Identifying and Categorizing Disaster-Related Tweets

Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, Ken Anderson
University of Colorado, Boulder, CO 80309
[kest1439, mpaul, mpalmer, palen, kena]@colorado.edu

Abstract

This paper presents a system for classifying disaster-related tweets. The focus is on Twitter data generated before, during, and after Hurricane Sandy, which impacted New York in the fall of 2012. We propose an annotation schema for identifying relevant tweets as well as the more fine-grained categories they represent, and develop feature-rich classifiers for relevance and fine-grained categorization.

1 Introduction

Social media provides a powerful lens for identifying people's behavior, decision-making, and information sources before, during, and after wide-scope events, such as natural disasters (Becker et al., 2010; Imran et al., 2014). This information is important for identiying what information is propagated through which channels, and what actions and decisions people pursue. However, so much information is generated from social media services like Twitter that filtering of noise becomes necessary.

Focusing on the 2012 Hurricane Sandy event, this paper presents classification methods for (i) filtering tweets relevant to the disaster, and (ii) categorizing relevant tweets into fine-grained categories such as preparation and evacuation. This type of automatic tweet categorization can be useful both during and after disaster events. During events, tweets can help crisis managers, first responders, and others take effective action. After the event, analysts can use social media information to understand people's behavior during the event. This type of understanding is of critical importance for improving risk communication and protective decision-making leading up to and during disasters, and thus for reducing harm (Demuth et al., 2012).

Our experiments show that such tweets can be classified accurately, and that combining a variety of linguistic and contextual features can substantially improve classifier performance.

2 Related Work

2.1 Analyzing Disasters with Social Media

A number of researchers have used social media as a data source to understand various disasters (Yin et al., 2012; Kogan et al., 2015), with applications such as situational awareness (Vieweg et al., 2010; Bennett et al., 2013) and understanding public sentiment (Doan et al., 2012). For a survey of social media analysis for disasters, see Imran et al. (2014).

Closely related to this work is that of Verma et al. (2011), who constructed classifiers to identify tweets that demonstrate situational awareness in four datasets (Red River floods of 2009 and 2010, the Haiti earthquake of 2010, and Oklahoma fires of 2009). Situational awareness is important for those analyzing social media data, but it does not encompass the entirety of people's reactions. A primary goal of our work is to capture tweets that relate to a hazard event, regardless of situational awareness.

2.2 Tweet Classification

Identifying relevant information in social media is challenging due to the low signal-to-noise ratio. A number of researchers have used NLP to address this challenge. There is significant work in the medical domain related to identifying health crises and events in social media data. Multiple studies have been done to analyze flu-related tweets (Culotta, 2010; Aramaki et al., 2011). Most closely related to our work (but in a different domain) is the flu classification system of Lamb et al. (2013),

which first classifies tweets for relevance and then applies finer-grained classifiers. They build classifiers using syntactic and Twitter-specific features to detect awareness versus infection, self versus others, and whether tweets are relevant to the flu or not. This last classification is very similar to this work, albeit for a different kind of event.

Similar systems have been developed to categorize tweets in more general domains, for example by identifying tweets related to news, events, and opinions (Sankaranarayanan et al., 2009; Sriram et al., 2010). Similar classifiers have been developed for sentiment analysis (Pang and Lee, 2008) to identify and categorize sentiment-expressing tweets (Go et al., 2009; Kouloumpis et al., 2011).

3 Data

3.1 Collection

In late October 2012, Hurricane Sandy generated a massive, disperse reaction in social media channels, with many users expressing their thoughts and actions taken before, during, and after the storm. We performed a keyword collection for this event capturing all tweets using the following keywords from October 23, 2012 to April 5, 2013:

DSNY, cleanup, debris, frankenstorm, garbage, hurricane, hurricanesandy, lbi, occupysandy, perfectstorm, sandy, sandycam, stormporn, superstorm

22.2M unique tweets were collected from 8M unique Twitter users. We then identified 100K users with a geo-located tweet in the time leading up to the landfall of the hurricane, and gathered all tweets generated by those users creating a dataset of 205M tweets produced by 92.2K users. We randomly selected 100 users from approximately 8,000 users who: (i) tweeted at least 50 times during the data collection period, and (ii) posted at least 3 geo-tagged tweets from within the mandatory evacuation zones in New York City. It's critical to filter the dataset to focus on users that were at high risk, and this first pass allowed us to lower the percentage of users that were not in the area and thus not affected by the event. Our dataset includes *all* tweets from these users, not just tweets containing the keywords. Seven users were removed for having predominately non-English tweets. The final dataset contained 7,490 tweets from 93 users, covering a 17 day time period starting one week before landfall (October 23rd to November 10th). Most tweets were irrelevant: Halloween, as well as the upcoming presidential election, yielded a large number of tweets not related to the storm, despite the collection bias toward Twitter users from affected areas.

3.2 Annotation Schema

Tweets were annotated with a fine-grained, multi-label schema developed in an iterative process with domain experts, social scientists, and linguists who are members of our larger project team. The schema was designed to annotate tweets that reflect the attitudes, information sources, and protective decision-making behavior of those tweeting. This schema is not exhaustive—anything deemed relevant that did not fall into an annotation category was marked as **Other**—but it is much richer than previous work. Tweets that were not labeled with any category were considered irrelevant (and as such, considered negative examples for relevance classification). Two additional categories, reporting on family members and referring to previous hurricane events, were seen as important to the event, but were very rare in the data (34 of 7,490 total tweets). The categories identified and annotated are as follows: Tweets could be labeled with any of the following:

Sentiment Tweets that express emotions or personal reactions towards the event, such as humor, excitement, frustration, worry, condolences, etc.

Action Tweets that describe physical actions taken to prepare for the event, such as powering phones, acquiring generators or alternative power sources, and buying other supplies.

Preparation Tweets that describe making plans in preparation for the storm, including those involving altering plans.

Reporting Tweets that report first-hand information available to the tweeter, including reporting on the weather and the environment around them, as well as the observed social situations.

Information Tweets that share or seek information from others (including public officials). This category is distinct from Reporting in that it only includes information received or request from outside sources, and not information perceived first-hand.

Movement Tweets that mention evacuation or sheltering behavior, including mentions of leaving, staying in place, or returning from another lo-

Category	Count	% tweets	Agreement
Relevance			
Relevance	1757	23.5%	48.6% (κ=.569)
Fine-Grained Annotations			
Reporting	1369	77.9%	80.2% (κ=.833)
Sentiment	786	44.7%	71.8% (κ=.798)
Information	600	34.1%	89.8% (κ=.934)
Action	295	16.8%	72.5% (κ=.827)
Preparation	188	10.7%	41.1% (κ=.565)
Movement	53	3.0%	43.3% (κ=.600)

Table 1: The number and percentage of tweets for each label, along with annotator agreement.

cation. Tweets about movement are rare, but especially important in determining a user's response to the event.

3.3 Annotation Results

Two annotators were trained by domain experts using 726 tweets collected for ten Twitter users. Annotation involved a two-step process: first, tweets were labeled for relevance, and then relevant tweets were labeled with the fine-grained categories described above. The annotators were instructed to use the linguistic information, including context of previous and following tweets, as well as the information present in links and images, to determine the appropriate category. A third annotator provided a deciding vote to resolve disagreements.

Table 1 shows the label proportions and annotator agreement for the different tasks. Because each tweet could belong to multiple categories, κ scores were calculated based on agreement per category: if a tweet was marked by both annotators as a particular category, it was marked as agreement for that category. Agreement was only moderate for relevance ($\kappa = .569$). Many tweets did not contain enough information to easily distinguish them, for example: "*tryin to cure this cabin fever!*" and "*Thanks to my kids for cleaning up the yard*" (edited to preserve privacy). Without context, it is difficult to determine whether these tweeters were dealing with hurricane-related issues.

Agreement was higher for fine-grained tagging ($\kappa = .814$). The hardest categories were the rarest (Preparation and Movement), with most confusions between Preparation, Reporting, and Sentiment as shown by the confusion matrix in Table 2.[1]

[1]Dataset available at https://github.com/kevincstowe/chime-annotation

	Act	Inf	Mis	Mov	Pre	Rep	Sen
Act	95	0	0	0	0	11	3
Inf	3	308	1	0	0	7	7
Misc	0	0	3	0	0	4	2
Mov	0	0	0	13	4	5	3
Prep	6	0	0	2	44	30	10
Rep	2	9	1	2	9	517	17
Sent	11	8	1	1	2	31	245

Table 2: The annotation confusion matrix.

4 Classification

We trained binary classifiers for each of the categories in Table 1, using independent classifiers for each of the fine-grained categories (for which a tweet may have none, or multiple).

4.1 Model Selection

Our baseline features are the counts of unigrams in tweets, after preprocessing to remove capitalization, punctuation and stopwords. We initially experimented with different classification models and feature selection methods using unigrams for relevance classification. We then used the best-performing approach for the rest of our experiments. 10% of the data was held out as a development set to use for these initial experiments, including parameter optimization (e.g., SVM regularization).

We assessed three classification models that have been successful in similar work (Verma et al., 2011; Go et al., 2009): support vector machines (SVMs), maximum entropy (MaxEnt) models, and Naive Bayes. We experimented with both the full feature set of unigrams, as well as a truncated set using standard feature selection techniques: removing rare words (frequency below 3) and selecting the n words with the highest pointwise mutual information between the word counts and document labels.

Each option was evaluated on the development data. Feature selection was substantially better than using all unigrams, with the SVM yielding the best F1 performance. For the remaining experiments, SVM with feature selection was used.

4.2 Features

In addition to unigrams, bigram counts were added (using feature selection described above), as well as:

- The **time** of the tweet is particularly relevant to the classification, as tweets during and after the event are more likely to be relevant than

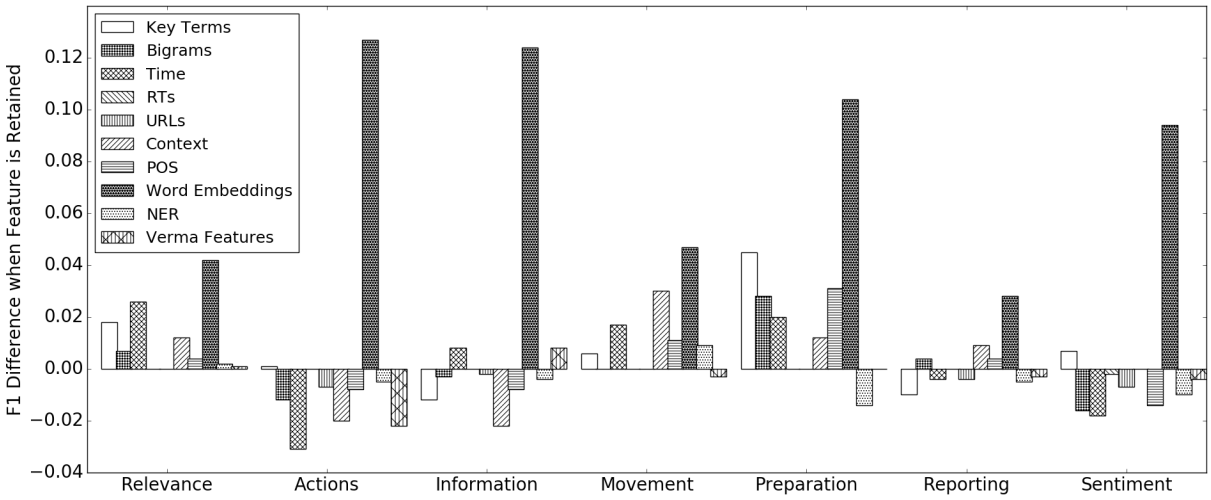

Figure 1: Negated difference in F1 for each feature removed from the full set (positive indicates improvement).

those before. The day/hour of the tweet is represented as a one-hot feature vector.

- We indicate whether a tweet is a **retweet** (RT), which is indicative of information-sharing rather than first-hand experience.
- Each **URL** found within a tweet was stripped to its base domain and added as a lexical feature.
- The annotators noted that **context** was important in classification. The unigrams from the previous tweet and previous two tweets were considered as features.
- We included n-grams augmented with their **part-of-speech** tags, as well as **named entities**, using the Twitter-based tagger of Ritter et al. (2011).
- **Word embeddings** have been used extensively in recent NLP work, with promising results (Goldberg, 2015). A Word2Vec model (Mikolov et al., 2013) was trained on the 22.2M tweets collected from the Hurricane Sandy dataset, using the Gensim package (Řehůřek and Sojka, 2010), using the C-BOW algorithm with negative sampling (n=5), a window of 5, and with 200 dimensions per word. For each tweet, the mean embedding of all words was used to create 200 features.
- The work of Verma et al. (2011) found that formal, objective, and impersonal tweets were useful indicators of situational awareness, and as such developed classifiers to tag tweets with four different categories: formal vs informal, subjective vs objective, personal vs impersonal, and situational awareness vs not. We used these four **Verma** classifiers to tag our Hurricane Sandy dataset and included these tags as features.

	Baseline			All Features			Best Features		
	F1	P	R	F1	P	R	F1	P	R
Relevance	.66	.80	.56	.71	.81	.64	**.72**	.79	.66
Actions	.26	.44	.19	.39	.46	.35	**.41**	.42	.40
Information	.33	.57	.24	.48	.57	.41	**.49**	.50	.49
Movement	.04	.04	.04	.07	.10	.07	**.08**	.10	.07
Preparation	.30	.44	.23	.36	.41	.32	**.36**	.38	.35
Reporting	.52	.76	.40	.73	.71	.75	**.75**	.71	.80
Sentiment	.37	.64	.26	**.53**	.58	.49	.52	.52	.52

Table 3: Results for relevance and fine-grained classification.

4.3 Classification Results

Classification performance was measured using five-fold cross-validation. We conducted an ablation study (Figure 1), removing individual features to determine which contributed to performance. Table 3 shows the cross-validation results using the baseline feature set (selected unigrams only), all features, and the best feature set (features which had a significant effect in the ablation study). In all categories except for Movement, the best features improved over the baseline with $p < .05$.

4.4 Performance Analysis

Time, context, and word embedding features help relevance classification. Timing information is helpful for distinguishing certain categories (e.g., Preparation happens before the storm while Movement can happen before or after). Context was

	Verma Acc	Ext. Acc	Verma F1	Ext. F1
SA	.845	**.856**	.423	**.551**

Table 4: Verma Comparison

also helpful, consistent with annotator observations. A larger context window would be theoretically more useful, as we noted distant tweets influenced annotation choices, but with this relatively small dataset increasing the context window also prohibitively increased sparsity of the feature.

Retweets and URLs were not generally useful, likely because the information was already captured by the lexical features. Part-of-speech tags yielded minimal improvements, perhaps because the lexical features critical to the task are unambiguous (e.g., "hurricane" is always a noun), nor did the addition of features from Verma et al. (2011), perhaps because these classifiers had only moderate performance to begin with and were being extended to a new domain.

Fine-grained classification was much harder. Lexical features (bigrams and key terms) were useful for most categories, with other features providing minor benefits. Word embeddings greatly improved performance across all categories, while most features had mixed results. This is consistent with our expectations of latent semantics : tweets within the same category tend to contain similar lexical items, and word embeddings allow this similarity to be captured despite the limited size of the dataset.

The categories that were most confused were Information and Reporting, and the categories with the worst performance were Movement, Actions, and Preparation. Movement simply lacks data, with only 53 labeled instances. Actions and Preparation contain wide varieties of tweets, and thus patterns to distinguish them are sparse. More training data would help fine-grained classification, particularly for Actions, Preparation, and Movement.

Classification for Reporting performs much better than others. This is likely because these tweets tend to fall into regular patterns: they often use weather and environment-related lexical items like "wind" and "trees", and frequently contain links to images. They also are relatively frequent, making their patterns easier to identify.

4.5 Performance in Other Domains

To see how well our methods work on other datasets, we compared our model to the situational awareness classification in the Verma et al. (2011) datasets described above. We replicated the original Verma et al. (2011) model with similar results, and then adjusted the model to incorporate features that performed positively from our experiments to create an 'extended' model. This entailed adding the mean word embeddings for each tweet as well as adjusting the unigram model to incorporate only key terms by PMI. They report only accuracy, which our system improves marginally, while making this modifications greatly improved F1, as shown in table 4.

5 Conclusion

Compared to the most closely related work of Verma et al. (2011), our proposed classifiers are both more general (identifying all relevant tweets, not just situational awareness) and richer (with fine-grained categorizations). Our experimental results show that it is possible to identify relevant tweets with high precision while maintaining fairly high recall. Fine-grained classification proved much more difficult, and additional work will be necessary to define appropriate features and models to detect more specific categories of language use. Data sparsity also causes difficulty, as many classes lack the positive examples necessary for the machine to reliably classify them, and we continue to work on further annotation to alleviate this issue.

Our primary research aims are to leverage both relevance classification and fine-grained classification to assist crisis managers and first responders. The preliminary results are show that relevant information can be extracted automatically via batch processing after events, and we aim to continue exploring possibilities to extend this approach to real-time processing. To make this research more applicable, our future research goals follow two main paths : the implementation of a real-time processing system that can provide accurate classification during an event, rather than after, and the application of the current results to additional datasets from other domains.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1576.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *ACM International Conference on Web Search and Data Mining (WSDM)*, pages 291–300.

K J Bennett, J M Olsen, S Harris, S Mekaru, A A Livinski, and J S Brownstein. 2013. The perfect storm of information: combining traditional and non-traditional data sources for public health situational awareness during hurricane response. *PLoS Curr*, 5.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Julie L. Demuth, Rebecca E. Morss, Betty Hearn Morrow, and Jeffrey K. Lazo. 2012. Creation and communication of hurricane risk information. *Bulletin of the American Meteorological Society*, 93(8):1133–1145.

Son Doan, Bao Khanh Ho Vo, and Nigel Collier. 2012. An analysis of Twitter messages in the 2011 Tohoku earthquake. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, volume 91 LNICST, pages 58–66.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Processing social media messages in mass emergency: A survey. *arXiv preprint arXiv:1407.7071*.

Marina Kogan, Leysia Palen, and Kenneth M Anderson. 2015. Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. In *ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541.

Alex Lamb, Michael J Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. 2011. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *ICWSM*.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In *CHI*.

Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59.