Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters

Saptarshi Ghosh
Department of CST,
IIEST Shibpur, India
sghosh@cs.iiests.ac.in

Kripabandhu Ghosh
Indian Statistical Institute,
Kolkata, India
kripa.ghosh@gmail.com

ABSTRACT

The FIRE 2016 Microblog track focused on retrieval of microblogs (tweets posted on Twitter) during disaster events. A collection of about 50,000 microblogs posted during a recent disaster event was made available to the participants, along with a set of seven practical information needs during a disaster situation. The task was to retrieve microblogs relevant to these needs. 10 teams participated in the task, submitting a total of 15 runs. The task resulted in comparison among performances of various microblog retrieval strategies over a benchmark collection, and brought out the challenges in microblog retrieval.

CCS Concepts

•**Information systems** → **Query reformulation;**

Keywords

FIRE 2016; Microblog track; Microblog retrieval; Disaster

1. INTRODUCTION

Microblogging sites such as Twitter (https://twitter.com) have become important sources of *situational information* during disaster events, such as earthquakes, floods, and hurricanes [2, 11]. On such sites, a lot of content is posted during disaster events (in the order of thousands to millions of tweets), and the important situational information is usually immersed in large amounts of general conversational content, e.g., sympathy for the victims of the disaster. Hence, automated IR techniques are needed to retrieve specific types of situational information from the large amount of text.

There have been few prior attempts to develop IR techniques over microblogs posted during disasters, but there has been little effort till now to develop a benchmark dataset / test collection using which various microblog retrieval methodologies can be compared and evaluated. The objectives of the FIRE 2016 Microblog track are two-fold – (i) to develop a test collection of microblogs posted during a disaster situation, which can serve as a benchmark dataset for evaluation of microblog retrieval methodologies, and (ii) to evaluate and compare the performance of various IR methodologies over the test collection. The track is inspired by the TREC Microblog Track [4] which aims to evaluate microblog retrieval strategies in general. In contrast, the FIRE 2016 Microblog Track focuses on microblog retrieval in a disaster situation.

In this track, a collection of about 50,000 microblogs posted during a recent disaster event was made available to the participants, along with a set of seven practical information needs that are faced in a disaster situation by the agencies responding to the disaster. Details of the collection are discussed in Section 2. The task was to retrieve microblogs relevant to the information needs (see Section 3. 10 teams participated in the track, submitting a total of 15 runs that are described in Section 4). The runs were evaluated against a gold standard developed by human assessors, using standard measures like Precision, Recall, and MAP.

2. THE TEST COLLECTION

In this section, we describe how the test collection for the Microblog track was developed. Following the Cranfield style [1], we describe the creation of topics (information needs), document set (here, microblogs or tweets) collection and relevance assessment to prepare the gold standard necessary for evaluation of IR methodologies.

2.1 Topics for retrieval

In this track, our objective was to develop a test collection to evaluate IR methodologies for extracting information (from microblogs) that can potentially help responding agencies to respond to a disaster situation such as an earthquake or a flood. To this end, we consulted members of some NGOs who regularly work in disaster-affected regions – such as, Doctors For You (http://doctorsforyou.org/) and SPADE (http://www.spadeindia.org/) – to know what are the typical information requirements during a disaster relief operation. They identified certain information needs such as *what resources are required / available* (especially medical resources), *what infrastructure damages are being reported*, *the situation at specific geographical locations*, *the ongoing activities of various NGOs and government agencies* (so that the operations of various responding agencies can be coordinated), and so on. Based on their feedback, we identified seven topics on which information needs to be retrieved during a disaster.

Table 1 states the seven topics which we have developed as a part of the test collection. These topics are written in the format conventionally used for TREC topics.[1] Each topic contains an identifying number (*num*), a textual representation of the information need (*title*), a brief description (*desc*) of the same and a more detailed narrative (*narr*) explaining what type of documents (tweets) will be considered relevant to the topic, and what type of tweets would not be considered relevant.

[1] trec.nist.gov/pubs/trec6/papers/overview.ps.gz

<num> **Number: FMT1** <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: FMT2** <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: FMT3** <title> **What medical resources were available** <desc> Identify the messages which give some information about availability of medicines and 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: FMT4** <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: FMT5** <title> **What were the requirements / availability of resources at specific locations** <desc> Identify the messages which describe the requirement or availability of resources at some particular geographical location. <narr> A relevant message must mention both the requirement or availability of some resource, (e.g., human resources like volunteers/medical staff, food, water, shelter, medical resources, tents, power supply) as well as a particular geographical location. Messages containing only the requirement / availability of some resource, without mentioning a geographical location would not be relevant.
<num> **Number: FMT6** <title> **What were the activities of various NGOs / Government organizations** <desc> Identify the messages which describe on-ground activities of different NGOs and Government organizations. <narr> A relevant message must contain information about relief-related activities of different NGOs and Government organizations in rescue and relief operation. Messages that contain information about the volunteers visiting different geographical locations would also be relevant. However, messages that do not contain the name of any NGO / Government organization would not be relevant.
<num> **Number: FMT7** <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 or Internet connectivity, etc. Generalized statements without reference to infrastructure resources would not be relevant.

Table 1: The seven topics (information needs) used in the track. Each topic is written following the format conventionally used in TREC tracks (containing a number, title, description and narrative). The task is to retrieve microblogs relevant to these topics.

2.2 Tweet dataset

We collected a large set of tweets related to the devastating earthquake that occurred in Nepal and parts of India on 25^{th} April 2015.[2] We collected tweets using the Twitter Search API [10], using the keyword 'nepal', that were posted during the two weeks following the earthquake. We collected only tweets in English (based on language identification by Twitter itself), and collected about 100K tweets in total.

Tweets often contain duplicates and near-duplicates since the same information is frequently retweeted / re-posted by multiple users [9]. However, duplicates are not desirable in a test collection for IR, since the presence of duplicates can result in over-estimation of the performance of an IR methodology. Additionally, the presence of duplicate documents also creates information overload for human annotators while developing the gold standard [3]. Hence, we removed duplicate and near-duplicate tweets using a simplified version of the methodologies discussed in [9], as follows.

Each tweet was considered as a bag of words (excluding standard English stopwords and URLs), and the similarity between two tweets was measured as the Jaccard similarity between the two corresponding bags (sets) of words. If the Jaccard similarity between two tweets was found to be higher than a threshold value (0.7), the two tweets were considered near-duplicates, and only the longer tweet (potentially more informative) was retained in the collection. After removing duplicates and near-duplicates, we obtained a set of *50,068 tweets*, which was used as the test collection for the track.

2.3 Developing gold standard for retrieval

Evaluation of any IR methodology requires a gold standard containing the documents that are actually relevant to the topics. As is the standard procedure, we used human annotators to develop this gold standard. A set of three human annotators were used, each of whom is proficient in English and is a regular user of Twitter, and has prior experience of working with social media content posted during disasters. The development of gold standard involved three phases.

Phase 1: Each annotator was given the set of 50,068 tweets,

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

and the seven topics (in TREC format, as stated in Table 1). Each annotator was asked to identify all tweets relevant to each topic, *independently*, i.e., without consulting the other annotators. To help the annotators, the tweets were indexed using the Indri IR system [8], which helped the annotators to search for tweets containing specific terms. For each topic, the annotators were asked to think of appropriate search-terms, retrieve tweets containing those search terms (using Indri), and to judge the relevance of the retrieved tweets.

After the first phase, we observed that the set of tweets identified to be relevant to the same topic by different annotators, was considerably different. This difference was because different annotators used different search-terms to retrieve tweets.[3] Hence, we conducted a second phase.

Phase 2: In this phase, for a particular topic, all tweets that were judged relevant by *at least one* annotator (in the first phase) were considered. The decision whether a tweet is relevant to a topic was finalised through discussion among all the annotators and mutual agreement.

Phase 3: The third phase used standard pooling [7] (as commonly done in TREC tracks) – the top 30 results of all the submitted runs were pooled (separately for each topic), and judged by the annotators. In this phase, all annotators were judging a common set of tweets, hence inter-annotator agreement could be measured. There was agreement among all annotators for over 90% of the tweets; for the rest, the relevance was decided through discussion among all the annotators and mutual agreement.

The final gold standard contains the following number of tweets judged relevant to the seven topics – FMT1: 589, FMT2: 301, FMT3: 334, FMT4: 112, FMT5: 189, FMT6: 378, FMT7: 254.

2.4 Insights from the gold standard development process

Through the process described above, we understood that for any of the topics, there are several tweets which are definitely relevant to the topic, but which were difficult to retrieve *even for human annotators*. This is evident from the fact that, many of the relevant tweets could initially be retrieved by only one out of the three annotators (in the first phase), but when the tweets were shown to the other annotators (in the second phase), they unanimously agreed that the tweet was relevant. These observations highlight the challenges in microblog retrieval.

Note that our approach for developing the gold standard is different from that used in TREC tracks, where the gold standard is usually developed by pooling few top-ranked documents retrieved by different submitted systems, and then annotating these top-ranked documents [7]. In other words, only the third phase (as described above) is applied in TREC tracks.

Given that it is challenging to identify many of the tweets relevant to a topic (as discussed above), annotating only a relatively small pool of documents retrieved by IR methodologies has the potential risk of missing many of the relevant documents which are more difficult to retrieve. We believe that our approach, where the annotators viewed the entire dataset instead of a relatively small pool, is likely to be more robust, and is expected to have resulted in development of a more complete gold standard which is irrespective of the performance of any IR methodology.

[3]Since the different annotators retrieved and judged very different sets of tweets, it is not meaningful to report inter-annotator agreement in this case.

3. DESCRIPTION OF THE TASK

The participants were given the tweet collection and the seven topics described earlier. It can be noted that the Twitter terms and conditions prohibit direct public sharing of tweets. Hence, only the tweet-ids[4] of the tweets were distributed among the participants, along with a Python script using which the tweets can be downloaded via the Twitter API.

The participants were invited to develop IR methodologies for retrieving tweets relevant to the seven topics. The participants were asked to submit a ranked list of tweets that they judge relevant to each topic. The ranked list was evaluated based on the gold standard (developed as described earlier) using the following measures: (i) *Precision at 20* (Prec@20), i.e., what fraction of the top-ranked 20 results are actually relevant according to the gold standard, (ii) *Recall at 1000* (Recall@1000), i.e., what fraction of all tweets relevant to a topic (as identified in the gold standard) is present among the top-ranked 1000 results, (iii) *Mean Average Precision at 1000* (MAP@1000), and (iv) *Overall MAP* considering the full retrieved ranked list. Out of these, we only report the Prec@20 and MAP measures (in the next section).

The track invited three types of methodologies – (i) Automatic, where both query formulation and retrieval are automated, and (ii) Semi-automatic, where manual intervention is involved in the query formulation stage (but not in the retrieval stage), and (iii) Manual, where manual intervention is involved in both query formulation and retrieval stages.

15 runs were submitted by the participants, out of which, one run was fully automatic, while the others were semi-automatic. The methodologies are summarized and compared in the next section.

4. METHODOLOGIES

Ten teams participated in the FIRE 2016 Microblog track. A summary of the methodologies used by each team is given in the next sub-section. Table 2 shows the evaluation performance of each submitted run, along with a brief summary. For each type, the runs are arranged in the decreasing order of the primary measure, i.e., *Precision@20*. In case of a tie, the arrangement is done in the decreasing order of *MAP*.

4.1 Method summary

We now summarize the methodologies adopted in the submitted runs.

- **dcu_fmt16**: This team participated from ADAPT Centre, School of Computing, Dublin City University, Ireland. It used WordNet[5] to perform synonym-based query expansion and submitted the following two runs:

    1. *dcu_fmt16_1*: This is an *Automatic* run (i.e. no manual step involved). First, the words in <*title*> and <*narr*> were considered, from which the

[4]Twitter assigns a unique numeric id to each tweet, called the tweet-id.

[5]https://wordnet.princeton.edu/

Run Id	Precision@20	MAP	Type	Method summary
dcu_fmt16_1	0.3786	0.1103	Automatic	WordNet, Query Expansion
iiest_saptarashmi_bandyopadhyay_1	0.4357	0.1125	Semi-automatic	Correlation, NER, Word2Vec
JU_NLP_1	0.4357	0.1079	Semi-automatic	WordNet, Query Expansion, NER, GloVe
dcu_fmt16_2	0.4286	0.0815	Semi-automatic	WordNet, Query Expansion, Relevance Feedback
JU_NLP_2	0.3714	0.0881	Semi-automatic	WordNet, Query Expansion, NER, GloVe, word bags split
JU_NLP_3	0.3714	0.0881	Semi-automatic	WordNet, Query Expansion, NER, GloVe, word bags split
iitbhu_fmt16_1	0.3214	0.0827	Semi-automatic	Lucene default model
relevancer_ru_nl	0.3143	0.0406	Semi-automatic	Relevancer system, Clustering Manual labelling, Naive Bayes classification
daiict_irlab_1	0.3143	0.0275	Semi-automatic	Word2vec, Query Expansion, equal term weight
daiict_irlab_2	0.3000	0.0250	Semi-automatic	Word2vec, Query Expansion, unequal term weights, WordNet
trish_iiest_ss	0.0929	0.0203	Semi-automatic	Word-overlap, POS tagging
trish_iiest_ws	0.0786	0.0099	Semi-automatic	WordNet, wup score, POS tagging
nita_nitmz_1	0.0583	0.0031	Semi-automatic	Apache Nutch 0.9, query segmentation, result merging
Helpingtech_1 (on 5 topics)	0.7700	0.2208	Semi-automatic	Entity and action verbs relationships, Temporal Importance
GANJI_1, GANJI_2, GANJI_3 (Combined) (on 3 topics)	0.8500	0.2420	Semi-automatic	Keyword extraction, Part-of-speech tagger, Word2Vec, WordNet, Terrier, Retrieval, Classification, SVM

Table 2: Comparison among all the submitted runs. Runs which attempted retrieval only for a subset of the topics are listed separately at the end of the table.

stopwords were removed. Thus the initial query was formed. Then, for each word in the query, the synonyms were added using WordNet, resulting in the expanded query. Retrieval was done from this expanded query using the BM25 model [6].

2. *dcu_fmt16_2*: This is a *Semi-automatic* run (i.e. manual step was involved). First an initial ranked list was generated using the original topic. From the top 30 tweets, 1-2 relevant tweets were manually identified and query expansion was done from these relevant tweets. The expanded query was further expanded using WordNet just as done for *dcu_fmt16_1*. This final expanded query was used for retrieval.

- **iiest_saptarashmi_bandyopadhyay**: This team participated from Indian Institute of Engineering Science and Technology, Shibpur, India. It submitted one *Semi-automatic* run described below:
  - *iiest_saptarashmi_bandyopadhyay_1*: Correlation between the topic words and the tweet was calculated and this value determined the relevance score for a given topic-tweet pair. The Stanford NER tagger[6] was used to identify the LOCATION, ORGANIZATION and PERSON names in the tweets. For each topic, some keywords were manually selected on which a number of tools (e.g., PyDictionary, NodeBox toolkit etc.) were used to find the corresponding synonyms, inflectional variants etc. The bag of words for each topic was further converted into a vector using Word2Vec package.[7] Finally, the relevance score was calculated from the correlation between the vector representations of the topic word bags and the tweet text.

- **JU_NLP**: This team participated from Jadavpur University, India. It submitted three *Semi-automatic* runs described as below:

  1. *JU_NLP_1*: This run was generated by using word embeddings. For each topic, relevant words were manually chosen and expanded using the synonyms obtained from NLTK WordNet toolkit. In addition, past, past participle and present continuous forms of verbs were obtained using the NodeBox library for Python. For the topics FMT5 and FMT6, location and organization information was extracted using Stanford NER tagger. GloVe[5] model was trained on the twitter collection. A tweet vector, as well as, a query vector was formed by taking the normalized summation of the vector (obtained from GloVe) of the constituent words. Then for each query-tweet pair,

[6]nlp.stanford.edu/software/Stanford-ner-2015-04-20.zip

[7]https://deeplearning4j.org/word2vec

the similarity score was calculated by the cosine similarity of the corresponding vectors.

2. *JU_NLP_2*: This run is similar to *JU_NLP_1* except that here word bags were split categorically and average similarity between the tweet vector and the split topic vectors was calculated.
3. *JU_NLP_3*: This is identical to *JU_NLP_2*.

- **iitbhu_fmt16**: This team participated from Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, India. It submitted one *Semi-automatic* run – *iitbhu_fmt16_1* described as follows:
  - *iitbhu_fmt16_1*: The Lucene[8] default similarity model, which combines Vector Space Model (VSM) and probabilistic models (e.g., BM25), was used to generate the run. StandardAnalyzer, which handled names and email address and lowercased each token, and removed stopwords and punctuations, was used. The query formulation stage involved manual intervention.
- **daiict_irlab**: This team participated from DAIICT, Gandhinagar, India and LDRP, Gandhinagar, India. It submitted two *Semi-automatic* runs described as follows:
  1. *daiict_irlab_1*: This run was generated using query expansion, where the 5 similar words and hashtags from the Word2vec model, trained on the tweet corpus, were added to the original query. Equal weight was assigned to each term.
  2. *daiict_irlab_2*: This run was generated in the same way as *daiict_irlab_1* except that different weights were assigned to the expanded terms than the original terms. More weights were assigned to the words like *required* and *available*. These terms were also expanded using WordNet.
- **trish_iiest**: This team participated from Indian Institute of Engineering Science and Technology, Shibpur, India. It submitted two *Semi-automatic* runs described below:
  1. *trish_iiest_ss*: The similarity score between a query and a tweet is the word-overlap between them, normalized by the query length. In each topic, the nouns, identified by the Stanford Part-Of-Speech Tagger, were selected to form the query. In addition, more weight is assigned on words like *availability* or *requirement*.
  2. *trish_iiest_ws*: For this run, *wup*[9] score is calculated on the synsets of each term obtained from WordNet.
- **nita_nitmz**: This team participated from National Institute of Technology, Agartala, India and National Institute of Technology, Mizoram. It submitted one *Semi-supervised* run described as below:
  - *nita_nitmz_1*: This run was generated on Apache Nutch 0.9. Search was done using the different combination of words present in the query. The results obtained from different combinations of query were merged.
- **Helpingtech**: This team participated from Indian Institute of Technology, Patna, Bihar, India and submitted the following *Semi-automatic* run (on 5 topics only):
  - *Helpingtech_1*: For each query, relationships entities and action verbs were defined through manual inspection. The ranking score was calculated on the basis of the presence of these pre-defined relationships in the tweet for a given query. More importance was given to a tweet which indicated immediate action than a one which indicated a proposed action for future.
- **GANJI**: This team participated from Èvora University, Portugal. It submitted three retrieval results (*GANJI_1, GANJI_2, GANJI_3*) for the first three topics only using *Semi-automatic* methodology, described below:
  - *GANJI_1, GANJI_2, GANJI_3 (combined)*: First, keyword extraction was done using Part-of-speech tagger, Word2Vec (to obtain the *nouns*) and WordNet (to obtain the *verbs*). Then, retrieval was performed on Terrier[10] using the BM25 model. Finally, SVM classifier was used to classify the retrieved tweets into *available*, *required* and *other* classes.
- **relevancer_ru_nl**: This team participated from Radboud University, the Netherlands and submitted the following *Semi-automatic* run:
  - *relevancer_ru_nl*: This run was produced by a tool *Relevancer*. After a pre-processing step, the tweet collection was clustered to identify *coherent* clusters. Each such cluster was manually labelled by some experts as relevant or non-relevant. This training data was used for Naive Bayes based classification. For each topic, the test tweets predicted as relevant by the classifier were submitted.

5. CONCLUSION AND FUTURE DIRECTIONS

The FIRE 2016 Microblog track successfully created a benchmark collection of microblogs posted during disaster events, and compared the performance of various IR methodologies over the collection.

In subsequent years, we hope to conduct extended versions of the Microblog track, where the following extensions can be considered:

- Instead of just considering binary relevance (where a tweet is either relevant to a topic or not), graded relevance can be considered, e.g., based on factors like how important or actionable the information contained in the tweet is, how useful the tweet is likely to be to the agencies responding to the disaster, and so on.

[8]https://lucene.apache.org/(2016,August20)

[9]http://search.cpan.org/dist/WordNet-Similarity/lib/WordNet/Similarity/wup.pm

[10]http://terrier.org

- The challenge in this year's track considered a static set of microblog. But in reality, microblogs are obtained in a continuous stream. The challenge can be extended to retrieve relevant microblogs dynamically, e.g., as and when they are posted.

It can be noted that even the best performing method submitted in the track achieved a relatively low MAP score of 0.24 (considering only three topics), which highlights the difficulty and challenges in microblog retrieval during a disaster situation. We hope that the test collection developed in this track will help development of better models for microblog retrieval in future.

Acknowledgements

The track organizers thank all the participants for their interest in this track. We also acknowledge our assessors, notably Moumita Basu and Somenath Das, for their help in developing the gold standard for the test collection. We also thank the FIRE 2016 organizers for their support in organizing the track.

6. REFERENCES

[1] C. Cleverdon. The cranfield tests on index language devices. In K. Sparck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[2] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 47(4):67:1–67:38, June 2015.

[3] J. Lin, M. Efron, Y. Wang, G. Sherman, and E. Voorhees. Overview of the TREC-2015 Microblog Track. Available at: https://cs.uwaterloo.ca/~jimmylin/publications/Lin_etal_TREC2015.pdf, 2015.

[4] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 Microblog Track. Available at: http://trec.nist.gov/pubs/trec20/papers/MICROBLOG.OVERVIEW.pdf, 2011.

[5] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[6] S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[7] K. Sparck Jones and C. van Rijsbergen. *Report on the need for and provision of an ideal information retrieval test collection.* Tech. Rep. 5266, Computer Laboratory, University of Cambridge, UK, 1975.

[8] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proc. ICIA*. Available at: http://www.lemurproject.org/indri/, 2004.

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

[10] Twitter Search API. https://dev.twitter.com/rest/public/search.

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