Microblog Retrieval in a Disaster Situation: A New Test Collection for Evaluation

Moumita Basu[1,2], Anurag Roy[1], Kripabandhu Ghosh[3],
Somprakash Bandyopadhyay[2], and Saptarshi Ghosh[4,1] ⋆

[1] Indian Institute of Engineering Science and Technology, Shibpur, India
[2] Indian Institute of Management, Calcutta, India
[3] Indian Institute of Technology, Kanpur, India
[4] Indian Institute of Technology, Kharagpur, India

Abstract. Microblogging sites are important sources of situational information during disaster situations. Hence it is important to design and evaluate Information Retrieval (IR) systems that retrieve information from microblogs during disaster situations. The primary contribution of this paper is to develop a test collection for evaluating IR systems for microblog retrieval in disaster situations. The collection consists of about 50,000 microblogs posted during the Nepal earthquake in April 2015, a set of five topics (information needs) that are practically important during a disaster, and the gold standard annotations of which microblogs are relevant to each topic. We also present some IR models that can be suitable in this evaluation setup, including a standard language model based retrieval, and word embedding based retrieval. We find that the term embedding based retrieval performs better for short, noisy microblogs.

Keywords: Microblog Retrieval, Test Collection, Disaster, Evaluation, Language Models, Word Embedding, word2vec

1 Introduction

Microblogging sites like Twitter and Weibo are vital sources of information during disaster [5, 16]. However, it has been observed that the important information is generally hidden among lots of conversational content (e.g., sympathy for the victims of the disaster). Therefore, automated IR methods need to be developed to extract those microblogs that contain specific types of situational information from the huge number of microblogs posted.

There have been efforts to extract specific types of microblogs (tweets) during disaster situations (see Section 2). However, all prior works have used their own datasets, and there has been little effort in developing a benchmark test collection for comparison and evaluation of various microblog retrieval methodologies. The main contribution of this work is to develop such a test collection.

We identified five practical information needs in a disaster scenario, in discussion with agencies who regularly participate in disaster relief operations (details in Section 3). We collected a set of about 50,000 distinct microblogs posted during a recent

⋆ Contact authors: Saptarshi Ghosh (`saptarshi.ghosh@gmail.com`), Kripabandhu Ghosh (`kripa.ghosh@gmail.com`)

disaster event – the Nepal earthquake in April 2015 – and following the Cranfield approach [2], employed human annotators to identify microblogs relevant to the said information needs (i.e., to develop the gold standard). Thus, the developed test collection can be used to calibrate microblog retrieval methodologies for addressing some practical information needs in a disaster scenario.

We also explore the use of two retrieval models (detailed in Section 4) on the developed dataset – (i) a standard language model, as implemented in the Indri IR system [12], and (ii) a word embedding based retrieval model (using word2vec [8]). We observe that word embedding based retrieval performs significantly better than the standard language model retrieval (Section 5). This superior performance of word embedding is primarily due to *context matching* between the query and the document (microblog), as opposed to keyword matching. The former is especially important for microblog retrieval due to the short, noisy nature of microblogs. We also compare between two strategies for query expansion via pseudo relevance feedback – the standard Rocchio expansion, and an expansion strategy based on word embeddings – and find that the two approaches perform comparably (with Rocchio performing slightly better).

To summarize, the present work has two contributions. First, we develop a test collection for evaluating microblog retrieval methodologies in the context of a disaster situation. Second, we establish that word embedding based retrieval is a promising approach for dealing with the short, noisy nature of microblogs.

2 Related Work

Developing test collections for evaluating IR strategies: The Text REtrieval Conference (TREC – `http://trec.nist.gov/`) was perhaps the first endeavor to present testbeds for standard evaluation of IR systems. They advocated the Cranfield style [2] which states that an IR test collection should comprise three components – (1) Text representations of information needs, called *topics*, (2) A static set of documents, and (3) Relevance status of the documents with respect to a query, called *relevance assessments*. We also adopt this style while preparing our dataset.

The TREC Microblog Track [6] focuses on the evaluation of microblog retrieval strategies in general. In contrast, in this paper, we look to evaluate IR systems on a test collection designed specifically for microblog retrieval in a real-life disaster situation.

Prior work on microblogs posted during disasters: There has been lot of recent interest in addressing various challenges on microblogs posted during disaster events, such as classification, summarization, event detection, and so on [5]. Some datasets of social media posts during disasters have also been developed [3], but they are primarily meant for evaluating methodologies for classification among different types of posts (and not for retrieval methodologies).

Few methodologies for retrieving specific types of microblogs have also been proposed, such as tweets asking for help, and tweets reporting infrastructure damage [1,15]. However, all such studies have used different datasets. To our knowledge, there is no standard test collection for evaluating strategies for microblog retrieval in a disaster scenario; this work attempts to develop such a test collection.

3 Developing the test collection

As stated earlier, our objective is to develop a test collection for microblog retrieval methodologies that can help agencies responding to a disaster situation. In this section, we describe how the test collection is developed.

3.1 Topics for retrieval

We consulted members of two NGOs[5] who regularly engage in post-disaster relief operations to know some of the typical information requirements during a disaster relief operation. Based on our discussions, we identified five critical topics (information needs) on which information needs to be retrieved, for efficient running of the relief operations. Table 1 states the five topics, in the format used traditionally for TREC topics. Each topic contains an identifier (num), and three fields ($title$, $desc$, and $narr$) describing the type of documents (microblogs) to be deemed relevant to the topic.[6]

Note that topic T3 (availability of medical resources) is intended to be a specific subset of topic T1 (availability of all resources). Similarly, topic T4 (requirement of medical resources) is intended to be a subset of topic T2 (requirement of all resources).

3.2 Tweet dataset

We considered a recent disaster event – the earthquake in Nepal on 25^{th} April 2015.[7] We collected tweets related to this event, through the Twitter Search API [14], using the keyword ‘nepal’. In total, we collected 100K English tweets (as identified by Twitter’s own language detection mechanism) posted during the couple of weeks after the earthquake.

It has been observed that multiple users on Twitter often re-post (retweet) the same information, possibly in slightly different language, leading to duplicate and near-duplicate tweets [13]. Duplicate documents are not desired in test collections for IR evaluation, as they can lead to over-estimation of the performance of IR methodologies, and also create information overload for human annotators (for developing the gold standard) [6]. Therefore, we removed duplicates / near-duplicates as follows.

We converted every tweet to a bag of words (after removing English stopwords and URLs), and the Jaccard similarity between the two bags (of words) was computed as the similarity between the two tweets. The tweets were considered in the chronological order in which they were posted, and each tweet was compared to the chronologically earlier tweets. If the similarity between two tweets was computed to be higher than a threshold value (chosen as 0.7), only the longer tweet was kept, assuming that the longer tweet will be the more informative one. After removing duplicates and near-duplicates, we got a set of *50,068 tweets*, which formed the test collection.

[5] Doctors For You (`doctorsforyou.org`) and SPADE (`www.spadeindia.org`)

[6] These topics have also been used in the FIRE 2016 track on Information Extraction from Microblogs Posted during Disasters [4].

[7] `https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake`

Table 1. Five topics portraying some critical information needs of agencies responding to disasters. Each topic is written in the format conventionally used in TREC tracks. These topics have also been used in the FIRE 2016 track on Information Extraction from Microblogs Posted during Disasters [4].

<num> **Number: T1** <title> **What resources were available** <desc> Identify the messages which describe the availability of some resources. <narr> A relevant message must mention the availability of some resource like food, drinking water, shelter, clothes, blankets, human resources like volunteers, resources to build or support infrastructure, like tents, water filter, power supply and so on. Messages informing the availability of transport vehicles for assisting the resource distribution process would also be relevant. However, generalized statements without reference to any resource or messages asking for donation of money would not be relevant.
<num> **Number: T2** <title> **What resources were required** <desc> Identify the messages which describe the requirement or need of some resources. <narr> A relevant message must mention the requirement / need of some resource like food, water, shelter, clothes, blankets, human resources like volunteers, resources to build or support infrastructure like tents, water filter, power supply, and so on. A message informing the requirement of transport vehicles assisting resource distribution process would also be relevant. However, generalized statements without reference to any particular resource, or messages asking for donation of money would not be relevant.
<num> **Number: T3** <title> **What medical resources were available** <desc> Identify the messages which give information about availability of medicines & other medical resources. <narr> A relevant message must mention the availability of some medical resource like medicines, medical equipments, blood, supplementary food items (e.g., milk for infants), human resources like doctors/staff and resources to build or support medical infrastructure like tents, water filter, power supply, ambulance, etc. Generalized statements without reference to medical resources would not be relevant.
<num> **Number: T4** <title> **What medical resources were required** <desc> Identify the messages which describe the requirement of some medicine or other medical resources. <narr> A relevant message must mention the requirement of some medical resource like medicines, medical equipments, supplementary food items, blood, human resources like doctors/staff and resources to build or support medical infrastructure like tents, water filter, power supply, ambulance, etc. Generalized statements without reference to medical resources would not be relevant.
<num> **Number: T5** <title> **What infrastructure damage and restoration were being reported** <desc> Identify the messages which contain information related to infrastructure damage or restoration. <narr> A relevant message must mention the damage or restoration of some specific infrastructure resources, such as structures (e.g., dams, houses, mobile tower), communication infrastructure (e.g., roads, runways, railway), electricity, mobile / Internet connectivity, etc. General statements without reference to infrastructure resources would not be relevant.

3.3 Developing gold standard for retrieval

We used three human annotators to develop the gold standard, each of whom is proficient in English and is a frequent user of Twitter, but none of whom is an author of this paper. Each annotator was given the set of tweets, and asked to identify all tweets relevant to each of the five topics. The tweets were indexed using the Indri IR system [12], which enabled the annotators to identify the tweets containing certain terms.

The gold standard was developed in two stages. In the *first stage*, each annotator identified relevant tweets *independently*, i.e., without consulting the other annotators. After the first phase, it was observed that, different annotators had judged quite different sets of tweets to be relevant to the same topic. This can be ascribed to the fact that different search-terms / queries were used by different annotators to retrieve tweets for the same topic. So, in the *second stage*, for a particular topic, all tweets that were judged relevant by at least one annotator were considered. The final set of tweets relevant to a topic was decided through mutual agreement among all three annotators.

The final gold standard contains the following number of tweets judged relevant to the five topics – T1: 589, T2: 301, T3: 334, T4: 112, T5: 254. Table 2 shows some examples of tweets that were judged relevant to each topic.

Note from Table 2 that keywords that are intuitively important for a particular topic are often *not* included in tweets that are relevant to the topic. For instance, many tweets relevant to topic T1 do *not* contain the term 'available' (or any of its morphological

Table 2. Examples of tweets judged relevant to each topic, by human annotators.

Topic T1: What resources were available
Indian railway send one lakh of water bottle to Nepal it will reach today night.
Hospital, Fire Brigade and Blood Banks #Nepal Eye Hospital 4250691 Tilganga Eye Hospital 4423684 #earthquake
If O+ve Blood is needed around Ilam, I am ready just mention. #NepalEarthquake
Topic T2: What resources were required
@nytimes plz send medicine and food packets to nepal if possible. #NepalEarthquake
Body bags, Tents, water, medicine, pain killers urgently needed in #earthquake stricken #Nepal
Nearly 1 million children require urgent humanitarian assistance after Nepal earthquake
Topic T3: What medical resources were available
Medicare National Hospital - Ambulance 4467067 Nepal Orthopaedic Hospital 4493725
#India to setup Field Hospital in #Nepal by tomorrow morning to provide medical facilities #NepalEarthquake
Dr. Madhur Basnet leading medical team going to remote villages of Gorkha dist which was epicenter of earthquake. His cell: [number]
Topic T4: What medical resources were required
There is shortage of Blood as well as oxygen cylinders...Nepal is in huge crisis.
List of meds needed in #nepal at Bir Hospital [url]
#Nepal #Earthquake Victims require #Orthopedic, Neuros, Anaesthetists & Paramedics. Plz volunteer.
Topic T5: What infrastructure damage and restoration were being reported
Kathmandu- Lamjung road cut off after earthquake. Follow live updates here: [url]
Historic Dharahara Tower in #Kathmandu, has collapsed #earthquake #Strength to the people affected.
Building Collapsed along with fallen electric pole in Golmadhi, Bhaktapur-7 [url]

variants), and many of the tweets relevant to topic T5 do *not* contain terms like ‘damage’ or ‘restore’. Hence, for a given topic, while retrieving some tweets might be relatively straightforward, many of the other tweets are more challenging to retrieve.

Note that our method for building the gold standard is different from the conventional pooling approach [11] that is used in TREC tracks, where some top-ranked documents retrieved by different systems, for a query, are pooled, and then only the documents appearing in the pool are considered for annotation. We believe that our approach, where the annotators had an opportunity to identify relevant tweets *from the whole collection* as opposed to a comparatively smaller pool, is likely to result in a more complete gold standard which is independent of the performance of the different competing IR methodologies. Hence, the developed gold standard can be utilized to estimate Recall of different methodologies.

3.4 The final test collection

The test collection developed as described above contains the 50,068 tweets, the five topics (information needs) and the gold standard specifying which tweets are relevant to each topic. The collection can be obtained via email from the contact authors.

4 Retrieval Methods

Now we describe some IR methodologies for retrieving microblogs relevant to the topics, and explore the empirical utility of the methodologies in contrast to each other, as tested on our dataset.

For a given topic, we consider three stages in the retrieval – (i) generating a query from the topic, (ii) retrieving and ranking microblogs with respect to the query, and (iii) expanding the query, and subsequently retrieving and ranking microblogs with respect to the expanded query.

We compare between two approaches for retrieving microblogs – (i) a baseline approach, using traditional IR techniques such as language model-based ranking (as implemented in the Indri system [12]) and Rocchio pseudo-relevance query expansion [7],

Table 3. Manually and automatically generated queries for the five topics. Each query is a set of unigrams selected from the text of the topics, and then stemmed. The terms in each query are written in alphabetical order.

Manually generated query	Automatically generated query
T1: What resources were available	
avail, blanket, cloth, distribut, filter, food, human, power, shelter, tent, transport, vehicl, volunt, water	avail, blanket, cloth, distribut, drink, process, shelter, transport, vehicl, volunt
T2: What resources were required	
blanket, cloth, filter, food, human, need, power, requir, shelter, tent, transport, vehicl, volunt, water	blanket, cloth, distribut, need, process, requir, shelter, transport, vehicl, volunt
T3: What medical resources were available	
ambul, avail, blood, doctor, equip, food, human, infant, item, medic, medicin, milk, staff	ambul, avail, blood, doctor, equip, infant, item, medic, medicin, milk, staff, supplementari
T4: What medical resources were required	
ambul, blood, doctor, equip, filter, medicin, need, requir, staff, tent, water	ambul, blood, doctor, equip, item, medic, medicin, requir, staff, supplementari
T5: What infrastructure damage and restoration were being reported	
connect, dam, damag, electr, hous, internet, mobil, mobile, railwai, restor, road, runwai, tower	commun, connect, dam, damag, electr, hous, internet, mobil, railwai, restor, road, runwai, specif, structur, tower

and (ii) a proposed approach which utilizes vector embedding of terms (word2vec [8]) for both ranking and query expansion.

4.1 Query generation from the topics

For a given topic, we consider the query to be a set of terms (unigrams) extracted from the text of the topic (stated in Table 1). We consider two approaches for extracting terms from the topic text.

Manual query generation: A human (an author of this paper) selected specific terms from the text of the topic, which are intuitively important and likely to be present in tweets relevant to the topic. Note that the annotators who prepared the gold standard were *not* consulted during the query generation process.

Automatic query generation: In this approach, terms are extracted automatically from the narrative part of the topics. Standard English stopwords are ignored, and terms which appear in the narrative of at least four out of the five topics (e.g., 'message', 'mention') are ignored. Additionally, the last sentence of the narrative, which mentions what type of tweets would *not* be relevant, is ignored. Next, an English part-of-speech tagger[8] is applied on the rest of the narrative text, and *nouns*, *verbs*, and *adjectives* (terms which are tagged as 'NN', 'VB', and 'JJ' by a POS tagger) are extracted.

For both the query generation methods, the selected terms are stemmed using the standard Porter stemmer, and the query is considered as a bag of the stemmed terms. Table 3 shows the manually generated and automatically generated queries (showing the terms obtained after stemming) for the five topics.

4.2 Microblog retrieval and ranking

We now describe two methodologies of retrieving microblogs for a given query. For both methodologies, the tweets are pre-processed by removing standard English stopwords, URLs and punctuation symbols, and all terms whose frequency in the entire

[8] The POS tagger included in the Python Natural Language Toolkit was used.

corpus is less than 5. The remaining terms are then stemmed using the standard Porter stemmer. Thus, each tweet is considered as a bag / set of terms (stemmed).

Retrieving microblogs using language modeling: We use the well-known Indri IR system [12], which implements a language modeling based retrieval model [9] as the baseline. All the tweets (after stopword removal and stemming) were indexed using Indri, and the query was used to retrieve and rank tweets using the default retrieval model of Indri.

Retrieving microblogs using word embeddings: We employ a word2vec [8] based retrieval model that is suitable for short documents like tweets.[9] We first trained word2vec over the pre-processed set of tweets. Specifically, the continuous bag of words model is used, along with Hierarchical softmax, and the following parameter values – Vector size: 2000, Context size: 5, Learning rate: 0.05. The word2vec model gives a vector (of dimension 2000, as decided by the parameters) for each term in the corpus, which we refer to as *term-vectors*.

The word2vec term-vector for a particular term is expected to capture the context in which the term has been used in the corpus. The term-vectors are additive, i.e., they can be added to capture the combined context of multiple terms [8]. Hence, for a given query, we derive a *query-vector* by summing the term-vectors of all terms in the query. Similarly, for each (pre-processed) tweet, we compute a *tweet-vector* by summing the term-vectors of all terms in the tweet. We expect the query-vectors and the tweet-vectors to capture the collective context of all the terms in the queries or the tweets.

For retrieving tweets for a given query, we compute the cosine similarity between the corresponding query-vector and each tweet-vector, and rank the tweets in decreasing order of the cosine similarity.

4.3 Query expansion

For query expansion, we consider two approaches for *pseudo (or blind) relevance feedback* [7] – (i) Rocchio expansion [7], and (ii) a method based on word2vec, as described below. For both approaches, we consider the 10 top-ranked tweets retrieved by the original query, and select $p = 5$ terms from these 10 top-ranked tweets to expand the query.

Rocchio expansion: We compute tf × idf Rocchio scores for each distinct term in the 10 top-ranked tweets retrieved by the original query, and the top p terms in decreasing order of Rocchio scores are used to expand the query.

Expansion using word2vec: After retrieving tweets relevant to the initial query, we identify a set of terms (within the 10 top-ranked tweets) that are most related to the context of the query, and use these terms to expand the query. As stated earlier, the query-vector is expected to capture the overall context of a query. Hence, we compute the cosine similarity of the query-vector with the term-vector of every distinct term in the top-ranked tweets, and select the p terms having the highest cosine similarity values.

We observed that the two query expansion strategies identify mostly different expansion terms for the same initial query and the same ranking model (details omitted

[9] The Gensim implementation for word2vec was used – `https://radimrehurek.com/gensim/models/word2vec.html`.

Table 4. Retrieval performance with initial queries (manual and automatic) and two ranking models – language model of Indri and word2vec-based model. The values for the performance measures have been averaged over the five topics. The performances by word2vec are statistically significantly better ($p < 0.05$) than that of Indri for both types of queries.

Query type	Ranking Model	Prec @20	Recall @1000	MAP @1000	MAP Overall
Manual	Indri	0.3900	0.5635	0.1285	0.1639
Manual	word2vec	**0.6700**	**0.6197**	**0.2343**	**0.2788**
Automatic	Indri	0.3000	0.4357	0.0891	0.1149
Automatic	word2vec	**0.4600**	**0.5591**	**0.1785**	**0.2242**

for brevity). The results of microblog retrieval using the different initial queries, ranking strategies, and query expansion strategies are reported in the next section.

5 Experimental results

This section discusses the performance of the IR methodologies described in the previous section. We consider the following measures to evaluate the performance of an IR methodology – (i) *Precision at 20* (Prec@20), (ii) *Recall at 1000* (Recall@1000), (iii) *Mean Average Precision at 1000* (MAP@1000), and (iv) *Overall MAP* considering the full retrieved ranked list.

5.1 Retrieval for initial queries

Table 4 compares the performance of the two ranking models – the language model-based ranking of Indri, and the word2vec-based ranking – for the two types of queries (manually generated and automatically generated). The values reported are averaged over all the five topics. For both types of queries, the word2vec-based model performs statistically significantly better than the baseline Indri model, at 95% confidence level (p-value < 0.05) by Wilcoxon signed-rank test [10]. In fact, even when the retrieval performances for individual topics are considered, the word2vec model out-performs the Indri model by a considerable margin for most of the topics.

Explaining the better retrieval by word2vec: There are two important factors that lead to the superiority of the word2vec retrieval model over the Indri model. *First*, the Indri model could *not* distinguish well between tweets which inform about requirements (topics T2 and T4), and tweets which inform about availability of resources (topics T1 and T3). Both these classes of tweets use similar terms (like ‘food’, ‘water’, ‘blood’), and we speculate that language modeling based approaches (as used by Indri) fail to distinguish between the two classes. Whereas, the word2vec model captures the overall context of a tweet, and hence could better distinguish between need and availability tweets. *Second*, we observe that a significant fraction of the microblogs relevant to a topic do *not* contain any of the terms in the corresponding query. The word2-vec based model could identify such tweets much better than the Indri model.

Table 5. Retrieval performance (averaged over the five topics) for queries expanded using two strategies – Rocchio and Word2Vec-based. For comparison, the performances with the initial queries (as was reported in Table 4) are also shown in the gray-colored rows.

Query type	Ranking Model	Query Expansion	Prec @20	Recall @1000	MAP @1000	MAP Overall
Manual	Indri	No expansion	0.3900	0.5635	0.1285	0.1639
Manual	Indri	Rocchio	0.3500	0.5532	0.1233	0.1598
Manual	Indri	word2vec	0.3900	0.5518	0.1193	0.1591
Manual	word2vec	No expansion	0.6700	0.6197	0.2343	0.2788
Manual	word2vec	Rocchio	**0.6500**	**0.6281**	**0.2441**	**0.2873**
Manual	word2vec	word2vec	0.6400	0.6080	0.2242	0.2689
Automatic	Indri	No expansion	0.3000	0.4357	0.0891	0.1149
Automatic	Indri	Rocchio	0.2800	0.4349	0.0842	0.1088
Automatic	Indri	word2vec	0.2700	0.4871	0.1006	0.1273
Automatic	word2vec	No expansion	0.4600	0.5591	0.1785	0.2242
Automatic	word2vec	Rocchio	**0.5000**	**0.5680**	**0.1838**	**0.2300**
Automatic	word2vec	word2vec	0.4900	0.5482	0.1792	0.2265

5.2 Retrieval for expanded queries

Finally, we compare the retrieval performance for the expanded queries, expanded either by the Rocchio strategy or by the word2vec-based strategy. Table 5 reports the retrieval performance for the different combinations, considering the two types of initial queries (manual / automatic), the two ranking models (Indri and word2vec) and the two query expansion strategies (Rocchio and word2vec). For easier comparison, the best performances obtained with the initial queries (as was reported in Table 4) are repeated in Table 5 (the gray-colored rows).

For the manually generated queries, the retrieval performance is better for the initial queries than for the expanded queries. This is possibly because the manually generated queries were verbose and already contained most of the relevant terms, which led to saturation, and further attempts of improvement by addition of more query-terms has resulted in query drift. In case of the automatic queries and the word2vec ranking model, the retrieval is better for the expanded queries than for the initial queries, thus demonstrating the utility of query expansion in finding relevant terms missing in the initial automatic queries.

We also see that, for a given query and ranking model, the Rocchio expansion strategy performs slightly better than the word2vec-based expansion strategy, though the differences are *not statistically significant* (p-value > 0.05). For both the automatic query and the manual query, the best retrieval is obtained by using the word2vec ranking model and Rocchio query expansion, as highlighted using boldface in Table 5.

6 Concluding Discussion

This work makes available to the community a novel test collection for evaluating microblog retrieval strategies for practical information needs in a disaster situation. Our experiments also demonstrate the value of context-based matching over keyword-based matching in microblog retrieval, as the former approach is better at retrieving relevant documents without any keywords appearing in the query (as is often the case for short microblogs).

The IR methodologies reported in this work achieved a precision of 0.670 and a relatively low MAP score of 0.278, which highlights the challenges in microblog retrieval during disasters. We believe that the contributed test collection will help the research community to develop better models for microblog retrieval in future.

Acknowledgement: The authors thank the anonymous reviewers whose suggestions helped to improve the paper. This research was partially supported by a grant from the Information Technology Research Academy (ITRA), DeITY, Government of India (Ref. No.: ITRA/15 (58)/Mobile/DISARM/05).

References

1. AIDR - Artificial Intelligence for Disaster Response, https://irevolutions.org/2013/10/01/aidr-artificial-intelligence-for-disaster-response/
2. Cleverdon, C.: The cranfield tests on index language devices. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
3. CrisisLex: Crisis-related Social Media Data and Tools, http://crisislex.org/
4. Ghosh, S., Ghosh, K.: Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters. In: Working notes for the 2016 Conference of the Forum for Information Retrieval Evaluation. CEUR Workshop Proceedings (2016)
5. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing Social Media Messages in Mass Emergency: A Survey. ACM Computing Surveys 47(4), 67:1–67:38 (Jun 2015)
6. Lin, J., Efron, M., Wang, Y., Sherman, G., Voorhees, E.: Overview of the TREC-2015 Microblog Track (2015)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
8. Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: NAACL HLT 2013 (2013)
9. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. ACM SIGIR. pp. 275–281 (1998)
10. Siegel, S.: Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill series in psychology, McGraw-Hill (1956)
11. Sparck Jones, K., van Rijsbergen, C.: Report on the need for and provision of an ideal information retrieval test collection. Tech. Rep. 5266, Computer Laboratory, University of Cambridge, UK (1975)
12. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proc. ICIA. Available at: http://www.lemurproject.org/indri/ (2004)
13. Tao, K., Abel, F., Hauff, C., Houben, G.J., Gadiraju, U.: Groundhog Day: Near-duplicate Detection on Twitter. In: Proc. World Wide Web (WWW) (2013)
14. Twitter Search API, https://dev.twitter.com/rest/public/search
15. Varga, I., et al.: Aid is out there: Looking for help from tweets during a large scale disaster. In: Proc. ACL (2013)
16. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In: Proc. ACM SIGCHI (2010)