# Tweedr: Mining Twitter to inform Disaster Response

Zahra Ashktorab Chris Brown Jit Nandi Aron Culotta

September 2, 2013

## 1 Introduction

In recent years, Twitter has become a major channel for communication during natural disasters. From #snowpocalypse to #Sandy, Twitter users have shared news, photos and observations from the midst of hurricanes, blizzards, earthquakes and other emergencies. First responders can utilize these streams of data generated by social media to find out where the disasters are happening and what specifically has been effected as a result of it. As with most social media conversations, informative signals are often drowned out by irrelevant and redundant noise. First responders struggle to glean actionable knowledge from the flood of tweets and status updates.

In order to solve this problem, we propose `tweedr`: Twitter for Disaster Response.[1] The goal of this tool is to extract information relevant for first responders from tweets generated during disasters. The `tweedr` pipeline consists of three main parts: classification, clustering and extraction. In the classification phase, we use a variety of classification methods (sLDA, SVM, and LogReg) to classify whether a tweet reports disaster damage or casualty information. In the clustering phase, we utilize filters to merge tweets that are similar to one another; and finally in the extraction phase, we extract specific tokens and phrases that report specific information about different classes of infrastructure damage, damage types, and casualties. Using these three phases, the Tweedr pipeline is able to extract actionable information from a stream of tweets during disasters.

## 2 Data

We identified 12 crisis events that occurred in North America since the founding of Twitter in 2006. We then constructed queries to collect relevant tweets from Gnip, a social media aggregation company. We constructed two types of queries: (1) keyword (**kw**) queries contain search terms and hashtags determined to be relevant based on a post-hoc analysis; (2) geographical queries (**geo**) consisting of a bounding box of coordinates around the epicenter of the event. Table 1 lists the number of tweets collected for each event.

We can see considerable variation in the number of messages for each crisis. This is in part explained by the popularity of Twitter overall, the number of people affected, and by Twitter usage in the affected area. Additionally, more recent events return more matches for the geographical queries – this follows from the increased usage of geolocation services on Twitter.

### 2.1 Data Annotation

To train and evaluate our automated methods, we must first collect human-annotated examples. We consider two tasks for annotation:

1. **Classification**: Does the tweet mention either specific infrastructure damage or human casualty? We treat this as a binary classification task. Positive examples include "10 injured in plant explosion" and "The windows are smashed at the Whole Foods on 1st"; however, "Hurricane Irene causes massive damage" would be a negative example, since it does not include specific, actionable damage information.

2. **Extraction**: For positive examples of the above, identify the tokens in the tweet corresponding to specific types of infrastructure damage or counts of the number of dead or injured. For example, in the tweet "Flooding bad up and down sides of Green River Rd," the token "Flooding" should be annotated as a damage type, and the tokens "Green River Rd" should be labeled as a road. The full ontology is listed in Table 3.

[1] `http://tweedr.dssg.io`

| **Event** | **N (kw)** | **N (geo)** | **total** | **keywords** |
|---|---|---|---|---|
| christchurch | 759,399 | 2,017 | 761,416 | #EQNZ,#CHCH,#NZquake,Christchurch |
| ike | 113,773 | 3,989 | 117,762 | Hurricane+Ike,Hurricane,Ike,Galveston,Houston |
| irene | 384,689 | 3,871,631 | 4,256,320 | #Hurricane,#Irene,#Tropics |
| moore | 842,318 | 1,579,056 | 2,421,374 | #mooretornado,#moore,newcastle |
| oklahoma | 1,697,891 | 6,209,510 | 7,907,401 | #oklahoma,#tornado,#oklahomatornado,#okwx,#okc |
| samoa | 235,208 | 2,016 | 237,224 | Samoa,tsunami,earthquake |
| slavelake | 45,072 | 1,281 | 46.353 | #SlaveLake,Slave+Lake |
| supertuesday | 22,205 | 2,201 | 24,406 | Super+Tuesday,Jackson,Memphis,supertuesday |
| tornado2011a | 476,730 | 2,293 | 479,023 | Tushka,oklahoma,okwx,arkansas,akwx,tornado |
| tornado2011b | 52,201 | 2,326 | 54,527 | #alwx,#okwx,#txwx,#tristatewx,tornado,#ALNeeds, #ALHaves,#WeAreAlabama |
| vtech | 16,652 | 2,869 | 19,521 | #vatech,#virginiatech,#hokies,#vtech,#vt |
| westtx | 651,045 | 178,846 | 829,891 | #WestExplosion,#WestTX |
| **Total** | **5,297,183** | **11,858,035** | **17,155,218** | |

Table 1: Number of tweets collected by event. We query for tweets both by keyword (**kw**) and geographical bounding box (**geo**).

Since not all data can be labeled manually, we sample a small subset. Half of the tweets are selected uniformly at random from each event; the remaining half are sampled from tweets matching a set of keywords heuristically determined to be relevant to our task.[2] We do this to mitigate the class imbalance problem (i.e., most tweets are not relevant to infrastructure damage or casualties).

We sampled 1,049 tweets of the resulting tweets, of which 793 were labeled as positive examples. We then annotate the extraction labels for each positive example.

# 3 Experiments

## 3.1 Classification

We compare a number of standard classification algorithms, including $K$-nearest neighbors, decision trees, naïve Bayes, and logistic regression, as implemented in the `scikit-learn` Python library.[3] We also compare with supervised latent Dirichlet allocation [1], for which we create a Python wrapper of the R `lda` package.[4]

Table 2 displays the accuracy of each method. Logistic regression appears to be the most reliable across several accuracy measures.

[2] The keywords are: bridge, intersection, car, bus, truck, vehicle, evacuation, evacuate, fire, police, institution, wind, impact, injured, damage, road, airplane, hospital, school, home, building, flood, collapse, death, casualty, missing.

[3] `http://http://scikit-learn.org/`

[4] `http://cran.r-project.org/web/packages/lda/`

## 3.2 Extraction

For extraction, we use a conditional random field (CRF) [10], as implemented by the `CRFsuite` toolkit.[5]. We consider several different types of features for our CRF. For each token in a tweet, we inspect capitalization, pluralization, whether it is numeric or includes a number, whether it is part of a determined lexicon of transportation types or building types, hypernyms, ngrams, and part of speech tags. To obtain precision, recall, and F1-score values, we split the data using two methods. In Table 3, we use 10-folds cross validation. Additionally, we split the data by disaster, training on the labeled data from our top 5 disaster and testing on the sixth. The disasters we trained on this second method include: Joplin, Irene, Samoa, Christchurch, Tornado2011b, and Oklahoma. By splitting the training and testing data on disaster, we can test the accuracy of our classifier on unseen disasters. In Table 4, we show the overall average for the CRF for performing on an unseen disaster.

As seen by Table 3, our entity extraction classifier performs well (obtains an F1-score above .5) on predicting missing persons, religious institutions, electricity loss, hospital and health infrastructures, death/casualties, and wind/projectile damage. However, it does not predict fires and homes/residential infrastructures as accurately as the aforementioned labels. Furthermore, due to the nature of content in tweets, there is insufficient labeled data for certain labels and thus precision, recall and F1-scores could not be obtained. We also evaluated our CRF across

[5] `http://www.chokkan.org/software/crfsuite/`

| Method | F1 | Pr | Re | Acc | AUC |
|---|---|---|---|---|---|
| **LogReg** | **.65 ± .07** | .78 ± .08 | .57 ± .09 | **.86 ± .03** | **.88** |
| **NB** | .63 ± .06 | .55 ± .07 | **.75 ± .09** | .80 ± .03 | .84 |
| **DTree** | .54 ± .09 | **.93 ± .07** | .39 ± .09 | .85 ± .02 | .69 |
| **KNN** | .51 ± .04 | .83 ± .10 | .38 ± .04 | .84 ± .02 | .73 |
| **sLDA** | .50 ± .07 | .42 ± .06 | .65 ± .15 | .70 ± .05 | .77 |

Table 2: Damage/casualty classification results (with standard deviations). **Pr:** precision, **Re:** recall, **Acc:** accuracy, **AUC:** area under the ROC curve.

disasters to evaluate how it performed on disasters it had not seen yet. The results were promising for some disasters, yielding promising F1-score values for four of the six disasters evaluated. The results are reflected in Table 4. Additionally, the confusion matrix in Figure 1 shows some misclassification between wind damage and death/casualties – both types of messages often contain numerical tokens (e.g., "100s people wounded" versus "100s of downed trees").

### 3.3 Clustering

We consider two different methods of approximate string matching: Bloom filters [2] and SimHash [3]. We have not yet performed any quantitative evaluation of these approaches – we leave this for future work.

## 4 Related Work

There has been growing interest in using social media for situational awareness during a disaster [6, 4, 8, 9, 5]. Our work builds upon that of Imran et al. [5], who also use a CRF to extract damage and casualty information from tweets. Our main contributions are experiments with an expanded ontology, feature set, and set of events.

## 5 Conclusions and Future Work

We have outlined initial experiments using `Tweedr` to extract relevant information from tweets during a disaster. Additional experiments are needed to understand the behavior of these methods in real-world, dynamic environments. Also, given the low frequency of relevant tweets, methods designed for high class imbalance [7] may be useful here.

| Label | F1 | Pr | Re |
|---|---|---|---|
| missing persons | 0.86 | 0.80 | 1.00 |
| religious institution | 0.68 | 0.63 | 0.75 |
| other damage | 0.63 | 0.52 | 1.00 |
| other building | - | - | - |
| electricity loss | 0.88 | 0.81 | 1.00 |
| snow | - | - | - |
| evacuation center | - | - | - |
| hospital/health services | 0.76 | 0.71 | 0.96 |
| other transportation | - | - | - |
| intersection | - | - | - |
| road | - | - | - |
| bridge | - | - | - |
| homes/residential | 0.47 | 0.33 | 1.00 |
| fire | 0.33 | 0.20 | 1.00 |
| school | - | - | - |
| fire/police department | - | - | - |
| building collapse | - | - | - |
| injured person | - | - | - |
| bus | - | - | - |
| airplane | - | - | - |
| death/casualties | 0.60 | 0.44 | 0.96 |
| wind/projectile damage | 0.75 | 0.65 | 0.90 |
| flood | - | - | - |
| vehicular impact | - | - | - |
| **Average** | **0.47** | **0.32** | **0.88** |

Table 3: Extraction F-score, Precision, and Recall values for each label. Missing values occur because many label types in our ontology did not appear in the data.

## 6 Acknowledgments

This work was performed during the 2013 Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago, in partnership with the Qatar Computational Research Institute. We are grateful to Gnip for providing access to the historical tweets used in this analysis, as well as to all the 2013 DSSG Fellows who helped with data annotation.

| Disaster | F1 | Pr | Re |
|---|---|---|---|
| Joplin | 0.79 | 0.65 | 1.00 |
| Irene | 0.13 | 0.11 | 0.714 |
| Samoa | 1.00 | 1.00 | 1.00 |
| Christchurch | - | - | - |
| Tornado 2011b | 1.00 | 1.00 | 1.00 |
| Oklahoma | 0.44 | 0.29 | 1.00 |
| **Average** | **0.60** | **0.49** | **0.77** |

Table 4: Extraction F-score, Precision, and Recall obtained by training on 5 disasters and testing on the sixth. This assesses the ability of the algorithm to generalize to new disasters.

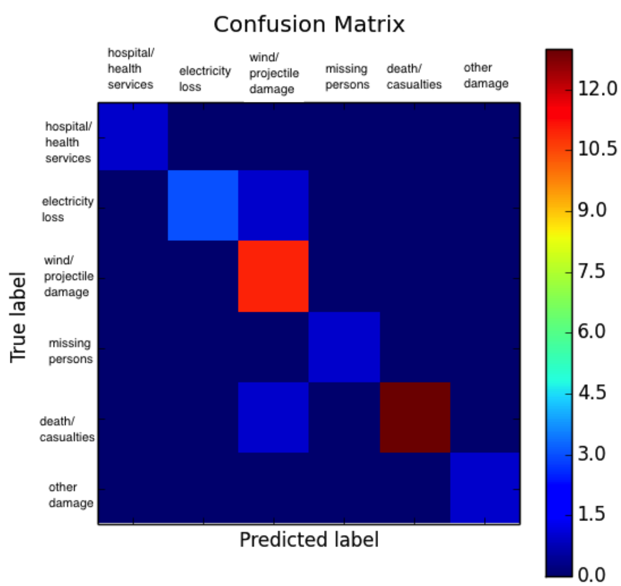


Figure 1: Confusion Matrix of predicted labels using 10-folds cross validation.

# References


[1] D. M. Blei and J. D. McAuliffe. Supervised topic models. arXiv e-print 1003.0783, Mar. 2010.

[2] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422426, July 1970.

[3] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, STOC '02, page 380388, New York, NY, USA, 2002. ACM.

[4] F. Cheong and C. Cheong. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *PACIS'11*, pages 46–46, 2011.

[5] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, page 10211024, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[6] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. TweetTracker: an analysis tool for humanitarian and disaster relief. In *ICWSM'11*, 2011.

[7] W.-J. Lin and J. J. Chen. Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, page bbs006, Mar. 2012. PMID: 22408190.

[8] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue. A demographic analysis of online sentiment during Hurricane Irene. In *NAACL-HLT Workshop on Language in Social Media*, 2012.

[9] P. Meier, C. Castillo, M. Imran, S. M. Elbassuoni, and F. Diaz. Extracting information nuggets from disaster-related messages in social media. In *10th International Conference on Information Systems for Crisis Response and Management*, 2013.

[10] C. Sutton and A. K. McCallum. *An introduction to conditional random fields.* Now Publishers, Hanover, MA, 2012.