Practical Extraction of Disaster-Relevant Information from Social Media

Muhammad Imran*
University of Trento
imran@disi.unitn.it

Shady Elbassuoni
American University of Beirut
se58@aub.edu.lb

Carlos Castillo
Qatar Computing Research Institute
chato@acm.org

Fernando Diaz
Microsoft Research
fdiaz@microsoft.com

Patrick Meier
Qatar Computing Research Institute
pmeier@qf.org.qa

ABSTRACT

During times of disasters online users generate a significant amount of data, some of which are extremely valuable for relief efforts. In this paper, we study the nature of social-media content generated during two different natural disasters. We also train a model based on conditional random fields to extract valuable information from such content. We evaluate our techniques over our two datasets through a set of carefully designed experiments. We also test our methods over a non-disaster dataset to show that our extraction model is useful for extracting information from socially-generated content in general.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

Social Media; Information Filtering; Information Extraction

1. INTRODUCTION

Microblogging platforms have become an important way to share information on the Web, especially during time-critical events such as natural and man-made disasters. In recent years, Twitter has been used to spread news about casualties and damages, donation efforts and alerts, including multimedia information such as videos and photos [1, 3]. Given the importance of on-topic tweets for time-critical situational awareness, disaster-affected communities and professional responders may benefit from using an automatic system to extract relevant information from Twitter.

We propose a two-step method for disaster-related information extraction: (i) classification of tweets and (ii) extraction from tweets. The classification step is based on our earlier work [8]; the extraction step is the focus of this paper. Both steps are done using off-the-shelf free software [6, 7], yielding a system that is easy to implement and that according to our experiments has good performance.

*Work done while the author was at QCRI.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

The rest of the paper is organized as follows. Section 2 describes our information-extraction method, which is evaluated in Section 3. Section 4 shows that our method can be applied also in non-disaster settings. In Section 5, we briefly outline related works, and conclude in Section 6.

2. DESCRIPTION OF OUR APPROACH

This section describes the classification and extraction steps of our method. For clarity of the exposition and concreteness, we begin by describing the datasets we use.

2.1 Datasets

We use two datasets related to recent emergencies:

Joplin 2011: 206,764 tweets collected during the tornado that struck Joplin, Missouri (USA) on May 22, 2011. Researchers at the University of Colorado at Boulder collected the dataset through Twitter's API using the hashtag[1] `#joplin`.

Sandy 2012: 140,000 tweets collected during the Hurricane Sandy, that hit Northeastern US on Oct 29, 2012. The dataset was collected using the hashtags `#sandy`, `#nyc`.

2.2 Classification

As the messages generated during a disaster are extremely varied, an automatic system needs to start by filtering out messages that do not contribute to valuable information. These include those that are entirely of personal nature and those not relevant to the crisis at hand. Specifically, we start by separating messages into two main classes classes:

- **Personal:** if a message is only of interest to its author and her immediate circle of family/friends and does not convey any useful information to people who do not know its author.
- **Informative:** if the message is *informative* (of interest to other people beyond the author's immediate circle).
- **Other**: if the message is not related to the disaster.

Furthermore, we differentiate between two types of informative messages: direct, i.e., written by a person who is a direct eyewitness of what is taking place or indirect, when the message repeats information reported by other sources.

Once we detect informative tweets, we classify them into the following classes (details on the choice of this ontology can be found in [8]):

[1]These hashtags are mostly announced by the crisis management authorities at the time of an incident.

Table 1: Type-dependent instructions given to the assessors for the extraction phase, and example (in boldface) of the extracted part.

Type	**Instruction**: Copy-paste the word/phrase that ...	**Example**
Caution or advice: All	... warns about a potential hazard or advices what to do	.@NYGovCuomo orders **closing of NYC bridges**. Only Staten Island bridges unaffected at this time. Bridges must close by 7pm. #Sandy #NYC.
Information source: Photos/videos	... indicates what the contents of a photo/video are about	RT @NBCNewsPictures: **Photos of the unbelievable scenes left in #Hurricane #Sandy's wake** http://t.co/09U9L5rW #NYC #NJ
People: missing or lost people found	... indicates who is missing or has been found	rt @911buff: public help needed: **2 boys 2 & 4 missing** nearly 24 hours after they got separated from their mom when car submerged in si. #sandy #911buff
Casualties and damage: Infrastructure	... names a structure, road, service, line, etc. that is not working or has been damaged	RT @TIME: NYC building had numerous construction complaints before **crane collapse** http://t.co/7EDmKOp3 #Sandy
Casualties and damage: Injured or dead	... indicates who has (or how many people have) been injured or dead	At least **39** dead millions without power in Sandy's aftermath. http://t.co/Wdvz8KK8
Donations: Requests money/goods/services	... indicates what (money, goods, work, free services, etc.) is being requested as a donation	**400 Volunteers are needed** for areas that #Sandy destroyed.
Donations: Offers money/goods/services	... indicates what (money, goods, work, free services, etc.) is being offered as a donation	I want to volunteer to **help the hurricane Sandy victims**. If anyone knows how I can get involved please let me know!
People: Celebrities/authorities	... names a celebrity or authority that reacts to the event or visits the area	**V.P. candidate Ryan** attends a food drive in Wisconsin for victims of Hurricane Sandy. PO-35WE on BitCentral.

- **Caution and Advice:** if a message conveys/reports information about some warning or a piece of advice about a possible hazard of an incident.
- **Casualties and Damage:** if a message reports the information about casualties or infrastructure damage done by an incident.
- **Donations** of money, goods or services: if a message speaks about goods or services offered or needed by the victims of an incident.
- **People** missing, found, or seen: if a message reports about a missing or found person affected by an incident, or reports the reaction or visit of a celebrity.
- **Information Sources:** if a message points to information sources, photos, videos; or mentions a website, TV or radio station providing extensive coverage.
- **Other:** other types of informative messages.

As we describe in our previous work [8], a set of multi-label classifiers were trained to automatically classify a tweet into one or more of the above classes. Naïve Bayesian classifiers are used as implemented in Weka [7]. Our classifiers use a rich set of features including word unigrams, bigrams, Part-of-Speech (POS) tags and others. Our feature set contains as well as a set of binary features (for example, whether a tweet contains a URL, an emoticon, a hashtag, etc) and scalar features (such as the tweet length). The training data for our classifiers were obtained by manually classifying a set of tweets using crowdsourcing via provider Crowdflower[2]. We obtained about 2,000 labels for the Sandy dataset, and about 4,400 for the Joplin dataset.

2.3 Extraction

Once a tweet has been classified into one of the above classes, class-relevant information can be extracted for further analysis. For example, for a *casualty and damage* tweet, the number of casualties or the name of the infrastructure that was damaged can be identified.

[2]http://www.crowdflower.com

We treated the task of detecting class-relevant information as a sequence labeling task. A tweet is considered a sequence of word tokens. In a sequence labeling task, each token is algorithmically labeled as part of a subsequence of target information or as unrelated to such information. In the example of the first tweet in Table 1, the tokens "closing", "of", "NYC", and "bridges" are labeled as positive (part of the target information), while the rest of the tokens are labeled as negative. An example is shown below –note that the period (".") is also a token:

```
... orders closing of NYC bridges . Only Staten ...
      -      +     +  +     +    -   -     -
```

We use conditional random fields, a machine learned sequence labeling algorithm, for our task [9]. A conditional random field (CRF) is a probabilistic model which, in our task, predicts the label of each token ("+" or "-") given both information endogenous to the token (e.g. 'token is a number', 'token is the word *bridges*') as well as information exogenous to the token (e.g. 'token is preceded by the word *closing*'). CRFs have been applied successfully in the past to other information extraction tasks [10].

We use ArkNLP, an implementation of CRFs and a set of features known to be effective for NLP tasks on Twitter data [6]. In practice, we simply change the training data of ArkNLP to conform to what we described above, and execute it without further modifications.

Crowdsourcing task. During the crowdsourcing task for extraction, we show to the assessors each tweet and the type (and sub-type, if available) determined during the classification phase. We use an instruction that is specific to each sub-type, as listed in the "instruction" column of Table 1.

The workers were shown a tweet, this instruction, and an empty text input field, and were asked to copy-paste a word or short phrase from the tweet conveying the specified information. We did not accept any training example in which the segment extracted by the crowdsourcing worker was not contained in the original tweet.

3. EXPERIMENTAL RESULTS

Metrics. We evaluate our system by comparing its output with the responses provided by humans. We **train** our system on a part of the human-provided labels, and **test** the system on the remaining part. There are two aspects we measure that are related to the sensitivity and the specificity of our system.

Detection rate (analogous to statistical sensitivity, or recall) measures the fraction of examples in which humans found a relevant piece of information, and our system also found something, even if that something is incorrect.

Hit ratio (analogous to one minus the specificity, or precision) measures the fraction of examples for which our system found something, and that something could be considered correct by humans. We consider the output correct if it overlaps in at least one word with the given human label.

Metrics example. An example can illustrate these metrics. Suppose the input and output are as follows:

	Input	Output
a	There were **12** injured	<empty>
b	A **bridge has collapsed**	bridge
c	**10 volunteers** needed	needed

In this case, the detection rate is 66%, given that in two ($\{b, c\}$) of the 3 examples our system detected something. The hit ratio is 50% given that only in one of the two (b) the output overlaps with the target extraction in the input.

General results. Table 2 shows the results of our various experiments, where we selected the largest classes we had available: caution and advice, casualties and damage: infrastructure, and donations. In general, and similarly to precision-recall trade-offs observed in information retrieval systems, often a higher detection rate is associated to a lower hit ratio and viceversa.

There are four blocks that study different scenarios. Let us focus for now on the first row of each block, where *Train* is "All" and *Test* is "All".

The first two blocks measure the performance of our system on JOPLIN and SANDY data. The detection rate is higher for JOPLIN (78%) than for SANDY (41%). The hit ratio is also higher for JOPLIN (90%) than for SANDY (78%). This points out that the second dataset is more challenging to our system than the first one. However, in both cases the hit ratio is rather high, indicating that when our system extracts some part of the tweet, it is often the correct part.

The third block measures the performance of a hypothetical system trained on data from JOPLIN, and then tested on data from SANDY. This is usually referred to as an *adaptation* or *transfer* scenario. We can observe that compared to an scenario where we would train on data from SANDY, the detection rate drops dramatically (11% vs 41%), while the hit ratio is not affected significantly (78% vs 79%).

The most affected class of tweets are the ones providing caution and advice, which seem to be quite event-specific. On the other hand, the performance for the donation-related tweets is the least affected among the three classes, indicating that the words and phrases used to describe do not vary as much as for the other classes from one event to another.

In the fourth block, we consider an adaptation scenario in which a limited amount of new data (from SANDY) is incorporated into the training. This simulates a case in which we wait for a few hours before generating an output, in order to obtain some labeled examples about the new event. The performance is higher than in the previous case, with a detection rate of 21% and a hit ratio of 81%.

This last result shows that we can incrementally improve our model to work better whenever we need to use it on a new disaster.

Table 2: Performance of the information extraction phase for several configurations of training and testing set. "All" means no distinction between categories. The second and fourth columns show the number of tweets in the training and test data respectively.

Train		Test		Detected	Det. rate	Hit ratio
Train on 66% of Joplin, Test on 33% of Joplin						
All	338	All	169	131	78%	90%
All	338	C&A	130	109	84%	93%
All	338	Infra.	4	3	75%	33%
All	338	Dona.	34	25	74%	92%
C&A	260	C&A	130	118	91%	95%
Infra.	10	Infra.	4	1	25%	0%
Dona.	69	Dona.	34	16	47%	81%
Train on 66% of Sandy, Test on 33% of Sandy						
Train		Test		Detected	Det. rate	Hit ratio
All	397	All	198	82	41%	79%
All	397	C&A	69	27	39%	74%
All	397	Infra.	93	71	76%	83%
All	397	Dona.	35	23	66%	83%
C&A	139	C&A	69	26	38%	85%
Infra.	187	Infra.	93	50	54%	80%
Dona.	72	Dona.	35	12	34%	83%
Train on 100% of Joplin, Test on 100% of Sandy						
Train (Joplin)		Test (Sandy)		Detected	Det. rate	Hit ratio
All	507	All	595	66	11%	78%
All	507	C&A	208	4	2%	100%
All	507	Infra.	280	24	9%	71%
All	507	Dona.	107	38	36%	82%
C&A	390	C&A	208	2	1%	100%
Infra.	14	Infra.	280	44	16%	73%
Dona.	103	Dona.	107	52	49%	90%
Train on 100% Joplin + 10% of Sandy, Test on 90% of Sandy						
Train (Joplin+)		Test (Sandy-)		Detected	Det. rate	Hit ratio
All	568	All	534	112	21%	81%
All	568	C&A	187	9	5%	100%
All	568	Infra.	251	64	25%	80%
All	568	Dona.	96	39	41%	79%
C&A	411	C&A	187	18	10%	71%
Infra.	43	Infra.	251	106	42%	83%
Dona.	114	Dona.	96	46	48%	89%

Detailed results. In each block, the first row reports the detection rate and the hit ratio when we train *a single model* over all the tweets in our training set and we test it over all the tweets in our test set regardless of the tweets' respective classes. In the next three rows we disaggregate this setting for each class in the testing part. Finally, in the last three rows we show the performance when we train *three different models*, one for each class, and test it only over tweets of the same class.

The results indicate that class-specific models may lead to improvements in performance for some classes but not for others. The class-specific models are particularly helpful for the caution and advice class of tweets, and yield improvements in the detection rate for the SANDY dataset in the case of donation-related tweets. There are no consistent gains for the tweets related to infrastructure damage, except when training on JOPLIN and testing on SANDY.

4. GENERALIZATION TO OTHER EVENTS

A robust approach should generalize to a variety of scenarios, including non-disaster related events. In this section we briefly discuss a set of experiments on a non-disaster dataset corresponding to a sports match. The dataset, which consists of 72,000 tweets, was collected using *Twitter Streaming API* using `#cricket`, `#indvspak`, `#indvpk` hashtags during a Cricket match between Pakistan and India on January 6th, 2013.

Crowdsourcing task. We label the data using the same procedure as for our other datasets. In the first task, which comprised of 2,000 unique tweets, we asked workers to label an individual tweet to (i) separate informative tweets from personal and (ii) for an informative tweet specify what information it conveys.

We used six classes that are domain-dependent and correspond to events during a cricket match: boundary, score, over, dismissal, ball and other[3]. In the second task, which comprised of 631 informative tweets, the workers were presented type and sub-type of a tweet and asked to copy-paste a word or short-phrase using a type-dependent instruction.

Experimental results. Table 3 shows the results of various experiments on this dataset. The first two rows are scenarios where a single model is created, and the remaining correspond to multiple class-specific models. When trained over the whole training set and tested on the whole test set, we observe a relatively low detection rate. This can be improved if we incorporate examples in which more than one type of information is present in a given tweet, as shown in the second row. We can also see significant improvements in hit ratio for all the class-specific models.

Table 3: Results with cricket data.

	Training cases	Testing cases	Detection rate	Hit ratio
All	321	161	43%	95%
All (multiple labels)	321	161	51%	95%
Score	129	66	65%	98%
Other	100	51	76%	92%
Dismissal	63	31	81%	88%
Boundary	18	8	88%	100%
Ball	6	3	100%	100%
Over	5	2	50%	100%

5. RELATED WORK

During emergencies social media platforms such as Facebook and Twitter distribute up-to-date situational awareness information (e.g., damage, causalities etc.) in all forms (e.g., photos, videos etc.) [2, 3]. Cameron et al. [4] describe a platform for emergency situation awareness, that detects incidents using burst keyword detection and classifies interesting tweets using an SVM classifier. However, identification of on-topic informative messages and extraction of actionable information pose serious challenges due to the noisy and unstructured nature of Twitter's data. Most previous works were based on standard machine learning methods which typically trained on formal news-text and performed poorly for an extremely informal source like Twitter [5].

In this paper we used the classification-extraction approach presented in our previous work [8], adapting in a simple and straightforward manner the Twitter-specific part-of-speech tagger ArkNLP to our task [6].

[3]http://en.wikipedia.org/wiki/Cricket

6. CONCLUSIONS AND FUTURE WORK

We have presented a practical system that can extract disaster-relevant information from tweets. According to extensive experiments on two different datasets, our approach can detect from 40% to 80% of the tweets containing this type of information, and generate an output that is correct 80% to 90% of the time.

This tweet-level extraction is in our opinion key to being able to extract reliable high-level information. Observing, for instance, that a large number of tweets in similar locations report the same infrastructure as being damaged, may be a strong indicator that this is indeed the case.

Please contact authors for inquiries about data availability.

Acknowledgments. Sincere thanks to Kate Starbird and Project EPIC at University of Boulder, Colorado, for sharing tweet-ids of the JOPLIN dataset.

7. REFERENCES

[1] Cynthia D. Balana. Social media: major tool in disaster response, 2012.

[2] Fredrik Bergstrand and Jonas Landgren. Information sharing using live video in emergency response work. In *Proc. of ISCRAM*. Citeseer, 2009.

[3] Heather Blanchard, Andy Carvin, Melissa Elliott Whitaker, and Merni Fitzgerald. The case for integrating crisis response with social media. White Paper, American Red Cross, 2012.

[4] Mark A Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proc. of WWW*, pages 695–698. ACM, 2012.

[5] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proc. of HLT*, pages 80–88, 2010.

[6] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proc. of HLT*, pages 42–47, Stroudsburg, PA, USA, 2011.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[8] Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM*, Baden-Baden, Germany, 2013.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289, 2001.

[10] Fuchun Peng and Andrew McCallum. Information extraction from research papers using conditional random fields. *IP&M*, 42(4):963 – 979, 2006.