# Extracting Situational Information from Microblogs during Disaster Events: a Classification-Summarization Approach

Koustav Rudra
IIT Kharagpur, India

Subham Ghosh
IIT Kharagpur, India

Niloy Ganguly
IIT Kharagpur, India

Pawan Goyal
IIT Kharagpur, India

Saptarshi Ghosh
MPI-SWS, Germany
IIEST Shibpur, India

## ABSTRACT

Microblogging sites like Twitter have become important sources of real-time information during disaster events. A significant amount of valuable *situational information* is available in these sites; however, this information is immersed among hundreds of thousands of tweets, mostly containing sentiments and opinion of the masses, that are posted during such events. To effectively utilize microblogging sites during disaster events, it is necessary to (i) extract the situational information from among the large amounts of sentiment and opinion, and (ii) summarize the situational information, to help decision-making processes when time is critical. In this paper, we develop a novel framework which first classifies tweets to extract situational information, and then summarizes the information. The proposed framework takes into consideration the typicalities pertaining to disaster events where (i) the same tweet often contains a mixture of situational and non-situational information, and (ii) certain numerical information, such as number of casualties, vary rapidly with time, and thus achieves superior performance compared to state-of-the-art tweet summarization approaches.

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval]: selection process; H.3.5 [On-line Information Services]: Web-based services

**Keywords:** Disaster events; Twitter; situational information; classification; summarization.

## 1. INTRODUCTION

In recent years, microblogging sites such as Twitter have become important sources of real-time information, especially during disaster events. Several recent research studies [1,12,17,19,26,28,30] have shown the importance of microblogging sites in enhancing situational awareness [20] during such events.

In a disaster situation, various types of information are posted by users in huge volume and at rapid rates, which include situational information, sentiment (e.g., sympathy for those affected by the disaster) and personal opinion (e.g., on the adequacy of relief operations). While different types of information have different utilities, *situational information* – information which helps the concerned authorities to gain a high-level understanding of the situation during disasters (including the actionable items [26] such as number of affected people) – is critical for the authorities to understand the situation and plan relief efforts accordingly. Hence it is important to develop automated methods to *extract microblogs / tweets which contribute to situational information* [27].[1] A related, yet different, challenge is to deal with the rapid rate at which microblogs are posted during such events – this calls for *summarization of the situational information*. Further, some of the situational information, such as the number of casualties or injured / stranded persons, changes rapidly with time, asking for special treatment. Since time is critical in a disaster situation, these tasks have to be performed in *near real-time*, so that the processed information is readily available to the authorities.

In this work, we observe that the tweets posted during disaster events have certain specific traits, which can be exploited for the above tasks. For instance, most of the important information is centered around a limited set of specific words, which we call *content words* (verbs, nouns, numerals). It is beneficial to focus on these content words while summarizing the situational tweets. Furthermore, a significant fraction of tweets posted during disasters have a mixture of situational and non-situational information within the same tweet (e.g., *'ayyo! not again! :( Blasts in Hyderabad, 7 Killed: tv reports'*). 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 summarization.

The present work proposes a novel classification-summarization framework for extracting situational information from microblog streams posted during disaster scenarios. We develop a classifier which uses 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. We

[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.


*CIKM'15,* October 19–23, 2015, Melbourne, VIC, Australia.
 ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806485.

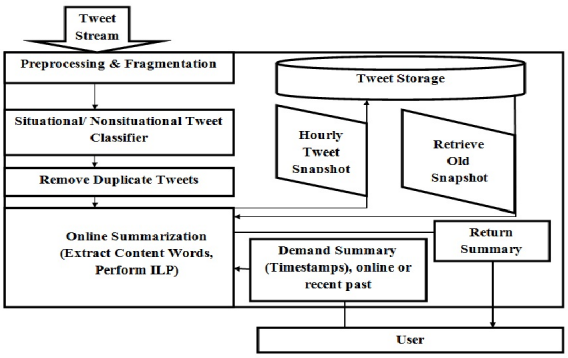



**Figure 1: Overview of the classification and summarization methodology proposed in this paper.**

then propose a novel content-word based summarization approach (COWTS) to summarize the situational tweet stream (Section 5) by optimizing the coverage of important content words in the summary, using an Integer Linear Programming (ILP) framework. 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 5.3).

Figure 1 gives an overview of our approach. First, the tweets are preprocessed and fragmented based on end-markers such as '!' and '?'. The fragmented tweets are classified to extract situational tweets. The situational tweet stream (after removing duplicate tweets) is then summarized using the content word based approach. To enable real-time summarization of long tweet streams (e.g., during disasters such as floods and typhoons, which can span several days), we maintain hourly snapshots of the tweets and the summaries generated previously, so that specific parts of the tweet stream can be summarized quickly when the user demands.

Experiments conducted over tweet streams related to four diverse disaster events show that the proposed classification model outperforms a vocabulary based approach [27] for various in-domain and cross-domain settings. The classification-summarization model also surpasses various state-of-the-art tweet summarization approaches [7, 8, 21, 31] in terms of ROUGE-1 Recall and F-scores over all the datasets (Section 6). Additionally, the proposed scheme is also deployed on tweets related to a recent (at the time of writing the paper) disaster event – the Nepal earthquake in April 2015 [11] – and it is found that the proposed scheme performs significantly better than several state-of-the-art summarization approaches.

As a final contribution, we make the tweet-ids of the tweets related to all these disaster events publicly available to the research community at **`http://cse.iitkgp.ac.in/~krudra/disaster_dataset.html`**.

## 2. RELATED WORK

Microblogging sites are serving as useful sources of situational information during disaster events [1, 12, 17, 19, 26, 28, 30]. However, for practical utility, such situational information has to be extracted from among a lot of conversational content, and summarized in near real-time. This section briefly discusses some recent studies on classification and summarization of tweets.

**Classification of tweets during disaster events:** Several studies have attempted to extract situational information during disaster events [26, 27]. Specifically, Verma *et al.* [27] observed that situational tweets are written in a more formal, objective, and impersonal linguistic style as compared to non-situational tweets, and used bag-of-words classifier models to classify tweets based on these features. However, as reported by Verma *et al.* themselves [27], this approach is heavily dependent on the vocabulary of a specific event, and does not work well in the practical cross-domain scenario where the classifier is trained on tweets of some past events and is used to classify tweets of a new disaster event. To overcome the limitations of bag-of-words model, this study uses low-level lexical and syntactic features of tweets to develop an event-independent classifier for situational and non-situational tweets, that outperforms the bag-of-words model.

**Tweet summarization:** Most of the prior research on tweet summarization has focused on summarizing sets of tweets, e.g., tweets posted during a sports event [2, 8, 22]. 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 [13, 21, 31]. Shou *et al.* [21] proposed a scheme based on first clustering similar tweets and then selecting few representative tweets from each cluster, finally ranking these according to importance via a graph-based approach (LexRank) [4]. Olariu *et al.* [13] proposed a graph-based abstractive summarization scheme where bigrams extracted from the tweets are considered as the graph-nodes. Osborne *et al.* [14] proposed a real event tracking system using greedy summarization. Along with standard summarization approaches, a few recent studies [7] have also focused specifically on summarization of tweets posted during disaster events. The studies in the TREC temporal summarization track [24] also attempt to summarize information related to events such as disasters, but the focus is on summarization of sentences rather than tweets (which are likely to be written more informally).

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

## 3. DATASET

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

### 3.1 Disaster events

We considered tweets posted during the following disaster events –
(i) **HDBlast** – two bomb blasts in the city of Hyderabad, India
(ii) **SHShoot** – an assailant killed 20 children and 6 adults at the Sandy Hook elementary school in Connecticut, USA
(iii) **UFlood** – devastating floods and landslides in the Uttaranchal state of India, and
(iv) **THagupit** – a strong cyclone code-named Typhoon Hagupit hit Philippines.
Note that the events are widely varied, including both man-made and natural disasters occurring in various regions of the world.

Hence, the vocabulary / linguistic style in the tweets can be expected to be diverse as well.

We collected relevant tweets posted during each event through the Twitter API [25] using keyword matching. For example, the keywords 'Hyderabad', 'bomb' and 'blast' were used to identify tweets related to the HDBlast event, while the keywords 'Sandyhook' and 'shooting' were used to collect tweets related to the SHShoot event. For each event, we selected the first 5,000 English tweets in chronological order. A tweet was considered to be in English if at least half of the words in the tweet (after removing @mentions and URLs) appeared in a standard English dictionary (similar to the approach in [6]). We make the tweet-ids of these tweets publicly available to the research community at `http://cse.iitkgp.ac.in/~krudra/disaster_dataset.html`.

## 3.2 Types of tweets posted during disasters

As stated earlier, tweets posted during a disaster event include both tweets contributing to situational awareness, and non-situational tweets. We employed human volunteers to observe the different categories of situational and non-situational tweets, and to annotate the tweets (details in Section 4). The different categories of tweets observed (which agrees with prior works [17]) are as follows. Some example tweets of each category are shown in Table 1.

**Situational tweets:** Tweets which contain situational information are primarily of the following two types: (i) *Status updates* – updates such as the number of casualties, and the current situation in various regions affected by the disaster, and (ii) *Helping relief operations* – information that can immediately help relief operations, e.g., phone numbers of nearby hospitals.

**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 incidents can be prevented in future, (ii) *Event analysis* – post-analysis of how and why the disaster occurred, findings from police investigation in case of man-made disasters, and (iii) *Charities* – 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 information from 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 used human annotators to obtain this gold standard (details below). During annotation, it was observed that *a significant number of tweets posted during disaster events contain a mixture of situational and non-situational information*. Table 2 shows some examples of such tweets. This observation motivated us to preprocess the tweets in order to identify the different fragments, and then process the fragments separately for classification and summarization steps. This preprocessing stage is described next.

## 4.1 Preprocessing of Tweets

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

(1) We use the Twitter-specific part-of-speech (POS) tagger [15] 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, and so on. We ignore the Twitter-specific words (that are assigned tag 'G' by the POS tagger [15]) because they represent abbreviations, foreign words, and symbols which do not contribute to meaningful information.

(2) We apply standard preprocessing steps like case-folding and lemmatization. Further, we use a standard abbreviation dictionary to replace contracted forms (such as *ppl, abt, shld, cud*) with their expanded versions. This is necessary since tweets often contain abbreviations due to the 140-character limit on their length. We also maintain uniformity across different representations of numeric information (e.g. '7' and 'seven').

(3) Subsequently, we focus on particular end-markers (e.g., '!', '.', '?') to split a tweet into multiple fragments. In case of the '.' end-marker, we break a tweet into two consecutive fragments if both the fragments possess their own verb; this prevents splitting a tweet at unnecessary breakpoints, such as the '.' in the tweet 'Upd Msg #31, Tropical Storm - Hagupit, NW Pacific Ocean, JTWC . [url]'.

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.

## 4.2 Establishing Gold Standard

For training the classifier, we considered 1000 randomly selected tweet fragments related to each of the four events. Three human volunteers independently observed the tweet fragments, deciding whether they contribute to situational awareness.[2] Before the annotation task, the volunteers were acquainted with some examples of situational and non-situational tweets identified in prior works [27, 28]. 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 THagupit events respectively. From each of these four datasets, we next selected an equal number of tweet-fragments that were judged non-situational, in order to construct balanced training sets for the classifier.

## 4.3 Classification features and performance

Prior research [27] 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 low-level lexical and syntactic features, as listed in Table 3, 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 text, such as subjective words, modal verbs, queries and intensifiers.

We use a Support Vector Machine (SVM) classifier – specifically, the LIBSVM package [3] with the default RBF kernel – to classify the fragmented tweets into two classes based on the features described in Table 3. We compare our classifier with a standard Bag-of-Words (BOW) model similar to that in [27], where the same SVM classifier is used considering as features – the frequency of every distinct unigram and bigram, POS tags, presence of strongly subjective words, and presence of personal pronouns.

We compare the performance of the two feature-sets (using the same classifier) under two scenarios — (i) *in-domain classification*,

[2]All volunteers are regular users of Twitter, have a good knowledge of the English language, and none of them is an author of this paper.

**Table 1: 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)** | | |
| Status updates | THagupit | 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] |
| Help relief operations | UFlood | call bsnl numbers 1503, 09412024365 to find out last active location of bsnl mobiles of missing persons in uttarakhand |
| | 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 loosing all faith in humanity. |
| | THagupit | 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 $2 to the school and victims |
| | THagupit | 1$ usd for a cause-super-typhoon hagupit, i'm raising money for eye care global fund, click to donate [url] |

**Table 2: Examples of tweets containing multiple fragments, some of which convey situational information while the other fragments are conversational in nature (in blue text).**

| |
|---|
| 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! |

where the classifier is trained and tested with the tweets of the same event using 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 4 shows the accuracies of the classifier using bag-of-words model (BOW) and the proposed features (PRO) on the fragmented tweets.

**In-domain classification**: BOW model performs well in the case of in-domain classification (diagonal entries in Table 4) due to the uniform vocabulary used during a particular event. However, the proposed features significantly outperform the BOW model. The result is specially significant since it shows that good classification can be achieved even without considering the event-specific words.

**Cross-domain classification**: The non-diagonal entries of Table 4 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 classification model performs much better than the BOW model in such scenarios, since it is independent of the vocabulary of specific events.

**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 5. Comparing the classification accuracies in Table 4 (on preprocessed and fragmented tweets) and Table 5 (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 also helps in improving information coverage during the summarization process.

Thus the proposed classification scheme based on lexical and syntactic features performs significantly better than word-based classifiers [27] under various experimental settings. However, since the best achieved classification accuracy is still around 80%, a question naturally arises as to whether the 20% mis-classification would substantially impact the subsequent summarization step. We shall discuss the effect of mis-classification on summarization in Section 6.

## 4.4 Applying classifier on future disaster events

The good cross-domain performance of the proposed classification scheme (as stated above) implies that the selected low-level 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 patterns, we expect that the accuracy of the classifier will not vary significantly based on the size and diversity of training set (e.g., if multiple past disasters of various types are used to train the classifier).

To demonstrate this, we performed another set of experiments taking THagupit (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 combined the remaining two / three events for training. The classifier achieved accuracy values of 79.24%, 79.47%, 78.97% and 80.46% 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 of 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 one or more past disaster events, and then deployed to classify tweets posted during future events.

# 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) nouns (e.g., names of places, important context words like people, hospital etc.),

**Table 3: Lexical and syntactic features used to classify between situational and non-situational tweets.**

| Feature | Explanation |
|---|---|
| Fraction of subjective words | Fraction of words listed as strongly subjective in a subjectivity lexicon for tweets [29]. Expected to be higher in non-situational tweets. |
| Count of personal pronouns | Number 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. |
| Count 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?'). |
| Count 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'). |
| Count of wh-words | Number of words such as 'why', 'when', etc. Expected to be higher in non-situational tweets containing queries of people, e.g., '*Why* don't you submit your coalgate scam money to disaster'. |
| Count of intensifiers | Number of frequently used intensifiers [18], more used in non-situational tweets to boost sentiment, e.g., 'My heart is *too* sad', 'Hyderabad blasts are *so* saddening'. |
| Count of non-situational words | We identify a set of words which appear *only* in non-situational tweets across all events, such as 'donate', 'condemn'. Then we count the number of such event-independent non-situational keywords. |

**Table 4: Classification accuracies of SVM on tweet-fragments, using (i) bag-of-words features (BOW), (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 | | THagupit | |
| | BOW | PRO | BOW | PRO | BOW | PRO | BOW | PRO |
| HDBlast | 68.509% | **78.245%** | 57.682% | **73.540%** | 56.845% | **83.162%** | 52.515% | **77.257%** |
| UFlood | 56.704% | **75.600%** | 61.928% | **76.142%** | 55.715% | **78.111%** | 53.485% | **78.476%** |
| SHShoot | 54.385% | **77.163%** | 57.139% | **74.428%** | 65.162% | **86.458%** | 55.897% | **78.697%** |
| THagupit | 51.052% | **76.923%** | 52.862% | **73.667%** | 54.361% | **84.574%** | 67.549% | **81.898%** |

and (iii) main verbs (e.g., 'killed', 'injured', 'stranded'). We refer to these terms as *content words*. This section describes our proposed methodology, which we call COWTS (COntent Word-based Tweet Summarization).

## 5.1 Need for disaster-specific summarization approach

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

To demonstrate this, we compare tweet streams posted during disaster events with those posted during three political, sports, and technology-related events; these streams were made publicly available by [21]. Figure 2 plots the variation in the number of distinct content words seen across the first 5,000 tweets in these three tweet streams, as well as the situational tweet streams posted during three disaster events. It is evident that the number of distinct content words increases very slowly in case of the disaster events. We find that this is primarily due to (i) presence of huge number of retweets or near-duplicates of few important tweets, and (ii) presence of large number of tweets giving latest updates on some specific contexts, such as the number of people killed or stranded. This leads to heavy usage of some specific content-words (primarily, verbs) – such as 'killed', 'injured' and 'stranded' – and rapidly changing numerical information in the context of these content-words.

The above observations indicate that summarizing situational information in disaster scenarios requires a different approach, as compared to prior approaches developed for other types of events.

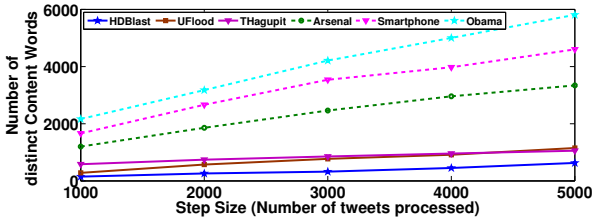



**Figure 2: Variation in the number of distinct content words with the number of tweets in chronological order, shown for disaster events (lower three curves), and other types of events (upper three curves).**

Hence, we (i) remove duplicate and near-duplicate tweets (using the techniques developed in [23]), (ii) focus on the content words during summarization (as described in Section 5.2), and (iii) adopt specific strategies for the heavily-repeated content words associated with frequently changing numerical information (described in Section 5.3).

## 5.2 Content word based summarization

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

**Table 5: Classification accuracies of SVM on raw tweets, using (i) bag-of-words features (BOW), (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 | | THagupit | |
| | BOW | PRO | BOW | PRO | BOW | PRO | BOW | PRO |
| HDBlast | 62.308% | **75.862%** | 50% | **67.254%** | 49.936% | **75.297%** | 50.110% | **76.051%** |
| UFlood | 50.063% | **70.498%** | 58.969% | **68.513%** | 50% | **77.976%** | 49.963% | **74.258%** |
| SHShoot | 48.375% | **71.295%** | 50% | **68.010%** | 64.369% | **78.273%** | 50.063% | **75.728%** |
| THagupit | 50% | **70.846%** | 49.963% | **68.010%** | 50.110% | **78.273%** | 56.847% | **81.717%** |

**Table 6: Notations used in the summarization technique**

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

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

We use an Integer Linear Programming (ILP)-based technique [16] to optimize the coverage of the content words. Table 6 states the notations used. The summarization is achieved by optimizing the following ILP objective function:

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

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \quad (2)$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m] \quad (3)$$

$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n] \quad (4)$$

where the symbols are as explained in Table 6. The objective function considers both the number of tweets included in the summary (through the $x_i$ variables) as well as the number of important content-words (through the $y_j$ variables) included. The constraint in Eqn. 2 ensures that the total number of words contained in the tweets that get included in the summary is at most the desired length $L$ (user-specified) while the constraint in Eqn. 3 ensures that if the content word $j$ is selected to be included in the summary, i.e., if $y_j = 1$, then at least one tweet in which this content word is present is selected. Similarly, the constraint in Eqn. 4 ensures that if a particular tweet $i$ is selected to be included in the summary, i.e., if $x_i = 1$, then the content words in that tweet are also selected.

We use GUROBI Optimizer [5] 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.

**Table 7: 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 |

## 5.3 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 7 shows how, during the HDBlast event, the reported number of victims / injured persons changed during a period of only seven minutes. Since such information is important and time-varying, we attempt to process such actionable information separately from summarizing the rest of the information. To our knowledge, none of the prior works on processing tweet streams during disaster events have attempted to deal with such rapidly changing (or even conflicting) information.[3]

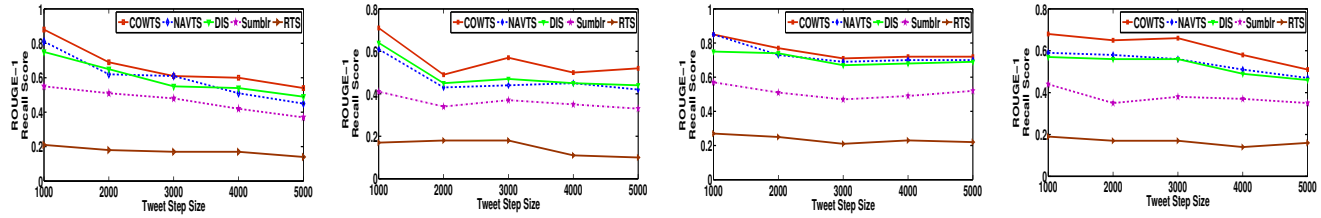
Specifically, we consider particular disaster-specific key verbs like 'kill', 'die', 'injure', 'strand', etc., and report the different numerical values attached to them, coupled with the number of tweets reporting that number. For instance, considering the tweets in Table 7, the information forwarded would be: 'seven people killed' is supported by two tweets, while 'ten killed' and 'fifteen killed' is supported by one tweet each.

**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 *'PM visits blasts sites in hyderabad, three days after two powerful bombs killed'* is *not* related with the verb *'killed'*, as opposed to the number 'seven' in the tweet *'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. We used the POS tagger and dependency parser for tweets [9] to capture this information. If a numeral is directly associated with a main verb (i.e., if an edge exists between numeral and the verb in the dependency

[3]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.

**Table 8: Comparison of ROUGE-1 F-scores (with classification, twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard rouge stemming(-m) and stopwords(-s) option) for COWTS (the proposed methodology) and the four baseline methods (RTS, NAVTS, DIS, and Sumblr) on the same situational tweet stream, at breakpoints 1K, 2K, 3K, 4K and 5K tweets.**

| Step size | ROUGE-1 F-score | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HDBlast | | | | | UFlood | | | | | SHShoot | | | | | THagupit | | | | |
| | COWTS | RTS | NAVTS | DIS | Sumblr | COWTS | RTS | NAVTS | DIS | Sumblr | COWTS | RTS | NAVTS | DIS | Sumblr | COWTS | RTS | NAVTS | DIS | Sumblr |
| 0–1000 | **0.88** | 0.21 | 0.81 | 0.75 | 0.55 | **0.71** | 0.17 | 0.61 | 0.64 | 0.41 | **0.85** | 0.27 | **0.85** | 0.75 | 0.57 | **0.68** | 0.19 | 0.59 | 0.57 | 0.44 |
| 0–2000 | **0.69** | 0.18 | 0.62 | 0.65 | 0.51 | **0.49** | 0.18 | 0.43 | 0.45 | 0.34 | **0.77** | 0.25 | 0.73 | 0.74 | 0.51 | **0.65** | 0.17 | 0.58 | 0.56 | 0.35 |
| 0–3000 | **0.61** | 0.17 | **0.61** | 0.55 | 0.48 | **0.57** | 0.18 | 0.44 | 0.47 | 0.37 | **0.71** | 0.21 | 0.69 | 0.67 | 0.47 | **0.66** | 0.17 | 0.56 | 0.56 | 0.38 |
| 0–4000 | **0.60** | 0.17 | 0.51 | 0.54 | 0.42 | **0.50** | 0.11 | 0.45 | 0.45 | 0.35 | **0.72** | 0.23 | 0.70 | 0.67 | 0.49 | **0.58** | 0.14 | 0.51 | 0.49 | 0.37 |
| 0–5000 | **0.54** | 0.14 | 0.45 | 0.49 | 0.37 | **0.52** | 0.10 | 0.42 | 0.44 | 0.33 | **0.72** | 0.22 | 0.70 | 0.68 | 0.52 | **0.51** | 0.16 | 0.47 | 0.46 | 0.35 |




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

**Figure 3: ROUGE-1 recall scores of the summaries of different events, generated by the proposed methodology (COWTS) and the four baseline methods, at breakpoints 1K, 2K, 3K, 4K and 5K tweets.**

tree), we associate that numeral with the verb (e.g., 'seven' with 'killed' in *'seven killed in hyderabad blast'*). The list of living-entity objects for disaster specific verbs was pruned manually from the exhaustive list obtained from Google syntactic $n$-grams (details omitted for brevity).[4] The performance of our methodology is discussed in the next section.

# 6. EXPERIMENTAL RESULTS

This section compares the performance of the proposed framework 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 Experiment settings: baselines & metrics

We considered the four disaster events described in Section 3 for the experiments. For each dataset, we considered the first 5000 tweet fragments in chronological order, extracted situational tweet-fragments using our classifier, and passed the situational tweets to the summarization modules. We considered five breakpoints at 1K, 2K, 3K, 4K and 5K tweets, i.e., the summaries were demanded at the corresponding time-instants.

**Establishing gold standard summaries:** At each of the breakpoints, three human volunteers (same as those involved in the classification stage) individually prepared summaries of length 30 tweets from the situational tweets. To prepare the final gold standard summary at a certain breakpoint, we first chose those tweets which were included in the individual summaries of all the volunteers, followed by those tweets which were included by the majority of the volunteers. Thus, we create a single gold-standard summary containing 30 tweets for each breakpoint, for each dataset.

**Baseline approaches:** We compare the performance of our proposed summarization scheme with that of four prior approaches, as described below. These include both generic tweet summarization approaches and disaster-specific approaches.

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

(i) **Sumblr:** the online tweet summarization approach by Shou *et al.* [21] is taken as the first baseline 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.
(ii) **RTS:** the Real-Time Summarization approach by Zubiaga *et al.* [31]. We consider all words after removing hashtags, URLs, and Twitter-specific words, compute their weights, and prepare the final summary based on the ranking methodology of [31].
(iii) **DIS:** the methodology proposed by Kedzie *et al.* [7], meant for summarizing news articles posted during disaster events. In our experiment, we apply their technique over the processed tweet streams. DIS is a semi-supervised method, requiring some prior knowledge of what to include in the summary; for this purpose, we consider some of the tweets that were selected in multiple summaries generated by human volunteers (as described above).
(iv) **NAVTS:** Since COWTS considers nouns, numerals and main verbs as content words, a question arises as to whether this choice of content words is prudent. To verify this, we devise a competing baseline where noun, verbs and adjectives are taken as content words; these parts of speech were found to be important for tweet summarization (not online) in a prior study by Khan *et al.* [8].
We applied COWTS and all the above baseline methods on the same situational tweet stream (obtained after classification), and retrieved summaries of the same length, i.e. the number of words present in the 30 tweets of the gold standard summary for a certain breakpoint (described earlier). *To maintain fairness, the same situational tweet stream (after classification) was input to all the summarization approaches.*

**Evaluation metrics:** We used the standard ROUGE [10] metric for evaluating the quality of the summaries generated. Due to the informal nature of tweets, we actually considered the *recall and F-score* of the ROUGE-1 variant.

## 6.2 Performance Comparison

Table 8 and Figure 3 give the ROUGE-1 F-scores and recall for the five algorithms for the four datasets, at the various breakpoints respectively. It is evident that COWTS performs significantly better

**Table 9: Summary of length 100 words, generated from the first 300 situational tweets of the THagupit dataset by (i) COWTS (proposed methodology), (ii) DIS proposed in [7]. The tweets have been case-folder to lower case.**

| Summary by COWTS | Summary by DIS |
|---|---|
| check out other emergency hotlines here, [url] | jtwc upgrades tropical depression of 22w to tropical storm 22w , hagupit , [url] |
| save and share, ndrrmc hotlines 0929 3356079 0917 5294438 0936 9108694 0929 3356077 0906 2042096 | latest, ts hagupit has to intensify as a severe tropical storm in 24 hours & in 36 hours it will turn as typhoon |
| wp tropical storm hagupit advisory, 4, 45 kt/52 mph winds, 5.9 n 149.1 e, moving, wnw at 16 kt/18 mph tropics | Typhoon hagupit will bring heavy rain, destructive winds and storm surges close to the eye local warnings [url] |
| pagasa forecast models show either to make landfall or veer toward japan | class at all levels declared suspended by mayor dominador agahan of almeria on friday , december 5 |
| jtwc sees becoming a super typhoon in 48 hours still split on track [url] | the hawaii-based joint typhoon warning center of the us navy on wednesday said hagupit will become a super typhoon in 48 hours |
| ndrrmc, dswd has 100,000 prepared family food packs regional offices have 30,000 in case of immediate response | now a typhoon 1,670 km east of mindanao and closing [url] |
| ndrrmc and pagasa released a list of areas deemed critical as typhoon approaches the country [url] | ruby, international name hagupit , intensifies into a tropical storm according to the us navy's joint typhoon [url] |
| as of december 3, 2014 at 5:04 pm class suspensions have been announced [url] | typhoon hagupit is a 100 knot , 115 mph , storm better hope gfs is right ecmwf could create devastating flooding [url] |
| if hagupit continues under projected track, an est 4.5 million people within the 65 km radius may be affected in 14 provinces in 4 regions | latest nasa satellite tracker typhoon hagupit curve north near philippine [url] |

than all the baseline approaches. For instance, mean scores indicate an average improvement of more than 60% in terms of F-scores over Sumblr [21] which is a general-purpose (i.e., not disaster-specific) summarization scheme. The proposed methodology also performs better than the disaster-specific summarization technique DIS [7] in all cases – on an average, we obtain improvement of 20% for F-scores and 12% for recall over DIS. Further, the higher F-scores for COWTS than those for NAVTS indicate that our selected content words lead to better summarization. Figure 3 shows that the trend holds 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 9 shows summaries generated by COWTS and DIS (both disaster-specific methodologies) from the same tweet stream – the first 300 situational tweets posted during the THagupit event. The two summaries are quite distinct, with no tweet in common. We find that the summary returned by COWTS is more informative, and contains crucial information about hotline numbers, food packs, critical areas and information necessary for volunteers. On the other hand, the summary returned by DIS mostly contains the same information (about the nature or intensity of the storm) expressed in various ways.

**Time taken for summarization:** Since time is critical during disaster events, it is important that the summaries are generated in real-time. Hence, we analyze the execution times of the various techniques. At the breakpoints of 1K, 2K, 3K, 4K and 5K tweets, the proposed COWTS method takes 5.953, 8.084, 10.295, 12.627 and 15.135 seconds on average (over the four datasets) respectively to generate summaries. The time taken increases sub-linearly with the number of tweets and is comparable to those taken by the RTS and NAVTS baselines (on the same situational tweet streams), and significantly better than those taken by Sumblr and DIS. Specifically, DIS requires more time due to computation of large similarity matrices and execution of affinity propagation algorithm, whereas Sumblr requires large time due to the LexRank graph generation.

**Benefit of classification before summarization:** We verified that separating out situational tweets from non-situational ones significantly improves the quality of summaries. Considering all the four events together, the mean ROUGE F-score at breakpoint 1000 for COWTS was 0.61 *without* prior classification (i.e., when all tweets were input to the summarizer) as compared to 0.78 after classification. Table 10 gives the mean F-score of COWTS on classified and unclassified tweets, averaged over all the four events.

**Effect of misclassification on summary recall:** As stated in Section 4, the proposed classifier achieved around 80% accuracy in classifying between situational and non-situational tweets. We now investigate how the 20% error in classification affects the subsequent summarization of situational information.

Evidently, errors where a situational tweet is misclassified as non-situational are far more critical since they may impact the recall of the subsequent summarization step. We find that out of all classification errors, such errors account for only 8.09%, 10.94%, 7.25% and 9.69% for the four datasets respectively (in the order stated in Table 4). Thus, very few situational tweets are actually left out from the summarization process due to mis-classification.

We further checked what fraction of content-words are really missed out due to misclassification. Across all the four datasets, more than 84% of the content-words present in the *mis-classified tweets* are also covered by the correctly classified situational tweets; this implies that only a small fraction of the content-words are missed in the stream sent for summarization.

**Effect of choice of content words:** Choosing what type of words to focus on is important for achieving good summarization of tweet streams, as also observed in [8]. As stated in Section 5, we considered three types of content words – numerals, nouns, and verbs. From the comparison between COWTS and NAVTS, it has already been established that our choice of content words achieves better summarization for tweets posted during disaster events, than the information words proposed in [8].

We now analyze whether all the three chosen types of content words are effective for summarization. For this, we analyze the quality of the summaries generated in the *absence* of one of these types of content words. Figure 4 compares the F-scores (averaged over all four datasets) considering all three types of content words, with those obtained considering any two types of content words. It is clear that all three types of content words are important for the summarization quality, numerals and nouns being the most important ones (since the numeral-noun combination outperforms the

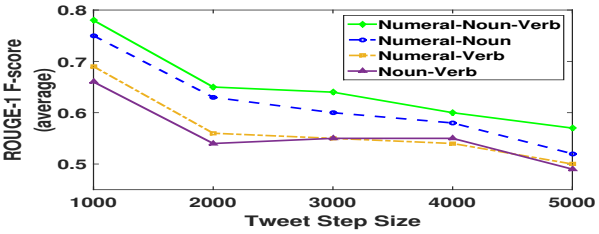
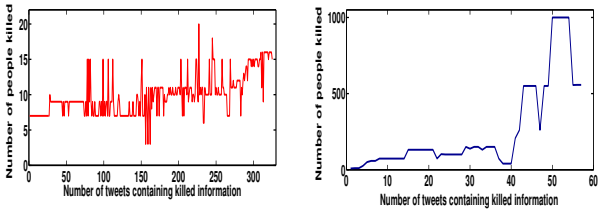



**Figure 4: Effect of individual types of content words on summary**




**(a) HDBlast** **(b) UFlood**

**Figure 5: Variation in the reported number of people killed, during two disaster events. The $x$-axis represents the sequence of tweets which contain such information.**

other 2-combinations). Side by side, as a sanity check, we have also included adjectives among the content words and run COWTS; we observed that 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).

**Handling frequently changing numerals:** Figure 5 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. 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 5.3). Specifically, we check what fraction of such numerical information could be correctly associated with the corresponding key verb. We compared the accuracy of our algorithm with a simple baseline algorithm where numerals occurring within a window of 3 words on either side of the verb were 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* – this shows the effectiveness of our strategy in extracting frequently changing numerical information.

## 6.3 Application on future disaster events

We envisage that the classification-summarization framework developed in the present work 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 final section, we demonstrate the utility of the framework by deploying it on tweets posted during a more recent disaster event – the earthquake in Nepal in April 2015 [11].

We collected related tweets using keyword matching, and then considered the first 1,000 matching tweets in chronological order. We trained our classifier model on the HDBlast dataset, then used the classifier to extract situational tweets, and then summarized the situational tweet stream. We also used the baseline methods NAVTS, Sumblr, and RTS on the same tweet stream in similar settings; however, the DIS method was not used since it is a semi-supervised method, requiring some prior knowledge about what to include in the summary.

**Table 10: ROUGE-1 F-score of COWTS on classified and unclassified tweets, averaged across four events.**

| Step size | ROUGE-1 F-score | |
|---|---|---|
| | Classified | Unclassified |
| 1000 | 0.78 | 0.61 |
| 2000 | 0.65 | 0.52 |
| 3000 | 0.64 | 0.48 |
| 4000 | 0.60 | 0.43 |
| 5000 | 0.57 | 0.45 |

**Table 11: Ranking by 5 volunteers for the summaries generated by various methods on the Nepal earthquake dataset.**

| Evaluator | Ranking | | | |
|---|---|---|---|---|
| | COWTS | NAVTS | Sumblr | RTS |
| 1 | 1 | 2 | 3 | 4 |
| 2 | 1 | 2 | 2 | 3 |
| 3 | 1 | 2 | 3 | 4 |
| 4 | 1 | 2 | 3 | 4 |
| 5 | 1 | 2 | 3 | 4 |

Since for this dataset, we do not have any ground truth summary, we performed a manual evaluation of the summaries generated by all the different methods. Five human volunteers were asked to rank the summaries (anonymized, i.e., the volunteers were not told which summary was generated by which methodology) based on their informativeness. We tabulate the rankings given by the volunteers in Table 11. It is evident that all the volunteers ranked the summary produced by COWTS as the most informative.

Further, in order to test the scalability of COWTS, we considered the first 20,000 tweets related to the event in chronological order, and then summarized the situational tweet stream at the two breakpoints – 10,000 and 20,000 tweets. COWTS took 32.130 and 47.412 seconds respectively to summarize the tweets at the two breakpoints, and all the five human evaluators expressed satisfaction at the content of the summary. This experiment shows that COWTS is able to summarize tweet streams posted during new disaster events satisfactorily, and in near real-time.

## 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 COWTS. 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 and RTS [21, 31] do not discriminate among different types of POS tags. Additionally, RTS [31] suffers from potential redundancies – this methodology ranks tweets and selects the top-ranking ones for the summary, which may lead to redundancy if most of the top-ranking tweets have similar content. Sumblr [21] maintains clusters of related information and finally chooses one top scoring tweet from each cluster. If the desired summary length is not equal to the number of clusters, then it needs to be decided as to which clusters are important and should be selected for preparing the final summary. Similar types of tweet selection problem also arises in case of DIS [7], when we have large number of exemplar tweets. To resolve such issues, focusing on particular POS-tags and ILP-based technique (as used in COWTS) proves to be very handy.

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

# 7. CONCLUSION

This paper 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) distinct lexical and syntactic features present in tweets can be used to separate out situational and non-situational tweets, which leads to significantly better summarization, (iii) content words are especially significant for summarization of disaster-specific tweet streams, and (iv) special arrangements are needed to deal with a small set of actionable keywords which have numerical qualifiers. We develop a domain-independent classifier which performs better than domain-dependent bag-of-words technique, and an ILP-based summarization framework that out-performs other summarization methods in summarizing the situational tweets.

We had several realizations during the course of this work. For instance, whereas some disasters are instantaneous and span short time durations (such as bomb blast, or shooting incidents), 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 (e.g., the last few hours) as well as historical summaries (e.g., the last week). Both these types of summaries can be generated by a minor modification of the underlying data structures of the present scheme. 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 hope to formalize this in more detail in our future work, which includes deploying a live system. Further, the module which we develop to handle continuous updates of the actionable numerical items shows that conflicting numbers often get posted at the same time, and a robust technique needs to be developed to differentiate between rumours and authentic information. This would be another potential future work.

As a final note, we believe that our work is significant especially in developing countries, where government-sponsored sophisticated systems to monitor situational updates in disaster scenario is largely missing.

**Acknowledgement:** This research was 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 was supported by a fellowship from Tata Consultancy Services, and S. Ghosh was supported by a fellowship from the Alexander von Humboldt Foundation. K. Rudra would also like to thank Google and Flipkart for their travel support.

# 8. REFERENCES

[1] M. A. Cameron, R. Power, B. Robinson, and J. Yin. Emergency Situation Awareness from Twitter for Crisis Management. In *Proc. Conference on World Wide Web (WWW)*, 2012.

[2] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proc. AAAI ICWSM*, 2011.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[4] G. Erkan and D. R. Radev. LexRank:Graph-based lexical centrality as salience in text summarization. *Artificial Intelligence Research*, 22:457–479, 2004.

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

[6] A. Hannak, E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald. Tweetin' in the Rain: Exploring societal-scale effects of weather on mood. In *Proc. AAAI ICWSM*, 2012.

[7] C. Kedzie, K. McKeown, and F. Diaz. Summarizing Disasters Over Time. In *Proc. Bloomberg Workshop on Social Good (with SIGKDD)*, 2014.

[8] M. A. H. Khan, D. Bollegala, G. Liu, and K. Sezaki. Multi-Tweet Summarization of Real-Time Events. In *Proc. IEEE Socialcom*, 2013.

[9] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith. A Dependency Parser for Tweets. In *Proc. EMNLP*, 2014.

[10] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out (with ACL)*, 2004.

[11] 2015 Nepal earthquake – Wikipedia, April 2015. http://en.wikipedia.org/wiki/2015_Nepal_earthquake.

[12] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami. Safety information mining – what can NLP do in a disaster. In *Proc. IJCNLP*, 2011.

[13] A. Olariu. Efficient online summarization of microblogging streams. In *Proc. EACL*, 2014.

[14] M. Osborne et al. Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media. In *Proc. ACL*, 2014.

[15] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. NAACL-HLT*, 2013.

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

[17] Y. Qu, C. Huang, P. Zhang, and J. Zhang. Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In *Proc. ACM CSCW*, 2011.

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

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. World Wide Web Conference (WWW)*, 2010.

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

[21] L. Shou, Z. Wang, K. Chen, and G. Chen. Sumblr: Continuous summarization of evolving tweet streams. In *Proc. ACM SIGIR*, 2013.

[22] H. Takamura, H. Yokono, and M. Okumura. Summarizing a document stream. In *Proc. ECIR*, 2011.

[23] K. Tao, F. Abel, C. Hauff, G.-J. Houben, and U. Gadiraju. Groundhog Day: Near-duplicate Detection on Twitter. In *Proc. World Wide Web Conference (WWW)*, 2013.

[24] TREC Temporal Summarization, 2015. http://www.trec-ts.org/.

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

[26] I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. D. Saeger. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proc. ACL*, 2013.

[27] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *Proc. AAAI ICWSM*, 2011.

[28] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proc. ACM SIGCHI*, 2010.

[29] S. Volkova, T. Wilson, and D. Yarowsky. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. In *Proc. ACL*, 2013.

[30] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using Social Media to Enhance Emergency Situation Awareness. *IEEE Intelligent Systems*, 27(6):52–59, 2012.

[31] A. Zubiaga, D. Spina, E. Amigo, and J. Gonzalo. Towards Real-Time Summarization of Scheduled Events from Twitter Streams. In *Proc. ACM Hypertext*, 2012.