ABSTRACT

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

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

These large volume of situational tweet streams are scattered across various humanitarian categories like 'infrastructure damage', 'missing or found people' etc. We also observe that each of these humanitarian categories contain information about various small scale sub-events like 'airport shut', 'building collapse' etc. We develop a noun-verb pair based method to detect sub-events which are more explainable compared to random collection of words. It is observed that different stakeholders are looking for different kinds of summaries like overall high level summary, humanitarian category based summary etc during disaster. To satisfy their needs, we develop an ILP-based generic summarization technique which combines information about sub-events, content words, and humanitarian categories to generate summaries from various perspectives.

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

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