TAQE: Tweet Retrieval Based Infrastructure Damage Assessment During Disasters

Preprint · December 2019

CITATIONS
0

READS
247

5 authors, including:

Shalini Priya
Indian Institute of Technology Patna
7 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Manish Bhanu
Indian Institute of Technology Patna
8 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Sourav Dandapat
Indian Institute of Technology Patna
14 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Kripabandhu Ghosh
Indian Statistical Institute
53 PUBLICATIONS **360** CITATIONS
SEE PROFILE

TAQE: Tweet Retrieval Based Infrastructure Damage Assessment During Disasters

Shalini Priya
Indian Institute of Technology Patna, India
shalini.pcs16@iitp.ac.in

Manish Bhanu
Indian Institute of Technology Patna, India
manish.pcs16@iitp.ac.in

Sourav Kumar Dandapat
Indian Institute of Technology Patna, India
sourav@iitp.ac.in

Kripabandhu Ghosh
TCS Research, Pune, India
kripa.ghosh@gmail.com

Joydeep Chandra
Indian Institute of Technology Patna, India
joydeep@iitp.ac.in

***Abstract*—Twitter is an active communication channel for the spreading of updated information in emergency situations. Retrieving specific information related to infrastructure damage offers the situational views to the concerned authorities, who can take necessary action to disburse help. However such usages of Twitter demand significant accuracy of the retrieved information. Previous techniques on IR have not been able to capture the semantic variations satisfactorily in the tweets, due to low content quality, vocabulary gap and consequently have failed to yield considerable performance. This has left ample scope for further improvement in this area of research. There are two major contributions of our work: (1) developing a relevant tweet retrieval framework that provides information about infrastructure damage and (2) assignment of a relative damage score to the affected regions so that the severity of the damage can be assessed. Our proposed technique involves a novel split-query based mechanism with topic aligned query expansion (TAQE) to retrieve relevant tweets that are subsequently used for measuring the infrastructure damage across different locations. We report empirical results on multiple crisis related datasets to establish the efficacy of our approach to these events at different locations. Empirical validation of our proposed approach on manually annotated ground-truth data reveals considerably better performance metrics in terms of precision, recall, bpref, and MAP over several state-of-the-art techniques.**

***Index Terms*—Tweets retrieval, Disaster events, Infrastructure Damage, Situational information**

I. Introduction

Twitter provides situational information on a wide range of social activities that are helpful during crisis events such as earthquakes, floods, cyclones, landslides, wildfire, etc. [1, 2, 3]. Efficiently monitoring of such social media contents is not only useful in assessing the extent of damages at different locations, but it also provides information about the need and availability of resources at the affected locations. However, the major problem with social media content is that they are unstructured i.e unorganized and can be non-textual or textual. The useful information often lay hidden within a large volume of irrelevant and noisy contents. Hence identifying and extracting the relevant tweets that provide critical information about damages remains a key task in early damage assessment. Several works have been directed towards identifying and categorizing informative tweets into broad range of classes like requirement, availability, rescue and recovery information and so on. However, information regarding infrastructure damage and estimating the scale of destruction at affected locations is of special importance as they can guide towards quick rescue and relief operations that can save several lives. However, most of the existing techniques (like TweetXplorer [4, 5]) that specifically retrieve situational tweets do not consider such location specificity. Furthermore, certain well-known methods use crowdsourced techniques (like AIDR [6]). These techniques are heavily data-driven and consequently consume significant time before achieving reasonable accuracy. Although certain efforts have been made towards providing abridged situational information from tweet streams [7], they are usually insufficient to determine the severity of the damage. For example, damage to a bridge can have a much higher impact if the bridge is the only connection between two different places. Hence, it is essential to build an automated tool that can not only capture the information regarding infrastructure damage at specific locations without potential human intervention but can also estimate the damage magnitude by utilizing tweets and hence provide a holistic measure of the severity of the damage.

However, there exist numerous challenges in retrieving important information out of tweets. Other than limited word length, duplication, noise (misspelled and slangs) and use of informal terms in the tweets, different representations of the same facts make the convergence of their interpretations difficult. Due to these factors, prior exact (like [8, 9]) or partial matching techniques (like [10, 11]) that rely on the choice of the keywords for query generation and expansion suffer from poor performance. Learning based techniques like [5, 12, 13] that learn to identify damage related tweets require sizable training dataset with a balanced distribution of all the damage classes. However, the crisis-informatics community lacks large volume of labeled data making the learning based techniques difficult to apply. The gaps of these systems thus motivate the need for further improvement in the retrieval mechanisms.

Further, our observations indicate major differences in the semantic patterns of the tweets reported from different locations (as can be visualized in Fig. 1) [14]. Hence, models that

(a) (b)

Fig. 1: The figures show the word clouds of the tweets reported from two different locations.

use word embeddings [5, 11] or are pre-trained on a previous disaster event [6, 15] do not perform well on a current event. Existing learning based approaches that are meant for domain adaptation often fail to handle this semantic difference of tweets across the crisis events [12]. Moreover, certain biases, like large difference in tweet volume originating from different locations, can also affect the accuracy of models used for predicting the extent of damage. For example, populated and developed locations may supposedly generate more number of intense tweets in comparison to the remote and lesser populated locations with much higher damage. During the Nepal earthquake[1], a considerably large number of initial tweets reported damages to the Indian side as compared to Nepal [16], even though Nepal was the worst hit. In this work, we address two major issues with the works mentioned above. We initially propose a novel probabilistic *Topic Aligned Query Expansion* technique (TAQE) over a set of initial query words that effectively handle the keyword diversity of the tweets to retrieve relevant infrastructure damage tweets. Unlike several existing techniques [10, 13], our proposed approach does not rely on any labeled training data. We subsequently apply certain measures on these damage tweets to identify highly damaged locations. The novelty of our proposed approach lies in the techniques used for identifying the appropriate initial keywords for query generation along with the topic alignment based keyword selection for query expansion as well as the method used for damage score assessment. The generation of initial query keywords is a semi-automatic process using LDA topic modeling. Wikipedia pages related to major earthquakes are selected for topic modeling because they cover all the topics including infrastructure damage in terms of details and consequences of the earthquake. An exact match Boolean retrieval technique [17] is initially used to retrieve tweet set (seed tweets) that contain at least one keyword from each of the topic buckets. TAQE expands the seed query set using a pseudo-relevance feedback mechanism [18], where the topically similar terms co-occurring with the seed query set are identified based on a probabilistic ranking model. The keywords with the maximum probability score are selected for the feedback. The expanded query set is used for further retrieval of tweets. Investigations on these tweets reveal that there exists a strong correlation between the sentiment score of the tweets and the seriousness of the damage around the location. We further exploit this relation to calculate the damage score at different locations. Empirical investigations on TAQE reveal a considerable improvement in the retrieved results as compared to various state-of-the-art techniques including recent deep learning based models [13]. The results indicate that the proposed approach can handle the semantic variation of the tweets considerably better than the baselines, without using any training data, unlike traditional machine learning and deep learning based approaches. We also validate the generality of the proposed approach by applying the proposed approach to an Italy earthquake and an Indonesian tsunami dataset. Further, the proposed damage score can also overcome the volume biases problem stated above in assessing the damage at a location. The relative damage scores estimated using our method for the locations show significant correlation with the actual damages as reported by government organizations (USAID) and USGS Shakemap for both Nepal and Italy earthquakes.

The organization of the paper is as follows. We next discuss the related literature in section II. A formal definition of the problem is outlined in section III. Section IV outlines some of the preliminary steps of our approach followed by the methodology in section V. Description of dataset and details of ground truth data is provided in section VI. We discuss our experimental results and analysis of the same in section VII. Finally, we conclude our study with future directions in section VIII.

II. Related Work

Several recent works have investigated the usefulness of social media during disaster situations [19, 2, 20]. Studies have explored different techniques for leveraging social media information in situational awareness including relief, warning, response, etc., in an impending crisis [21, 22]. However, these purposes can be better served by monitoring the behavioral aspect of contents posted by users during disaster [23, 24, 25], detecting or predicting the users' location from their contents posted [26, 27]. With the help of these pieces of information deploying advanced mechanisms that can identify the specific informative tweets from the incoming stream [5, 28] can be more effective. Empirical investigations suggest that during crisis people post various kinds of information and from heterogeneous sources by sharing content about natural hazards and human-induced disaster [29, 22]. The most useful and credible information generally originates from the people who have personally sensed the impact of crisis or have direct access to those information [23].

However, in such situations, location identification of the user becomes essential for the usefulness and actionability of the information. Moreover, it is essential to gather situational information to calibrate the contents usually posted [30] on social media. A major challenge to gather these informations

[1]https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake

is in geo-spotting the users posting the tweets, though many existing works claim to perform satisfactorily in this research area [31]. Techniques like personal observation filtering [32], particle filtering [33] and use of pre-loaded data from online mapping services and volunteered geographic information sources [34] have been proposed. However, only few tweets out of massive tweet sets are useful during crisis, so these works were carried for the identification of informative tweets and relevant information out of it. Studies also report that sentiment and subjectivity have a major impact on the tweets considered as informative during natural disaster [35]. For decades, a lot of research work engaging pattern matching [36], NLP and deep learning techniques [37], have focused on retrieval of important information concerning natural disasters incidents and mishaps. Purohit et al. [36] used a collection of regular expressions to automate the identification of resource requests and helps. A similar approach, with an extended set of patterns, was adopted to retrieve tweets of special categories [38]. A model trained on conditional random fields to cover specific information from tweets was proposed by Imran et al. [37]. Recent efforts have been made towards the application of neural network techniques over microblogs for information retrieval. However, these models demand a large volume of data for training. Tweets from different locations vary semantically, hence such IR approaches find it difficult to extract meaningful information in real-time from tweets stream. Authors in [6] proposed a tool that required manual annotation using crowd-sourced mechanism. These approaches incur heavy time costs and also require a good number of tweets from each location for their better result performances. Palshikar et al. [39] used a semi-supervised word based model to quickly classify tweets in an incoming stream as disaster related or non-disaster related.

Several unsupervised ranking methods have also been proposed to extract informative tweets related to needs and availability. Singla et al. [8] adopted ranking models like vector space model, BM25, etc. In a recent work [10] the authors constructed models for finding the needs and availability tweets, where query expansion was done using Rocchio expansion and Word2vec [40] similarity between the tweet vector and the query vectors. Word2vec based techniques produced more accurate results than pattern matching techniques [36, 38] because word embedding techniques were more effective in capturing the semantics of the text [10, 30] to identify the tweets relevant to infrastructure damage. In a more recent work [41], Basu et al. employed two methods to match the pertinent "need-tweets" and "availability-tweets": common noun overlap and semantic matching (as harnessed by word embedding models). While the local (trained on the current dataset) Word2vec embeddings produced the best results in the Nepal dataset, the common noun overlap based method produced better results in the Italy dataset. To the best of our knowledge, limited research has been done on classification [30] or retrieval [14] of the tweets related to infrastructure damage. In [42], the authors consider a set of textual features to train a SVM classifier for classifying the damage related tweets. Two binary Support Vector Machines are trained using word-embeddings as features. However as highlighted in [30] the complexity of these models for training as well as classification grows linearly with the number of features. Moreover the time spent to use the NLP tools used for enriching the data adds additional complexity to it. One of our recent works [14, 3] uses a split-query based approach coupled with pseudo-relevance feedback for retrieving tweets reporting on infrastructure damage. In this work, the initially retrieved tweet set are ranked and highly co-occurring terms are considered for query expansion. However, their results indicate that sufficient scope of improvement in these techniques remains so as to achieve reasonable accuracy. Our proposed work in the current paper attempts to work towards this.
Next, we discuss the objective and our proposed methodology in detail.

III. Problem outline

In this section, we provide a brief overview of the problem we look to address in the scope of the paper. We assume a stream of tweets arriving concerning a disaster event. We design our problem to follow two major objectives using these tweets that are as follows:
Objective 1: Identify the tweets that represent and provide information about infrastructure damage caused by the disaster. We also provide a query based ranking of these tweets depending upon their relevance with respect to certain queries representing infrastructure damage.
Objective 2: Utilize those relevant tweets for estimating the severity of damage at different locations.

These broader goals require addressing certain key issues like generating an appropriate initial query, handling the noise and keyword diversity in the retrieved tweets to identify the most relevant ones and subsequently use these tweets to identify keywords to expand the query set to retrieve further relevant tweets. Finally, the ranked tweets are used to derive the damage score of a location. In the next section, we describe the methodology that is used to address these tasks identified and finally to achieve our goals.

IV. Preliminary steps

In this section we describe two pre-processing steps that are essential to carry out the proposed methodology, i) location identification of the tweet, ii) obtaining an initial optimal query keyword set.
Location identification: Since identifying the disaster affected locations from the tweets is important for our approach, we use a popular technique [43] that was specifically proposed for this objective. The proposed approach showed that named entity recognizers when trained on microblog data can identify disaster locations and points of interest with considerably high accuracy. Using *Standford NER* they reported high accuracy and $F1$ score (around 0.9). We trained the Standford NER using tweets in our dataset that explicitly mention the location information. We next describe the method used to generate the initial query for the tweet retrieval.

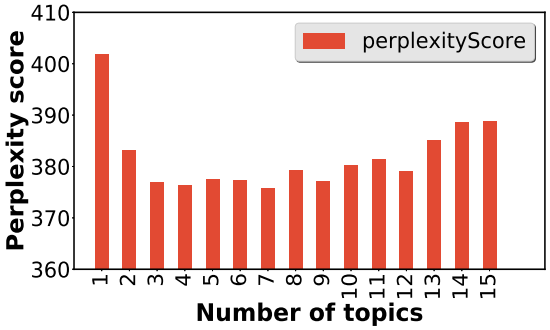


Fig. 2: Figure shows the perplexity score of different LDA model i.e with different number of topics. Least perplexity score is obtained with 7 topics.

Query Generation: To obtain our initial query keywords, we use a semi-automated approach. We collect Wikipedia pages related to major earthquakes and prepare a document with words appearing in those pages[2]. The Wikipedia page of major crisis events covers many primary aspects background of the earthquake, casualties, damages, rescue and aftermaths of the events. In other words, each Wikipedia page is a collection of all crisis-related topics that include infrastructure damage too. It's worth noting that we include Wikipedia pages of tsunami events as well while generating the query keywords for retrieving the tweets relevant to infrastructure damage of tsunami datasets. On the words collected from those documents, we apply a topic based clustering (Latent Dirichlet Allocation (LDA)) [44] to obtain the major topics and consequently the keywords in those topics. Topic models are algorithms for discovering the main themes that pervade a large collection of documents. Perplexity is the metric to evaluate the goodness-of-fit of the LDA model and lower the perplexity better the model [44]. We determined the optimal set of hyperparameters by testing the performance of our LDA models for different parameter combinations. We used the Tree of Parzen Estimators (TPE) algorithm[3] [45] to search the parameter space and minimize the perplexity score. The parameter space is defined as follows: number of topics in range(1, 16), maximum iteration in range (100, 200, 500), random state from (1,1000), learning method as "batch" or "online", batch size in the range of (64, 128) and number of words in range(5,40). The minimum perplexity score is obtained at number of topics(n) as 7 and count of words in each topic (w) as 30, hence considered as the best topic model. We show the perplexity score value for the different number of topics in Fig. 2.

We show the word cloud representation of all the keywords generated using the best LDA model in Fig. 3(a). After scrutinizing the topics manually by the authors, the topics which are found to be most relevant to our objective are used for creating an initial set of query. The word cloud of those relevant topic keywords is shown in Fig. 3(b). Using a set of human annotators, the words which are related to infrastructure or damage are only kept in the initial query set while rest are discarded in order to reduce the query set. To avoid any biases in the categorization of the keywords that may result from the knowledge of the tweets, the manual annotators are only provided with the keywords and have no idea about the tweets. After having manually reduced query set, we segregate it into two sets. Using NLTK parts-of-speech tagger, we put the words tagged as noun (NN, NNS etc) into "object wordset" while those tagged as verbs (VB, VBZ etc.) are collected in "feature wordset". After dividing the single query set into two mutually exclusive query sets, we also add the stemmed words (using porter stemmer using NLTK library[4]) of each query word to the query set itself to ensure capturing relevant tweets in their root form.

The size of the feedback term set can impact the system performance; hence we later check the performance of the proposed model for different feedback sizes, the results of which are discussed in section VII-D. The next section discusses our proposed approach.

V. Proposed Approach

In this section, we describe our proposed topic aligned query expansion technique for identifying the relevant tweets. We subsequently use these tweets to estimate the damage score across different locations identified from the tweets. However, the query expansion technique requires an initial set of tweets from which the expanded query keywords are identified. We next describe the method to extract these tweets from the initial query keywords followed by our query expansion approach.

A. Seed tweet creation

We use the initial query set generated in the previous step to retrieve a set of most relevant tweets that we term as the *seed tweet set*. This set needs to be highly relevant with respect to the desired context. This is because relevant keywords occurring in these tweets with high topical relevance would be subsequently identified as query keywords in the expanded query set. To generate the seed tweet set, we use a Boolean retrieval technique [17] to create an initial set of relevant tweets using the query keyword set. Major challenge we face is the removal of noise from the tweets to meet our objective. Noises present in these tweets like misspelled words are corrected using TextBlob[5]. From this initial tweet set, we identify those tweets as seed tweets that have at least one exact match from each of the two wordsets — object and feature. Thus a seed tweet contains at least one word from object wordset as well as one from feature wordset. The presence of keywords from both these wordsets enhances the possibility of the tweet being an infrastructure damage tweet.
The primary reason behind this exact match model can be better understood with this example tweet. For example, the tweet "#prayforNepal Earthquake destroyed the country. Let's cherish what we have. Because you never know when will

[2]https://en.wikipedia.org/wiki/Lists_of_earthquakes
[3]We used the implementation of the TPE algorithm in the hyperopt package.
[4]http://www.nltk.org/howto/stem.html
[5]http://textblob.readthedocs.io/en/dev/quickstart.html

(a) (b)

Fig. 3: Fig. 3(a) shows all the top 30 keywords of 7 topics generated using LDA and Fig. 3(b) shows the keyword set of specific topic that is considered to be related to infrastructure damage.

we lose them" contains the word *destroy* but does not reflect any infrastructure damage. Another tweet "We need to be pro-active. Building structure and material should be quake resistant and stop fiddling with nature." informs the presence of infrastructure term but cannot be considered as relevant to the query. On the other hand the tweet "Before and after photographs: Devastating earthquake ruined important monuments in Nepal." contains both object word (*ruined*) and feature word (*monuments*) and can be considered as relevant to infrastructure damage. Although there are several cases where this assumption fails, we subsequently handle them by doing certain error analysis in terms of structural refinements and are discussed next in section V-C. Thus the mathematical formulation of our model for identifying the seed tweets would be given as follows: if Q_f, Q_o and D are set of words in feature wordset, object wordset and a tweet respectively, then for the tweet to qualify as seed tweet we should have

$$Q_f \cap D \neq \phi \text{ and } Q_o \cap D \neq \phi \quad (1)$$

However, these seed tweets are not exhaustive and hence we use these tweets to identify relevant query terms that can be used to retrieve the relevant tweets which could have been missed otherwise. We use the seed tweets for query expansion as described next.

B. Topic aligned query expansion(TAQE)

Our proposed query expansion mechanism (TAQE) follows a two-stage process that initially identifies those relevant terms in the seed tweet set that most frequently co-occurs with the initial query keywords and subsequently selects a subset of these terms that are best topically aligned with respect to the query.

Co-occurrence based keyword selection: In this step, we identify the top-κ co-occurring pairs of words, occurring in the seed tweet set $\mathscr{R} = \{R_1, R_2, \ldots, R_n\}$, whose one term belongs to either of the query wordset (object or feature). We denote these words as $W = \{w_1, w_2, \ldots w_n\}$. The keyword set W contains only the additional words that are not included in the initial object and feature wordset. The value of κ is set to 20 that we determine empirically, the details of which is discussed later in section VII-D.

The problem with this step is that due to the keyword diversity in the tweets several unrelated keywords may seem to be related due to their high co-occurrence and hence also get included in the expanded query set [46]. Such inclusions are likely to lower the retrieval quality. For example, while experimenting with Nepal earthquake dataset we observed, the inclusion of few top co-occurring terms like 'rebuild', 'need', 'people' etc. adds a huge number of tweets that are irrelevant to the objective. These examples can be seen in Table I.

A count of the total false positive tweets[6] generated by adding different number of the top co-occurring terms can be seen in Table II.

The objective of the second step is to address this precise problem by considering the topic relevance of the keywords before selecting them. We briefly describe this step below.

TABLE I

Nepalis, r w/o water & electricity. Water is essential to be supplied to the **affected people** in Nepal.
Nepal's central bank announced a low-cost loan to help **rebuild homes** lost in the earthquake @2% Interest #Nepalâ€¦
Good initiative. Corporate contributing with **home** is **need** of the hour #earthquake

TABLE II

Number of top co-occurring terms (κ)	Number of False Positive Tweets
5	755
10	780
15	670
20	523
25	920
30	5900

[6]Tweets identified as relevant but are actually irrelevant according to the ground truth.

Topic alignment based keyword selection: In this step, we select a subset of the words in W based on the relevance of the tweets in which they appear, with respect to the initial query Q. If the initial query set contains m keywords given as $\{q_1, q_2, \ldots, q_m\}$, then for a word $w_k \in W$ that appears in a tweet R_i, we determine a query sensitive importance score that considers the relevance of both R_i with respect to Q as well as the word w_k with respect to R_i. We use a probabilistic model to derive the query sensitive importance score. Let $P(w_k|R_i)$ denote the probability of generating a word w_k from tweet R_i and $P(R_i|Q)$ denote the probability of generating the keywords in R_i from the given query set Q. We introduce a term scoring function, $\mathscr{F}(w_k, R, Q)$ that is represented as:

$$\mathscr{F}(w_k, R, Q) = \sum_{i=1}^{n} P(w_k|R_i)P(R_i|Q) \quad (2)$$

$$\sim \sum_{i=1}^{n} P(w_k|R_i)P(Q|R_i)P(R_i) \quad (3)$$

(By Bayes' rule)

We adopt the premise of query likelihood model [47] where the words in Q are assumed to appear independent of each other in R_i. Hence the query likelihood $P(Q|R_i)$ can be written as

$$P(Q|R_i) = \prod_{j=1}^{m} [P(q_j|R_i)]P(R_i) \quad (4)$$

Using equation (4), equation (3) can be re-written as

$$\mathscr{F}(w_k, R, Q) = \sum_{i=1}^{n} P(w_k|R_i) \prod_{j=1}^{m} [P(q_j|R_i)](P(R_i))^2 \quad (5)$$

We use the formulation (5) to rank the words $w_k \in W$ obtained in the previous step. We use Dirichlet smoothing [48], so that the term likelihood may not over-fit the data for the query words that are not present in the relevant tweet. The keywords whose ranking score is greater than a threshold (say τ) are finally considered as possible keywords for query expansion. Here τ is decided based on the normalized probability values of words. For the words in the initial query set (both in the feature as well as to object wordsets), we calculate the normalized probability values of these words that are obtained from the term scoring function (eq 5) and only those words are retained in the wordset whose normalized values are greater than 0.3. This threshold value is determined based on manual observation of their relevance with respect to the disaster. For example, words like 'footage', 'report', 'video' etc., were found to have normalized scores less than 0.3, which do not seem relevant with respect to infrastructure damage. The selected keywords are subsequently tagged as nouns and verbs and are added in the query set as object and feature word sets respectively. This process enriches the query set and also captures location-specific keywords. To retrieve more relevant infrastructure damage tweets using the enriched query set, we apply the split-query based approach that has been used to retrieve the seed tweets (as discussed in section V-A). We later show that the proposed term ranking mechanism used for query expansion outperforms several state-of-the-art techniques in retrieving relevant tweets for a given query set. We next investigate some of the failure cases in the retrieval process and outline how these failures are handled.

C. Error Analysis and Refinement of the Results

Based on the observations made from few retrieved tweets, we discovered that a small fraction of tweets (approx 2% of overall tweets) express uncertainty in the disaster. For example tweets like *i'm not sure if any buildings have been destroyed, but i'm very scared* and *Rigid buildings are more likely to collapse during quake* contain both object and feature keywords together in the same tweet and hence are selected as relevant although they do not provide any reliable information about infrastructure damage. We identify these uncertain tweets based on an uncertainty cue corpora that are built by leveraging some of the existing state-of-the-art techniques. Most of the existing techniques are built upon the uncertainty corpora (BioScope Corpus [49], WikiWeasel [50]). Besides these existing corpora, a new corpus hUnCertainty was annotated for semantic and discourse-level uncertainty cues for natural social media text in [51]. A tweet was considered to be uncertain if it contained at least one uncertainty cue and was subsequently considered to be irrelevant. Some specific tweets as *Earthquake barely damaged my house* have negative adverbs that are irrelevant to our task. These negative adverbs in the tweets are tackled with the lexicon of negative adverbs[7]. Such tweets are filtered out to enhance the accuracy of the proposed technique. We next outline the approach used to derive the damage scores of the locations from the retrieved tweets.

D. Damage score of a location

In this step, we use the extracted seed tweets to derive a damage score for a location. The damage score of a tweet set is a possible indication of the extent of damages at a place. The damage score is based on the total fraction of tweets mentioning an event that contains both infrastructure and damage keywords present in the object and feature word sets [14]. Thus if Q_f and Q_o represents the query keywords in the feature and object word sets respectively, and D represent the set of keywords in a tweet, then the initial significance score of each tweet is given as

$$I_D = \frac{(|Q_f \cap D|)}{|Q_f|} \frac{(|Q_o \cap D|)}{|Q_o|} \quad (6)$$

However, as stated earlier the mere presence of object and feature keywords in a tweet may not be a strong indicator of the damage and may not indicate equal severity. For example, "The building was really beautiful as it seems from aerial view of the damage". Although the tweet contains both infrastructure and damage keywords but the tweet does not strongly indicate the extent of damage. Several IR works have reported that documents in which the query terms are closer

[7]https://www.thefreedictionary.com/Negative-Adverbs.htm

are more likely to be relevant to the query [52]. In the previous example the context of the sentence changes, if two query terms *building* and *damage* are far from each other. Thus, based on the existing works [52], considering the distance between the object and feature keywords may provide stronger evidence of the damage. Hence, we assign more weights to the feature and object keywords that are closely placed. Based on this concept we introduce a *proximity based measure of the damage score* where the final significance score, I_p, of each tweet is estimated by the following formula:

$$I_p = I_D \left(1 - \frac{P_D}{n+1}\right)$$

Here, n is the number of tokens in the tweet. P_D is the distance between the keywords matched from the keyword sets. Note that if we find more matches between the query sets, we consider minimum distance as a final P_D.

To estimate the severity of damage across locations, we derive a damage score for each location. Earlier works have shown that disaster-affected areas are characterized by more negative sentiment tweets from that location [35]. Hence we use this observation to derive the damage score of a location. We consider the tweets mapped to a particular location based on our implemented location estimation technique and derive their sentiment score using VADER sentiment analyzer. This sentiment analyzer performs better on shorter text than other sentiment analysis tools [53]. VADER produces four sentiment scores for each tweet – positive, negative, neutral and compound score. We consider a tweet as negative if its negative score is greater than the positive score. Assuming for location L, N_L represents the set of negative tweets and I_L represents the set of negative tweets related to infrastructure damage, then the infrastructure damage score for location L is defined as

$$D_L = \frac{|I_L|}{|N_L|}\left(\sum_{I \in I_L} I_P\right) \tag{7}$$

Thus D_L depends upon both the number of negative sentiment tweets that are related to infrastructure damages as well as the fraction of these tweets with respect to the total number of negative tweets. This fraction reduces the bias in the damage score that may be induced in highly populated areas where the total number of negative tweets generated may be much higher as compared to less populated areas.

We next show the importance of considering the negative sentiment tweets to determine the infrastructure damage score in equation 7. It has been observed that regional sentiment bias exists during the reporting of crisis events in social media, due to which people from different parts of the world react differently during crisis [54]. Hence we do a generic investigation with tweets set from different countries. We study the tweets shared by [55], that contains the tweets of different crisis event including California wildfire, Hurricane Irma, Mexico earthquake and Srilanka flood. Each tweet is labeled into one of the following specific categories: informative, dead or injured, infrastructure damage, non-informative, missing people, rescue, affected, vehicle damaged. Each tweet is labeled by the manual annotators and label information of each tweet is available with the dataset. We find the mean negative sentiment score of each class and build a heatmap to show the intensity of the negative sentiment of each class. We show a heatmap in Fig. 4, which shows the infrastructure class has a very high correlation with the negative sentiment score (only less than the injured class). This justifies the use of negative tweets along with their significance score in determining the damage score. We later empirically show in the next section that the damage score of a location correlates with the actual damages in that area.

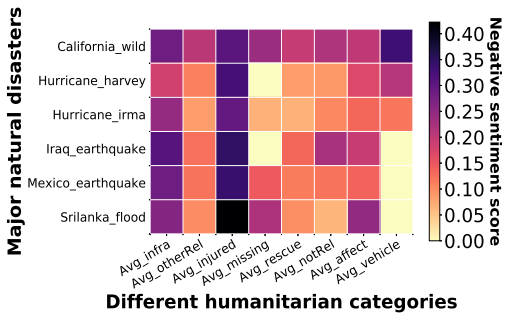


Fig. 4: On the heatmap, x-axis shows 8 categories of the crisis tweet and y-axis shows six major natural disasters. Avg_infra represents average infrastructure damage reported from different crisis events.

VI. Experimental Evaluation

In this section, we initially describe the dataset along with the ground truth data used in our experiments. Subsequently, the baseline techniques used for comparison are outlined.

A. Dataset Preparation

We use two datasets for our experiments: Nepal 2015 and Italy 2016 earthquake dataset. We sequentially discuss the details of both the datasets.

Nepal 2015 earthquake: This dataset is obtained from the organizers of SMERP 2018, on request [56]. This dataset contains all the information related to the tweets and was collected during the Nepal earthquake[8]. These crisis related tweets were collected by the organizers of SMERP 2018 using the Twitter Search API. The total number of tweets that are used for the experiment is 50,068. However, the provided dataset *does not* contain any ground truth regarding infrastructure damage. So, we prepared the ground truth for our purpose (as described in the following section).

Italy 2016 earthquake: We obtained this dataset, on request, from the organizers of SMERP 2017 [57]. Similar to the above mentioned dataset it contains tweets that were collected during the earthquake in Italy[9]. A set of 63152 tweets was

[8]https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake
[9]https://en.wikipedia.org/wiki/August_2016_Central_Italy_earthquake

TABLE III: Summary of datasets related to three recent disaster events

Dataset	Total tweets	Number of distict tweet	Number of Infrastructure damage tweets
Neapl Earthquake	100k	50,018	838
Italy Earthquake	180k	63152	1980
Indonesia Tsunami	400k	35138	509

obtained which we subsequently use for our experiments. Both the supplied datasets do not contain any duplicate or near-duplicate tweets. Out of total Italy earthquake tweets around 5% of the total tweets were not in the English language. So we use *py-translate* python library to convert the tweet text to English and add it with the English tweets.

Indonesia Tsunami 2018: We collect tweets made during the Indonesian Tsunami[10] using the query keywords like "indonesia tsunami", "volcano indonesia", "Krakatoa", "krakatau" through the Twitter Search API. We use relevant manually curated keyword search that has been used in relevant popular state-of-the-arts in [10, 5]. After removing duplicates and near-duplicates, a set of 35138 tweets out of 400K was obtained which we use for our experiments. The process involved in removing the duplicates is the same as used in [10]. We consider only tweets in English, as identified by the Twitter language identification system.

The NLTK library is further used to refine the tweet text by discarding punctuations and stopwords. Also, we removed re-tweet tags, URLs and user mentions. Necessary stemming is also done on the keywords using Porter Stemmer. Table III shows the statistics of all the three datasets. However, these datasets *do not* contain any ground truth specifically regarding infrastructure damage. So, we generated the ground truth data from these datasets through a manual process that we describe next.

B. Ground Truth Data

Manual annotation is done for all the tweets present in the datasets to prepare the ground truth data. Two professionals with English knowledge were selected as annotators who had used Twitter earlier. Each annotator was first asked to identify infrastructure damage tweets separately, i.e., without consulting each other. The inter-rater agreement was measured using Cohen kappa coefficient. We observed the Cohen kappa coefficient values of 84%, 82% and 78% for Nepal, Italy and Indonesia datasets respectively. This indicates high agreement among the annotators for all the datasets. The disagreements were resolved through mutual discussions.

Table IV shows some example tweets that are considered as infrastructure damage tweets posted during the Nepal earthquake, Italy earthquake and Indonesia Tsunami.

C. Baseline Techniques

In order to evaluate our proposed method, we compared our method with the following baseline methods on our dataset for accurate infrastructure damage tweets retrieval:

[10]https://en.wikipedia.org/wiki/2018_Sulawesi_earthquake_and_tsunami

TABLE IV: Examples of Infrastructure damage tweets posted during Nepal earthquake. Italy Earthquake and Indonesia Tsunami

Nepal Earthquake 2015 dataset
@JimatYearZero: Earthquake in Nepal destroys some very important and historic places. MAny people are dead and in need. Please help if…
Deupur Gairibisauna of Kavre- no houses left, no food and livestock trapped under debris. plz go for help is anyone is …
Italy Earthquake 2016 dataset
Strong, shallow M6.2 earthquake NW of Rome in central Italy. Preliminary reports of widespread structural damage.
Houses and old school castles came crashing down near #roma #earthquake #terremoto #amatrice
Indonesia Tsunami 2018 dataset
@the_hindu: #Indonesia homes were heavily damaged when the #tsunami struck. In pictures: https://t.co/t6FRfAmqZE
Watch horrific video showing moment tsunami waves wiped out a stage as band played at a concert in Indonesia https://t.co/MS0Mh6VYNr
Death toll rising after Indonesia 'volcano' tsunami destroys hundreds of homes and resorts https://t.co/0RBPfpT93A

Okapi BM25: This method used in [8] retrieves tweet set using BM25 values on initial query keywords. The keywords in the initial query are generated using the procedure as explained in section IV.

BM25 with Synonyms: Expansion of initial query set is carried by adding the synonyms of words present in our initial query section IV in [8]. Next, the relevant tweets are retrieved using Okapi BM25 score using the expanded query set.

Word2vec: As proposed in [10], a Word2vec model is trained on the tweets in our dataset provided. The training specifics are the same as that of [10] for constructing the Word2vec model on tweets. We obtain the tweets vector (also query vector) using sum of term-vectors for terms present in tweets and the query. These tweets and query vectors are normalized by their respective lengths. To rank each tweet, a value is assigned to it based on the cosine similarity of that tweet vector and query vector. Next, the following two methods are used for query expansion:

- *Query Expansion using Word2vec:* For each distinct term in the dataset we calculate the cosine similarity value between term vector and an initial query vector and assign it to the term. Top 5 high value terms are used for query expansion.
- *Query expansion using Rocchio score:* After getting the top 10 highest valued tweets using cosine similarity with the initial query, we find distinct terms from those tweets and calculated *tf-idf Rocchio scores* for each of these terms. The top 5 highest valued terms are used for query expansion.

Both these methods are used as baselines for comparing our proposed method.

Convolutional neural network (CNN): Widely used CNN based learning technique for crisis tweets classification as proposed in [13], is also used to evaluate our method performance. We used three-layered CNN model with ReLU, softmax function and 500 dense nodes, batch size of 4 and epochs count of 100. For our word embedding, pre-trained

GloVe "glove.twitter.27B.zip"[11] vectors are used. Five-fold cross-validation is applied during training.

Split-Query without Query Expansion: To evaluate the performance of the co-occurrence based feedback approach used for the expansion of the query, we introduce a baseline with our approach of split-query based tweet retrieval with no feedback (query expansion) technique.

Split-Query with Co-occurring Feedback (SQC): This approach was also proposed in our earlier paper [14], where splitted queries were expanded using highly co-occurring terms. We also use this method as one of the baselines for our investigation. It is worth noting here that the initial query keywords are the same as the currently proposed technique so that the result based on methodological differences can be compared.

VII. Experimental Results and Insights

We start with a discussion on the performance measures used for the evaluation followed by the observations and insights.

A. Performance Measures

Different performance metrics such as Precision@k, Recall, $F1$ measure, Mean average precision and bpref are used for comparing the performance of TAQE with different state-of-the-arts [17]. Next, we briefly define the performance measures. If $\mathscr{R}$ is a ranked list of tweets obtained using any given method, then

Precision (P) is the fraction of the number of actual infrastructure damage tweets that are retrieved by the approach over the total number of retrieved relevant infrastructure damage tweets considered. *Precision@k* ($P@k$) is the precision calculated when the first top k retrieved relevant infrastructure damage tweets are only considered.

$$Precision@k = \frac{number\ of\ relevant\ tweets\ retrieved}{total\ number\ of\ tweets\ retrieved} \quad (8)$$

Recall (R) it is the ratio of infrastructure damage tweets retrieved to the total number of infrastructure damage tweets $\mathscr{R}$.

$$recall = \frac{number\ of\ relevant\ tweets\ retrieved}{number\ of\ relevant\ tweets\ in\ collection} \quad (9)$$

$F1$-measure It is the harmonic mean of the precision and recall.

$$F1 = \frac{2 * P * R}{P + R} \quad (10)$$

Mean Average Precision (MAP): Let $rel(i)$ be the relevance of the i_{th} tweet in the ranked list $\mathscr{R}$, the *average precision* (AP) is calculated as $\frac{\sum_{i=1}^{|\mathscr{R}|} P@i \times rel(i)}{|\mathscr{R}|}$, providing the measure of top-k retrieved tweets' relevance. Mean Average Precision (*MAP*) is the mean of the average precision of each query score from a set of given queries. It can measure the quality of ranking for different recall levels.

[11] https://nlp.stanford.edu/projects/glove/

TABLE V: This table shows some of the example tweets retrieved by our proposed technique for Nepal, Italy earthquake and Indonesia tsunami and is labelled as *Yes*. Few other tweets that are correctly rejected by our approach is also shown and is labelled as *No*.

Example Tweets retrieved from Nepal earthquake dataset	**Labels**
Pashupatinath temple in Nepal suffers damages but survives the earthquake	**Yes**
Officials: Walls cracked and power lines downed in Papua New Guinea during magnitude 7.5 earthquake; small tsunami	**Yes**
" Days after Nepal's devastating earthquake, a teen survivor is rescued from the rubble of a seven-story building	**Yes**
@BBSBhutan the videos you have been showing of swimming pool and building collapse after earthquake are not of Nepal. can you Verify it	**No**
imagine your town is demolished by an earthquake and the only road in and out can t carry heavy plant	**No**
Example Tweets retrieved from Italy earthquake dataset	**Labels**
italy earthquake leaves dead towns in ruins biggest damage appears to be in amatrice and accumoli in lazi	**Yes**
bbc news earthquake causes houses to crash down hillside in italy village	**Yes**
6.2 earthquake hit north of #Rome.Ppl trapped under structural damage, no other details yet re casualties.	**Yes**
Photo of damage from tonight's #earthquake in #Amatrice. Rigid buildings are more likely to collapse during quake	**No**
"@tonyrobinsonn here we felt the earthquake, but fortunately there are no damage to houses and people. Thank you for your message!"	**No**
Example Tweets retrieved from Indonesia Tsunami dataset	**Labels**
Horrific Moment Tsunami Waves Wiped Out A Stage As Band Played At A Concert InÂ Indonesia https://t.co/ZorOgnknSb	**Yes**
Circumstances are unavoidable and unexpected. So many people have lost their homes in the devastating Indonesia Tsunami	**Yes**
a Tsunami has hit the coast of Indonesia killing at least 222 people and injuring hundreds more.	**No**
concert hit by Tsunami Indonesia seventeenband SEVENTEEN PrayForAnyer TsunamiSelatSunda"	**No**

B-pref: Bpref measures how early the i^{th} relevant document is retrieved ahead of the irrelevant documents.

$$bpref = \frac{1}{R}\sum_{r}(1 - \frac{|n\ ranked\ higher\ than\ r|}{min(R,N)}) \quad (11)$$

where R, N, r are the numbers of judged relevant document, non-relevant document, relevant document retrieved respectively; n is the member of the first R relevant retrieved document.

B. Performance comparison with the baselines

Next, we compare the performance of our proposed methodology with six other baselines discussed in section VI. This comparison is shown in Table VI. It is observed, for all the performance metrics, our proposed methodology outperforms the baselines.

P@k values: The comparison of $P@k$ (here $k = 1000$) values are shown in table VI. As observed, for all the datasets (Nepal, Italy, and Indonesia) our proposed technique considerably outperforms the baselines. As observed, Word2vec based baselines with both variations perform poorly with respect to Precision. However, BM25 and its variation show better Precision in comparison to Word2vec approaches. TAQE

outperforms significantly over the best performing baseline with an improvement of around 13% in the Precision values in case of Nepal, around 7% in case of Italy and 34% in case of Indonesia dataset. Short length, semantic variation, and noise are the possible reasons for the poor Precision values observed in case of Word2vec approaches, as these characteristics make the training of the embedding model difficult. However, we observe better precision in the case of Word2vec when Word2Vec expansion is used rather than Rocchio expansion. This may be due to the fact that Rochhio expansion adds the term with higher tf-idf score ignoring its relevance with the query set, while Word2vec expansion adds the term to the original query based on the embedding similarities.

Recall and $F1$-measure: We discover that TAQE shows considerable improvement over the baselines in terms of recall as well as $F1$ for Nepal, Italy and Indonesia datasets. While comparing with the BM25 based baselines, we observe that the Recall and $F1$ values using our proposed technique are respectively 14% and 20% higher for Nepal-quake, 12% and 10% higher for the Italy-quake and 26% and 31% higher for Indonesia-tsunami dataset. Further, TAQE uses a two-stage query expansion that results in considerably higher Recall value (0.83) as compared to the case when no query expansion is involved (0.31) for the Nepal-quake dataset . Similar trends are observed for the Italy-quake dataset too. This highlights the importance of our language modeling and co-occurrence based query expansion.

Table V shows example tweets relevant to infrastructure damage that could only be retrieved by our proposed approach for all the three datasets. We also show some of the example tweets that are correctly discarded by our approach. A closer look at these tweets reveals that certain tweets do not contain keywords like "destroy" or "collapse" that are present in the query are still retrieved using our proposed approach. On the other hand, certain tweets that may contain both the object and feature keywords, are discarded being contextually irrelevant to the infrastructure damage. In addition, the examples also include some tweets that appear to be uncertain and hence are also being discarded using our technique. This clearly establishes the utility of our proposed approach in identifying pertinent tweets on infrastructure damage. Table VIII shows some of the example tweets that are extracted as relevant by the BM25, Word2vec, and SQC approach but are not relevant according to ground truth and are rejected by our proposed approach. Table VII shows the number of false positive tweets identified by state-of-the-art techniques including our proposed technique. The number of false positive tweets extracted using TAQE is significantly lesser than the state-of-the-art techniques. We observe that the number of false positive tweets in case of Italy earthquake is relatively higher than CNN but it's significantly lesser than the other methods.

MAP scores: MAP scores provide a measure of the quality of ranking. As observed in Table VI, the ranking quality of our proposed approach is considerably better than all other baselines. The MAP value is calculated for top 1000 tweets. A very low MAP score of Word2vec based methods indicates the poor ranking quality for both the datasets. MAP score using BM25 without feedback performs better than the Word2vec baselines for both the dataset. TAQE achieves an improvement ranging from 21% over the baselines for Nepal and Italy dataset while the corresponding improvement is around 70% for Indonesia tsunami.

B-pref: Bpref validates whether the infrastructure damage tweets are ahead of the non-infrastructure damage tweets in the ranked list. TAQE outperforms the other techniques in terms of this parameter also, where the improvement ranges from $7-12\%$ depending upon the methods and datasets.

Comparison with CNN: After applying the CNN on Nepal, Italy and Indonesia datasets, the obtained $F1$ scores are 0.40, 0.24, 0.20 respectively. Because of the huge difference between the number of infrastructure damage and non-infrastructure damage tweets, the CNN model ends up considering more irrelevant tweets as relevant thereby lowering the precision and subsequently less $F1$ score as compared to TAQE.

Thus our empirical investigations reveal that our proposed

TABLE VI: Performance comparison of baselines and our proposed technique.

Comparison with the baselines on Nepal earthquake dataset						
Ranking Model	Expansion	P@1000	Recall	$F1$	MAP	Bpref
Okapi BM25	No	0.28	0.30	0.29	0.10	0.19
Okapi BM25	Synonyms	0.31	0.42	0.35	0.13	0.23
Word2vec	Word2vec	0.22	0.25	0.23	0.04	0.13
Word2vec	Rochhio	0.06	0.48	0.10	0.003	0.03
Split-query	No	0.48	0.31	0.37	0.20	0.19
Split-query(SQC)	Co-occurring term + Synonyms	0.52	0.61	0.56	0.30	0.41
Split-query(TAQE)	**Topically aligned co-occurring terms**	**0.609**	**0.83**	**0.70**	**0.38**	**0.47**
Comparison with the baselines on Italy earthquake dataset						
Okapi BM25	No	0.18	0.11	0.14	0.081	0.18
Okapi BM25	Synonyms	0.20	0.13	0.15	0.056	0.18
Word2vec	Word2vec	0.1	0.16	0.12	0.012	0.06
Word2vec	Rochhio	0.034	0.02	0.025	0.0015	0.023
Split-query	No	0.14	0.20	0.17	0.01	0.01
Split-query(SQC)	Co-occurring term + Synonyms	0.24	0.69	0.35	0.09	0.177
Split-query(TAQE)	**Topically aligned co-occurring terms**	**0.26**	**0.79**	**0.39**	**0.09**	**0.19**
Comparison with the baselines on Indonesia earthquake dataset						
Okapi BM25	No	0.05	0.10	0.06	0.02	0.047
Okapi BM25	Synonyms	0.06	0.11	0.07	0.02	0.05
Word2vec	Word2vec	0.1	0.15	0.12	0.012	0.06
Word2vec	Rochhio	0.031	0.16	0.052	0.001	0.017
Split-query	No	0.12	0.23	0.15	0.02	0.02
Split-query(SQC)	Co-occurring term + Synonyms	0.17	0.38	0.24	0.02	0.070
Split-query(TAQE)	**Topically aligned co-occurring terms**	**0.26**	**0.52**	**0.35**	**0.09**	**0.29**

TABLE VII: Number of false positive tweets obtained using baselines.

Approach	Nepal Earthquake	Italy Earthquake	Indonesia Earthquake
SQC	755	19534	1041
BM25	5432	11315	5094
W2V	2567	2964	3110
CNN	883	1191	709
TAQE	470	2474	287

TABLE VIII: Example tweets retrieved using the baselines (BM25, Word2vec, SQC) but rejected by TAQE.

Example tweets retrieved as relevant using state-of-the-art	Baselines	TAQE
Glad to know my family is okay. Can't believe all the damage the earthquake caused	BM25(Yes)	No
@MrWrestling97 yep , the earthquake hits my town too but without damage	BM25(Yes)	No
@JessicaDHarvard: I want to take this moment to pray for people kids families that are without today.without food family homes warmth	BM25(Yes)	No
Italy: Drone captures earthquake-devastated Pescara del Tronto	Word2vec(Yes)	No
Italy earthquake: 10yo girl pulled alive from Pescara del Tronto rubble 17 hours after disaster	Word2vec(Yes)	No
@Chutkla: I dont mind corruption but they are passing non-earthquake resistant houses in seismic zone IV & V. This will kill people.	Word2vec(Yes)	No
Earthquake didn't kill those thousands, the buildings did #NepalEarthquake	Word2vec(Yes)	No
Following the devastating #earthquake in #Nepal please #support GAN's 'Rebuild Nepal' #fund to support reconstruction	SQC(Yes)	No
@MamataOfficial: Today I visited a hospital in Siliguri to meet people with #earthquake injuries amd trauma	SQC(Yes)	No

approach outperforms all the other baselines in terms of almost all the performance measures. It is evident that across all methodologies for earthquake datasets, the performances are considerably better over the Nepal earthquake dataset than over the Italy-quake dataset. Although we are unable to discover a precise reason for the same, however, we believe that this may be due to the presence of multi-lingual tweet in significant proportion in the Italy earthquake dataset. The efficiency of the language conversion library may have played a role in the slightly lower performance of TAQE. As these findings also reveal that the performance of SQC and BM25 techniques are relatively better as compared to the other baselines, so we use these techniques for comparing certain other performance aspects like the possible applicability on the real-time dataset.

C. Performance on a Tweet Stream

We also investigate the possible applicability of our approach a live stream of tweets where they arrive sequentially in a time-ordered fashion and the entire tweet set is not available at the beginning. We compare the performance of TAQE with BM25 and SQC with increasing dataset size for different time windows.

Performance gain with increasing dataset size: Starting with very low dataset size (10% of the total tweets sorted in order of arrival time), the size of the dataset is increased gradually by including tweets according to their arrival time. Using all the available dataset, we subsequently compare the performance of TAQE with SQC and the BMS (BM25 based) baselines with respect to five performance measures, (*a*) Precision@1000, (*b*) Recall, (*c*) $F1$-measure, (*d*) MAP and (*e*) B-pref. Figures 5, 7 and 8 show the performance comparison for the Nepal, Italy and Indonesia dataset respectively. We observed that for all these measures TAQE consistently performs better as compared to the other two approaches. The F1-measure for the BMS approach is the lowest for low data sizes and gains rapidly with increasing size of the available data, however, the performance of TAQE remains consistently higher than both these baselines.

Performance changes over delayed data: We next investigate the changes in the performance of the proposed approach considering the tweets arriving on each day from the time of occurrence of the event. The objective is to investigate the variation in performance when the entire tweet set from the beginning is not available. Fig. 6 shows the performance of TAQE as compared to SQC and BMS for all the earlier performance measures for the Nepal dataset. For each of the approaches, the performance varies depending upon the number and nature of the tweets that arrive daily. However, overall the performance of TAQE remains high over all the other approaches for all the parameters. These results indicate that our proposed approach provides noteworthy improvement in the performance over our previously proposed approach (SQC) [14]. However, considering the very low F1 scores that are achieved for certain days that are far apart from the day of occurrence, it is likely that applying these approaches to tweets that arrive much later will not be productive.

TABLE IX: Effect of number of co-occurring terms as feedback on our proposed technique

Number of top co-occurring terms (κ)	**Precision@1000**	**Recall**	**F1-measure**	**MAP**	**B-pref**
3	0.521	0.68	0.58	0.41	0.49
5	0.55	0.70	0.61	0.370	0.473
10	0.606	0.833	0.701	0.37	0.47
20	0.609	0.835	0.704	0.381	0.48
30	0.172	0.813	0.283	0.081	0.135
40	0.60	0.67	0.63	0.40	0.50

D. Determining the co-occurring set size κ

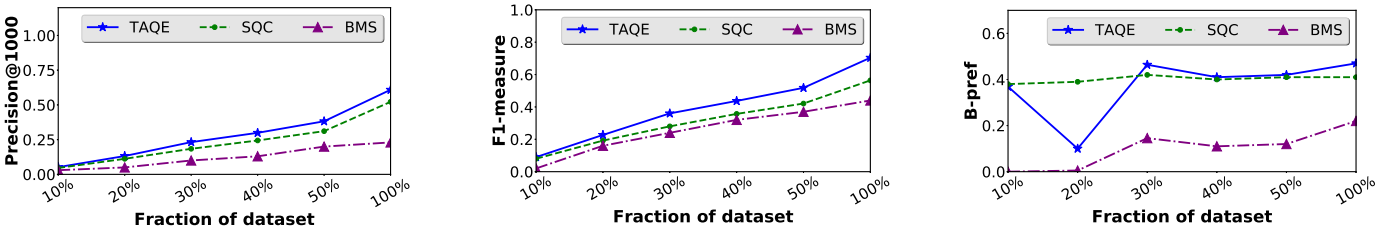
To determine the number of co-occurring keywords, κ, that should be considered in the initial query expansion step, we varied the number of co-occurring keywords from 3 to 40. For each value of κ, we subsequently apply TAQE on the Nepal dataset to retrieve the relevant tweets and observe the performance in terms of all the previously mentioned measures. As shown in Table IX, the best $F1$ score is obtained when the value of κ is considered as 20. Although the recall increases beyond a κ value of 25, the precision drops rapidly. This result is intuitive as increasing κ helps in capturing more relevant tweets thereby increasing the recall, but also simultaneously increases the fraction of false positives and hence the precision drops. The MAP and bpref scores also get affected as with increasing values of κ, less relevant tweets may get captured over the relevant ones. Thus considering these results a value of κ between 10 and 20 seems to work best. Hence for our investigations we consider $\kappa = 20$.

We next investigate the accuracy of the proposed damage score measure.

E. Accuracy of damage score

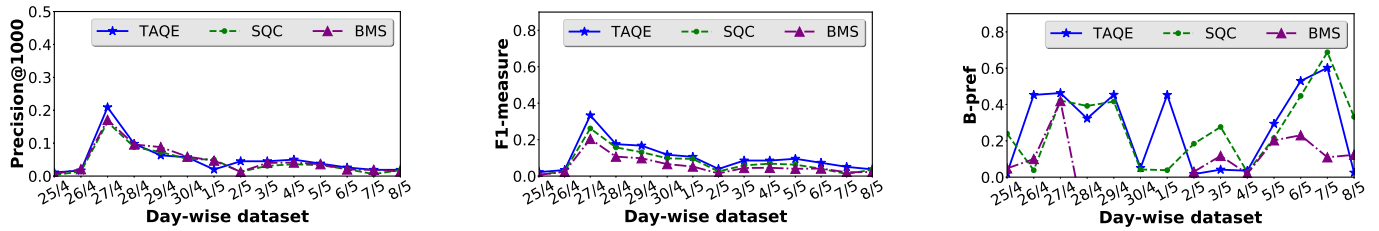
Next, we compare the estimated damage score at locations (eqn 7) using our proposed approach with the actual infrastructure damages, as reported by "United States Agency for International Development (USAID[12])" for different locations in Nepal. This report is also used in [14] to verify the gravity of damage at locations. Information regarding the number

[12]https://www.usaid.gov/sites/default/files/documents/1866/05.18.15-USAID-DCHANepalEarthquakeMap.pdf



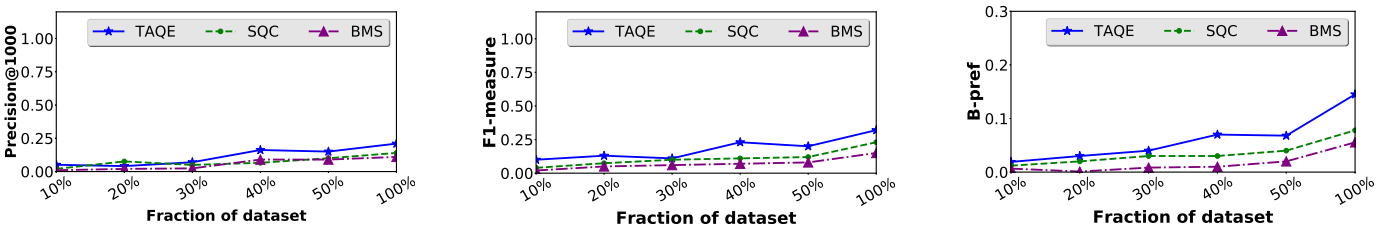
(a) Precision@1000 values (b) F1-measure values (c) B-pref values

Fig. 5: Effects on performance with increasing dataset size on Nepal earthquake dataset. BMS represents BM25 ranking model with synonyms for expansion. TAQE represent proposed technique i.e. split-query ranking model with probabilistic term for expansion and SQC represents split-query ranking model with the co-occurring term for expansion.



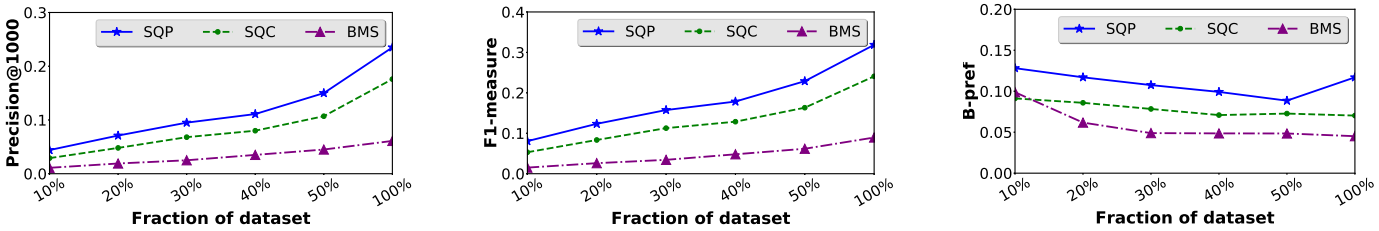
(a) Precision@1000 values (b) F1-measure values (c) B-pref values

Fig. 6: Effects on performance with increase in time on Nepal earthquake dataset. BMS represents BM25 ranking model with synonyms. TAQE represent proposed technique i.e. split-query ranking model with probabilistic term for expansion and SQC represents split-query ranking model with the co-occurring term for expansion.



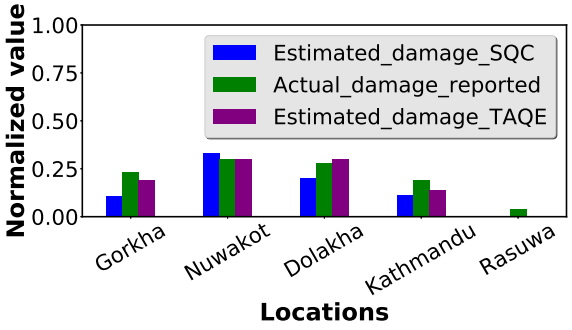
(a) Precision@1000 values (b) F1-measure values (c) B-pref values

Fig. 7: Effects on performance with increasing dataset size on Italy earthquake dataset. BMS represents BM25 ranking model with synonyms. TAQE represent proposed technique i.e. split-query ranking model with probabilistic term for expansion and SQC represents split-query ranking model with the co-occurring term for expansion.

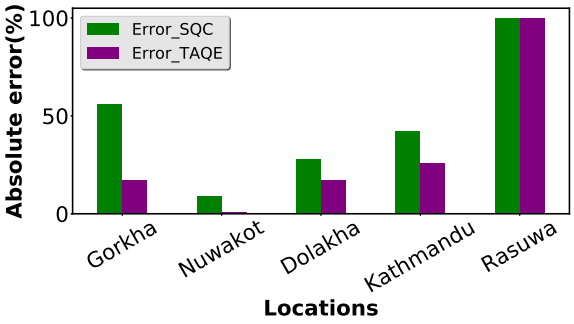


(a) Precision@1000 values (b) F1-measure values (c) B-pref values

Fig. 8: Effects on performance with increasing dataset size on Indonesia Tsunami dataset. BMS represents BM25 ranking model with synonyms. TAQE represent proposed technique i.e. split-query ranking model with probabilistic term for expansion and SQC represents split-query ranking model with the co-occurring term for expansion.



(a)



(b)

Fig. 9: Fig. 9(a) Compares the damage scores with actual infrastructure damages and a baseline at various regions of Nepal. The number of actual infrastructure damages is obtained from USAID. Both damage score and actual damages are normalized by their maximum values. Fig. 9(b) compares the absolute percentage error between the actual and TAQE/SQC technique.

(a)

(b)

Fig. 10: Fig. 10(a) shows the damaged location with their severity on the heatmap, whose score is estimated using our proposed technique. Fig. 10(b) shows the damaged location with their severity as reported in The New York Times.

of damaged public buildings in the USAID report gives the measure of the actual infrastructure damage. Since these values are not on the same scale, we normalize them with their corresponding highest values to keep the values within $[0, 1]$. The correlation between the actual damages reported and the derived damage scores can be visualized in Fig. 9(a). As observed, Nuwakot was severely damaged, and the tweets reporting about that region had much higher negative sentiment as well as tweet significance score as compared to the rest of the regions. Hence the damage score for Nuwakot was much higher as compared to the rest.

We also investigate the difference in actual and estimated values for the TAQE and SQC approaches, in terms of the absolute percentage error. The absolute percentage error is given as, $Absolute\ \%\ error =$

$$\frac{|Actual\ damage\ reported - Estimated\ damage\ reported|}{Actual\ damage\ reported} * 100$$

We observe from Fig. 9(b), absolute percentage error is minimum for TAQE for all the locations except Rasuwa. However, Rasuwa being a remote location was not much reported through tweets and hence none of the methods could identify its actual severity. Hence, for Rasuwa the error is nearly same for both the techniques.

We observe similar results for the Italy earthquake dataset as well. We validated the predicted damage scores with reports published in the New York Times about the incident[13]. It was also verified from USGS Shakemap[14] reported in Wikipedia[15]. We show the predicted results as well as the actual damages using two separate heatmaps shown in Fig. 10(a) and 10(b) respectively. Color intensity represents the severity of damage at that location. It is observed from Fig. 10(b), Amatrice was the epicentre and was highly damaged. Nearby locations to the Amatrice were also damaged but the impact was less. Other locations like Norcia, Accumoli, Rome, San Marino, Perigua were some locations that were also damaged. Similar locations are predicted based on the damage scores as shown in Fig. 10(a). Thus, our proposed technique can efficiently identify the epicentre as well as the less damaged locations.

[13]https://www.nytimes.com/2016/08/25/world/europe/italy-earthquake.html

[14]https://en.wikipedia.org/wiki/August_2016_Central_Italy_earthquake#/media/File:Shakemap_Earthquake_24_Aug_2016_Italy.jpg

[15]https://en.wikipedia.org/wiki/August_2016_Central_Italy_earthquake

VIII. Conclusion

In this paper, we proposed a split-query with topic aligned query expansion approach (*TAQE*) for retrieving infrastructure damage tweets from the tweet stream. Our approach can identify infrastructure damage tweets with high accuracy as compared to state-of-the-art baselines for all the datasets, thereby helping an early assessment of the damages. We also propose a technique to use the retrieved tweets to measure the infrastructure damage severity at different locations. Investigations of our approach on empirical data of Italy and Nepal earthquake reveals that the predicted damage score correlates well with the actual reported damages. As *TAQE* performs significantly better for earthquake and tsunami datasets in terms of all the performance metrics, this shows the generality of TAQE and we believe it can be applied over other datasets as well. We further expect to enhance the performance of retrieval by including multi-modal information like images and videos that are present in many of the tweets.

ACKNOWLEDGMENT

The first author of this paper is supported by the Government of India under Visvesvaraya Ph.D. scheme of Department of Electronics and Information Technology.

References

[1] S. Priya, R. Sequeira, J. Chandra, and S. K. Dandapat, "Where should one get news updates: Twitter or reddit," *Online Social Networks and Media*, vol. 9, pp. 17–29, 2019.

[2] M. Mendoza, B. Poblete, and I. Valderrama, "Nowcasting earthquake damages with twitter," *EPJ Data Science*, vol. 8, no. 1, p. 3, 2019.

[3] S. Priya, S. Singh, S. K. Dandapat, K. Ghosh, and J. Chandra, "Identifying infrastructure damage during earthquake using deep active learning,"

[4] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding twitter data with tweetxplorer," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), pp. 1482–1485, ACM, 2013.

[5] P. Khosla, M. Basu, K. Ghosh, and S. Ghosh, "Microblog retrieval for post-disaster relief: Applying and comparing neural ir models," *arXiv preprint arXiv:1707.06112*, 2017.

[6] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "Aidr: Artificial intelligence for disaster response," in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, (New York, NY, USA), pp. 159–162, ACM, 2014.

[7] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing situational and topical information during crises," *arXiv preprint arXiv:1610.01561*, 2016.

[8] R. Singla, S. Modha, P. Majumder, and C. Mandalia, "Information extraction from microblog for disaster related event.," in *SMERP@ ECIR*, pp. 85–92, 2017.

[9] I. P. Temnikova, C. Castillo, and S. Vieweg, "Emterms 1.0: A terminological resource for crisis tweets.," in *ISCRAM*, 2015.

[10] M. Basu, K. Ghosh, S. Das, R. Dey, S. Bandyopadhyay, and S. Ghosh, "Identifying post-disaster resource needs and availabilities from microblogs," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 427–430, ACM, 2017.

[11] M. Basu, A. Shandilya, P. Khosla, K. Ghosh, and S. Ghosh, "Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 604–618, 2019.

[12] F. Alam, S. Joty, and M. Imran, "Domain adaptation with adversarial training and graph embeddings," *arXiv preprint arXiv:1805.05151*, 2018.

[13] D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[14] S. Priya, M. Bhanu, S. K. Dandapat, K. Ghosh, and J. Chandra, "Characterizing infrastructure damage after earthquake: A split-query based ir approach," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 202–209, IEEE, 2018.

[15] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: Survey summary," in *Companion of the The Web Conference 2018 on The Web Conference 2018*, pp. 507–511, International WWW Conferences Steering Committee, 2018.

[16] L. Thapa, "Spatial-temporal analysis of social media data related to nepal earthquake 2015," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, p. 567, 2016.

[17] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, vol. 39. Cambridge University Press, 2008.

[18] B. Xu, H. Lin, Y. Lin, L. Yang, and K. Xu, "Improving pseudo-relevance feedback with neural network-based word representations," *IEEE Access*, vol. 6, pp. 62152–62165, 2018.

[19] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian, "Rapid assessment of disaster damage using social media activity," *Science Advances*, vol. 2, no. 3, 2016.

[20] T. Li, N. Xie, C. Zeng, W. Zhou, L. Zheng, Y. Jiang, Y. Yang, H.-Y. Ha, W. Xue, Y. Huang, S.-C. Chen, J. Navlakha, and S. S. Iyengar, "Data-driven techniques in disaster information management," *ACM Comput. Surv.*, vol. 50, pp. 1:1–1:45, Mar. 2017.

[21] T. H. Nazer, G. Xue, Y. Ji, and H. Liu, "Intelligent disaster response via social media analysis a survey," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 46–59, 2017.

[22] K. Rudra, A. Sharma, N. Ganguly, and M. Imran, "Classifying and summarizing information from microblogs during epidemics," *Information Systems Frontiers*, pp. 1–16, 2018.

[23] S. Kumar, X. Hu, and H. Liu, "A behavior analytics approach to identifying tweets from crisis regions," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, (New York, NY, USA), pp. 255–260, ACM, 2014.

[24] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," *IEEE Transactions on Computational Social Systems*, 2018.

[25] J. Sampson, F. Morstatter, R. Zafarani, and H. Liu, "Real-time crisis mapping using language distribution," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pp. 1648–1651, IEEE, 2015.

[26] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu, "Whom should i follow?: identifying relevant users during crises," in *Proceedings of the 24th ACM conference on hypertext and social media*, pp. 139–147, ACM, 2013.

[27] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *Visual analytics science and technology (VAST), 2011 IEEE conference on*, pp. 181–190, IEEE, 2011.

[28] C. Havas, B. Resch, C. Francalanci, B. Pernici, G. Scalia, J. L. Fernandez-Marquez, T. Van Achte, G. Zeug, M. R. R. Mondardini, D. Grandoni, *et al.*, "E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time," *Sensors*, vol. 17, no. 12, p. 2766, 2017.

[29] A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 994–1009, ACM, 2015.

[30] M. Avvenuti, S. Cresci, F. Del Vigna, T. Fagni, and M. Tesconi, "Crismap: a big data crisis mapping system based on damage detection and geoparsing," *Information Systems Frontiers*, pp. 1–19, 2018.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 919–931, April 2013.

[32] Y. Zhang, C. Szabo, Q. Z. Sheng, and X. S. Fang, "Snaf: Observation filtering and location inference for event monitoring on twitter," *World Wide Web*, vol. 21, no. 2, pp. 311–343, 2018.

[33] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 919–931, 2013.

[34] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 9–17, 2014.

[35] D. Buscaldi and I. Hernandez-Farias, "Sentiment analysis on microblogs for natural disasters management: a study on the 2014 genoa floodings," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1185–1188, ACM, 2015.

[36] H. Purohit, C. Castillo, F. Diaz, A. Sheth, and P. Meier, "Emergency-relief coordination on social media: Automatically matching resource requests and offers," *First Monday*, vol. 19, no. 1, 2013.

[37] M. Imran, P. Meier, and K. Boersma, "The use of social media for crisis management," *Big Data, Surveillance and Crisis Management*, p. 2, 2017.

[38] I. P. Temnikova, C. Castillo, and S. Vieweg, "Emterms 1.0: A terminological resource for crisis tweets.," in *ISCRAM*, 2015.

[39] G. K. Palshikar, M. Apte, and D. Pandita, "Weakly supervised and online learning of word models for classification to detect disaster reporting tweets," *Information Systems Frontiers*, pp. 1–11, 2018.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[41] M. Basu, A. Shandilya, K. Ghosh, and S. Ghosh, "Automatic matching of resource needs and availabilities in microblogs for post-disaster relief," in *Companion Proceedings of the The Web Conference 2018*, WWW '18, (Republic and Canton of Geneva, Switzerland), pp. 25–26, International World Wide Web Conferences Steering Committee, 2018.

[42] M. Avvenuti, S. Cresci, F. Del Vigna, and M. Tesconi, "Impromptu crisis mapping to prioritize emergency response," *Computer*, vol. 49, no. 5, pp. 28–37, 2016.

[43] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Proceedings of the 22nd international conference on world wide web*, pp. 1017–1020, ACM, 2013.

[44] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[45] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," *arXiv preprint arXiv:1806.03713*, 2018.

[46] T. Miyanishi, K. Seki, and K. Uehara, "Improving pseudo-relevance feedback via tweet selection," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 439–448, ACM, 2013.

[47] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,

SIGIR '98, (New York, NY, USA), pp. 275–281, ACM, 1998.

[48] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 2, pp. 179–214, 2004.

[49] R. Farkas, g. Szarvas, I. Hegedus, A. Almasi, V. Vincze, R. Ormandi, and R. Busa-Fekete, "Semi-automated construction of decision rules to predict morbidities from clinical texts," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 601–605, 2009.

[50] "Cross-genre and cross-domain detection of semantic uncertainty," *Computational Linguistics*, vol. 38, no. 2, pp. 335–367, 2012.

[51] V. Vincze, *Uncertainty detection in natural language texts*. PhD thesis, szte, 2015.

[52] J. Zhao and Y. Yun, "A proximity language model for information retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 291–298, ACM, 2009.

[53] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science*, vol. 5, no. 1, pp. 1–29, 2016.

[54] K. S. Smith, R. McCreadie, C. Macdonald, and I. Ounis, "Regional sentiment bias in social media reporting during crises," *Information Systems Frontiers*, pp. 1–13, 2018.

[55] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *AAAI Conference on Web and Social Media (ICWSM)*, (Stanford, California, USA), AAAI, AAAI, June 2018.

[56] M. Moens, G. J. F. Jones, S. Ghosh, D. Ganguly, T. Chakraborty, and K. Ghosh, "WWW'18 workshop on exploitation of social media for emergency relief and preparedness: Chairs' welcome & organization," in *Companion WWW 2018, Lyon , France, April 23-27, 2018*, pp. 1609–1611, 2018.

[57] S. Ghosh, K. Ghosh, T. Chakraborty, D. Ganguly, G. Jones, and M.-F. Moens, "First International Workshop on Exploitation of Social Media for Emergency Reliefand Preparedness (SMERP)," in *Proceedings of the 39th European Conference on IR Research – J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193*, ECIR 2017, pp. 779–783, Springer International Publishing AG, 2017.

Shalini Priya is a research scholar student in the department of Computer Science and Engineering at IIT, Patna. Her current research interest includes Online Social Network, Data mining and Information Retrieval.

Manish Bhanu is a research scholar student in the department of Computer Science and Engineering at Indian Institute of Technology, Patna. He has worked with LIAAD, INESEC TEC, Porto, Portugal. Her current research interest includes Intelligent Transportation Systems, Data mining and Information Retrieval.

Dr. Sourav Kumar Dandapat is an assistant Professor in the department of Computer Science and Engineering of IIT Patna, India. He has completed his PhD in 2015, from Computer Sc. and Engg. department of IIT Kharagpur, India. His current research interest includes Online Social Network, Mobile Computing, Human Computer Interaction.

Kripabandhu Ghosh received the Ph.D. degree from the ISI, Kolkata, India. He has been an International Scholar at KU Leuven, Leuven, Belgium. He was a Post-Doctoral Researcher at the Department of Computer Sc. and Engg., IIT Kanpur, India. Currently he is a researcher at Tata Research Development and Design Centre, TCS, Pune. His research interests include information retrieval, data mining, Natural Language Processing etc.

Joydeep Chandra is an Assistant Professor in the Department of Computer Sc. and Engg. at Indian Institute of Technology, Patna, India. He received his PhD from IIT, Kharagpur, India in 2012. He was also a research fellow at the Chair of Systems Design at ETH Zurich. His research interest includes modeling of social networks, studying diffusion of information and identifying influentials. Application domains include Journalism, Disaster, Healthcare and Crimes on the Web.