Overview of the FIRE 2018 track: Information Retrieval from Microblogs during Disasters (IRMiDis)

Moumita Basu[1,2], Saptarshi Ghosh[1,4], and Kripabandhu Ghosh[3]

[1] Indian Institute of Engineering Science and Technology, Shibpur, India
[2] University of Engineering & Management, Kolkata, India
[3] Indian Institute of Technology, Kanpur, India
[4] Indian Institute of Technology, Kharagpur, India
saptarshi@cse.iitkgp.ernet.in

Abstract. In countries like the US, European countries, Australia and Japan, user-generated content from microblogging sites is extensively used for crowdsourcing actionable information during disasters. However, there has been limited work in this direction in India. Moreover, there has been a limited attempt to verify the credibility of the information extracted from microblogs from other reliable sources. To this end, the FIRE 2018 Information Retrieval from Microblogs during Disasters (IRMiDis) track focused on the identification of factual or fact-checkable tweets and supporting news article for each fact-checkable tweets. The data consists of around $50,000$ microblogs (tweets) from Twitter and $6,000$ news articles, that were posted during the Nepal earthquake in April 2015. There were two tasks. The first task (Task 1) was to identify factual or fact-checkable tweets and the second task (Task 2) was to identify supporting news articles for fact-checkable tweets.

Keywords: FIRE 2018; Microblog track; Multi-source data; Disaster

1 Introduction

Microblogging sites like Twitter are increasingly being used for aiding relief operations during various mass emergencies. However, critical actionable information is often immersed in the deluge of insignificant conversational contents. Hence, automated methodologies are needed to extract the important information from microblogs during such an event [4]. Moreover, messages posted on microblogging sites often contain rumors and overstated facts. In such situations, identification of factual or fact-checkable tweets, i.e., tweets that report some relevant and verifiable fact (other than sympathy or prayer) is extremely important for effective coordination of post disaster relief operations. Additionally, cross verification of such critical information is a practical necessity to ensure the trustworthiness. Online news articles are more reliable source of information than microblogs. Hence, the credibility of the information extracted from microblogs can be verified from other reliable sources like online news articles. Thus, automated IR

fact-checkable
ibnlive:Nepal earthquake: Tribhuvan International Airport bans landing of big aircraft [url]
@mashable some pictures from Norvic Hospital *A Class Hospital of nepal* Patients have been put on parking lot.
@siromanid: Many temples in UNESCO world heritage site Bhaktapur Durbar Square have been reduced 2 debris after recent earthquake [url]
non-fact-checkable
Students of Himalayan Komang Hostel are praying for all beings who lost their life after earthquake!!! Please do...[url]
We humans need to come up with a strong solutions to create earthquake proof zone's.
Shocked to oversee the outcome of Massive earthquake..., let's create a Help wave in support to the affected people..

Table 1. Examples of fact-checkable and non-fact-checkable posted during a recent disaster event (2015 Nepal earthquake).

techniques are needed to identify, process and verify the credibility of information from multiple sources.

To address the aforesaid issues, we organized the FIRE 2018 IRMiDis task. The track had two tasks, as described below.

Task 1: Identifying factual or fact-checkable tweets: Here the participants needed to develop automatic methodologies for identifying fact-checkable tweets. This is mainly a classification problem, where tweets are classified into two classes fact-checkable tweets and non-fact-checkable tweets. However, apart from classification, the problem of identifying fact-checkable tweets can also be viewed as a pattern matching problem or an IR problem. Table 1 shows some examples of fact-checkable tweets and non-fact-checkable tweets from thedataset that consists of about 50,000 tweets posted during the 2015 Nepal earthquake (details of dataset given in Section 2).

Task 2: Identification of supporting news articles for fact-checkable tweets: A fact-checkable tweet is said to be supported/verified by a news article if the same fact is reported by both the media. In this task, the participants were asked to develop methodologies for matching fact-checkable tweets with appropriate news articles. Table 2 shows some examples of fact-checkable tweets and extracts from news article that verifies/supports the fact reported in the tweet, posted during the 2015 Nepal earthquake from the dataset that was made available to the participants (described in the next section).

For each fact-checkable tweet, participants should report - (i) the supporting news article id, and (ii) the particular sentence in the news article, which supports the fact-checkable tweet. It should be noted that many of the fact-checkable tweets might not have supporting news articles in the dataset.

Examples of Fact-checkable tweet	**Headline of news article**	**Extract from supporting news article**	**Url of news site**
ibnlive:Nepal earthquake: Tribhuvan International Airport bans landing of big aircraft [url]	President Sirisena expresses condolences to earthquake victims in Nepal	Tribhuvan International Airport in Katmandu, Nepal has been closed for all commercial flights. Only flights carrying relief are allowed into its runways.	http://newsfirst.lk/english/2015/04/president-sirisena-expresses-condolences-to-earthquake-victims-in-nepal/91798
#Nepal #Earthquake day four. Slowly in the capital valley Internet and electricity being restored. A relief for at least some ones	Protests over poor relief as Nepal toll crosses 5,000 (Roundup)	Four days after the deadly quake, more shops opened here and traffic returned to Kathmandu's roads. Authorities also restored electricity while telephones began to function in more areas.	http://www.business-standard.com/article/news-ians/protests-over-poor-relief-as-nepal-toll-crosses-5-000-roundup-115042901022_1.html
Many temples in UNESCO world heritage site Bhaktapur Durbar Square have been reduced 2 debris after recent earthquake [url]	Nepal earthquake: Over 1,800 dead	Historical monuments such as Dharhara and Basantapur Durbar Square and Patan Durbar Square have been completely destroyed by the tremors	http://www.business-standard.com/article/news-ians/nepal-earthquake-over-1-800-dead-115042600075_1.html
@MSF_canada: UPDATE: We're now sending 8 teams to Nepal including highly skilled emergency surgical teams[url]	US Pledges $1 Million, Relief Teams to Nepal After Earthquake	Doctors Without Borders sent eight medical teams and four arrived the same day. The teams include a surgical team composed of eight highly skilled MSF staff members to set up surgical units and mobile clinics	http://www.breitbart.com/national-security/2015/04/26/us-pledges-1-million-relief-teams-to-nepal-after-earthquake/

Table 2. Examples of fact-checkable tweets and corresponding news article that verifies/supports the fact, posted during the 2015 Nepal earthquake

2 The test collection

In this track, our motivation was to develop a test collection containing microblogs for evaluating–

- Methodologies for identifying specific type of actionable situational information – factual or fact-checkable tweets, and
- Methodologies for identification of supporting news articles for fact-checkable tweets

The detail description of the test collection development procedure of IRMiDis track is described in this section.

2.1 Multi-source dataset

In the present task, we included both microblogs (tweets) and news articles in our dataset. We reused the tweet collection of $50,018$ English tweets related to the Nepal earthquake that occurred on 25^{th} April 2015[5] developed and released as part of the same track in FIRE 2016 [3]. This collection is also utilized to evaluate several IR methodologies developed by ourselves and other researchers [1, 2].

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

Additionally, we introduced a collection of $6,000$ news articles, that were posted during the Nepal earthquake in April 2015. We used Radian6 tool[6] to search for news articles posted during the two weeks after the earthquake, using the query term ‘nepal’. The dataset contains tweets/articles in English only.

2.2 Developing gold standard for identification of fact-checkable tweets

We employed pooling for the current task. We pooled top 100 results from each run and involved a set of three human annotators having proficiency in English, who are regular users of Twitter, and had previous experience of working with social media content posted during disasters. We asked the annotators to judge the fact-checkability of the tweets and independently. Annotators observed that there were different types of fact-checkable tweets, based on how definite the factual information reported in the tweet is. Hence we decided to adopt a graded gold standard. The graded gold standard development process is as described below –
Annotators were asked to grade the fact-checkable tweets in three levels and scores were assigned as 1, 2, 3 depending on the definiteness of facts reported in the fact-checkable tweets.

- **Grade 1:** Grade ‘1’ depicts tweets are containing factual information but without Nepal-related information i.e., about some location outside Nepal
- **Grade 2:** Grade ‘2’ signifies fact-checkable tweets containing Nepal-related information. However, the factual information is generic and very definite i.e specific resource name, quantity, organizations are not reported by the tweet
- **Grade 3:** Grade ‘3’ signifies highly fact-checkable tweets having specific a reference of source, location, organization, quantity, resource name etc.

The grade of the rest of the tweets is assigned as 0, that signifies tweets are non-fact-checkable. The final set of graded tweets in different categories was decided through a mutual agreement among all three annotators.

The summary of the numbers of graded fact-checkable tweets present in the final gold standard is reported in Table 3 along with the example of each category of tweets.

2.3 Identification of supporting news articles for fact-checkable tweets

In Task-2 of IRMiDis track this year, only one run was submitted. Thus, pooling was employed on only one run to create the gold-standard. By checking the overlapping with the gold standard developed as a part of Task 1 (Identifying fact-checkable tweets) it was noticed that the supporting new-articles were reported against 40 correctly identified fact-checkable tweets (according to the gold standard of Task 1) only. Hence, human assessors were employed and asked

[6] https://socialstudio.radian6.com

Category of tweets	**Count**	**Examples**
Grade 1	99	@ndtv 5.1 Magnitude Earthquake 10 Km From Mirik, WB, India at 6pm 70 KM from siliguri, siligurians are not safe #saveslg
Grade 2	254	Nepal earthquake: Aid material, doctors and relief workers from near countries began arriving [url]
Grade 3	968	@ArtofLivingABC Distributing food in Tudikhel - Relief work for Nepal #earthquake by #ArtofLiving [url]

Table 3. Summary of the gold standard used in FIRE 2018 IRMiDis Task 1

to manually inspect each of the relevant fact-checkable tweets and corresponding matching news-article reported by the run and decide whether matching is correct or not. Accordingly, the evaluation metrics were calculated.

3 Task 1: Identifying fact-checkable tweets

In IRMiDis track this year 6 teams participated in Task 1 and nine *automatic* and three *semi-automatic* runs were submitted. The different methodologies developed by the participating teams are summarized and described in the following sub-section.

3.1 Methodologies

- **MIDAS**: This team from Indraprastha Institute of Information Technology Delhi (IIIT-D) submitted the one *semi-automatic* and one *automatic* run. For both the runs, tweets were pre-processed by removing punctuations, stop-words and emojis. Hence, pre-processed tweets were POS tagged.
  - *MIDAS_1 (automatic)*: The run used proper nouns and numbers present in the tweets as features. Factuality score was calculated by the average of two scores namely PROPN and NUM, where PROPN is the proper noun count in each tweet and NUM is the count of numbers present in each tweet. The scores were normalized by dividing each of the counts by maximum count of the corresponding features across the dataset.
  - *MIDAS_2 (semi-automatic)*: Around 1500 tweets were manually labeled to train the classifier, features were extracted using cbow and bi-gram models then fastText classification algorithm was used to classify the tweets. Tweets were ranked according to the confidence score provided by the classifier.
- **FAST-NU**: This team partook from, FAST National University Karachi Campus, Pakistan. It formulated the task as an Information Retrieval problem and submitted three *automatic* runs. It used the set of 6000 news articles introduced as a part of the dataset in the current track. It considered both

string similarity(α) and cosine similarity(β) to rank the tweets. It used Aho-Corasick algorithm to compute string similarity(α). Details of runs were illustrated as below:

- *FAST_NU_Run1* : Computed α between the tweet and news headlines and β between news content and tweets
- *FAST_NU_Run2* : This run computed α between hashtags extracted from tweets and news headlines
- *FAST_NU_Run3* : Combination of previous two approaches as described above. A tweet was considered as a factual tweet if it has a supporting news article both from *FAST_NU_Run1* and *FAST_NU_Run2*.

– **UEM-DataMining-CSE** : This team from University of Engineering and Management, Kolkata, India, submitted two *automatic* runs. Tweets were pre-processed and POS tagged [7]to extract proper nouns from the tweets. Both the runs were generated by using SVM classifier with linear kernel to classify the tweet. Used bag-of-words as feature extraction algorithm.

- *UEM_DataMining_CSE_run1*: SelectKbest feature selection algorithm was used to select top perfoming 10000 proper nouns as features.
- *UEM_DataMining_CSE_run2* : Used TfidfVectorizer algorithm to select top-ranked 6000 proper nouns (according to tf*idf score) as features .

– **iitbhu_irlab_irmidis_hm**: This team is from Indian Institute of Technology (BHU) Varanasi, India. It submitted the one *automatic* run described as follows:

- *iitbhu_irlab_irmidis_hm_r1*: This run trained word2vec model with $50,000$ pre-processed tweets in default settings. Created a tf*idf based ranked list of terms from 84 ground truth tweets provided as a part of the dataset in present task. Hence, a weighted function of tf*idf score and word-embedding was used to rank the tweets.

– **DAIICT-Hildesheim**: This team participated from, Hildesheim University, Germany and Dhirubhai Ambani Institute of Information and Communication Technology, Gujrat, India. It submitted three*automatic* runs. Tweets were pre-processed by removing @string_value, RT, and URLs. Among these first two runs (DAIICT-Hildesheim-mod1-sif, DAIICT-Hildesheim-mod1-nosif): used Recursive Neural Network based approach to obtain the semantic label of the tweets using Stanford semantic analysis library [5]. Word embeddings for first two runs were created by training the model with Nepal earthquake dataset. Afterwards, the term vectors of the proposed model were replaced by the term vectors obtained from the pre-trained model (by Google-News dataset), if any term was co-occuring in both the models. Cosine similarities between fact-checkable tweets (provided as labels with the dataset) and the rest were used to rank the tweets.

- *DAIICT-Hildesheim-mod1*: Sentence vector was computed by taking sum of all the term vectors present in a sentence and then dividing by the length of the sentence to take the average. Hence, first principle component is multiplied with each sentence vector.

[7] https://gate.ac.uk/wiki/twitter-postagger.html

  - *DAIICT-Hildesheim-mod1*: Sentence vector is calculated in the same way as of the first run(DAIICT-Hildesheim-mod1-sif).
  - *DAIICT-Hildesheim-mod3* (semi-automatic): This run used a Convolution Neural Network based classifier. CNN was intialized with the GloVe pre-trained vectors. The classifier was trained with 1700 tweets labeled as subjective/non-fact-checkable tweets and 2000 tweets labeled as objective/fact-checkable tweets with 10-fold cross validation.

- **iitbhu_irlab_irmidis_ab**: This team participated from Indian Institute of Technology (BHU) Varanasi, India. It submitted one *Semi-Automatic* run described as follows:
  - *iitbhu_irlab_irmidis_ab_2* (semi-automatic): This run trained a doc2vec model for representing each tweet as a 50 dimension vector. Manually observed the datasets and randomly labeled few fact-checkable tweets. Used crystallization of the dataset. SVM Classifier is used to classify the tweets.

3.2 Evaluation Measures and Result

The performance of the methodologies submitted to the Task 1 of FIRE 2018 IR-MiDis track are illustrated in this section. We considered NDCG as the primary measure for evaluation. Ranking of runs are based on NDCG scores. However, we noted the following measures as well to evaluate the performance – (i) **Precision at 100 (Precision@100)**: what fraction of the top ranked 100 results are actually relevant according to the gold standard, i.e., what fraction of the retrieved tweets are actually fact-checkable tweet (ii) **Recall at 1000 (Recall@1000)**: fraction of relevant tweets (according to the gold standard) that are in the top 1000 retrieved tweets (iii) **NDCG at 100 (NDGC@100)**: considering ranking upto top 100 retrieved tweets (iv)) **NDCG (NDCG Overall)**: considering the full retrieved ranked list Table 4 reports the retrieval performance for all the submitted runs in Task 1. Each of the measures (i.e. Precision@100, Recall@1000, NDCG@100, NDCG Overall) are reported.

It is observed that simple NLP and classification based approaches performed better than the other methodologies based on word-embeddings as is evident from the Table 4.

4 Task 2: Identification of supporting news articles for fact-checkable tweets

In Task 2, one team participated and one *semi-automatic* run was submitted. Description of the run is as follows–

Run Id	Type	Precision @100	Recall @1000	NDCG @100	NDCG Overall	Method summary
MIDAS_1	Automatic	0.8800	0.1292	0.5649	0.6835	POS tagging, Normalized sum of proper noun count (PROPN) & number count (NUM)
FAST_NU_Run2	Automatic	0.7000	0.0885	0.5723	0.6676	String similarity between hashtags & news headlines , Cosine similarity, AhoCorasick algorithm
UEM_DataMining_CSE_run2	Automatic	0.6800	0.1427	0.5332	0.6396	POS tagging, TfidfVectorizer, SVM classifier (linear kernel)
UEM_DataMining_CSE_run1	Automatic	0.6400	0.1069	0.5237	0.5276	POS tagging, SelectKbest feature selection algorithm, SVM classifier (linear kernel)
iitbhu_irlab_irmidis_hm_r1	Automatic	0.9300	0.1938	0.8645	0.4532	tf*idf score & Word embedding using word2ec
FAST_NU_Run1	Automatic	0.2700	0.0670	0.2009	0.3208	String similarity between tweets & news headlines, Cosine similarity, AhoCorasick algorithm
FAST_NU_Run3	Automatic	0.2800	0.0566	0.1785	0.3105	Combining methology of FAST_NU_Run1 & FAST_NU_Run2
DAIICT-Hildesheim-mod1	Automatic	0.1500	0.0670	0.0930	0.1330	Stanford semantic analysis library Word embeddings,First principle component
DAIICT-Hildesheim-mod2	Automatic	0.0100	0.0670	0.0033	0.1271	Stanford semantic analysis library Word embeddings
DAIICT-Hildesheim-mod3	Semi-Automatic	0.4000	0.2002	0.4021	0.7492	GloVe pre-trained vector, CNN based Classifier 10-fold cross validation
MIDAS_2	Semi-Automatic	0.9600	0.1148	0.6007	0.6899	feature extraction using cbow and bi-gram models, fastText classifier
iitbhu_irlab_irmidis_ab_2	Semi-Automatic	0.3900	0.0447	0.3272	0.6200	Word embedding using doc2vec model, Crystallization,SVM Classifier

Table 4. Comparison among all the submitted runs in Task 1 (Identifying fact-checkable tweets). The primary measure for graded relevance is NDCG. Hence, the table is sorted according to the NDCG measure

4.1 Methodology

iitbhu_irlab_irmidis_hm : This team is from Indian Institute of Technology (BHU) Varanasi, India. It submitted one *Semi-Automatic* run described as follows: It utilised Apache Lucene, a open source Java-based text search engine library[8].News articles and tweets were pre-processed by stopwords, hashtags and addressing removal, stemming (porter stemmer) and case-folding. Then, headline and the first three sentences of each news article were combined to form the test documents and each pre-processed tweet was used as a query. Tweets were categorized according to the score returned by Lucene search engine.

4.2 Evaluation Measures and Result

Only one run was submitted in Task 2 and the run could retrieve only 40 fact-checkable tweets according to the gold standard developed for Task 1. We employ pooling, though on only one run. Thus the human annotators only checked the relevance of the news articles retrieved for these 40 tweets. Thus, for each of the 40 fact-checkable tweets identified, total how many news articles were identified, and out of that how many were judged to be correct (i.e, the news article sentence that was retrieved actually verified the information contained in the tweet) needed to be evaluated. Hence, we have evaluated the run according to the measure Precision@N described as below:
Precision@N: for each fact-checkable tweet, out of N retrieved supporting articles, how many are correctly identified.
The performance of the submitted run is evaluated as **0.9378 (Precision@N)**. It is evident that term-based matching could produce good result. However, it is to be noted that the result is evaluated only for 40 fact-checkable tweets. Hence, it may be concluded that those fact-checkable tweets were easy to match. For rest of the tweets other methodologies needs to be explored.

5 Conclusion and Future Directions

The FIRE 2018 IRMiDis track successfully created a benchmark collection of fact-checkable tweets posted during disaster events with graded relevance. The track also compared the performance of various methodologies in identifying fact-checkable tweets and matching the fact-checkable tweet with supporting news articles. We hope that the test collection developed in this track will help the research community in the development of a better model for retrieval and matching in future. Moreover, Task 2 did not have much participation this year. Hence, the problem of matching the fact-checkable tweet with supporting news articles needs to be explored more in subsequent years.

[8] https://lucene.apache.org/

Acknowledgements

The track organizers thank all the participants for their interest in this track. We acknowledge our annotators for their sincere effort in developing the ground truth. We also thank the FIRE 2018 organizers for their support in organizing the track.

References

1. Basu, M., Ghosh, K., Das, S., Dey, R., Bandyopadhyay, S., Ghosh, S.: Identifying Post-Disaster Resource Needs and Availabilities from Microblogs. In: Proc. ASONAM (2017)
2. Basu, M., Roy, A., Ghosh, K., Bandyopadhyay, S., Ghosh, S.: Microblog retrieval in a disaster situation: A new test collection for evaluation. In: Proc. Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP) co-located with European Conference on Information Retrieval. pp. 22–31 (2017)
3. Ghosh, S., Ghosh, K.: Overview of the FIRE 2016 microblog track: Information extraction from microblogs posted during disasters. In: Working notes of FIRE 2016 (2016)
4. 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)
5. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (October 2013), http://www.aclweb.org/anthology/D13-1170