Tweedr: Mining Twitter to Inform Disaster Response

Zahra Ashktorab
University of Maryland, College Park
parnia@umd.edu

Christopher Brown
University of Texas, Austin
chrisbrown@utexas.edu

Manojit Nandi
Carnegie Mellon University
mnandi@andrew.cmu.edu

Aron Culotta
Illinois Institute of Technology
aculotta@iit.edu

ABSTRACT

In this paper, we introduce Tweedr, a Twitter-mining tool that extracts actionable information for disaster relief workers during natural 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 logistic regression) to identify tweets reporting damage or casualties. In the clustering phase, we use filters to merge tweets that are similar to one another; and finally, in the extraction phase, we extract tokens and phrases that report specific information about different classes of infrastructure damage, damage types, and casualties. We empirically validate our approach with tweets collected from 12 different crises in the United States since 2006.

Keywords

Social network analysis, text mining, social media, disaster response

INTRODUCTION

In recent years, Twitter has become a major channel for communication during natural disasters. Twitter users have shared news, photos and observations from the midst of hurricanes, blizzards, earthquakes and other emergencies (Cheong and Cheong, 2011; Kumar, Babier, Abbasi, and Liu, 2011; Mandel, Culotta, Boulahanis, Stark, Lewis, Rodrigue, 2012; Meier, Castillo, Imran, Elbassuoni, and Diaz, 2013). First responders can use these streams of data generated by social media to find out where the disasters are happening and what specifically has been affected as a result of it. As with most social media conversations, informative signals are often overwhelmed by irrelevant and redundant noise. First responders struggle to glean actionable knowledge from the large volume of tweets and status updates. In order to effectively extract information relevant to disaster relief workers, 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 in real time as well as enable analysis after the disaster has occurred. 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 logistic regression) to classify whether a tweet reports disaster damage or casualty information. In the clustering phase, we use filters to merge tweets that are similar to one another; and finally in the extraction phase, we extract 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.

RELATED WORK

There has been growing interest in using social media for situational awareness during disasters (Cheong and Cheong, 2011; Kumar, Babier, Abbasi, and Liu, 2011; Mandel, Culotta, Boulahanis, Stark, Lewis, Rodrigue, 2012; Meier, Castillo, Imran, Elbassuoni, and Diaz, 2013). Cheong et al. use social network measures such as betweenness and global centrality to identify important clusters and individuals during the 2010-2011 Queensland floods in Australia (Cheong and Cheong, 2011). Kumar et al. analyze the geo-location and specific keywords in a tweet to help first responders gain situational awareness during disasters (Kumar, Babier, Abbasi, and Liu, 2011). Furthermore, Mandel et al. use machine learning classification methods to examine tweets

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

during Hurricane Irene and conclude that the number of Twitter messages correlate with the peaks during the hurricane and that the degree of concern in Twitter messages is dependent on location (Mandel, Culotta, Boulahanis, Stark, Lewis, Rodrigue, 2012). This information can help disaster responders identify where the most help is required during hurricanes. Meier et al. extract "nuggets" of information from disaster (Meier, Castillo, Imran, Elbassuoni, and Diaz, 2013). They utilize machine learning to tools and classify tweets as disaster tweets and specifically extract "caution and advice" tweets, casualty and damage nuggets, donation and offer nuggets, and information source nuggets. Imran et al. use a conditional random field (CRF) to extract damage and casualty information from tweets. The ontology listed in Table 1 builds upon the ontology in the work of Imran et al. (Imran, Elbassuoni, Castillo, Diaz, and Meier, 2013). Our expanded ontology draws from common name of infrastructures in the areas affected by the natural disasters and damage types incurred during natural disasters. Furthermore, we have incorporated the extraction in a pipeline through an application programming interface available at https://github.com/dssg/tweedr. We have added additional features to the extraction phase as well.

While the aforementioned research addresses the information need and situational awareness during natural disasters, current literature lacks a cohesive pipeline that takes into consideration all of the facets of data extraction. In this work, we introduce a cohesive pipeline that extracts relevant information for disaster relief workers through a pipeline, which consists of: classification, extraction, and clustering.

Event	N (kw)	N (geo)	total	keywords
christchurch	757,382	-	757,382	#EQNZ,#CHCH,#NZquake,Christchurch
ike	109,784	-	109,784	Hurricane+Ike,Hurricane,Ike,Galveston,Houston
irene	381,378	3,868,320	4,249,698	#Hurricane,#Irene,#Tropics
moore	840,013	1,576,751	2,416,764	#mooretornado,#moore,newcastle
oklahoma	1,695,156	6,206,775	7,901,931	#oklahoma,#tornado,#oklahomatornado,#okwx,#okc
samoa	233,192	-	233,192	Samoa,tsunami,earthquake
slavelake	43,791	-	43,791	#SlaveLake,Slave+Lake
supertuesday	20,004	-	20,004	Super+Tuesday,Jackson,Memphis,supertuesday
tornado2011a	474,437	-	474,437	Tushka,oklahoma,okwx,arkansas,akwx,tornado
tornado2011b	49,875	-	49,875	#alwx,#okwx,#txwx,#tristatewx,tornado,#ALNeeds, #ALHaves,#WeAreAlabama
vtech	13,783	-	3,783	#vatech,#virginiatech,#hokies,#vtech,#vt
westtx	647,876	175,677	823,553	#WestExplosion,#WestTX
Total	**5,266,671**	**11,827,523**	**17,094,194**	

Table 2: Number of tweets collected by event. We query for tweets both by keyword (kw) and geographical bounding box (geo)

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 aggregator. 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) consist of a bounding box of coordinates around the epicenter of the event. In future work, we aim to develop a more robust method of determining the size of geographical bounding box around the epicenter of the event. Table 2 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.

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 1.

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

EXPERIMENTS

Classification

We compare a number of standard classification algorithms, including K-nearest neighbors, decision trees, naive Bayes, and logistic regression, as implemented in the scikit-learn Python library.[3] We represent each document with a standard unigram feature vector. We also compare with supervised latent Dirichlet allocation, for which we create a Python wrapper of the R LDA package[4] (Blei and McAuliffe, 2010). In each of our classifications, logistic regression appears to be the most reliable across several accuracy measures. Our results can be seen in Table 3.

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 3: Damage/casualty classification results (with standard deviations) using 10-fold cross-validation. Pr: precision, Re: recall, Acc: accuracy, AUC: area under the ROC curve.

Extraction and Clustering

For extraction, we use conditional random fields (CRF) (Sutton and McCallum, 2012), as implemented by the CRFsuite toolkit (Okazaki, 2007).[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, WordNet hypernyms, ngrams, and part of speech tags. To obtain precision, recall, and F1-score values, we split the data using two methods. In Table 1, we use 10-fold cross validation. Additionally, we split the data by disaster, training on the labeled data from our top five disasters 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 between distinct disasters, we can test the accuracy of our classifier on unseen disasters, and even unseen disaster types. In Table 4, we show the overall average performance of the CRF on an unseen disaster.

As seen in Table 1, our entity extraction classifier performs well (obtains an F1-score above 0.5) 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

[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://scikit-learn.org

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

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

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 disasters to evaluate how it performed on disasters it had not seen yet. The results were promising for some disasters, yielding high F1-scores for four of the six disasters evaluated; however, more labeled data is needed to estimate generalization accuracy. The results are reflected in Table 3. 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").

For clustering, we consider two different methods of approximate string matching: Bloom filters and SimHash (Bloom, 1970; Charikar, 2002). In future work, we plan on doing quantitative comparison between these two methods.

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. These metrics assess the ability of the algorithm to general to new disasters.

CONCLUSION

Our aim was to extract meaningful information from tweets during natural disasters. We can use machine learning tools to extract valuable information from noisy social media data. Our results are promising in that they demonstrate that it is possible to extract nuggets of information from heterogeneous Twitter data using a relatively small set of labeled data. 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. Using a combination of classification techniques, conditional random fields for extraction, and clustering, we were able to extract informative and potentially actionable information from tweets produced during natural disasters. In future work, given the low frequency of relevant tweets, methods designed for high class imbalance may be useful here (Lin and Chen, 2012). Furthermore, while our pipeline performs well for the labels in Figure 1, our future work will focus on tweets that are within the domain of labels that did not appear in the set of tweets examined in this work. There is also room for the exploration of conflicting labels (i.e. missing persons or death/casualties) and determining which label is more accurate for a given tweet given the features of the document.

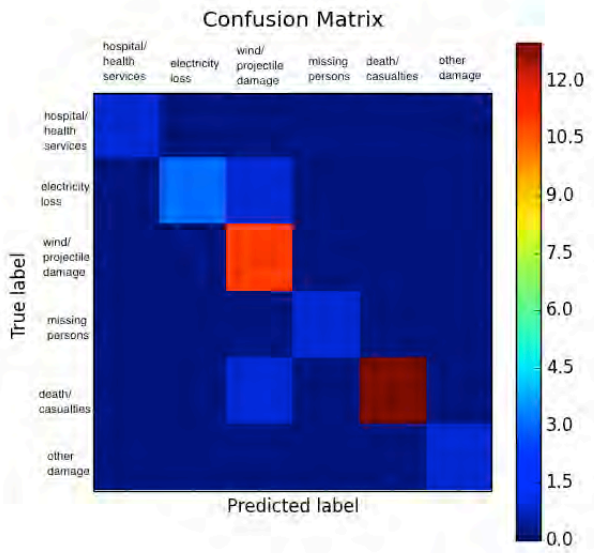

Figure 1: Confusion Matrix of predicted labels using 10-folds cross validation.

ACKNOWLEDGEMENTS

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 thank Patrick Meier and Carlos Castillo (QCRI) for helpful guidance and discussions. 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.

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. N. Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). 2007.

11. C. Sutton and A. K. McCallum. An introduction to conditional random fields. Now Publishers, Hanover, MA, 2012.