Microblog Retrieval for Disaster Relief: How To Create Ground Truths?

Ribhav Soni and Sukomal Pal

IIT(BHU) Varanasi
{ribhav.soni.cse13,spal.cse}@iitbhu.ac.in

Abstract. Microblogging services like Twitter are an important source of real-time information during disasters and can be utilized to aid rescue, relief and rehabilitation efforts. The focus of this work is on the creation of gold standard data for automatic retrieval of helpful tweets. Using various experiments on the gold standard data prepared in the FIRE 2016 Microblog Track [3], we show that the gold standard data prepared in [3] missed many relevant tweets. We also demonstrate that using a machine learning model can help in retrieving the remaining relevant tweets by training an SVM model on a subset of the data and using it to get the most useful tweets in the entire dataset. We obtain high precision and recall even with very little training data, which makes such a model suitable for use in a real-time disaster situation.

Keywords: Crisis Informatics, Disaster, Emergency, Hazards, Microblog Retrieval, Social Media, Text Categorization.

1 Introduction

Social media is a very useful resource for obtaining real-time information during disasters. Traditional media like television, newspaper, etc. have limited use for aiding in disaster relief due to their slow updates, and may even be unavailable due to the disaster event. In such situations, social media presents valuable information to aid in disaster relief and rehabilitation with very little time overhead [1].

Twitter in particular is especially suited for extracting details and first-hand accounts within moments of an event, anywhere in the world [6], and can thus be exploited for help in relief work. However, it also involves challenges of filtering out information about the crisis situation that is not useful for relief efforts, including tweets expressing shock, condolences, opinion, etc. Some tweets that are not useful for disaster relief efforts are shown in Table 1.

The FIRE 2016 Microblog Track [3] focused on comparing different IR methodologies for retrieval in such scenario, and led to the creation of a benchmark collection of ground truth data for such tasks. However, based on our experiments, we argue that the ground truth annotation exercise missed up to four times as many tweets as were found. This represents a significant loss of information that could potentially be very useful in a disaster situation. Also, since the accuracy of

Table 1. Some examples of tweets that are not useful for disaster relief efforts

Tweet Text
RT @tarsem_insan:,@Gurmeetramrahim Guru ji #MSGHelpEarthquakeVictims I m also Shocked!!!,hearing #earthquake #MSGHelpEarthquakeVictims
RT @vrinda_90:,really sad to hear about d earthquake. praying for all the ppl who suffered,& lost their loved ones. hope they get all the h
The Government is,so quick to help earthquake victims but why are they so reluctant to our own,farmers needs?
Haven't studied anything coz of earthquake and have to go for exam.
RT @guthali2:,Imagine Kejriwal were the PM in Nepal Earthquake situation, " Hum kuch,nai kar sakte hai jee, army president ke neeche hai".

gold standard data is crucial for evaluation and comparison of retrieval systems, it may lead to weaker systems being ranked above better systems.

First, we manually labeled a small, random subset of the data and found that many relevant tweets were missing from the gold standard in [3]. We then proceeded to train an SVM model on a subset of the data, and used it to retrieve 100 tweets with the highest confidence score of the trained model. We found that, averaged across all topics, only less than half of the relevant tweets among those were identified in the gold standard in [3].

We also performed bootstrapping on the labeled random subset to estimate the number of relevant tweets in the entire collection, and obtained about 5 times the relevant tweets from the gold standard in [3]. Also, we trained our SVM model on small fractions of the training data, and obtained high precision and recall even with very little training data, which shows that such a model can be used effectively in disaster situations with very low time overhead.

The rest of this paper is organized as follows. We first describe the data used in Section 2, our experiments and results in Section 3, and discussion and future work in Section 4.

2 Data

We used the dataset provided by the organizers of the FIRE 2016 Microblog Track [3]. The data was a collection of 50,068 tweets posted during the earthquake in Nepal in 2015 [1].

Organizations involved in relief work during disasters need specific, actionable information to help in the relief efforts. Thus, a set of seven specific information needs were identified by the authors in [3] after consulting members of such organizations.

The task in [3] involved retrieving tweets relevant to each of these seven information needs, expressed as topics in TREC format. The seven topics are listed in Table 2.

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

Table 2. The seven topics (information needs) used in FIRE 2016 Microblog Track [3]

< *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 flter, 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.

The gold standard preparation in [3] involved three phases, which can be briefly summarized as follows.

1. Three annotators independently tried to search for relevant tweets using intuitive keywords, after all tweets were indexed using Indri.

2. All tweets identified by at least one of the three annotators in Phase 1 were considered and their relevance annotation finalized by mutual discussion among the annotators.

3. Standard pooling was employed, taking the top 30 results from each run and deciding on their relevance.

The initial collection by the authors of [3] consisted of about 100,000 tweets, and the final dataset of 50,068 tweets was obtained by removing duplicate tweets (tweets with similarity greater than a threshold). The collection still included many tweets that were not duplicates but expressed almost the same information. All such instances were classified as relevant in the annotation exercise.

3 Experiments and Results

3.1 Exhaustive labeling on a small, random subset

A set of 700 tweets was randomly chosen, and relevance was judged for each tweet in the set separately for each of the seven topics. Within the random sample, the number of relevant tweets identified in the gold standard in [3] and those identified by exhaustive labeling are given in Table 3.

Table 3. Number of tweets in the sample of 700 tweets identified in the gold standard in [3] and in manual labeling by us

Topic	Gold Standard	Manual Labeling
FMT1	7	43
FMT2	4	12
FMT3	5	10
FMT4	1	4
FMT5	4	9
FMT6	5	53
FMT7	3	28
Any of the topics	22	105

As we can see, within the random sample, the number of relevant tweets identified by our exhaustive annotation was about 5 times of that identified in the gold standard in [3].

3.2 Bootstrapping to estimate the number of relevant documents in the entire collection

After exhaustively labeling the random sample of 700 tweets, we used Bootstrapping [2] for estimating the number of relevant tweets in the whole collection. Bootstrapping is a resampling method that involves random sampling with replacement, so we generated 1000 samples, each of size 700 tweets, from our sample of 700 tweets with replacement. The number of relevant tweets in each sample was computed, and then its average was taken across all 1000 samples. The resulting number of tweets, divided by the sample size, was taken to be an estimate for the fraction of relevant tweets in the entire collection. We thus estimated the number of relevant tweets in the collection of 50,068 tweets to be about 7,520 tweets (i.e., 15.02% of the tweets).

On the contrary, only 1,565 relevant tweets (3.13% of the tweets) were identified in the gold standard in [3]. This represents a loss of about 6,000 useful tweets missed by the annotators in [3].

3.3 Machine Learning for automatic filtering of tweets

We trained machine learning models for automatic classification of tweets into topics, with the aim of automatically retrieving the most useful tweets that may have been missed in the annotation exercise in [3]. As one tweet can be relevant to multiple topics, we applied supervised machine learning models separately for each topic, thus training a total of seven binary classifiers.

We used Support Vector Machines (SVM) for our classification task, as they have been found to be among the best models for text classification [4] [5]. We used the implementation of LinearSVC (SVM with linear kernel) in the scikit-learn machine-learning library [7].

Training data As seen in Table 3, we could identify at most only 53 relevant tweets for one topic out of a sample of 700 tweets. Thus, the classification task is highly skewed, with non-relevant tweets forming a large majority.

To overcome the problems associated with such skewed classification, we used undersampling, i.e., we balanced the training data by taking only as many non-relevant tweets as we had relevant tweets.

Besides the positively labeled tweets that we labeled from our sample of 700 tweets, we also had the set of relevant gold standard tweets from [3] to use for our machine learning task. Table 4 lists the final number of labeled tweets that we used for each of the topics. (Our number of gold standard tweets are slightly less than in the original gold standard because we could not download about 500 tweets from the original collection from twitter due to those tweets getting deleted in the meantime. Also, the number of relevant tweets from the two sources, manual labeling by us of the sample of 700 tweets and gold standard in [3], do not add up perfectly, because some tweets are common between them.)

We applied minimal preprocessing on the tweets. The only operation that we applied was the removal of hashtag symbols (retaining the attached text).

We randomly divided the available training data into 70% for training and 30% for testing, for each topic.

Table 4. Number of relevant tweets for each topic (from a combination of manual labeling and the gold standard), with the same number of non-relevant tweets added to make the data balanced

Topic	Manual Labeling	Gold Standard	Total relevant	Non-relevant tweets added	Total labeled examples used
FMT1	43	579	615	615	1230
FMT2	12	290	298	298	596
FMT3	10	334	336	336	672
FMT4	4	110	113	113	226
FMT5	9	187	192	192	384
FMT6	53	373	421	421	842
FMT7	28	253	278	278	556

Feature Extraction Scikit-learn's CountVectorizer was used to extract token counts with a bag-of-words model. We experimented using (1) unigram features only, and (2) both unigram and bigram features, and got better results using unigram features only. We thus used only unigram features for all our remaining experiments. Also, no stemming or stopword removal was done, and tokenization of tweets was done by extracting words of at least 2 letters.

Then, TfidfTransformer was used to convert the raw counts to tf-idf weights. Thus, a bag-of-words model with unigram features of tf-idf weights was used.

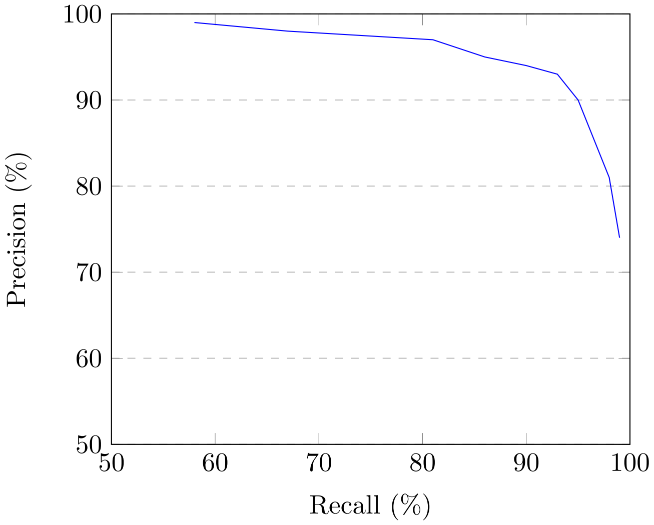
Each experiment was carried out for 100 iterations with random partitions of the data in each iteration to training (70%) and test sets (30%), and the average of all performance metrics for the 100 iterations was taken.

Results The performance of the classifiers based on various metrics are shown in Table 5. The precision-recall curve of the classifier for topic FMT1 is also shown.

Table 5. Peformance of the seven binary classifiers based on various metrics (all in percentage)

Classifier for	Accuracy	Accuracy for +1	Accuracy for -1	Precision	Recall	F1 score
FMT1	92.72	92.83	92.64	92.56	92.83	92.67
FMT2	93.15	92.81	93.55	93.45	92.81	93.09
FMT3	95.23	93.99	96.48	96.35	93.99	95.14
FMT4	91.94	90.68	93.32	93.06	90.68	91.74
FMT5	90.91	88.47	93.46	92.95	88.47	90.57
FMT6	90.01	89.06	91.05	90.88	89.06	89.91
FMT7	91.22	90.49	92.04	91.89	90.49	91.13

Precision-Recall curve for the SVM classifier for topic FMT1



3.4 Classification performance with number of examples

We tested the performance of our classifiers when using only a fraction of the available data. For each classifier and each given fraction of data, we randomly took a subset of the usable data for 100 iterations, and took the average of the performance scores for the classifier on the 100 iterations. The F1 scores of the classifiers with varying fractions of the data are shown in Table 6.

Table 6. F1 scores of the classifiers with varying percentage of available labeled examples used (all in percentage)

Percentage of labeled examples used	FMT1	FMT2	FMT3	FMT4	FMT5	FMT6	FMT7
10	81.58	76.18	79.7	69.96	64.24	73.35	66.12
20	85.89	84.34	87.9	70.33	74.47	82.37	77.54
30	88.03	86.78	90.64	78.71	80.72	85.72	82.48
40	89.15	88.67	92.08	82.79	83.28	86.59	84.32
50	90.19	90.51	93.07	85.89	86.26	87.92	86.83
60	90.64	90.66	93.56	87.99	87.32	88.08	87.72
70	91.52	91.42	94.26	88.84	88.64	89.13	88.32
80	91.74	92.22	94.43	90.49	89.52	89.71	89.67
90	92.36	92.68	94.77	91.32	89.96	89.54	90.65
100	92.67	93.09	95.14	91.74	90.57	89.91	91.13

3.5 Retrieving most relevant tweets in the entire collection

We used the trained classifiers to retrieve the 100 most relevant tweets for each topic in the entire dataset by taking the 100 tweets with the maximum confidence scores of each classifier.

We manually checked the sets of 100 tweets corresponding to the seven topics to determine how many of them were actually relevant, and how many of the relevant ones were identified by the gold standard in [3]. The results of this exercise are shown in Table 7.

Table 7. Number of tweets out of 100 that were actually relevant, and among them the number of tweets that were identified in the gold standard in [3]

Topic	Actually relevant	Marked in Gold Standard	Percentage of relevant tweets marked in Gold Standard
FMT1	80	43	53.75
FMT2	73	48	65.75
FMT3	92	57	61.96
FMT4	62	33	53.23
FMT5	65	22	33.85
FMT6	84	23	27.38
FMT7	94	32	34.04
	78.57 %		47.14 %

4 Discussion and Future Work

We showed that the gold standard annotation exercise in [3] missed many relevant tweets, even with a three-phase approach. Some major reasons why this happened may be:

1. Tweets are very short and noisy, and often relevant tweets do not contain the terms/keywords that one might intuitively expect for a given topic. Thus, the annotators could not find all relevant tweets using keyword searches in Phase 1.

2. Pooling works only when the number of participating systems is large, and the systems are diverse. Unlike tracks on TREC, the number of participants in [3] was not large, and so standard pooling employed in Phase 3 also failed to find all relevant tweets. ([9] studies the reliability of pooling, and concludes that it is reliable if the depth of the pool is deep enough, i.e., many of the top results from all systems are taken into account, which is true for TREC with a depth of top 100 documents from each participating system, but taking only top 30 documents as was done in [3] may not have been enough.)

Since exhaustive annotation is not possible for the complete collection, to find relevant tweets in the remaining collection, a machine learning model as presented in this paper can be trained and used on the remaining data to retrieve the tweets with the highest confidence scores, and then manual confirmation of the relevance can be carried out for as many tweets as annotator time permits.

Another approach could be to exhaustively annotate a small random subset of the data, and then use keywords of the relevant-marked tweets to query into the entire collection, to retrieve relevant tweets in the remaining collection. This is one future possibility for us to experiment with.

Some of the relevant tweets that were missed in the creation of gold standard in [3] are listed in Table 8.

Table 8. Some tweets that were missed in the gold standard in [3] but were found by our ML models

Topic	Tweet Text
FMT1	Earthquake Relief Distribution: Distributed Relief materials to the earthquake victims of Tukcha-1 (Pandy-Rai... http://t.co/0VlGHeFF4p
	Delhi Govt has decided to send 25,000 packets of food and 25,000 pouches of drinking water as immediate relief for the people in Nepal
FMT2	RT @worldtoiletday: Nepal earthquake: Urgent need for water, #sanitation and food: http://t.co/uOb6Hq81pY #NepalEarthquake @UNICEF @UN_Water
	UN agency stresses urgent funding needs to get food to earthquake victims http://t.co/xkn26ab08h
FMT3	RT Bloodbanks #Nepal Hospital and Research Centre 4476225 Norvic Hospital 4258554 #NepalEarthquake #MNTL #India
	WOREC and NAWHRD team are mobilized to Kavre and Bhaktapur districts to provide relief to the earthquake victims and survivors.
FMT4	RT @FocusNewsIndia: #NepalEarthquake — #Nepal PM Sushil Koirala requests for urgent blood donation for victims rescued from #earthquake htt
	#Nepal #Earthquake: Death toll could reach 10,000, says PM Sushil Koirala — Appeals for foreign supplies of tents and medicines.
FMT5	Tomorrow, We are moving to Hansapur VDC of Gorkha District to provide relief materials to the earthquake... http://t.co/GYZiT3eyip
	At Shanupalati village, Barabise district. Please retweet. Free Clinic Nepal earthquake Relief.
FMT6	RT @HDLindiaOrg: #RSS sends 20k swamsewaks to Nepal. GOI sent 4 Tonnes relief material, Team of doctors, NDRF, JCBs, food, water, medicines
	#ArtofLiving Nepal Centre providing shelter to 100's of ppl. Volunteers providing food & water #NepalEarthquakeRelief http://t.co/15RmABe2vO
FMT7	RT @PDChina: The rubble of Hanumndhoka Durbar Square, a @UNESCO world #heritage site, was badly damaged by earthquake in Kathmandu http://t
	Historic Dharahara tower collapses in Kathmandu after earthquake http://t.co/ZeovAnQESi

We were able to achieve reasonably high F1 scores for our classifiers even with a training size of a few hundred examples (Table 6). This shows that automatic text classification is a viable approach to extract useful information from tweets during times of disasters, since a few hundred examples can easily be annotated in a short amount of time. It may also be fruitful to train supervised machine learning models in advance for different types of disaster situations, and use them in times of disaster until newly annotated data is obtained.

To improve on the machine learning model, some avenues to explore are:

- using more features, including word embeddings, spatio-temporal features, linguistic features (as used in [8]), etc.
- employing better preprocessing techniques, like using twitter-specific spelling correction, expanding common twitter abbreviations, better data cleaning, etc.

5 Acknowledgements

We thank the anonymous reviewers for their thorough comments.

References

1. Internet becomes a lifeline in nepal after earthquake. http://www.computerworld.com/article/2914641/internet/internet-becomes-a-lifeline-in-nepal-after-earthquake.html, accessed: 2017-03-16
2. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
3. Ghosh, S., Ghosh, K.: Overview of the fire 2016 microblog track: Information extraction from microblogs posted during disasters. Working notes of FIRE pp. 7–10 (2016)
4. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. pp. 137–142. Springer (1998)
5. Khan, A., Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. Journal of advances in information technology 1(1), 4–20 (2010)
6. Mills, A., Chen, R., Lee, J., Raghav Rao, H.: Web 2.0 emergency applications: How useful can twitter be for emergency response? Journal of Information Privacy and Security 5(3), 3–26 (2009)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
8. Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., Ghosh, S.: Extracting situational information from microblogs during disaster events: a classification-summarization approach. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 583–592. ACM (2015)
9. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 307–314. ACM (1998)