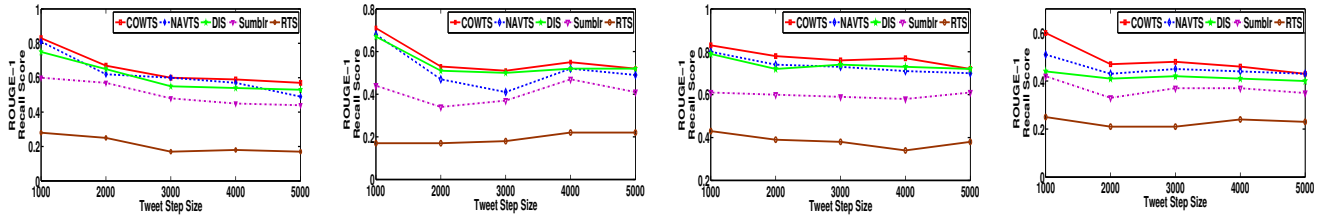
Table 8: Comparison of ROUGE-1 F-scores (with classification, twitter specific tags, emoticons, hashtags, mentions, urls, removed and standard rouge stemming(-m) and stopwords(-s) option) for COWTS (the proposed methodology) and the four baseline methods (RTS, NAVTS, DIS, and Sumblr) on the same situational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

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	**0.83**	0.28	0.81	0.75	0.60	**0.71**	0.17	0.68	0.67	0.44	**0.83**	0.43	0.80	0.79	0.61	**0.60**	0.25	0.51	0.48	0.42
0–2000	**0.67**	0.25	0.62	0.65	0.57	**0.53**	0.17	0.47	0.51	0.34	**0.78**	0.39	0.74	0.72	0.60	**0.47**	0.21	0.43	0.41	0.33
0–3000	**0.60**	0.17	**0.60**	0.55	0.48	**0.51**	0.18	0.41	0.50	0.37	**0.76**	0.38	0.73	0.74	0.59	**0.48**	0.21	0.45	0.42	0.37
0–4000	**0.59**	0.18	0.57	0.54	0.45	**0.55**	0.22	0.52	0.52	0.47	**0.77**	0.34	0.71	0.73	0.58	**0.46**	0.24	0.44	0.41	0.37
0–5000	**0.57**	0.17	0.49	0.53	0.44	**0.52**	0.22	0.49	0.52	0.41	**0.72**	0.38	0.70	0.72	0.61	**0.43**	0.23	0.43	0.40	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 technique (COWTS) and the four baseline techniques, at breakpoints 1K, 2K, 3K, 4K and 5K tweets.

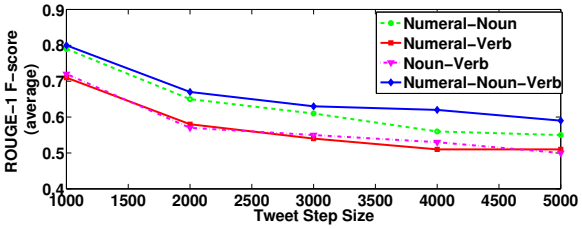
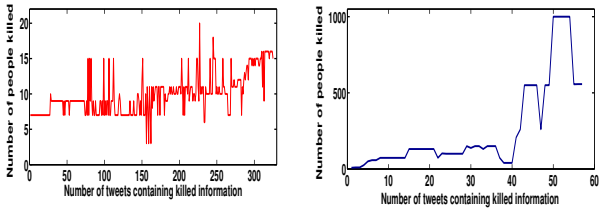


Figure 4: Effect of individual types of content words on summary



(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.

ered 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 [9]. 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) obtained 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 other 2-combinations). Side by side, as a sanity check, we have also include varbs and adjectives in the content word list and run COWTS - 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 attached 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. For example, it is always not immediately clear whether a tweet tweeting about increase in number of death is anamolous or real as several tweets appearing *after* this tweet report lower number of deaths. Thus, we realized that early, correct and automatic detection of the amount of casuality is a more involved problem which needs to be tackled in a future work.

We now evaluate the performance of our algorithm to relate 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