randomly sampled communal tweets (as judged by our classifier) and the users who posted them (referred to as communal users) during the same event.



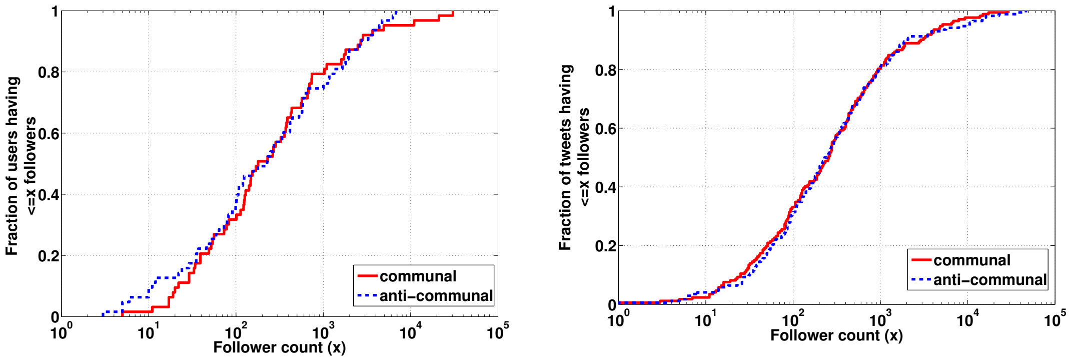
(a) NEQuake (b) GShoot

Figure 6.3: Comparing the popularity of communal and anti-communal tweets – communal tweets are much more retweeted than anti-communal tweets.

Do anti-communal tweets get similar exposure as communal tweets? As earlier, we measure the exposure or popularity of a tweet by its retweet-count. Figure 6.3 shows the distributions of retweet-count of communal and anti-communal tweets posted during two of the disaster events. We observe that anti-communal tweets are much *less retweeted* compared to communal tweets. We obtain a similar observation across all the events.

We next investigate why anti-communal tweets get less popularity compared to communal tweets. Our first intuition was that the users who post communal tweets might be more popular than the ones who post anti-communal tweets. To verify this, we compare the distributions of follower counts of users who post communal tweets and users who post anti-communal tweets during the same event. Figure 6.4 shows the comparison for two events (similar results were obtained for all other events). It is clear that both sets of users have very similar follower counts. Thus, variation in user-popularity cannot explain why anti-communal tweets get lower exposure than communal tweets.

We find that the number of distinct users who post anti-communal tweets is much lesser than the number of users who post communal tweets. As a result, other users receive much lesser exposure to such tweets. We believe that an effective way of countering communal content would be to automatically identify anti-communal



(a) NEQuake (b) GShoot

Figure 6.4: Comparing the popularity of users who post communal tweets and those who post anti-communal tweets – both types of users have similar follower-count distributions.

tweets, and to promote such tweets by getting more and more users (preferably popular users) to retweet them. Additionally proper wording of tweets are also necessary to make them popular. In future, we will try to promote and increase the popularity of such anti-communal tweets.

6.6 Conclusion

In this work we try to characterize communal tweets posted during disaster scenario and analyze the users involved in posting such tweets. We propose an event-independent classifier which can be used to filter out communal tweets early. We also find that communal tweets are retweeted heavily and posted by many popular users mostly belonging to news media and politics domain. Users involved in initiating and promoting communal contents form a strong social bond among themselves. Additionally, most of the users get angry suddenly due to such kind of event and express their hates towards specific religious communities involved in the event. We observe that, during a disaster, some users also post anti-communal content asking people to stop spreading communal posts and it is necessary to counter the potential adverse effects of communal tweets. We have proposed an event-independent classifier to identify such anti-communal tweets. However, we have found that such anti-communal tweets are retweeted much less compared to the communal tweets and they are also very few in number compared to the communal tweets. Our proposed communal tweet classifier can be used as an early warning signal to identify communal tweets, and then celebrities, political personalities can be made aware of the situation and requested to post anti-communal tweets so that such tweets get higher exposure.

Chapter 7

Conclusion and Future Work

In present times, social media have become an important source of real-time information during disaster. With social network getting accessible on mobile phones, the rate at which tweets are posted soon after disaster has increased exponentially. However, this information not only carries situational updates but also personal opinions and sentiments of common masses. Hence, segregation of large volume of situational information and summarization of those information is necessary to produce real-time updates. Extraction of information from different languages is also helpful to cover more diverse and a new set of situational updates. In general, disaster creates a panic among common masses and taking advantage of such a situation, some people try to disrupt communal harmony by posting communal tweets targeting specific religious communities. In this thesis, we have developed methods to solve the above stated problems by analyzing the needs of different stakeholders (NGOs, government, common people, rescue team etc.) in disaster scenario. Here, we summarize and reiterate the contributions of the thesis, show the utility of proposed techniques and develop a future road map to further carry forward this research work.

7.1 Summary of contributions

The major contributions of this thesis are as follows:

1. Developing a vocabulary independent situational tweet classifier and integer linear programming (ILP) based extractive summarization technique which maximizes the coverage of content words (Chapter 3).
2. Developing a dependency parser based method which extracts direct objects of some specific verbs (‘kill’, ‘die’ etc.) to handle fast changing numerical information (Chapter 3).
3. Developing classification-summarization framework for situational updates posted in regional languages like Hindi (Chapter 3).
4. Developing a noun-verb pair based sub-event detection approach and ILP-based extractive summarization technique which maximizes the coverage of content words and sub-events to produce summaries from different perspectives (overall high level, humanitarian class specific summary, etc.) (Chapter 4).
5. Developing an ILP-based abstractive summarization technique which first generates paths from tweets using a word graph and covers important paths and content words to produce the final summary (Chapter 5).
6. Developing a communal tweet classifier, characterizing their users, identifying the process in which communal tweets become popular, and proposing a way to counter their adverse effects (Chapter 6).

7.1.1 Classifying and summarizing situational information from Twitter

Microblogging sites like Twitter provide important situational updates but this information is hidden in the deluge of non-situational tweets which mostly contain personal opinions and sentiments of masses. Sometimes, a single tweet contains both situational and non-situational information and fragmenting them based on sentence end-markers ('.','!','?') helps to separate situational and non-situational parts. Considering huge volume of information posted during disaster, it is necessary to develop some automated method to separate situational tweets. We develop a situational tweet classifier which depends on low-level lexical features like presence of exclamation, question marks, personal pronouns, strong subjective words etc. These low level features are helpful in capturing generic patterns users follow in posting non-situational tweets. Hence, our proposed classifier is independent of the vocabularies used during a particular disaster and can be directly deployed over any future disaster event. Overall, we achieve an average F-score of 0.80 and fragmentation helps to improve the classification accuracy.

After classifying situational tweets, we develop an integer linear programming (ILP) based extractive summarization technique which exploits specific traits of tweets posted during disaster. We observe that some particular words, i.e., nouns, verbs, numerals (*content words*) capture most of the situational information and they grow very slowly in case of disaster compared to any other real life event like sports, music, politics etc. Hence, capturing these content words can provide a good information coverage. We propose an ILP-based technique which maximizes the coverage of content words to produce final summary. Our proposed summarization technique produces better summaries compared to other state of the art real-time summarization approaches.

We observe that some situational updates like information about missing, killed, died or stranded people are changing rapidly. Hence, we develop a dependency relation based method where we utilize the direct objects of disaster-specific verbs (e.g., 'kill' or 'injure') to continuously update important, time-varying actionable items such as the number of casualties. Our proposed verb to numeral association method achieves precision of 0.95.

Finally, we notice that in India, many situational updates are available only in local languages (Hindi). Hence, we extend our classification-summarization framework to Hindi tweets. Our Hindi situational tweet classifier obtains an accuracy of 0.78 and summarization results are also far better compared to other available methods. To our knowledge, this is the first attempt to extract and summarize situational Hindi tweets.

7.1.2 Uncovering small scale sub-events and summarizing information from Twitter

In Chapter 3, we develop a two class classifier which classifies situational and non-situational tweets. However, we observe that situational tweets contain information from different humanitarian classes like 'infrastructure damage', 'missing or trapped people', 'shelter and services', 'volunteering services', and so on. These high level information classes in turn consist of many small scale sub-events. Identifying humanitarian classes and sub-events present in those classes helps users to get a clear picture of the situation. Hence, first we classify situational tweets into different information classes using AIDR [56]. Next, we develop a dependency parser based approach which extracts noun-verb pairs from each of the humanitarian classes to represent sub-events. Crowdsource based evaluation reveals that sub-events detected by our proposed method are well organized, highly useful, and easier to follow by disaster responders compared to random bag of words identified by traditional clustering based sub-event detection techniques.

In the second step, we develop an ILP-based generic summarization method which maximizes the coverage of sub-events and content words. We show that our proposed approach can be tuned to generate different kind of situational summaries like — (i). high level update, (ii). humanitarian class specific update, (iii). missing person update etc. Side by side, highlighting class and sub-event information along with the tweets helps users to comprehend the summary more easily. We have tested our method over five natural and two man-made disaster events. In all the cases, proposed method performs better (6% - 30%) compared to other approaches.

7.1.3 Combining information from related tweets and generating abstractive summaries

In Chapter 3 and Chapter 4, we develop ILP-based techniques to generate extractive summaries by maximizing different factors like presence of content words, sub-events etc. However, we observe that many related tweets are posted during disaster which contain same information with little variations. Hence, we develop an abstractive summarization method which works in two steps — (i). we extract a set of important tweets based on our proposed extractive summarization technique (Chapter 3) (ii). after extracting important tweets, we build a word graph to generate paths by combining information from related tweets. Next, we develop an ILP-based method to produce abstractive summaries by maximizing the coverage of important paths and content words. Our proposed abstractive summarization method achieves better information coverage and diversity compared to extractive summarization methods.

7.1.4 Analyzing non-situational content during disaster

Most of the common people use non-situational tweets as a medium of expression of their opinions, sentiments, grievances and so on. However, we observe that few people use this disaster situation and microblogging platforms to attack particular religious communities. These *communal tweets* are potentially dangerous to the society. We develop a rule based classifier to identify communal tweets posted during disaster. We test the performance of our proposed classifier over five recent disaster events and obtain an average F-score of 0.90. After identifying communal content, we characterize their users. We discover many interesting facts like — (i). people post communal content even during natural disasters like earthquake and flood, (ii). popular politicians, tv reporters are also involved in posting such communal content along with common people, (iii). while most of the people get exasperated by sudden occurrence of disaster event, it is a common practice of few users (5 - 10%) to post communal content in a regular fashion, (iv). a few users also post *anti-communal* content which asks people not to spread communal hatred among masses, (v). users mostly prefer vernacular languages such as Hindi over English to post communal tweets, negative sentiments, and slangs.

As a final note, we have uploaded relevant codes of this thesis in the github repository (`https://github.com/krudra/koustav_phdthesis_2018`). We make the tweet-ids of the tweets related to disaster events publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterSummarizer/dataset.html`, `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`.

7.2 Directions for future work

Research works presented in this thesis open a lot of potential future directions. In this section, we describe some of them:

Extracting situational information from social networks beyond Twitter: In this thesis, we primarily work on the data obtained from Twitter. However, situational updates during disasters are available from other popular social media like Facebook, Google+ etc. Humanitarian organizations and disaster experts check the consistency of information from different social media before producing final reports [55]. All the media do not contain exactly the same information; some information missing in one medium may be available in others. Hence, it is straightforward and fruitful future work to obtain data from different sources before classifying and summarizing them. It will help to increase information coverage and diversity. Apart from that, time is critical in such scenarios. Hence, if some medium provides some situational updates earlier compared to others, we can directly use that to provide fast timely updates.

Extracting situational information from languages other than English: In this thesis, we mainly extract and summarize situational information from English tweets. Later on we observe that majority of people living in distant rural areas post information in their local languages (Hindi, Telugu, Marathi, Tamil, etc.) and some of these regional tweets contain exclusively new information which is not available in the English ones. Out of these vernacular languages, majority of the content posted is in Hindi. Hence, in this thesis, we extend our proposed classification-summarization framework also to Hindi tweets. However, we realize that extraction of information only from English and Hindi tweets is not enough for diverse countries like India where most of the useful information about an event is available in the long tail [2]. It is necessary and useful to extend the classification-summarization framework to other regional languages such as Marathi, Bengali etc. Side by side, combination of different languages might also bring some insight on the event, for instance if information is confirmed or not, if some details are not similar, etc. However, in order to combine information from multiple languages, we have to convert all the information into one common language. In recent times, researchers put lot of effort in cross-lingual content analysis [13, 140], machine translation [36, 139] etc. In future, we will try to incorporate such techniques to automatically extract and compare information posted in multiple languages. There exist lots of challenges in extending the framework to regional resource-poor languages. We list some of these challenges below.

1. In the classification of situational tweets, we need several dictionaries like list of modal verbs, subjective words, intensifiers etc. It is difficult to collect such dictionaries for resource-poor languages. This limitation is likely to affect precision and accuracy of the classification phase.
2. Because of the non-availability of Twitter specific tools for resource-poor languages (such as POS tagger, parser), the tools built for the formal texts have to be used, which can affect the detection of content words, and in-turn the outcome of the summarization method.

We have to address the above mentioned challenges before applying the proposed classification-summarization framework to local resource-poor languages.

Incorporating importance of humanitarian categories: In Chapter 6, we develop a generic ILP-based method which combines information about humanitarian categories, sub-events and content words to summarize situational tweets. Our method assigns uniform weights to each of the humanitarian categories like ‘infrastructure damage’, ‘missing person’ etc. However, we observe that natural disasters such as floods and hurricanes span much longer time periods. In such long ranging disasters, importance of different categories varies across different days. Incorporating the importance quotient of different humanitarian categories in the summarization formula will produce high quality informative summary. Along with high level summaries, disaster experts may also focus into the summaries of some dominant humanitarian categories based on their importance quotient.

Handling rumours: During crisis events, which include natural emergencies such as earthquakes, tsunami and cyclones, as well as man-made emergencies such as bomb blasts, and riots – a lot of valuable information is available via online social media [111, 132]. However, not all information obtained through online social networks are trustworthy [43, 44]. For instance, during Boston marathon blast in April, 2013, 29% of most viral content being discussed on Twitter were rumours and fake contents, 51% were generic comments and rest were true information [43]. Rumours are posted not only by common people, rather reputed and verified users also sometimes inadvertently post such misinformation [43]. In this thesis, we observe that numerical information about injured, missing people fluctuates and varies at a quick pace. Out of these values some information may be incorrect. We did not verify the authenticity of an information in this thesis. It may be interesting to check the authenticity / credibility of a tweet before forwarding the tweet to classification-summarization framework. In recent times, researchers explored interaction chain (reply, retweet) of tweets to judge the accuracy of the content. There is always a tradeoff between the authenticity of the content of a tweet and its detection time. On one hand, we can rely only on information posted by renowned news channels. However, local people start to post updates soon after the disaster and if we have to wait for news channels then important updates may be lost. In contrary, we can explore interaction pattern, information about the user who posted a tweet to check the authenticity of the content and it will take some time. Hence, detection of rumors in *real-time* is a challenging task which we will try to address in future.

Promoting anti-communal content: In Chapter 6, we discover that along with communal tweets a small fraction of users also post anti-communal content. Anti-communal tweets are less popular compared to communal ones. Prior researches showed that different factors like wording [122], popularity of the user who posted the tweet etc. affect the popularity of tweets. Mentioning popular users like politicians, celebrities in anti-communal tweets may help in getting high exposures. However, overloading users with information does not help in tweet popularity [136]. We have to find a suitable way to promote such anti-communal posts. These observations also raise many intriguing social questions like 'interaction between communal and anti-communal users', 'demographic biases' etc. We can address these issues in future.

Developing an end to end system to assist end-users: In this thesis, we propose different methods to address broadly the following sets of problems — (i). classification and summarization of situational tweets, (ii). extraction of small scale sub-events, (iii). identification of communal tweets. We believe that our work is significant especially in developing countries, where government-sponsored sophisticated systems to monitor situational updates in disaster scenario is largely missing. In future, we will try to deploy a system based on the approaches proposed in this thesis so that it can be practically used for any future disaster event.

Extracting information from images during crisis: During crisis situation information is also available in the form of images along with texts (tweets). With the increased usage of smartphones people tend to post lots of images which can describe the severity of damages and the present situation in the crisis region. In recent times researchers have shown a lot of interest in extracting information from images [9, 78]. It will be better if we can extract information from multimodal sources (both texts and images) during crisis. We can use image captioning, image text retrieval to enrich information set. Multimodal data analysis will provide us a way to compare information retrieved from both texts and images and estimate the importance as well as credibility of such dataset.

Extracting information during epidemics: During disease outbreaks, information posted on microblogging platforms such as Twitter by affected communities provide rapid access to diverse and useful insights helpful to understand various aspects of the outbreak. Research studies conducted with formal health organizations have shown the utility of such health related information on Twitter for quick response [29, 53, 66, 95].

To effectively utilize the tweets posted for any type of response efforts or decision-making processes, fast processing and analysis of raw tweet stream is necessary during an epidemic situation. During an epidemic, various types of information, including disease-related updates and personal opinions are posted by users in huge volume and at rapid rates. This online content contains valuable but multi-dimensional information like 'disease sign and symptoms', 'prevention mechanisms', 'transmission medium','death reports' etc. To make it presentable to health experts, these tweets first need to be automatically classified into different informative categories (e.g., symptom reports, prevention, treatment, etc). The number of messages classified in each category are still quite large and beyond the scope of human processing. It may be interesting to develop a classification-summarization framework for the tweets posted during epidemics.

7.3 Final words

In recent times, microblogging platforms like Twitter appear to be very effective source of real-time information during disaster. However, different kinds of information are posted by users and all of them are not useful. With the increased usage of social media, needs and requirements of different stakeholders are changing over time. Some communities try to extract useful situational information to support victims, provide reliefs to the affected people. On the other hand, some users try to use this panic situation to spread religious hatred. The studies presented in this thesis have tried to provide solution to some of the fundamental problems that arise during disaster. However, information content of different social media, requirements of end users etc., are changing day by day which will raise new sets of challenges and further research is required to address those challenges.

Bibliography

[1] Dhekar Abhik and Durga Toshniwal. Sub-event detection during natural hazards using features of social media data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 783–788. ACM, 2013.

[2] Puneet Agarwal, Rajgopal Vaithiyanathan, Saurabh Sharma, and Gautam Shroff. Catching the long-tail: Extracting local news events from twitter. In *ICWSM*, 2012.

[3] Nasser Alsaedi, Pete Burnap, and Omer Rana. Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, 17(2):18, 2017.

[4] Imran Awan and Irene Zempi. 'i will blow your face off'—virtual and physical world anti-muslim hate crime. *British Journal of Criminology*, page azv122, 2015.

[5] Norhidayah Azman, David Millard, and Mark Weal. Patterns of implicit and non-follower retweet propagation: Investigating the role of applications and hashtags. 2011.

[6] Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. Hindi subjective lexicon : A lexical resource for hindi polarity classification. In *Proc. LREC*, Austin, Texas, USA, May 2012.

[7] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-document abstractive summarization using ilp based multi-sentence compression. In

Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[8] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proc. ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

[9] Melissa Bica, Leysia Palen, and Chris Bopp. Visual representations of disaster. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1262–1276. ACM, 2017.

[10] Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J Passonneau. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1587–1597, 2015.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[12] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[13] Chloé Braud, Ophélie Lacroix, and Anders Søgaard. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 237–243, 2017.

[14] Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 2013.

[15] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11, 2016.

[16] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):1–14, 2014.

[17] Peter Burnap and Matthew Leighton Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, 7:223–242, 2015.

[18] Dongfeng Cai, Yonghua Hu, Xuelei Miao, and Yan Song. Dependency grammar based english subject-verb agreement evaluation. In *PACLIC*, pages 63–71. Citeseer, 2009.

[19] 2015 San Bernardino, California attack, December 2015. `https://en.wikipedia.org/wiki/2015_San_Bernardino_attack`.

[20] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency Situation Awareness from Twitter for Crisis Management. In *Proceedings of the International Conference Companion on World Wide Web (WWW)*, pages 695–698. ACM, 2012.

[21] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159, 2015.

[22] Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576, 2013.

[23] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. AAAI ICWSM*, May 2010.

[24] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[25] Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proc. AAAI ICWSM*, pages 340–348, 2011.

[26] Irfan Chaudhry. # hashtagging hate: Using twitter to track racism online. *First Monday*, 20(2), 2015.

[27] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[28] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 484–494, 2016.

[29] M. De. Choudhury. Anorexia on tumblr: A characterization study. In *Proc. ACM Digital Health*, pages 43–50. ACM, 2015.

[30] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[31] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proc. WWW Companion*, 2015.

[32] Gunes Erkan and Dragomir R. Radev. LexRank:Graph-based lexical centrality as salience in text summarization. volume 22, pages 457–479, 2004.

[33] Katja Filippova. Multi-sentence compression: finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 322–330. Association for Computational Linguistics, 2010.

[34] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proc. COLING*, pages 340–348, 2010.

[35] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3):10–14, 2011.

[36] Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 123–135, 2017.

[37] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *EMNLP*, pages 1602–1613, 2014.

[38] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah Smith, A. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc. ACL*, 2011.

[39] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[40] Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Bernhard Scholkopf. Quantifying information overload in social media and its impact on social contagions. In *Proc. AAAI ICWSM*, 2014.

[41] Edel Greevy and Alan F Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM, 2004.

[42] Aditi Gupta and Ponnurangam Kumaraguru. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? Technical report, IIITD-TR-2011-005, 2012.

[43] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter. In *eCrime Researchers Summit (eCRS), 2013*, pages 1–12. IEEE, 2013.

[44] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.

[45] Vishal Gupta. Hybrid algorithm for multilingual summarization of hindi and punjabi documents. In *Mining Intelligence and Knowledge Exploration*, pages 717–727. Springer, 2013.

[46] Gurobi – The overall fastest and best supported solver available, 2015. `http://www.gurobi.com/`.

[47] 2015 Gurdaspur attack – Wikipedia, July 2015. `https://en.wikipedia.org/wiki/2015_Gurdaspur_attack`.

[48] Typhoon Hagupit – Wikipedia, December 2014. `http://en.wikipedia.org/wiki/Typhoon_Hagupit`.

[49] Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. Tweetin' in the Rain: Exploring societal-scale effects of weather on mood. In *Proc. AAAI ICWSM*, 2012.

[50] 2015 Harda train derailment – Wikipedia, August 2015. `http://en.wikipedia.org/wiki/2015_Harda_accident`.

[51] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[52] Hindi parser and pos-tagger, 2015. `http://sivareddy.in/downloads/`.

[53] C. M. Homan, N. Lu, X. Tu, M. C. Lytle, and V. Silenzio. Social structure and depression in trevorspace. In *Proc. CSCW*, pages 615–625. ACM, 2014.

[54] Hyderabad blasts – Wikipedia, February 2013. `http://en.wikipedia.org/wiki/2013_Hyderabad_blasts`.

[55] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: a survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[56] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial intelligence for disaster response. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 159–162. International World Wide Web Conferences Steering Committee, 2014.

[57] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting Information Nuggets from Disaster-Related Messages in Social Media. In *Proc. ISCRAM*, 2013.

[58] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proc. LREC*, 2016.

[59] Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*, pages 31–39, 2014.

[60] 2014 India-Pakistan floods – Wikipedia, April 2015. `http://m.thehindu.com/news/national/flood-alert-in-kashmir-as-jhelum-crosses-danger-mark/article7352963.ece`.

[61] Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications.* Transaction Publishers, 1966.

[62] Chris Kedzie, Fernando Diaz, and Kathleen McKeown. Real-time web scale event summarization using sequential decision making. In *IJCAI*, 2016.

[63] Chris Kedzie, Kathleen McKeown, and Fernando Diaz. Predicting Salient Updates for Disaster Summarization. In *Proc. ACL*, 2015.

[64] Atif Khan, Naomie Salim, and Yogan Jaya Kumar. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747, 2015.

[65] Muhammad Asif Hossain Khan, Danushka Bollegala, Guangwen Liu, and Kaoru Sezaki. Multi-Tweet Summarization of Real-Time Events. In *Socialcom*, 2013.

[66] N. A. Kinnane and D. J. Milne. The role of the internet in supporting and informing carers of people with cancer: a literature review. *Supportive Care in Cancer*, 18(9):1123–1136, 2010.

[67] Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D Manning. Named entity recognition with character-level models. In *Proc. HLT-NAACL 2003-Volume 4*, pages 180–183. Association for Computational Linguistics, 2003.

[68] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A Dependency Parser for Tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[69] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[70] Chen Li, Xian Qian, and Yang Liu. Using supervised bigram-based ilp for extractive summarization. In *ACL (1)*, pages 1004–1013, 2013.

[71] Wei Li. Abstractive multi-document summarization with semantic information extraction. In *EMNLP*, pages 1908–1913, 2015.

[72] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out, ACL*, pages 74–81, 2004.

[73] Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. Toward abstractive summarization using semantic representations. 2015.

[74] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proc. AAAI ICWSM*, 2014.

[75] Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. # isisisnotislam or# deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, pages 95–106. ACM, 2016.

[76] Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. Detecting flames and insults in text. 2008.

[77] Suman Maity, Anshit Chaudhary, Shraman Kumar, Animesh Mukherjee, Chaitanya Sarda, Abhijeet Patil, and Akash Mondal. Wassup? lol : Characterizing out-of-vocabulary words in twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW '16 Companion, pages 341–344, New York, NY, USA, 2016. ACM.

[78] Yelena Mejova, Sofiane Abbar, and Hamed Haddadi. Fetishizing food in digital age:# foodporn around the world. In *ICWSM*, pages 250–258, 2016.

[79] Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Degeneracy-based real-time sub-event detection in twitter stream. In *Proc. AAAI ICWSM*, pages 248–257, 2015.

[80] Rada Mihalcea and Paul Tarau. TextRank:Bringing order into texts. In *Proc. EMNLP*, pages 404–411, 2004.

[81] 2015 Nepal earthquake – Wikipedia, April 2015. `http://en.wikipedia.org/wiki/2015_Nepal_earthquake`.

[82] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety information mining – what can nlp do in a disaster –. In *Proc. International Joint Conference on Natural Language Processing (IJCNLP)*, 2011.

[83] Minh-Tien Nguyen, Asanobu Kitamoto, and Tri-Thanh Nguyen. Tsum4act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction. In *Proc. PAKDD*, 2015.

[84] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM, 2012.

[85] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385, 2010.

[86] Andrei Olariu. Hierarchical clustering in improving microblog stream summarization. In *Proc. CICLing*, pages 424–435, 2013.

[87] Andrei Olariu. Efficient online summarization of microblogging streams. In *Proc. EACL(short paper)*, pages 236–240, 2014.

[88] Miles Osborne and Mark Dredze. Facebook, twitter and google plus for breaking news: Is there a winner? In *ICWSM*, 2014.

[89] Miles Osborne, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, Tom Jackson, Fabio Ciravegna, and Ann OBrien. Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media. In *Proc. ACL*, 2014.

[90] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[91] Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.

[92] 2015 Paris attacks, November 2015. `https://en.wikipedia.org/wiki/November_2015_Paris_attacks`.

[93] Daraksha Parveen and Michael Strube. Multi-document Summarization Using Bipartite Graphs. In *Proc. TextGraphs Workshop on Graph-based Methods for Natural Language Processing*, pages 15–24, October 2014.

[94] Daraksha Parveen and Michael Strube. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In *IJCAI*, pages 1298–1304, 2015.

[95] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proc. Icwsm*, volume 20, pages 265–272, 2011.

[96] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[97] Nick Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *null*, pages 235–241. IEEE, 2007.

[98] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. Can twitter replace newswire for breaking news? In *Seventh international AAAI conference on weblogs and social media*, 2013.

[99] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. Automatic sub-event detection in emergency management using social media. In *Proc. WWW*, pages 683–686. ACM, 2012.

[100] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. Social media for crisis management: clustering approaches for sub-event detection. *Multimedia Tools and Applications*, 74(11):3901–3932, 2015.

[101] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212, 2007.

[102] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 25–34. ACM, 2011.

[103] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, and D. Crystal. *A comprehensive grammar of the English language*, volume 397. Cambridge University Press, 1985.

[104] Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[105] Ashwin Rajadesingan. *Sarcasm detection on Twitter: A behavioral modeling approach*. Arizona State University, 2014.

[106] Siva Reddy and Serge Sharoff. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In *Proc. International Workshop On Cross Lingual Information Access*, pages 11–19. Asian Federation of Natural Language Processing, November 2011.

[107] Alan Ritter, Sam Clark, and Oren Etzioni. Named entity recognition in tweets: an experimental study. In *Proc. EMNLP*, pages 1524–1534, 2011.

[108] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[109] Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *EMNLP*, pages 1131–1141, 2016.

[110] Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Saptarshi Ghosh. Characterizing communal microblogs during disaster events. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 96–99. IEEE, 2016.

[111] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. World Wide Web Conference (WWW)*, pages 851–860, 2010.

[112] Sandy Hook Elementary School shooting – Wikipedia, December 2012. `http://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting`.

[113] N. B. Sarter and D. D. Woods. Situation awareness: a critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1(1):45–57, 1991.

[114] Naveen Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Inferring Who-is-Who in the Twitter Social Network. In *Proc. WOSN Workshop*, 2012.

[115] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. Sumblr: Continuous summarization of evolving tweet streams. In *Proc. ACM SIGIR*, 2013.

[116] Leandro Silva, Mainack Mondal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*, 2016.

[117] Szymkiewicz-Simpson coefficient, 2017. `https://en.wikipedia.org/wiki/Overlap_coefficient`.

[118] Reem Suwaileh, Maram Hasanain, and Tamer Elsayed. Light-weight, conservative, yet effective: scalable real-time tweet summarization. TREC, 2016.

[119] W. t. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *Proc. IJCAI*, pages 1776–1782, 2007.

[120] H. Takamura, H. Yokono, and M. Okumura. Summarizing a document stream. In *Proc. ECIR*, 2011.

[121] Lynda Tamine, Laure Soulier, Lamjed Ben Jabeur, Frederic Amblard, Chihab Hanachi, Gilles Hubert, and Camille Roth. Social media-based collaborative information access: Analysis of online crisis-related twitter conversations. In *ACM 27th Conference on Hypertext & Social Media*, 2016.

[122] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation:topic- and author-controlled natural experiments on twitter. In *Proc. ACL*, 2014.

[123] Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ujwal Gadiraju. Groundhog Day: Near-duplicate Detection on Twitter. In *Proc. Conference on World Wide Web (WWW)*, 2013.

[124] C. Thaokar and L. Malik. Test model for summarizing hindi text using extraction method. In *IEEE Conference on Information Communication Technologies*, pages 1138–1143, April 2013.

[125] Twitter – data statistics, 2017. `https://www.omnicoreagency.com/twitter-statistics/`.

[126] Twitter – user statistics, 2017. `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`.

[127] REST API Resources, Twitter Developers, 2017. `https://dev.twitter.com/docs/api`.

[128] North India floods – Wikipedia, June 2013. `http://en.wikipedia.org/wiki/2013_North_India_floods`.

[129] Istvan Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proc. ACL*, 2013.

[130] Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth M. Anderson. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *Proc. AAAI ICWSM*, 2011.

[131] Sarah Vieweg, Carlos Castillo, and Muhammad Imran. Integrating social media communications into the rapid assessment of sudden onset disasters. In *Social Informatics*, pages 444–461. Springer, 2014.

[132] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proc. ACM SIGCHI*, 2010.

[133] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. In *Proc. ACL (Vol2: Short Papers)*, 2013.

[134] Marilyn A Walker, Owen Rambow, and Monica Rogati. Spot: A trainable sentence planner. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[135] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1331–1340. International World Wide Web Conferences Steering Committee, 2013.

[136] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proc. ACM International Conference on World Wide Web (WWW)*, pages 1331–1340, 2013.

[137] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27:1301–1314, 2015.

[138] Matthew Williams and Olivia Pearson. Hate crime and bullying in the age of social media. 2016.

[139] Shuangzhi Wu, Dongdong Zhang, Nan Yang, Mu Li, and Ming Zhou. Sequence-to-dependency neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 698–707, 2017.

[140] Ruochen Xu and Yiming Yang. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1415–1425, 2017.

[141] Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. A preliminary study of tweet summarization using information extraction. *NAACL 2013*, page 20, 2013.

[142] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proc. WWW*, pages 1445–1456. ACM, 2013.

[143] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using Social Media to Enhance Emergency Situation Awareness. *IEEE Intelligent Systems*, 27(6):52–59, 2012.

[144] Q. Zhou, L. Sun, and J. Nie. Is sum: A multi-document summarizer based on document index graphic and lexical chains. In *Proc. DUC2005*, 2005.

[145] Arkaitz Zubiaga, Damiano Spina, Enrique Amigo, and Julio Gonzalo. Towards Real-Time Summarization of Scheduled Events from Twitter Streams. In *Hypertext(Poster)*, 2012.

Appendix A

List of all publications by the candidate during his PhD tenure (at the time of defence - 02/04/2018)

Following is a list of all the publications by the candidate including those on and related to the work presented in the thesis. The publications are arranged in chronological order and in four sections – (i). book chapter, (ii). journals, (iii). conferences, and (iv). workshops.

Book Chapter

Koustav Rudra, Abhijnan Chakraborty, Niloy Ganguly, Saptarshi Ghosh. "Understanding the Usage of Idioms in Twitter Social Network", Pattern Recognition and Big Data, World Scientific, pp. 767–788, February 2017, (Editors: Amita Pal, Sankar K Pal), (DOI: https://doi.org/10.1142/9789813144552_0024) (ISBN: 978-981-3144-54-5).

Journals

1. **Koustav Rudra**, Ashish Sharma, Niloy Ganguly, Muhammad Imran. “Classifying and Summarizing Information from Microblogs during Epidemics”, Special Issue on “Exploitation of Social Media for Emergency Relief and Preparedness” in the journal Information Systems Frontiers (Springer), 2018.

2. **Koustav Rudra**, Niloy Ganguly, Pawan Goyal, Saptarshi Ghosh. “Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters”, ACM Transactions on the Web (ACM TWEB), 2018.

3. **Koustav Rudra**, Ashish Sharma, Niloy Ganguly, Saptarshi Ghosh. ”Characterizing and Countering Communal Microblogs during Disaster Events”, IEEE Transactions on Computational Social Systems (IEEE TCSS), 2018, DOI: 10.1109/TCSS.2018.2802942.

Conferences

1. **Koustav Rudra**, Ashish Sharma, Niloy Ganguly, Muhammad Imran. “Classifying Information from Microblogs during Epidemics”, 7th ACM Digital Health Conference (DH 2017), London, July 2-5, 2017.

2. **Koustav Rudra**, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, Niloy Ganguly, paper on ”Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter?”, Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas, USA, November 1–5, 2016.

3. Rafiya Begum, Kalika Bali, Monojit Choudhury, **Koustav Rudra**, Niloy Ganguly. “Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments”, The 10th edition of the Language Resources and Evaluation Conference (LREC 2016), Slovenia, 23-28 May, 2016.

4. **Koustav Rudra**, Ashish Sharma, Niloy Ganguly, Saptarshi Ghosh. “Characterizing Communal Microblogs during Disaster Events”, The 2016

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016), San Francisco, CA, USA, August 18-21, 2016.

5. **Koustav Rudra**, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, Prasenjit Mitra. “Summarizing Situational Tweets in Crisis Scenario”, 27th ACM Conference on Hypertext and Social Media (HT 2016), Halifax, Canada, July 10-13, 2016.

6. **Koustav Rudra**, Subham Ghosh, Niloy Ganguly, Pawan Goyal, Saptarshi Ghosh. “Extracting Situational Information from Microblogs during Disaster Events: a Classification-Summarization Approach”, 23rd ACM International Conference on Information and Knowledge Management (CIKM 2015), Melbourne, October 19-23, 2015.

7. **Koustav Rudra**, Abhijnan Chakraborty, Manav Sethi , Shreyasi Das, Niloy Ganguly, Saptarshi Ghosh. “#FewThingsAboutIdioms: Understanding Idioms and its Users in the Twitter Online Social Network”, Pacific Asia Knowledge Data Discovery (PAKDD 2015), Vietnam, May 19-22, 2015.

Workshops

1. Prabhat Agarwal, Ashish Sharma, Jeenu Grover, Mayank Sikka, **Koustav Rudra**, and Monojit Choudhury. “I may talk in English but gaali toh Hindi mein hi denge : A study of English-Hindi Code-Switching and Swearing Pattern on Social Networks”, Social Networking Workshop, COMSNETS 2017, Bangalore, India, January 5-9.

2. **Koustav Rudra**, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. “Summarizing Situational and Topical Information during Crises”, The Fourth International Workshop on Social Web for Disaster Management, 2016 (SWDM 2016 — colocated with CIKM 2016), (arXiv preprint: https://arxiv.org/abs/1610.015610).

3. Anirban Sen, **Koustav Rudra**, Saptarshi Ghosh. “Extracting Situational

Awareness from Microblogs during Disaster Events", Social Networking Workshop, COMSNETS 2015, Bangalore, India, January 6-10.