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

KOUSTAV RUDRA, NILOY GANGULY, and PAWAN GOYAL,
Department of CSE, IIT Kharagpur, India
SAPTARSHI GHOSH, Department of CSE, IIT Kharagpur and Department of CST, IIEST Shibpur, India

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 sentiments and opinions, but also to *summarize* the large amounts of situational information posted in real-time. During disasters in countries like India, a sizable 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/Devanagari (the national language of India), and some of the information contained in such non-English tweets is 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 non-English tweets.

CCS Concepts: • **Information systems** → **Social networks**; **Information extraction**; **Clustering and classification**; **Summarization**; *Information retrieval*;

Additional Key Words and Phrases: Disasters, microblogs, twitter, situational tweets, classification, summarization, content words

ACM Reference format:
Koustav Rudra, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2018. Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters. *ACM Trans. Web* 12, 3, Article 17 (July 2018), 35 pages.
https://doi.org/10.1145/3178541

This work is an extended version of the paper by Rudra et al., Extracting Situational Information from Microblogs during Disaster Events: A Classification-Summarization Approach, Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM), 2015.
This research is partially supported by a grant from the Information Technology Research Academy (ITRA), DeITY, Government of India (Ref. No. ITRA/15 (58)/Mobile/DISARM/05). Additionally, K. Rudra is supported by a fellowship from Tata Consultancy Services.
Authors' addresses: K. Rudra, N. Ganguly, P. Goyal, and S. Ghosh, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur 721302, India; emails: koustav.rudra@cse.iitkgp.ernet.in, krudra5@gmail.com, {niloy, pawang, saptarshi}@cse.iitkgp.ac.in.

https://doi.org/10.1145/3178541

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. In particular, several recent studies have shown that microblogging sites play a key role in obtaining situational information during *disaster events* [2, 3, 14, 23, 31, 38, 49, 51, 54].

During a disaster event, various types of information, including situational updates, personal opinions (e.g., on the adequacy of relief operations), and sentiments (e.g., sympathy for those affected by the disaster) are posted by users in huge volume and at rapid rates. While different types of information have different utilities, *situational information*—information which helps the concerned authorities (e.g., governmental and non-governmental agencies) to gain a high-level understanding of the situation—is critical for the authorities to plan relief efforts [40]. Hence, it is important to develop automated methods to *extract microblogs /tweets which contribute to situational information* [24, 50].[1] A related, yet different, challenge is to deal with the rapid rate at which microblogs are posted during such events, which calls for *summarization of the situational information*. 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 [50] and summarizing such information [17, 24]. However, these prior works have certain limitations, as detailed in Section 2. For instance, most of the classifiers developed to distinguish 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 the Indian context where information posted on Twitter in Hindi and English, respectively, is often different, i.e., some information is present only in Hindi tweets and is not available via English ones (details in Section 3).

The present work proposes a novel framework for extracting and summarizing situational information from microblog streams posted during disaster scenarios. In brief, the tweets are first preprocessed and fragmented based on end-markers such as "!" and "?". The fragmented tweets are then classified to extract situational tweets, and the situational tweet stream is then summarized (after removing duplicate tweets). The proposed methodology takes advantage of some specific traits of tweet streams posted during disasters. Our major contributions are listed below.

(i) Analyzing tweets posted during several recent disaster events (a detailed description of the dataset is provided in Section 3), 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 performance.

[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*.

[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.

(ii) We develop a classifier using low-level lexical and syntactic features to distinguish between situational and non-situational information (Section 4). Using 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 [50] for in-domain and cross-domain settings.

(iii) We observe that most of the important information posted during disasters 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. Additionally, exploiting the semantic relation among the content words helps to further improve the quality of the final summary. We propose a novel content-word and semantic-relation–based summarization approach (SEMCOWTS) 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 5). The proposed approach surpasses various state-of-the-art tweet summarization approaches [18, 24, 41] in terms of ROUGE-1 recall and F-scores (Section 6). 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 5.5).

(iv) For Hindi tweets, we cannot directly use the 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 the classification phase are not available (sometimes not enriched) in Hindi. (ii) In the summarization phase, direct measurement of semantic similarity among Hindi content words is not possible. To solve the first problem, we develop necessary lexicons for Hindi. In the case of summarization, Hindi content words are converted to English using the Bing translator service[3] and then we use a standard English summarization framework. To the best of our knowledge, this is the first attempt to summarize tweets in regional languages such as Hindi. Experiments show that the proposed scheme performs significantly better than several state-of-the-art summarization approaches.

Note that our classification-summarization approach was first proposed in a prior study [36]. The present work extends our prior work as follows. First, we improve the methodology (COWTS) in [36] and show that the methodology proposed in the present work (SEMCOWTS) performs better than that in [36]. Second, we try to provide global as well as local location-specific updates about victims who were killed, stranded, trapped, died, and so on. Third, we develop the classification-summarization framework to handle tweets in Hindi, which shows the scope of extending the framework to other regional languages as well.

As a final contribution, we make the tweet-ids of the tweets related to several disaster events publicly available to the research community at http://www.cnergres.iitkgp.ac.in/disasterSummarizer/dataset.html.

[3] https://www.microsoft.com/en-us/translator/translatorapi.aspx.

2 RELATED WORK

Microblogging sites are serving as useful sources of situational information during disaster events [2, 3, 14, 23, 30, 31, 38, 49, 51, 54]. However, for practical utility, such situational information has to be extracted from among a lot of conversational and sentimental information, and summarized in near real-time. This section briefly discusses some recent studies on classification and summarization of tweets. We also discuss some of the prior work on processing non-English text, especially text in Hindi (Devanagari script).

2.1 Classification of Tweets during Disaster Events

Several studies have attempted to extract situational information during disaster events [15, 49, 50]. Most of these studies used classifiers based on bag-of-words models (i.e., classifiers which are trained on particular types of words extracted from the tweets themselves) to distinguish between tweets which contain situational information, and tweets which do not. However, this approach is heavily dependent on the vocabulary of a specific event, and does not work well in the practical scenario where the classifier is trained on tweets of some past events and is then used to classify tweets of a new disaster event [50]. To overcome these limitations, our prior work [36] proposed a classifier based on lexical features of tweets (see Section 4 for details). In the present work, we use the scheme in [36] for English tweets, and extend the scheme to use the classifier for Hindi tweets as well.

2.2 Tweet Summarization

Most of the prior research on tweet summarization has focused on summarizing a set of tweets, e.g., tweets posted during the course of a sports event [4, 18, 43]. 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 [25, 41, 53, 55]. For instance, Shou et al. [41] proposed a scheme based on first clustering similar tweets and then selecting few representative tweets from each cluster, finally ranking these tweets according to importance via a graph-based approach (LexRank) [5]. Olariu [25] proposed a graph-based abstractive summarization scheme where bigrams extracted from the tweets are considered as the graph nodes. Osborne et al. [26] proposed a real event tracking system using greedy summarization.

Along with standard summarization approaches, a few recent studies [17, 24, 35, 36] have also focused specifically on summarization of news articles and tweets posted during disasters. In particular, our prior work [36] proposed a classification-summarization technique to extract and summarize situational information from tweet streams.

The present work has two important advantages over these prior studies, including our prior work [36]. First, while [36] attempted to capture actionable information (such as number of casualties), they did not focus on *location-specific variances*—such as the number of casualties at different locations—which are especially important during disaster events spanning large geographical areas. For instance, it is important to know the number of casualties at different locations in order to plan the distribution of relief materials and personnel among the various disaster-affected areas. The methodology proposed in the present work separately identifies and summarizes time-varying, location-specific actionable information. Second, to our knowledge, all the prior works attempt to extract information only from English tweets. However, 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.

Table 1. Keywords Used to Collect Tweets for Different Disaster Events

Event	Keywords
HDBlast	"Hyderabad and bomb." "Hyderabad and bomb and blast," "#Hyderabadblast"
UFlood	"Uttarakhand and flood, "#Uttarakhandflood"
SHShoot	"Sandyhook and shoot," "#Sandyhookshooting"
Hagupit	"Typhoon and Hagupit," "#TyphoonHagupit"
NEquake	"Nepal and quake," "Nepal and earthquake," "#NepalEarthquake"
HDerail	"Harda and derail," "Harda and train and derail," "#Harda"

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.

2.3 Processing Devanagari Documents

There have been prior attempts to summarize Devanagari documents [45], and to develop basic natural language processing tools such as parts-of-speech (POS) taggers [33] and subjectivity lexicons [1] for Devanagari. 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 [10]. As such, research on processing of tweets written in regional languages such as Devanagari is still in its infancy, and to our knowledge, this study is the first systematic attempt to extract useful information from Devanagari tweets.

3 DATASETS

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

3.1 Disaster Events

We consider tweets posted during the following disaster events:

(1) **HDBlast**—two bomb blasts in the city of Hyderabad, India [13].
(2) **SHShoot**—an assailant killed 20 children and 6 adults at the Sandy Hook elementary school in Connecticut, USA [39].
(3) **UFlood**—devastating floods and landslides in the Uttaranchal state of India [48].
(4) **Hagupit**—a strong cyclone code-named Typhoon Hagupit hit Philippines [9].
(5) **NEquake**—a devastating earthquake in Nepal [22].
(6) **HDerail**—two passenger trains got derailed near Harda in India [11].

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 of the tweets can be expected to be diverse as well.

We collected relevant tweets posted during each event through the Twitter API [46] using keyword-based matching. Table 1 shows the keywords used to collect the tweets for the above disaster events.

Among the events listed above, we use the first four to develop and evaluate our classification-summarization framework. For each of these four events, we select approximately the first 5,000 English tweets in chronological order. We then use 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

Table 2. 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)

adapting it to Hindi tweets. For these two events, we collect both English and Hindi tweets using the Twitter API, by searching for tweets containing the hashtags #NepalEarthquake and #Harda.[4] For these two events, we gather 19,970 and 4,171 English tweets, and 6,349 and 1,083 Hindi tweets, respectively.

3.2 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 2 shows some sample Hindi tweets containing information that is not available in the English tweets for the same event. Similarly, Table 3 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, 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, and so forth, 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 derive 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 the same time span. First, we remove duplicate tweets from the Hindi dataset. After this step, we get 230 and 128 Hindi tweets for HDerail and NEquake events, respectively.[5] 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:

[4]Note that even tweets in other languages use English hashtags for greater visibility.
[5]Only situational tweets were considered for this analysis, as identified by the approach described later in the article.

Table 3. 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
HDerail	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

(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 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 [6] 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 appears in the English tweets; these disagreements were resolved through discussions among the annotators.

We find that 15.45% and 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 the 8.13% and 14.29% cases, the information is obtained earlier in Hindi tweets than the English tweets. These observations establish the need to process tweets in regional 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.

3.3 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 [31, 50] showed that situational tweets contain information about the current situation, whereas non-situational tweets mostly consist of opinions, sentiments, abbreviations, and so on. Recently, Imran et al. [16] 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

Table 4. 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]

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 the case of tweets related to donation or charities, as to whether they should be considered situational or otherwise. Prior works such as Qu et al. [31] 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 4.

3.3.1 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 [16]. Relief information includes information about helping organizations, necessary requirements of affected victims, phone numbers of nearby hospitals, and so on. Such information can immediately help in relief operations.

3.3.2 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

Table 5. 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!

from police investigation in the 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 4), and then summarizing the situational information (Section 5).

4 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 use human annotators to obtain this gold standard (details below). During annotation, it is observed that *a significant number of tweets posted during disaster events contain a mixture of situational and non-situational information.* Table 5 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 motivates us to identify different fragments of a tweet and process them separately for classification and summarization steps. This preprocessing stage is described next.

4.1 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 POS tagger [7] 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 [7] 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 4). 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 [21]. Table 6 shows examples of some modal verbs and their variations.

Table 6. 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," "wouldnt́," "wudnt," "wudnt́"

Table 7. 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

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 5).

(iii) Subsequently, we focus on particular end-markers (e.g., "!," "." "?") to split a tweet into multiple fragments. We use Twitter parts-of-speech tagger to identify these 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 7 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).

4.2 Establishing Gold Standard

For training the classifier, we considered 1,000 randomly selected tweet fragments related to each of the first four events described in Section 3. Three human volunteers independently observed the tweet fragments. All the volunteers are regular users of Twitter, have a good knowledge of the English and Hindi languages, and none of them is an author of this article. Before the annotation task, the volunteers were acquainted with some examples of situational and non-situational tweets identified in prior works [50, 51].

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 the same annotation process also for the 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 1,000 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

Table 8. Lexical and Syntactic 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 [52]. 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 colgate scam money to disaster."
Presence of intensifiers	Existence of frequently used intensifiers [32], 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 [37], 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."

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 a balanced training set for the raw tweets.

4.3 Classification Features and Performance

Prior research [50] 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 11 low-level lexical and syntactic features, as listed in Table 8, 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.

Table 9. 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

Table 10. Classification Accuracies of Support Vector Machine (SVM) Classifier on *Tweet Fragments*, Using (i) Bag-of-Words Features (BOW) and (ii) Proposed Lexical and Syntactic Features (PRO). 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	69.576%	**84.260%**	55.035%	**78.220%**	61.805%	**89.583%**	46.688%	**82.339%**
UFlood	55.769%	**82.451%**	63.349%	**79.609%**	61.458%	**89.814%**	50.220%	**81.456%**
SHShoot	55.769%	**83.052%**	49.882%	**79.859%**	75.454%	**90.042%**	49.889%	**80.242%**
Hagupit	51.201%	**77.283%**	49.765%	**75.644%**	63.310%	**86.458%**	71.524%	**85.862%**

We use a Support Vector Machine (SVM) classifier (with default RBF kernel) [29] to classify the fragmented tweets into two classes based on the features described in Table 8. Note that the first nine features in Table 8 have been taken from our prior work [36] and in the present work, we include two new features: (i) presence of religious terms and (ii) presence of slangs, to improve the performance of the classifier.

We compare our classifier with a standard bag-of-words (BOW) model similar to that in [50], where the same SVM classifier is used considering as features the frequency of every distinct unigram and bigram (Twitter-specific tags are removed using POS tagger), POS tags, count of strongly subjective words, and presence of personal pronouns. In the case of the BOW model, for each of the events, total feature space consists of a number of distinct unigrams, bigrams, POS tags, strong subjective word count, and personal pronouns. Table 9 shows the number of distinct unigrams, bigrams, total feature space, and 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. Table 10 shows the accuracies of the classifier using BOW model and the proposed features (PRO) on the fragmented tweets. 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 the rest of the events.

4.4 In-Domain Classification

The BOW model performs well in the case of in-domain classification (diagonal entries in Table 10) due to the uniform vocabulary used during a particular event. However, the proposed features significantly outperform the BOW model. The result is especially significant since it shows that a higher accuracy can be achieved even without considering the event-specific words.

Table 11. Classification Recall and F-scores of Support Vector Machine (SVM) Classifier on *Situational Tweet Fragments*, Using Proposed Lexical and Syntactic Features. Non-Diagonal Entries are for Cross-Domain Classification

Train set	**Test set**							
	HDBlast		UFlood		SHShoot		Hagupit	
	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score
HDBlast	0.85	0.84	0.75	0.77	0.87	0.89	0.77	0.81
UFlood	0.85	0.83	0.81	0.80	0.88	0.89	0.78	0.81
SHShoot	0.85	0.83	0.77	0.79	0.87	0.89	0.75	0.79
Hagupit	0.89	0.79	0.86	0.78	0.88	0.86	0.94	0.87

Table 12. Classification Accuracies of SVM on *Raw Tweets*, Using (i) Bag-of-Words Features (BOW) and (ii) Proposed Features (PRO). 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	66.916%	**81.899%**	52.576%	**76.112%**	59.794%	**81.890%**	49.875%	**77.431 %**
UFlood	58.111%	**80.984%**	59.939%	**77.062%**	61.161%	**82.118%**	50%	**80.548%**
SHShoot	50.265%	**78.723%**	50%	**75.058%**	70.845%	**84.738%**	50%	**76.059%**
Hagupit	52.925%	**78.590%**	52.810%	**75.409%**	54.441%	**79.612%**	60.954%	**79.667%**

4.5 Cross-Domain Classification

The non-diagonal entries of Table 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-scores of our proposed classification model over the situational tweets, in Table 11.

4.6 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 12. Comparing the classification accuracies in Table 10 (on preprocessed and fragmented tweets) and Table 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 6) that the preprocessing phase in turn helps in information coverage during the summarization process.

4.7 Feature Ablation

In this part, we try to judge the importance of individual features in the classification, through feature ablation experiments. Table 13 reports the in-domain and cross-domain accuracies of the situational tweet classifier for feature ablation experiments, averaged over all the datasets. The presence of numerals, pronouns, exclamation mark, subjective words, and non-situational words

Table 13. Feature Ablation Experiments for the Situational Tweet Classifier for Both In-Domain and Cross-Domain Scenarios. NONE Represents the Case When all the Features are 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

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 [50] under various experimental settings. However, since the best achieved classification accuracy is still around 80%, a question naturally arises as to whether the 20% misclassification would substantially impact the subsequent summarization step. We shall discuss the effect of misclassification on summarization in Section 6.

4.8 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 the 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 6.3, we will actually deploy the classifier trained on earlier events over tweets related to the two later events, NEquake and HDerail.

Table 14. Size of Different Devanagari Lexicons Developed as a Part of this Study

Subjective words	Pronouns	Modal verbs	Wh words	Intensifiers	Religious terms	Slangs	Nonsituational words
5670	95	56	36	42	26	20	377

4.9 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.

4.10 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 [12] which is designed for *formal* Devanagari text. Hence, we apply the following preprocessing 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 end-markers "!," "?," "|"

The lexical and syntactic features that are listed in Table 8 (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 [1]. All the other lexicons like pronouns, intensifier, wh-words, and so forth, are collected from Wikipedia and online sources. All these lexicons also contain many phonetic variations of a particular word (e.g., अपना, अपनी, अपने). The sizes of different lexicons are reported in Table 14. These lexicons can be downloaded and used for research purposes.[6]

We apply the same methodology as described in Section 4—tweets are partitioned into fragments, the features listed in Table 8 are computed for each fragment, and the fragments are then classified into situational or non-situational using a SVM classifier (with default RBF kernel).

4.11 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 good knowledge of the Hindi language and none of them is an author of this article—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 select an equal

[6] http://www.cnergres.iitkgp.ac.in/disasterSummarizer/dataset.html.

Table 15. Classification Accuracies of SVM on *Fragmented Hindi Tweets*, Using (i) Bag-of-Words Features (BOW) and (ii) Proposed Lexical and Syntactic Features (PRO)

Train set	**Test set**			
	NEquake		HDerail	
	BOW	PRO	BOW	PRO
NEquake	71.687%	**81.305%**	50%	**72.222%**
HDerail	50%	**77.935%**	62%	**74.222%**

number of tweet fragments that were judged non-situational, and construct balanced training sets for the classifier.

Similar to Section 4, we use SVM classifiers, and compare the performance of the proposed features with a bag-of-words model. Table 15 shows the in-domain accuracies (diagonal entries) and cross-domain accuracies (non-diagonal elements). It is seen that the proposed features lead to significantly better classification of Hindi tweets than the bag-of-words model, especially in the cross-domain scenarios.

From Table 15, we can see that in the case of Hindi tweet classification, we achieve low cross-validation accuracy compared to English tweets 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 want to consider only those tweets which are classified with a certain confidence level. For this, we test our proposed English and Hindi situational tweet classifiers on manually annotated datasets. We check 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 the remaining three cases, the precision, recall, and F-scores are comparable (F-score is around 0.84). For both English and Hindi tweets, we decide to set the confidence level to 0.8, i.e., we select only those SVM-classified situational messages for which the confidence of the classifier is $\geq$0.80.

5 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) locations (e.g., names of places), (iii) nouns (e.g., important context words like people, hospital), and (iv) main verbs (e.g., "killed." "injured," "stranded"). We refer to these terms as *content words*. We also consider *semantic relations among these content words* in order to group similar nouns and verbs into communities, which in turn helps to include a diverse set of content words in final summary. This methodology also enables the generated summary to cover information from various dimensions like relief, helpline, rescue efforts, caution messages, and so on. This section describes our proposed methodology, which we call SEMCOWTS.

5.1 Need for Disaster-Specific Summarization Approach

We observe a specific trend in the case of situational tweets posted during disaster events, which is very different from tweet streams posted during other types of events. As tweets are seen in chronological order, the number of *distinct content words* increases very slowly with the number of tweets, in the case of disaster events.

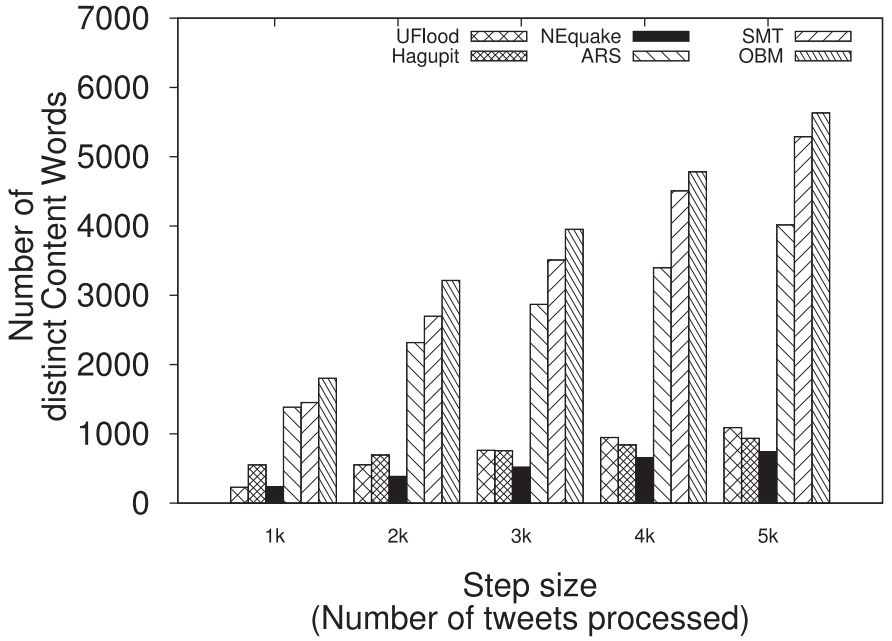


Fig. 1. Variation in the number of distinct content words with the number of tweets in chronological order, shown for disaster events (three left bars in each group), and other events (three right bars in each group (Tweets about Arsenal vs Chelsea match, Smartphone launch, Obama Healthcare Policy)).

To demonstrate this, we compare tweet streams posted during disaster events with those posted during three political, sports, and technology-related events; these streams were made publicly available by a previous study [41]. Figure 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 the case of the disaster events. We find that this is primarily due to (i) the presence of a huge number of retweets or near-duplicates of few important tweets, and (ii) the presence of a 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 [44], (ii) focus on some semantically related content words during summarization—the identification of which is described in Section 5.2, while the summarization framework is described in Section 5.3, and (iii) adopt specific strategies for the heavily repeated content words associated with frequently changing numerical information (described in Section 5.5).

5.2 Generating Semantic Content Words

In our prior work [36], we considered numerals, all nouns, and verbs as content words because information in a tweet is generally centered around these two core nuggets—nouns which represent the concepts, and verbs which represent the events. As mentioned here we divide nouns in two

parts: (a) place/location and (b) other important context nouns. To identify location information, we use various online sources.[7]

For the rest of the nouns, we observe that some nouns represent semantically the same concepts. For example, "blast" and "explosion" provide information regarding the same concept. There are also some verbs which denote the same event. Hence, in the present work, we attempt to group together semantically similar nouns and verbs into components/communities in order to reduce redundancy in the final summary and develop a well-defined semantic set of content words.

5.2.1 Extracting Concepts or Noun Communities. For this purpose, we (i) extract all the nouns from the dataset, (ii) construct an undirected weighted graph where the nodes are the nouns, (iii) add an edge between two nouns if their semantic similarity is $\geq$0.8 (we have used UMBC service [47] for measuring semantic similarity) and finally, (iv) identify different connected components in the graph where each of the components represents one noun community/concept.

5.2.2 Extracting events or verb communities. Verbs like "killed," "injured," and so on, represent important events during a disaster. We develop an undirected weighted graph using verbs as nodes, and edges constructed based on the UMBC service [47], and extract connected components (similar to noun communities) from the graph. At the end, each of the identified components represents one verb community/event.

Our objective is to combine concepts/events which are semantically similar, but are represented through synonyms. Semantic services return words which are not directly related to each other. For example, in the case of "forget," we get "neglect" (0.7), "miss" (0.7), "lose" (0.6), and similarly for "torture," we get "wound" (0.7), "injure" (0.6) (the values in parentheses represent similarity scores). However, formation of clusters with such related words does not satisfy our requirements. Rather, we find that such words are used in different context in other tweets. In general, the semantic set of a word w is reduced if threshold is too high, and many unrelated words are categorized as semantically similar when threshold is too low. For example, "end," "discontinue," and "stop" are semantically related to "terminate" with confidence score 0.9, and "break," "finalize," and "decide" are related with score 0.6. Setting threshold to 1.0 discards many similar words like "stop" and "end." On the other hand, many unrelated words like "decide" and "finalize" will appear if we set threshold to 0.6. We test the effect of this threshold by varying it between 0.9 and 0.7. In the case of 0.8, we obtain an average improvement of 0.8% and 2% compared to 0.9 and 0.7 thresholds, respectively, in terms of ROUGE-1 F-score [20] (details about ROUGE score computation are presented in Section 6). Hence, we set the semantic similarity threshold as 0.8.

After all the preprocessing described above, we get a list of **semantic content words** as follows: (i) numerals (e.g., number of casualties or affected people, or emergency contact numbers), (ii) locations (e.g., names of places), (iii) concepts (identified noun components), and (iv) events (identified verb components). We observe that important information during disasters gets posted around such **semantic content words**. In Table 16, we report the total number of nouns, verbs, concepts, and events obtained for each of the datasets.

5.3 Semantic Relation-Based Content Word 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 described earlier. 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

[7] http://www.zipcodestogo.com/Connecticut/, http://zip-codes.philsite.net/, http://www.nepalpost.gov.np/index.php/postal-codes-of-nepal, http://www.mapsofindia.com/pincode/.

Table 16. Statistics of Concepts, Events, Distinct Numerals, and Places Across Different Disaster Events

Event	#Nouns	#Verbs	#Concepts	#Events	#Numerals	#Places
HDBlast	309	155	265	103	94	19
UFlood	632	306	515	201	223	67
SHShoot	225	119	193	83	22	5
Hagupit	470	215	380	154	285	40
NEquake	844	368	654	227	300	54
HDerail	297	157	249	110	122	34

Table 17. 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 semantic content words in the n tweets
i	Index for tweets
j	Index for semantic 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 semantic content word j
$Length(i)$	Number of words present in tweet i
Score(j)	tf-idf score of semantic content word j
T_j	Set of tweets where semantic content word j is present
C_i	Set of semantic content words present in tweet i

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 centered around semantic content words, an effective way to attain good coverage of important information in the summary is by optimizing the coverage of *important semantic content words* in the tweets included in the summary. We use an ILP-based technique [28] to optimize the coverage of the important semantic content words. The ILP method proposed by Parveen et al. [28] assigns weights to the sentences/tweets, as obtained via the PageRank algorithm in the ILP framework. However, one of the objectives of our summarization technique is to generate the summary in real-time. Due to this, the PageRank-based method cannot be applied because the computation of PageRank does not scale over large datasets in real-time [41]. Hence, in our ILP framework, we provide weights to the semantic content words instead of tweets and try to maximize their coverage. Table 17 states the notations used.

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, and is given by

$$Score(j) = (1 + log(|T_j|)) * log(n/|T_j|). \quad (1)$$

However, in the case of concept one component may contain more than one noun. In such cases, we compute scores of individual nouns present in that component, take maximum value among

them, and set that score as the score of that component. same thing also holds for verb components/ events.

The summarization is achieved by optimizing the following ILP objective function:

$$max\left(\sum_{i=1}^{n} x_i \; + \; \sum_{j=1}^{m} Score(j).y_j\right) \tag{2}$$

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L, \tag{3}$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m], \tag{4}$$

$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n], \tag{5}$$

where the symbols are as explained in Table 17. 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 semantic content-words (through the y_j variables) included. The constraint in Equation (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 Equation (4) ensures that if the semantic 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 semantic content word is present is selected. Similarly, the constraint in Equation (5) ensures that if a particular tweet i is selected to be included in the summary, i.e., if $x_i = 1$, then the semantic content words in that tweet are also selected.

We use GUROBI Optimizer [8] 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.

5.4 Summarizing Hindi Tweets

We now describe how the summarization scheme is extended to summarize Hindi tweets, and the challenges therein. As mentioned in the previous section,the performance of our proposed summarization algorithm depends on three parameter: (i) extraction of content words, (ii) deriving semantic relations among the content words, and finally, (iii) developing semantic content words. Usability of the various summarization algorithms on Hindi tweets is limited by the unavailability of natural language processing tools for Hindi tweets.

5.4.1 Extraction of Content Words. To our knowledge, there does not exist any Twitter-specific part-of-speech tagger and semantic similarity measure for Hindi. Hence, we apply a standard Hindi POS tagger [12] to identify nouns and verbs. For English tweets, we use standard Twitter-specific POS tagger [7] having accuracy ≥90%. Hence, for English tweets we can detect content words with ≥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 (numerals, nouns, and verbs) are extracted from these tweets as marked by Hindi POS tagger [12]. 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 get annotated as content words. Hence, the limitation of POS tagging affects the performance of summarization of non-English tweets.

Table 18. 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

5.4.2 Deriving Semantic Relations among the Content Words. For finding semantic similarity among Hindi nouns and verbs, first we map them to corresponding English words using Bing translator service,[8] and then we measure semantic similarity between the English terms using the UMBC service [47]. After that, we apply a standard graph-based component detection method to the mapped nouns and verbs in order to extract noun and verb communities (as proposed in Section 5.2).

However, this Hindi-to-English conversion depends on the accuracy of Bing translator and it introduces errors in the conversion process. For this, in the next step, we are not able to find semantically similar words for 35% nouns and 31% verbs. In such cases, we are not able to compute semantic similarity; hence, we directly consider such words as content words (singleton component). In this way, we finally obtain **semantic content words** for Hindi. This limitation hampers the diversity of information in the final summary generated (if we were able to group semantically related nouns and verbs, then we can collect information in the final summary from various noun and verb communities).

5.5 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 18 shows how during the HDBlast event, the reported number of victims/injured persons changed during a period of only 7 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.[9]

Specifically, we consider particular disaster-specific key verbs like "kill," "die," "injure," and "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 18, 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.

5.5.1 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*

[8] https://www.microsoft.com/en-us/translator/translatorapi.aspx.

[9] 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.

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 *"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 the 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 [19] to capture this information. If a numeral is directly associated with a main verb (i.e., if an edge exists between the 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 is pruned manually from the exhaustive list obtained from Google syntactic *n*-grams.[10]

5.5.2 Assigning Locational Information to Key Verbs. 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 EXPERIMENTAL RESULTS

This section compares the performance of the proposed framework (SEMCOWTS) 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.

6.1 Experimental Settings: Baselines and Metrics

We consider the first four disaster events described in Section 3 for the experiments. For each dataset, we consider the first 5,000 tweet fragments in chronological order, extracted situational tweet-fragments using our classifier, and pass the situational tweets to the summarization modules. We consider two breakpoints at 2K, and 5K tweets, i.e., the summaries are demanded at the corresponding time-instants.

6.1.1 Establishing Gold Standard Summaries. At each of the breakpoints, three human volunteers (the 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 they 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

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

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.

6.1.2 Baseline Approaches. We compare the performance of our proposed summarization scheme with that of four prior approaches, which consist of recent unsupervised 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.

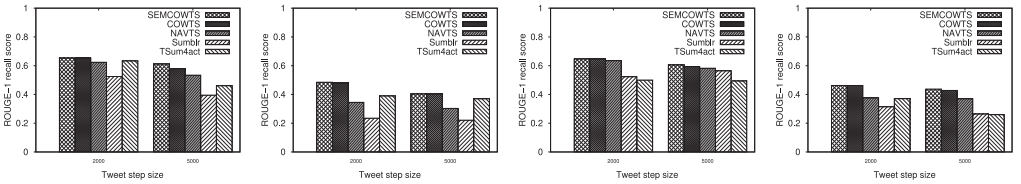
(i) **COWTS:** Disaster-specific tweet summarization approach developed in our prior work [36], where we considered all nouns, verbs, and numerals as content words, without attempting to extract the key content words (as described in Section 5.2).
(ii) **Sumblr:** The online tweet summarization approach by Shou et al. [41], 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) **TSum4act:** The methodology proposed by Nguyen et al. [24]. They prepare clusters of situational tweets using LDA, extract numerals, geo-location information, and events from tweets using the Twitter NER tool [34], construct a weighted graph among the tweets using cosine similarity as the edge weights, apply weighted PageRank [27], and finally select tweets based on Simpson's similarity measure from each cluster. In our case, we select 20 clusters for the LDA model and one tweet from each cluster until we cross the pre-specified word limit.
(iv) **NAVTS:** Since SEMCOWTS 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 nouns, 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. [18].

We apply SEMCOWTS 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) is 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 [7]—#(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.

6.1.3 Evaluation Metrics. We use the standard ROUGE [20] 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 are present in the candidate summary normalized by the count of unigrams present in the 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 are present in the 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.

Table 19. 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 SEMCOWTS (The Proposed Methodology) and the Four Baseline Methods (COWTS, NAVTS, Sumblr, and TSum4act) on the Same Situational Tweet Stream, at Breakpoints 2K, and 5K Tweets

Step size	**ROUGE-1 F-score**									
	HDBlast					UFlood				
	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act
0–2000	**0.7132**	0.6717	0.6478	0.5643	0.6580	**0.4778**	0.4717	0.3520	0.2582	0.3804
0–5000	**0.5898**	0.5854	0.5352	0.4207	0.4686	**0.4124**	0.3990	0.3095	0.2466	0.3693
Step size	**ROUGE-1 F-score**									
	SHShoot					Hagupit				
	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act
0–2000	**0.6508**	0.6324	0.6324	0.5361	0.5214	**0.4829**	0.4438	0.3646	0.3227	0.4460
0–5000	**0.6114**	0.5838	0.5838	0.5800	0.5000	**0.4577**	0.4405	0.3879	0.2846	0.2755



(a) **HDBlast** (b) **UFlood** (c) **SHShoot** (d) **Hagupit**

Fig. 2. ROUGE-1 recall scores of the summaries of different events, generated by the proposed methodology (SEMCOWTS) and the four baseline methods, at breakpoints 2K and 5K tweets.

6.2 Performance Comparison

Table 19 and Figure 2 give the ROUGE-1 recall and F-scores for the five algorithms for the four datasets, at breakpoints 2K and 5K, respectively. It is evident that SEMCOWTS performs significantly better than all the baseline approaches. For instance, mean scores indicate an average improvement of more than 40% in terms of F-score over Sumblr [41], 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 [24], and COWTS in all cases—on average, we obtain improvement of 24% and 3% for F-scores over TSum4act and COWTS, respectively. Further, the higher F-scores for SEMCOWTS than those for NAVTS indicate that our selected content words lead to better summarization. We also see that the better performance of SEMCOWTS 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 methodologies, Table 20 shows summaries of length 100 words, generated by SEMCOWTS and TSum4act (both disaster-specific methodologies) from the same tweet stream—at the 5K breakpoint during the UFlood event—and Table 21 shows the ground truth summary of 250 words from the same tweet stream—at the 5K breakpoint during the same UFlood event. The two summaries are quite distinct, with most of the tweets being different. We find that the summary returned by SEMCOWTS 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.

Table 20. Summary of 100 Words, Generated at 5K Breakpoint of the UFlood Dataset by (i) SEMCOWTS (Proposed Methodology) and (ii) TSum4act, another Disaster-Specific Summarization Methodology

Summary by SEMCOWTS	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, 5,000 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. 1,000 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.

Table 21. Ground Truth Summary of 250 Words, Generated at 5K Breakpoint of the UFlood Dataset

helpline numbers, 0135-2710334, 0135-2710335, 0135-2710233 Helpline number for pilgrims 0755-2556422. Uttarkashi, 01374-226126, Chamoli, 01372-251437 Tehri, 01376-233433, Rudraprayag 01732-1077, Nainital & around, 05946-250138. helpline numbers Army Medical Emergency numbers, 18001805558 18004190282 8009833388. Nearly 6,000 to 8,000 pilgrims stranded in Kedarnath, 2,500 stranded in Hemkunt Sahib. Army paratroopers reach Sonprayag, Gauri Kund in Kedarnath. 2,500 more troops, 14 choppers, 100 mountain rescue teams pressed into action for rescue and relief work. 30 IAF aircraft & choppers have airlifted 1,400 people from Uttarakhand to safety in 150 sorties. 15 more Paratroopers being loaded in Dhruv, at Jolly Grant Airport for induction to Kedarnath-Gaurikund. Roads from joshimath and Rudraprayag have opened but those in kedarnath are still trapped. No Kedarnath-Badrinath pilgrimage for three years. 3,000 still trapped near Ghangaria. Google launches Person Finder to help people. ITBP has also established helpline for info on UttarakhandFloods, 011-24362892, 0-9968383478. 163 people rescued from flooded areas in Uttarakhand, Himachal. Army Column along with a JCB is clearing landslide at Patal Ganga, Joshimath, Uttarakhand. IAF says 100 sorties conducted, 1,300 people rescued from Uttarakhand and other flood hit areas. 300 from Bihar missing in Uttarakhand. 14,000 people missing, 60,000 still stranded, 80,000 people have been evacuated till now. Minor earthquake 3.5 hits Pithoragarh district of Uttarakhand. 2,232 houses , 154 bridges, and 1,520 roads damaged. 4,200 villages in desperate need. Name/contact info of major hospitals in Dehradun. If missing in Uttarakhand used BSNL , call toll-free numbers 1503, 09412024365 to ID last active location. Oxfam to give dry food, blankets to Uttarakhand flood victims.

Table 22. ROUGE-1 F-scores of SEMCOWTS 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	0.7132	0.5879	0.5898	0.4449
UFlood	0.4778	0.4350	0.4124	0.3961
SHShoot	0.6508	0.5500	0.6114	0.4369
Hagupit	0.4537	0.4249	0.4577	0.3919

6.2.1 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 SEMCOWTS takes 7.759, and 9.562 seconds on average (over the four datasets), respectively, to generate summaries.[11] The time taken increases sub-linearly with the number of tweets and is slightly higher compared to that taken by the COWTS and NAVTS baselines due to component detection phase (on the same situational tweet streams), and significantly better than the time taken by TSum4act.

6.2.2 Benefit of Classification before Summarization. We verify that separating out situational tweets from non-situational ones significantly improves the quality of summaries. Considering all four events together, the mean ROUGE F-score at breakpoint 2,000 for SEMCOWTS is 0.4994 *without* prior classification (i.e., when all tweets are input to the summarizer) as compared to 0.5738 after classification. Table 22 gives the F-scores of SEMCOWTS on classified and unclassified tweets, for all four events at two breakpoints. As time progresses, a fraction of non-situational tweets is also increased in number which affects the summarization step to a great extent. As is evident from Table 22, F-scores at 5K are significantly low when we consider a whole set of tweets.

6.2.3 Effect of Misclassification on Summary Recall. As stated in Section 4, 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 *misclassified tweets* are also covered by the correctly classified situational tweets. On the other hand, if we consider semantic similarity measure, then we are able to cover 89.09% of the content-words present in the *misclassified tweets*. Table 23 shows coverage of content words present in misclassified situational tweets for all four events. In the latter case, we are able to capture semantically related words like ("picture," "image"), ("blast," "explosion"), which belong to the same component (as discussed in Section 5.2). This implies that only a small fraction of the content-words are missed out in the stream sent for summarization.

6.2.4 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 [18]. As stated in Section 5, we consider three types of content words : numerals, nouns, and verbs. From the

[11]We do not include time to compute semantic similarity using UMBC service. If we have that database, then we can avoid crawling and running time will increase in a linear fashion.

Table 23. Statistics of Coverage of Content Words Present in Misclassified Situational Tweets

Event	Coverage	Semantic coverage
HDBlast	84.48	87.55
UFlood	83.55	87.69
SHShoot	87.57	91.04
Hagupit	86.07	90.09

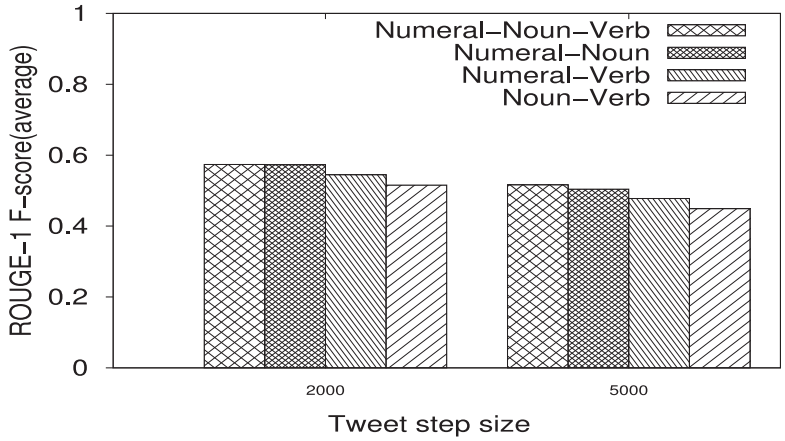


Fig. 3. Effect of individual types of content words on the summary.

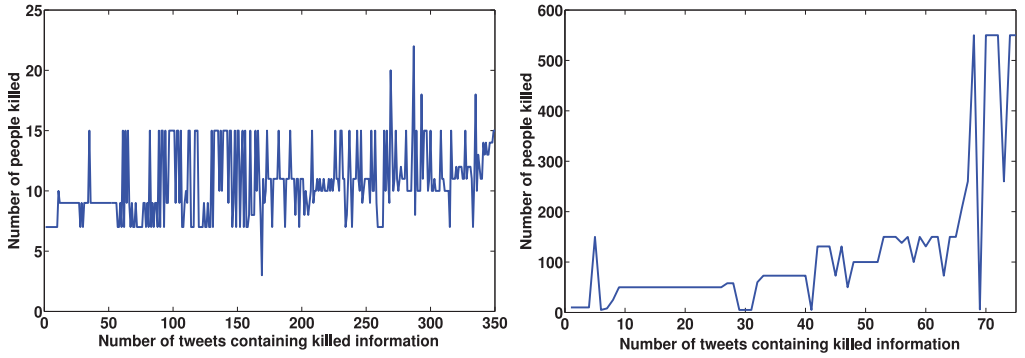
comparison between SEMCOWTS and NAVTS, it has already been established that our choice of content words achieves better summarization for tweets posted during disaster events than the information words proposed in [18].

We now analyze whether all 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 compares the F-scores (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). As a sanity check, we also include adjectives among the content words and run SEMCOWTS; however, the performance deteriorates noticeably.

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).

6.2.5 Handling Frequently Changing Numerals. Figure 4 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 5.5). Specifically, we check what fraction of such



(a) HDBlast (b) UFlood

Fig. 4. Variation in the reported number of people killed, during two disaster events. The x-axis represents the sequence of tweets which contain such information.

Table 24. 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 SEMCOWTS (the Proposed Methodology) and the Four Baseline Methods (COWTS, NAVTS, Sumblr, and TSum4act) on the Same Situational Tweet Stream, at Two Breakpoints ($\mathbf{B_1}$ = 10,000, 2,000 and $\mathbf{B_2}$ = 19,102, 4,361 for NEquake and HDerail, Respectively)

Step size	**ROUGE-1 F-score**									
	NEquake					HDerail				
	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act	SEMCOWTS	COWTS	NAVTS	Sumblr	TSum4act
B_1	**0.4049**	0.3650	0.3224	0.2160	0.3385	**0.4895**	0.4870	0.4554	0.4312	0.4381
B_2	**0.3753**	0.3500	0.2666	0.2000	0.3560	**0.4769**	0.4757	0.4433	0.4293	0.3929

numerical information can 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 three 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.

6.3 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, and then deploying it on tweets posted during the two most recent disaster events—NEquake (the earthquake in Nepal in April 2015) [22], and HDerail (train derailment at Harda, India in June 2015) [11].

6.3.1 Summarization of English Tweets. We directly use SEMCOWTS for summarizing the English tweets. We compute summaries at two breakpoints—in the middle of the stream B_1 (at 10,000 and 2,000 tweets for the NEquake and HDerail events, respectively) and at the end of the stream. Three human volunteers are used to prepare gold standard summaries at these breakpoints following the approach used in Section 6. Table 24 states the ROUGE-1 F-scores for SEMCOWTS and the

Table 25. Comparison of Unigram F-scores for SEMCOWTS (the Proposed Methodology) and Three Baseline Methods (COWTS, NAVTS, and Sumblr) on the Same Situational Hindi Tweet Stream

Event	ROUGE-1 F-score			
	SEMCOWTS	COWTS	NAVTS	Sumblr
NEquake	**0.5770**	0.5694	0.4700	0.3588
HDerail	**0.6602**	0.6469	0.6135	0.5392

four baseline strategies: COWTS, NAVTS, Sumblr, and TSum4act. It is evident that SEMCOWTS *has the highest F-score*. Further, SEMCOWTS 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.

6.3.2 Summarization of Hindi Tweets. We apply SEMCOWTS to situational Hindi tweets, and compare its performance with that of three baseline techniques COWTS, NAVTS, and Sumblr. We cannot apply the TSum4act method [24] due to unavailability of named entity recognizers for Hindi. Similar to earlier evaluation frameworks, three human volunteers are 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, and so forth as per the ROUGE-1 F-score [20].[12] The results are stated in Table 25; it is evident that SEMCOWTS outperforms the baseline approaches in terms of coverage and quality of the summaries. To give an idea, Table 26 shows summaries generated by SEMCOWTS, COWTS for the NEquake event and Table 27 shows ground truth summary for the same event.[13]

In Section 3, we have shown that some information is exclusively available in Hindi tweets, i.e., they are not present in English tweets. In this section, we have measured the *Szymkiewicz-Simpson Similarity* [42] between the content words present in Hindi summary and English summary. For NEquake and HDerail events. we obtain similarity scores of 0.14 and 0.20, respectively. These values indicate that Hindi and English summaries are quite different for both events.

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

6.4 Discussion on Performance

A deeper look at various baseline techniques helps us to understand their shortcomings and the reasons behind the superior performance of SEMCOWTS. COWTS considers all nouns and verbs but sometimes the same concept or event is represented by two different synonyms. Capturing and clubbing them helps to improve information coverage and summary quality. 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.

Out of the other baseline techniques, Sumblr [41] 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

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

[13]English translations of Table 26 and Table 27 are provided in Section 7.

Table 26. Summary of NEquake Dataset (Hindi) by (i) SEMCOWTS (Proposed Method) and (ii) COWTS, Another Disaster-Specific Summarization Method

Summary by SEMCOWTS	Summary by COWTS
आंध्र प्रदेश सरकार ने दिल्ली और हैदराबाद में नियंत्रण कक्ष स्थापित किये हैं दिल्ली :011-23782388 हैदरा। नेपाल अपडेट शाम 4:45 पर ग्लोबमास्टर 3 विमान 285 पैसेंजर्स को लेकर पालम एयरपोर्ट पहुंचेगा। नेपाल से लौटे पंजाबी तीर्थयात्रियों की अगवानी में नजर आए सीएम बादल। ट्रक में लगभग तीन हजार आईवी फ्ल्यूड की बोतले ड्रेसिंग का सामान भी काठमांडू भेजा गया है। 2 भारतीयों की मौत। अब तो जाग जा ऐ मुर्ख इंसान। नेपाल भूकंप रेस्क्यू ऑपरेशन को मैत्री नाम दिया गया। भूंकप की ताजा तस्वीरें देखें। मौत और आपदा सिर्फ एक सूचना नहीं है। नेपाल भूकंप के कारण बंद हुआ त्रिसुली हाईड्रोपावर का टनल 60 मजदूर फंसे। राजस्थान स्कूल की दीवार गिरने से लड़की मौत 8 बच्चे घायल। में हजारों लोगों के सिर से छिन गई छत खुले में बारिश औऱ ठंड से परेशान हैं लोग। ज्यादा से प्रधानमंत्री राहत कोष में दान दे। कोलकाता शॉपिंग मॉल में लगी आग फायर ब्रिगेड की 22 गाडियां मौके पर। बद्रीनाथ और जोशीमठ में भूस्खलन कई यात्री फंसे। 32 5606 वाल्वो बस उपलब्ध करायी टीम रविवार रात को गोरखपुर पहुंचेगी सोमवार सुबह काठमांडू। कुद्रत का बिछौना मर्ग का खिलौना मर्ग मौत। नेपाल में फंसी भारत अंडर 14 गर्ल्स फुटबॉल टीम। उत्तर प्रदेश में भूंकप के झटके दफ्तर घर छोड़ सड़कों पर निकले लोग। ताश के पत्तों की तरह बिखरा नेपाल। भूकंप से हिला हिमालय से निकाले गए 17 शव। नेपाल जाने हेतु डेरा सेवादार नीचे दिए लिंक को खोलें। नीतीश कुमार ने बताया बिहार में भूकंप से 50 की मौत पढ़ें खबर।	ऐसे दुखद समय में भी संघी नहीं सुधरेंगे। नेपाल में भारतीय दूतावास ने हेल्पलाइन नंबर जारी किए हेल्पलाइन नंबर +977-985-110-7021, +977-985-113-5141। नेपाल अपडेट शाम 4:45 पर ग्लोबमास्टर 3 विमान 285 पैसेंजर्स को लेकर पालम एयरपोर्ट पहुंचेगा। 32 5606 वाल्वो बस उपलब्ध करायी टीम रविवार रात को गोरखपुर पहुंचेगी सोमवार सुबह काठमांडू। उत्तर प्रदेश सरकार कंट्रोल रूम का फोन नं तथा फैक्स नं। राजस्थान स्कूल की दीवार गिरने से लड़की मौत 8 बच्चे घायल। भूंकप की ताजा तस्वीरें देखें। अगले 48 घंटे तक भारत से नेपाल जाने वाली कॉल फ्रीः। अब तो जाग जा ऐ मुर्ख इंसान। ट्रक में लगभग तीन हजार आईवी फ्ल्यूड की बोतले ड्रेसिंग का सामान भी काठमांडू भेजा गया है। कुद्रत का बिछौना मर्ग का खिलौना मर्ग मौत। नेपाल भूकंप के कारण बंद हुआ त्रिसुली हाईड्रोपावर का टनल 60 मजदूर फंसे। ज्यादा से प्रधानमंत्री राहत कोष में दान दे। के हृदयेश जोशी रोड के रास्ते पहुंचे काठमांडु दूर दराज़ के इलाकों से पहुंचायेगे हाल। कोलकाता शॉपिंग मॉल में लगी आग फायर ब्रिगेड की 22 गाडियां मौके पर। राहत का काम छोड़ अमेरिकी हेलिकॉप्टर को खोज हैं सैकड़ों नेपाली। नेपाल में फंसी भारत अंडर 14 गर्ल्स फुटबॉल टीम। पहले बारिश फिर ओले अब धरती क्यूँ डोले क्या हो रहा है भोले। नेपाल से लौटे पंजाबी तीर्थयात्रियों की अगवानी में नजर आए सीएम बादल। ताश के पत्तों की तरह बिखरा नेपाल। कैसे हैं ये बताए। आंध्र प्रदेश सरकार ने दिल्ली और हैदराबाद में नियंत्रण कक्ष स्थापित किये हैं दिल्ली :011-23782388 हैदरा।

cluster. Tweets within a cluster are ranked based on the 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 the importance of clusters is also necessary for preparing the final summary. Similar types of tweet selection problems also arise in case of TSum4act [24]. TSum4act [24] captures disaster-specific terms like numerals, events, noun-phrases, and locations but it has two limitations: (i) determiningthe importance of different clusters (same as Sumblr), (ii) PageRank-based iterative updates take a long time for large datasets which creates a bottleneck for real-time summarization. To resolve such issues, focusing on particular POS tags, similarity between terms, and ILP-based technique (as used in SEMCOWTS) 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 SEMCOWTS are that it is completely unsupervised and can be applied to any type of disaster event.

Table 27. Ground Truth Summary of 250 Words, Generated for the NEquake Event, in Hindi

आंध्र प्रदेश सरकार ने दिल्ली और हैदराबाद में नियंत्रण कक्ष स्थापित किये हैं दिल्ली :011–23782388 हैदरा। नेपाल अपडेट शाम 4:45 पर ग्लोबमास्टर 3 विमान 285 पैसेंजर्स को लेकर पालम एयरपोर्ट पहुंचेगा। ट्रक में लगभग तीन हजार आईवी फ्ल्यूड की बोतले ड्रेसिंग का सामान भी काठमांडू भेजा गया है। नेपाल भूकंप के कारण बंद हुआ त्रिसुली हाईड्रोपावर का टनल 60 मजदूर फंसे। राजस्थान स्कूल की दीवार गिरने से लड़की मौत 8 बच्चे घायल। बद्रीनाथ और जोशीमठ में भूस्खलन कई यात्री फंसे। 32 5606 वाल्वो बस उपलब्ध करायी टीम रविवार रात को गोरखपुर पहुंचेगी सोमवार सुबह काठमांडू। नेपाल में फंसी भारत अंडर 14 गर्ल्स फुटबॉल टीम। भूकंप से हिला हिमालय से निकाले गए 17 शव। नीतीश कुमार ने बताया बिहार में भूकंप से 50 की मौत पढ़ें खबर। अगले 48 घंटे तक भारत से नेपाल जाने वाली कॉल फ्रीः। होम मिनिस्ट्री ने हेल्पलाइन नंबर जारी किया है +91 11 2301 2113 +91 11 2301 4104 +91 11 2301 7905। नेपाल गृह मंत्रालय के मुताबिक मरने वालों का आंकड़ा 618 पहुंचा 125 भारतीयों के फंसे होने की। नेपाल में भारतीय दूतावास ने हेल्पलाइन नंबर जारी किए हेल्पलाइन नंबर +977–985–110–7021, +977–985–113–5141। 550 गुजराती भी फंसे हैं नेपाल में। इंडियन एयरफोर्स का विमान ग्लोबमास्टर सी नेपाल में फंसे 103 भारतीयों को लेकर दिल्ली के पालम एयरपोर्ट पहुंचा। नेपाल में फंसे हैं करीब तीन लाख पर्यटक। राहत और बचाव कार्य जारी नेपाल में फंसे 544 भारतीयों को भारतीय वायुसेना के विमान से स्वदेश लाया गया। ने 23 डॉक्टरों व 24 पैरा मेडिकल स्टाफ समेत 47 सदस्यीय चिकित्सा दल को नेपाल भेजा है।

In the case of Hindi tweets, we have less (varied) tweets compared to English. Hence, capturing important content words is relatively easy in the case of Hindi tweets. Due to this, our system obtains high ROUGE-1 F-scores (Table 25) for Hindi tweets in spite of such resource constraints.

As is evident from the discussion till now, there exist lots of challenges in extending the classification-summarization framework to Hindi tweets. We believe that such kinds of constraints will also be faced if the framework is to be applied to other local, resource-poor languages. We list some of these challenges below.

(1) In the classification phase, we need several dictionaries like list of modal verbs, subjective words, intensifiers, and so on. 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.
(3) In the summarization phase, we have to measure the semantic similarity between two nouns or verbs. While there are good tools available for this task in English, we have to take one of the following approaches for other languages.
    —We can convert words in other languages to English and then compute their similarities. However, the translation/conversion step can introduce errors, which are propagated in the pipeline, and finally hamper the summarization process.
    —We can develop algorithms for measuring semantic similarity of words in non-English languages, e.g., using neural network–based models. However, to develop such methods, we need large amounts of training data. These individual steps are crucial and time-consuming. We have to systematically resolve these problems to achieve the final goal of computing semantic similarity between two words in non-English languages.

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

7 CONCLUSION

This article presents 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 entire tweets, (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 develop a domain-independent classifier which performs better than the domain-dependent BOW technique. We also propose an ILP-based optimizing framework to summarize the situational tweets, which out-performs other summarization methods, in the case of English as well as Hindi tweets. This framework to work in Hindi, however, calls for certain preprocessing and resource building steps which is taken up systematically in this article. Lexicons generated for Hindi tweets are mainly collected from manually annotated tweets and web services.[14] However, this lexicon list is not an exhaustive one. In order to use proposed the classification-summarization framework over Hindi tweets with 100% efficiency, we need to develop a complete and exhaustive resource dictionary, and a proper Hindi-to-English conversion scheme for Hindi tweets. This is beyond the scope of this submission.

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. In such long-ranging disasters, users may be interested both in current summaries (say, the last few hours) as well as historical summaries (last week). A minor modification of the underlying data structures of the present scheme would solve the issue. The Content Word Dictionary, which maintains the content words as well as the rate at which they are appearing in the tweets, can be created for each epoch, and accordingly both recent as well as historical summaries can be obtained as per user-requirement. We will formalize this in more detail in our future work, which includes deploying a live system.

As mentioned, the module which we develop to handle continuous updates of the actionable numerical items shows that conflicting numbers are getting updated at the same time, and a robust technique needs to be developed to differentiate between spam/rumor and real information. In this part, after forwarding the information, we can take help fro other media sources to check their validity. This would be our immediate future work.

The methods proposed in this article have a lot of scope for future improvement. In the classification phase, we have used a traditional SVM classifier. We plan to apply neural network models to improve the classification performance of our situational tweet classifier. The usage of deep neural networks is not only to improve abstract metrics, but also can qualitatively change the type of errors made by the classifier. Additionally, we are also trying to use deep learning models to measure more accurate semantic similarity between words. In particular, we will attempt to develop methods to directly compute the semantic similarity between two non-English words, i.e., without converting them to English, which will help in reducing error due to translation in the summarization pipeline.

As a final note, we believe that the impact of our work is significant especially in emerging countries, where government-sponsored sophisticated systems to monitor situational updates in

[14] goo.gl/6vPOkT, goo.gl/wIS5yD.

disaster scenarios are largely missing, whereas processable information is available not only through English tweets but also in regional languages like Hindi.

REFERENCES

[1] Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for Hindi adjective polarity classification. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. ELRA, 1189–1196.

[2] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12 Companion)*. ACM, 695–698.

[3] Carlos Castillo. 2016. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations* (1st ed.). Cambridge University Press, New York.

[4] Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM'11)*. AAAI, 340–348.

[5] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22, 1 (Dec. 2004), 457–479.

[6] Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 3 (1973), 613–619.

[7] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2 (HLT'11)*. ACL, 42–47.

[8] Gurobi 2016. Gurobi—The state-of-the-art mathematical programming solver for prescriptive analytics. Retrieved July 2016 from http://www.gurobi.com.

[9] Hagupit-wiki 2014. Typhoon Hagupit—Wikipedia. Retrieved December 2014 from http://en.wikipedia.org/wiki/Typhoon_Hagupit.

[10] Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM'12)*. AAAI, 479–482.

[11] Harda-derailment-wiki 2015. 2015 Harda Train Derailment—Wikipedia. Retrieved August 2015 from http://en.wikipedia.org/wiki/2015_Harda_accident.

[12] Hindi-postagger 2015. Hindi Parser and POS-Tagger. http://sivareddy.in/downloads/.

[13] Hyderabad-blast-wiki 2013. Hyderabad Blasts—Wikipedia. Retrieved February 2013 from http://en.wikipedia.org/wiki/2013_Hyderabad_blasts.

[14] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys* 47, 4 (June 2015), 67:1–67:38.

[15] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)*. ACM, 159–162.

[16] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. ELRA, 1638–1643.

[17] Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2015. Predicting salient updates for disaster summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing (IJCNLP'15) (Volume 1: Long Papers)*. ACL, 1608–1617.

[18] Muhammad Asif Hossain Khan, Danushka Bollegala, Guangwen Liu, and Kaoru Sezaki. 2013. Multi-tweet summarization of real-time events. In *Proceedings of the 2013 International Conference on Social Computing*. IEEE Computer Society, 128–133.

[19] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. ACL, 1001–1012.

[20] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. ACL, 74–81.

[21] Suman Maity, Anshit Chaudhary, Shraman Kumar, Animesh Mukherjee, Chaitanya Sarda, Abhijeet Patil, and Akash Mondal. 2016. 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)*. ACM, 341–344.

[22] Nepal-quake-wiki 2015. 2015 Nepal Earthquake—Wikipedia. Retrieved April 2015 from http://en.wikipedia.org/wiki/2015_Nepal_earthquake.

[23] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining—What can NLP do in a disaster—. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*. AFNLP, 965–973.

[24] Minh-Tien Nguyen, Asanobu Kitamoto, and Tri-Thanh Nguyen. 2015. TSum4act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction. In *Proceedings of the 19th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'15)*. Springer International Publishing, 64–75.

[25] Andrei Olariu. 2014. Efficient online summarization of microblogging streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*. ACL, 236–240.

[26] 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 O'Brien. 2014. Real-time detection, tracking, and monitoring of automatically discovered events in social media. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*. ACL, 37–42.

[27] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford Infolab. http://ilpubs.stanford.edu:8090/422/.

[28] Daraksha Parveen and Michael Strube. 2014. Multi-document summarization using bipartite graphs. In *Proceedings of TextGraphs Workshop on Graph-based Methods for Natural Language Processing*. ACL, 15–24.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[30] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19, 1 (2013).

[31] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. 2011. 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'11)*. ACM, 25–34.

[32] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

[33] Siva Reddy and Serge Sharoff. 2011. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proceedings of the International Workshop On Cross Lingual Information Access*. AFNLP, 11–19.

[34] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNL'11)*. ACL, 1524–1534.

[35] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. 2016. Summarizing situational tweets in crisis scenario. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media (HT'16)*. ACM, 137–147.

[36] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM'15)*. ACM, 583–592.

[37] Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Saptarshi Ghosh. 2016. Characterizing communal microblogs during disaster events. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*. IEEE, 96–99.

[38] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, 851–860.

[39] Sandyhook-wiki 2012. Sandy Hook Elementary School Shooting—Wikipedia. Retrieved December 2012 from http://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting.

[40] Nadine B. Sarter and David D. Woods. 1991. Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology* 1, 1 (1991), 45–57.

[41] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: Continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, 533–542.

[42] Simpson 2017. Szymkiewicz-Simpson Coefficient. https://en.wikipedia.org/wiki/Overlap_coefficient.

[43] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. 2011. Summarizing a document stream. In *Proceedings of 33rd European Conference on IR Research (ECIR'11)*. Springer, Berlin, 177–188.

[44] Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ujwal Gadiraju. 2013. Groundhog day: Near-duplicate detection on twitter. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 1273–1284.

[45] Chetana Thaokar and Latesh Malik. 2013. Test model for summarizing Hindi text using extraction method. In *Proceedings of 2013 IEEE Conference on Information Communication Technologies (ICoICT'13)*. IEEE, 1138–1143.

[46] TwitterAPI 2015. REST API Resources, Twitter Developers. https://dev.twitter.com/docs/api.

[47] UMBC-semantic 2015. UMBC Semantic Similarity Service. http://swoogle.umbc.edu/SimService/.

[48] Uttarakhand-flood-wiki 2013. North India Floods—Wikipedia. Retrieved June 2013 from http://en.wikipedia.org/wiki/2013_North_India_floods.

[49] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13) (Volume 1: Long Papers)*. ACL, 1619–1629.

[50] Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth M. Anderson. 2011. Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM'11)*. AAAI, 385–392.

[51] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, 1079–1088.

[52] Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13) (Vol. 2: Short Papers)*. ACL, 505–510.

[53] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2015. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering* 27, 5 (2015), 1301–1314.

[54] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems* 27, 6 (2012), 52–59.

[55] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*. ACM, 319–320.

Received November 2016; revised October 2017; accepted December 2017