Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages

Soudip Roy Chowdhury
University of Trento, Italy
rchowdhry@disi.unitn.it

Muhammad Imran
University of Trento, Italy
imran@disi.unitn.it

Muhammad Rizwan Asghar
University of Trento, Italy
asghar@disi.unitn.it

Sihem Amer-Yahia
CNRS, LIG, France
Sihem.Amer-Yahia@imag.fr

Carlos Castillo
QCRI, Doha, Qatar
chato@acm.org

ABSTRACT

We present *Tweet4act,* a system to detect and classify crisis-related messages communicated over a microblogging platform. Our system relies on extracting content features from each message. These features and the use of an incident-specific dictionary allow us to determine the period type of an incident that each message belongs to. The period types are: pre-incident (messages talking about prevention, mitigation, and preparedness), during-incident (messages sent while the incident is taking place), and post-incident (messages related to the response, recovery, and reconstruction). We show that our detection method can effectively identify incident-related messages with high precision and recall, and that our incident-period classification method outperforms standard machine learning classification methods.

Keywords

microblogging, crisis informatics, disaster management, twitter data-analytics

INTRODUCTION

During crisis situations (e.g., when a flood breaks out, or a Tsunami strikes), people turn to social media platforms (such as, Twitter[1]) for exchanging information about the incident, or to express their views about, for instance, relief works and the government role in the response and recovery efforts **(**Palen et al., 2010). Often, messages, which are exchanged through the social media, contain valuable information that can contribute to a situational awareness during an incident if analyzed in a timely and efficient manner **(**Starbird et al., 2010; Latonero and Shklovski, 2010). A number of methods (e.g., Cataldi et al. 2010 among others) have been proposed to detect such situational awareness messages; however, they mostly rely on detecting frequency increase in the usage of a certain keyword or hashtag (also known as "trending topics"). Unfortunately, such methods may fail to produce satisfactory results for smaller scale incidents. Instead, in this paper, we look at properties that characterize a specific type of incident (e.g., typhoon, tornado, or earthquake) and perform automated filtering of messages based on those characteristics. Emergency management agencies, such as FEMA[2] and OCHA[3] recognize that emergency management has distinctive phases **(**Baird, 2010). In this paper, we describe a method for automatically classifying messages into **pre-incident**, **during-incident and post-incident classes**; the methods described in this paper are useful even for detecting period types of small-scale incidents.

Next, we discuss our data collection process and show the efficiency of our automated data-cleaning approach, whose goal is to verify if a tweet is related to a crisis or not. Then we show the result of our automated period-assignment algorithm, which classifies tweets according to the incident periods. We demonstrate the efficiency of our algorithms by cross-validating its results against the human annotated data and by comparing its accuracy

[1] *https://twitter.com*
[2] *http://www.fema.gov*
[3] *http://www.unocha.org/what-we-do/coordination-tools/cluster-coordination*

against other state-of-the-art algorithms. Finally we conclude the paper, with a vivid discussion on the Tweet4act system and by showing the future work directions.

CRISIS DATA COLLECTION AND PHASES

In this section, we seek to understand the characteristics of messages exchanged during different phases of a crisis situation. For this purpose, we collected data from Twitter for three recent natural disasters: the Joplin Tornado in Missouri, USA on May 22, 2011[4] consists of 1500 tweets, the Nesat Typhoon in Phillipines on September 27, 2011[5] consists of 500, and the Haiti Earthquake in Haiti on January 12, 2010[6] consists of 1500 tweets. These datasets were collected using the *Twitter Streaming API*[7] with appropriate *hashtags* (i.e., keywords) that are mostly announced by the crisis management authorities at the time of an incident.

Pre-incident messages. Before the Nesat Typhoon actually had struck Philippines on September 24, 2011, few of the messages that were communicated over the Twitter data stream, are shown below:
- New #tropical storm forms in the West #Pacific. #Nesat may hit the #Philippines & #China as a #typhoon next week
- typhoon, #philippines: Tracking Typhoon Nesat (Pedring), as it poses severe threat to Philippines … #nesat #pedring #20w

During-incident messages. During the Nesat Typhoon, few of the tweet communications were as follows:
- @YahooNews: Powerful #typhoon with winds up to 106 mph makes landfall in #Philippines as 100,000 ordered to flee homes...
- Typhoon #Nesat expected to reach #Baguio #Philippines 3 hours from now …

Post-incident messages. From the post-incident phase, we found few interesting tweets as follows:
- News5 Action Center is now accepting donations for the victims of Typhoon Pedring. Drop boxes are located @ TV5 Office
- Hagonoy appeals for 15,000 packs of relief goods. Contact Lizzie Fajardo mayors office …

From these messages, one may observe that people post warning and caution related messages before, causalities and damage related messages during and request for help or donations related messages after an incident. From these observations, we can infer that messages posted before, during, and after an incident are having unique characteristics. The main goals of our research can be summarized as: first, we want to identify messages related to an incident and second, to classify such *incident-related* messages with the corresponding periods (PRE, DURING, POST) of an incident. Figure 1 illustrates main components of our proposed system, which are described in the next two sections.

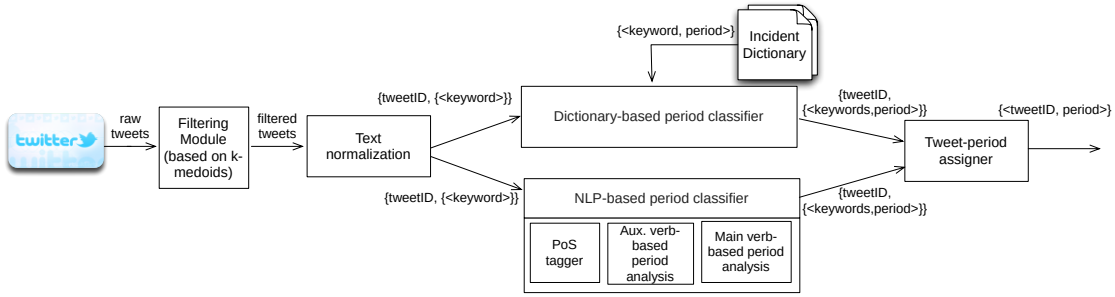


Figure 1: Overview of the Tweet4act System

[4] http://www.epacha.org/Pages/Tornadoes_USA_Missouri_05_2011.aspx

[5] http://www.bbc.co.uk/news/world-asia-pacific-15070550

[6] http://www.epacha.org/Pages/Earthquake_Haiti_2010.aspx

[7] https://dev.twitter.com/docs/streaming-apis

FINDING CRISIS-RELATED MESSAGES

The data collection procedures described in the previous section are keyword-based and do not guarantee that the obtained messages are incident-related. After analyzing the result of a crowd-sourced labeling task, we have identified that a considerable subset (5-13%) of the collected tweets is not incident-related.

Method: Outlier Detection

In order to filter the messages unrelated to incidents, we used an **outlier detection** method. Specifically, we applied a state-of-the-art clustering algorithm: *K-medoid* **(**Hodge and Austin, 2004)**.** We chose k-medoid over the standard k-mean based clustering algorithm because it is less susceptible to local minima, thus producing better quality clusters. Our filtering algorithm performs the following steps sequentially:

1. Normalize message text: remove the "*RT @username*" and "*@username*" prefixes.
2. Remove duplicate messages after normalization.
3. Remove all terms that appear in less than a fraction *s* of messages (we set $s = 0.05$).
4. Run the k-medoid clustering algorithm on each dataset.
5. Discard clusters having a negative number or zero as silhouette coefficient, where silhouette coefficient (L. Kaufman, and P. J. Rousseeuw, 1990) determines the quality of a cluster in terms of distance between points within a cluster and the distance with other clusters. The value of silhouette coefficient lies in the range (-1,1), while -1 indicates bad cluster and 1 indicates good quality cluster.
6. Select from each cluster the fraction *m* of messages closer to the medoid.

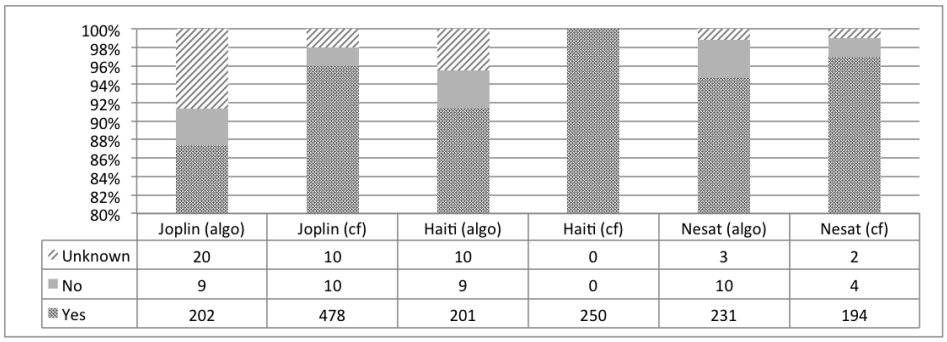
By implementing this outlier detection method, we managed to filter the top-*m* fraction of the most representative messages from each cluster. The accuracy of this method is described in the following section.

Experimentation

After normalization and removal of duplicates, we found 1,198, 1,167 and 373 unique messages in the Joplin, Haiti and Nesat datasets, respectively. To validate the correctness of our automated cleaning algorithm, we verified the filtered messages using the CrowdFlower platform[8], where we developed separate tasks for each dataset, each task containing the output of the automated cleaning algorithm.

Measuring precision. In order to test the precision, we asked crowd-sourcing workers to label 498 tweets identified as crisis-related by our method from the Joplin dataset, 250 from the Haiti dataset, and 200 for the *Nesat* dataset. Each task, which also consists of a set of correctly labeled tweets (i.e., golden data), simply asked workers to choose whether a tweet is incident-related or not and in total 3 trusted answers required to finalize a labeling decision. The distribution of the labels is shown in Figure 2 (see (cf) columns) for all three crowdsourcing tasks. For the Joplin task, the inter-annotator agreement (percentage of agreement between 3 workers answers on a particular label for a tweet) was 75.34%, whereas for the Haiti task it was 86.95%. The inter-annotator values show the agreement level among assessors for an assessable unit. Considering all the messages with a valid label (not "unknown'"), this yields a precision (% of tweets that the system could correctly classify with their expected labels) of 96%, 100%, and 97%, respectively.

[8] http://crowdflower.com



	Joplin (algo)	Joplin (cf)	Haiti (algo)	Haiti (cf)	Nesat (algo)	Nesat (cf)
Unknown	20	10	10	0	3	2
No	9	10	9	0	10	4
Yes	202	478	201	250	231	194

Figure 2: Distribution of labels unfiltered (algo), filtered (cf): crisis-related (Yes) and not-crisis-related (No).

Measuring recall. To calculate the recall of our k-medoid based filtering algorithm, we selected random samples of 231, 220, and 244 tweets from *Joplin*, *Haiti* and *Nesat* raw datasets, respectively (before applying the filter). Next, we manually labeled those messages as crisis-related or not crisis-related; the distribution of labels is shown in Figure 2 (see (algo) columns). Using this method, we found 30, 24, and 36 false negatives from our filtering method, i.e., tweets that should have passed the filtering but did not. Using the standard formula for recall, we found recall measures (% of the expected tweets labels that are correctly identified by our system) of 85%, 88% and 84% for three datasets, respectively.

PERIOD ASSIGNMENT

Method: The ***dictionary-based period classifier*** compares the words in each message against a dictionary of words known to be present in specific periods of a crisis-incident e.g., keywords such as "warning" and "alert" are typically found in the pre-incident period, keywords such as "now", "sweeps" are typically found during-incident, and keywords such as "aftermath", "donate" are found in post-incident messages. The ***natural language processing based period classifier module*** identifies the tense of the verbs (main, as well as auxiliary) in a message. Finally, we assign a score to each word in a message according to the following algorithm:

1. If the word is listed in the dictionary, add +1 to the period it is listed under and stop processing that word (i.e., if a verb is in the dictionary, we ignore it below).

2. If the word is an auxiliary verb, add +1 to the period it is associated to (e.g., could-PRE, are-DURING, did-POST etc).

3. If the tense of the main verb support the tense of the auxiliary verb (e.g., would be *coming*–PRE), add +0.5 to the corresponding period, respectively.

After this procedure, we sum up the scores of each period across all the words in the phrase and pick the period with the largest sum. In case of ties, PRE is preferred over DURING and POST, and DURING is preferred over POST. For instance, let us consider the following message: "NFL teams gathering supplies aid for tornado victims in Kansas Missouri (Morning Call) …". In this message, both words "aid" and "victim" are matched in the dictionary for the POST period. The verb "gathering" is in continuous form and contributes to the DURING period. In total, the message has +2 score for POST and +0.5 for DURING; hence, it is classified as POST.

EXPERIMENTAL RESULTS

Labeling. We labeled a set of messages classified as crisis-incident-related by the method described in the Section

Finding Crisis-related messages. We asked assessors to indicate the period (PRE, DURING, POST, UNKNOWN) for each message. 23%, 3%, and 11% of the data were labeled as PRE, 35%, 19% and 56% were labeled as DURING, 24%, 71% and 11% of the data were labeled as POST and rest 18%, 7% and 22% of the data were reported as UNKNOWN for the Joplin, Haiti and Nesat dataset, respectively.

Baselines. We compared our period detection algorithm against four text-based supervised classifiers including support vector machine (SVM), maximum entropy (MaxEnt), decision tree (Tree) and an ensemble classifier random forest (RF). We represented each message as a bag of *unigrams* and *bigrams* feature vector and used them for the classification criteria in each of the baseline approaches. To get the best performance from the baseline algorithms, we fine-tuned the parameters for each of these algorithms. We trained a specific model for each of the three corpora by using 80% of the data and validated the models over remaining 20% of the data by using 10-fold cross validation technique.

Results

The precision, recall and F_1 measures of SVM, MaxEnt, Tree and RF algorithms on our test datasets are shown in Table 1. Note the high variance of the performance of the SVM, Tree and RF algorithms due to their sole dependency on n-gram feature in the experiment setting. For Joplin dataset, our algorithm performs better than the baseline algorithms by a large margin (0.69 vs. 0.44), the baseline algorithms perform better for the Haiti Earthquake dataset by a smaller margin (0.71 vs. 0.80), and they tie for the Nesat Typhoon dataset (0.71).

Table 1: Tweet4act and standard text classification methods in terms of precision (P), recall (R), and F_1 measures.

Period	Tweet4act			SVM			MaxEnt			Tree			RF		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	Joplin Tornado														
PRE	0.33	0.85	0.48	0.00	0.00	0.00	0.43	0.21	0.28	0	0	0	0	0	0
DURING	0.88	0.89	0.89	0.32	0.91	0.47	0.35	0.55	0.43	0.3	0.73	0.43	0.32	0.1	0.48
POST	0.61	0.84	0.71	0.67	0.20	0.31	0.55	0.6	0.57	0.57	0.4	0.47	1.00	0.1	0.18
AVG.	0.61	0.86	0.69	0.33	0.37	0.39	0.44	0.45	0.42	0.29	0.38	0.45	0.66	0.37	0.33
	Haiti Earthquake														
PRE	0.63	1.00	0.77	1.00	0.67	0.80	1.00	1.00	1.00	1	0.67	0.8	1.00	0.33	0.5
DURING	0.72	0.97	0.83	0.75	0.6	0.67	0.67	0.80	0.73	0.6	0.6	0.6	1.00	0.4	0.57
POST	0.46	0.82	0.59	0.92	0.97	0.94	0.97	0.95	0.96	0.92	0.95	0.93	0.88	1.00	0.94
AVG.	0.60	0.87	0.71	0.89	0.74	0.80	0.88	0.91	0.89	0.84	0.74	0.78	0.96	0.58	0.67
	Nesat Typhoon														
PRE	0.36	1.00	0.53	1.00	0.50	0.67	1.00	0.50	0.67	0.33	0.25	0.28	1.00	0.5	0.67
DURING	0.94	0.94	0.94	0.79	1.00	0.88	0.81	1.00	0.90	0.71	0.77	0.74	0.79	1	0.88
POST	0.52	0.85	0.65	1.00	0.2	0.33	1.00	0.40	0.57	0	0	0	1.00	0.2	0.33
AVG.	0.61	0.93	0.71	0.93	0.57	0.62	0.94	0.63	0.71	0.35	0.34	0.51	0.93	0.57	0.63

RELATED WORK

Twitter for emergency response. The growth of microblogging has led to a vast body of research related to the real-time burst detection and event detection based on the tweet content analysis. Starbird et al. **(**Starbird et al., 2010) and Vieweg et al. (Vieweg et al., 2010) analyzed microblog usage and information lifecycles during crisis situations. Starbird et al., found that the information including geo-location, situational updates and warnings can contribute to *situation awareness* and are typically communicated during each crisis-incident on Twitter. Hughes and Palen **(**Hughes et al., 2009) examined Twitter usage during four high profile mass-convergence events and emergencies. They found that tweets sent during emergency events reveal features of information that can support information broadcasting and brokerage. The focus of most of these studies is characteristics from the content, type of message and/or sentiment expressed in the message, and some other on keyword

frequencies (Mathioudakis et al., 2010; Sayyadi et al., 2009). They do not attempt to determine if a message is incident-related or not or its period.

Period detection. Iyengar et al. (Iyengar et al., 2011) propose an approach to automatically classify tweets by periods. Their system applies supervised SVM classifier, and hidden markov model over unigrams and verb-based features. As experiment results show, our method performs better than this approach in most of the cases.

CONCLUSIONS

In this paper we showed how a platform like Tweet4act, which leverages other language property (tense) along with the n-gram can efficiently filter informative messages and is capable of detecting phases of crisis-incidents. In our future work we will be working towards extending the Tweet4act's incident dictionary to handle incidents other than natural disasters (e.g., hunger strike, disease break-out etc).

REFERENCES

1. M. E. Baird (2010) The "phases" of emergency management. In http://www.memphis.edu/cait/pdfs/Phases_of_Emergency_Mngt_FINAL.pdf, pages 1–42.
2. M. Cataldi, L. Di Caro, and C. Schifanella (2010) Emerging topic detection on twitter based on temporal and social terms evaluation. In Proceedings of the MDMKDD '10, pages 4:1–4:10, New York, USA. ACM.
3. V. Hodge and J. Austin (2004) A survey of outlier detection methodologies. Artif. Intell. Rev.,85–126, Oct.
4. Hughes and L. Palen (2009) Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3): pages 248–260.
5. A. Iyengar, T. Finin, and A. Joshi (2011) Content-based prediction of temporal boundaries for events in Twitter. In Proceedings of the Third IEEE International Conference on Social Computing.
6. M. Mathioudakis and N. Koudas (2010) Twittermonitor: trend detection over the twitter stream. In Proceedings of the SIGMOD '10, pages 1155–1158, New York, USA. ACM.
7. L. Palen, K. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald (2010). A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference.
8. K. Starbird, L. Palen, A. Hughes, and S. Vieweg (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pages 241–250. ACM.
9. H. Sayyadi, M. Hurst, and A. Maykov. "Event detection and tracking in social streams." Proceedings of International Conference on Weblogs and Social Media (ICWSM). 2009.
10. L. Kaufman, and P. J. Rousseeuw. "Finding groups in data: an introduction to cluster analysis" John Wiley and Sons, New York, (1990)
11. Latonero, Mark, and Irina Shklovski. ""Respectfully Yours in Safety and Service": Emergency Management & Social Media Evangelism." Proceedings of the 7th International ISCRAM Conference–Seattle. Vol. 1. 2010.