The Hurricane Sandy Twitter Corpus

Haoyu Wang
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
haoyuw@andrew.cmu.edu

Eduard Hovy
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
hovy@cmu.edu

Mark Dredze
Human Language Technology
Center of Excellence
Johns Hopkins University
Baltimore, MD 21211
mdredze@cs.jhu.edu

Abstract

The growing use of social media has made it a critical component of disaster response and recovery efforts. Both in terms of preparedness and response, public health officials and first responders have turned to automated tools to assist with organizing and visualizing large streams of social media. In turn, this has spurred new research into algorithms for information extraction, event detection and organization, and information visualization. One challenge of these efforts has been the lack of a common corpus for disaster response on which researchers can compare and contrast their work. This paper describes the Hurricane Sandy Twitter Corpus: 6.5 million geotagged Twitter posts from the geographic area and time period of the 2012 Hurricane Sandy.

Introduction

Preparing for and responding to natural disasters is a key function of public health agencies. Before an event, officials work to ensure that resources are in place to respond to likely disasters, and work towards response plans that can be activated in the event of a disaster. During an event, these officials work with other agencies to ensure public safety through the dissemination of critical information and identifying on the ground problems that need immediate attention. All of these roles depend on high quality information about what is happening in the community, which can include information about access to medical and food supplies, damage to public infrastructure, and reports of immediate danger to the public. Increasingly, public health officials are turning to social media for both the collection of information and the dissemination of updates to a community.

With respect to disaster situations, social media can and have been used in the following general ways:

- Some media, such as Twitter, serve as a free and open **distributed sensor network**. If one is able to identify from the tweet stream those tweets relevant to the situation, one may obtain local and timely information otherwise impossible to get. One may also learn about the onset of a disaster before news reaches from the traditional communications channels.
- Some media, including Twitter and even Facebook, serve as widely accessible and public news and **information dispersion mechanisms**. Agencies responding to disasters can post warnings, advisories, and other information to reach people they would otherwise have no access to.
- Most social media allow **post-hoc data collection** about the disaster, which enables moment-by-moment, local-specific, and participant-oriented situation analysis, furthering improved response management in the future.

Already several disaster response organizations are using social media as part of their response strategy. The American Red Cross relies on social media to both learn about what is happening during a disaster and to disseminate information. During Hurricane Sandy, the Red Cross helped people locate emergency shelters, find missing people, and dispel rumors all using social media[1]. Other organizations like Ushahidi's CrisisNET[2] and Humanity Road[3] provide tools for utilizing social media for disaster response. More generally, Google's Crisis Response project[4] enables individuals to collect and share information in the aftermath of a disaster. The United States Federal Emergency Management Agency (FEMA) has identified the use of social media as essential for future emergency management[5]. This requirement has similarly been expressed in the literature (Merchant, Elmer, and Lurie 2011).

Critical to all of these efforts are computational tools that can organize social media, identify critical pieces of information, and visualize developing trends. Without them, the flood of information is overwhelming and counterproductive. This is already an active area of research, including work in information extraction, event detection, geolocation, and HCI tools. For high-quality results, these research endeavors require appropriately tagged and organized social media data from disaster events. While individual research groups have each collected their own partial data sets from

[1] http://www.redcross.org/images/MEDIA_CustomProductCatalog/m22442828_Social_Media_-_Suzanne_Bernier_-_SB_Crisis_Consulting.pdf

[2] http://www.ushahidi.com/blog/product/crisisnet/

[3] http://humanityroad.org/

[4] https://www.google.org/crisisresponse/

[5] http://www.fema.gov/media-library-data/20130726-1816-25045-5167/sfi_report_13.jan.2012_final.docx.pdf

	Tokens	Types
Unigrams	63,898,947	4,227,741
Bigrams	57,344,203	17,228,767
Trigrams	50,902,601	30,138,172

Table 1: N-gram counts for tokens and types.

a variety of disasters, there does not exist a common corpus, focusing on a single clear disaster, on which multiple groups can compare their efforts. Such resources can be critical to developing a thriving research community.

In this paper we present the Hurricane Sandy Twitter Corpus, a collection of 6.5 million geotagged tweets that represent all geotagged tweets from the time and region impacted by Hurricane Sandy, the largest Atlantic hurricane on record. This paper summarizes properties of the data set and makes available of it.

Hurricane Sandy

Hurricane Sandy started as a 2012 late-season post-tropical cyclone in the Atlantic. It initially hit the Caribbean — Jamaica and Cuba— and then turned to the East Coast of the United States in late October. Sandy initially formed as a tropical wave on October 22 and was a Category 3 storm at its peak when it made landfall in Cuba on October 25. When it hit the northeastern United States early on October 29, it was the largest Atlantic hurricane on record. U.S. damage estimates are near $65 billion, making Sandy the second-costliest cyclone to hit the United States since 1900. At least 147 deaths were directly due to Sandy, with 72 of these fatalities occurring in the mid-Atlantic and northeastern US (Blake et al. 2013; Wikipedia 2014).

Corpus

Data Collection

Data were obtained by identifying all tweets from the Twitter firehose that matched a filter. The filter included all tweets from October 22nd, 2012 —the day Sandy was formed— until November 2nd, 2012 — the day that it dissipated (Blake et al. 2013). Additionally, tweets were only included if they were geotagged and located in Washington DC or one of 13 US states affected by Sandy: Connecticut, Delaware, Massachusetts, Maryland, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Rhode Island, South Carolina, Virginia, West Virginia. This filter was based on a set of bounding boxes that covered the desired area, which also covered small parts of adjacent states. No content based filter was applied; the corpus contains tweets both relevant and irrelevant to Hurricane Sandy. We chose to include all tweets as any filter for Sandy specific tweets would be imperfect. In fact, identifying tweets relevant to an event is an active research problem.

In total, the corpus contains 6,556,328 geotagged tweets, and requires 2.3GB of disk space compressed.

	Unique	Total
Tweets	-	6,556,328
Users	265,043	-
Links	499,570	539,515
Hashtags	417,813	1,632,536
Cited Users	960,174	3,523,251
Sources	551	-

Table 2: Twitter specific statistics.

Topic	Number of occurrences
#Sandy	21,338
#sandy	17,289
#oomf	11,732
#100ThingsAboutMe	7,051
#StudyingForTheSAT	5,698
#HurricaneSandy	5,638
#hurricanesandy	5,095
#halloween	4,962
#ToMyFutureSon	4,480
#jobs	4,424

Table 3: Twitter-specific statistics.

Corpus Statistics

To illustrate basic properties of the corpus, we present a variety of statistics of potential use to researchers.

Table 1 shows the basic n-gram statistics, including 1-,2-, and 3-grams. We report both the number of tokens and types.

Statistics specific to Twitter data are shown in Table 2. The dataset contains over a quarter of a million users. Replies are computed based on tweet metadata that indicates when a tweet replies to another tweet. Sources are the number of different applications used to post tweets, e.g., web interface, iOS app, etc. Links are based on the full URL in the tweet metadata, but may still be a shortened URL so the reported number of unique links is an upper bound.

Table 3 shows the 10 most common hashtags in the corpus (case-sensitive). Unsurprisingly, `#Sandy` `#sandy` `#hurricaneSandy` and `#hurricanesandy` are the most common. The other popular hashtags reflect other events happening at the same time (`#halloween`) or trending topics on Twitter: `#100ThingsAboutMe` and `#ToMyFutureSon`. `#oomf` means "One of my friends/followers".

Table 4 shows the 10 most sources (posting methods). Previous work (Petrovic, Osborne, and Lavrenko 2010a) has indicated that the Web is the most popular posting source for tweets. However, in our corpus, the most popular sources are mobile devices. There may be several reasons for this difference. First, we are only considering geotagged tweets, which will be dominated by mobile apps with GPS capabilities. Second, our data set is from 2012, two years after the earlier statistics were published. There may have been a shift in posting methods since 2010. Finally, many people in the areas affected by Hurricane Sandy lost power, and thus used mobile devices as their primary method of internet access.

Figure 1 shows the geographic distribution of the tweets

Source	Tweets
Twitter for iPhone	3,367,307
Twitter for Android	1,952,025
web	235,327
foursquare	226,696
Instagram	165,419
Twitter for Windows Phone	94,883
Tweetbot for iOS	83,001
Mobile Web	78,969
dlvr.it	68,052
TweetCaster for Android	59,209

Table 4: Sources (posting platforms) of Tweets.

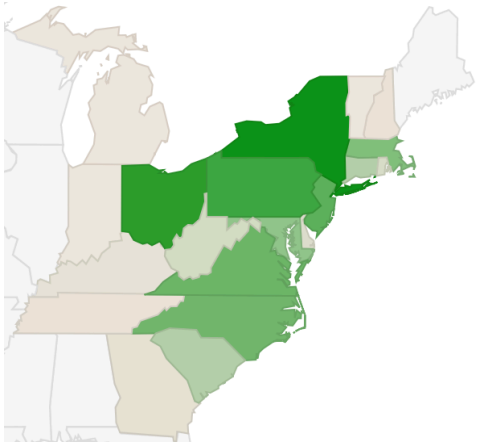

Figure 1: Geographic distribution of the corpus.

according to their geotags. All tweets are from the East Coast of the United States. The numbers of tweets for the five most common states in the corpus are shown in Table 5, along with their population in 2012.

The blue line in Figure 2 shows the number of tweets per day. However, since only a subset of the tweets are relevant to Hurricane Sandy the temporal pattern may not reflect trends in Sandy tweets. While identifying relevant tweets is a research challenge, we use a simple proxy: including only tweets that contain the word "sandy". The precision of this filter is likely high, but the recall is certainly poor. Nevertheless, we see a clear trend in these tweets (red line) as compared to Sandy-related events, which are overplayed on the figure. There is a significant uptick in tweets on October 28, when Sandy continued to move northeast, and a large peak on October 29, when Sandy made its sharp turn toward New Jersey. Figure 3 shows tweets with the keyword "sandy" by day for the five most common states. This demonstrates that the surge of October 29 —the day when the storm hit New York City— comes from New York State tweets.

Finally, in order to give a sense of the range of content in this corpus, we ran LDA (Blei, Ng, and Jordan 2003) using the Stanford Topic Modeling Toolbox[6] with 10 topics.

[6]http://www-nlp.stanford.edu/software/tmt/

State	Tweets	Population
New York	861,593	19,570,000
Pennsylvania	679,847	12,760,000
New Jersey	562,042	8,865,000
Virginia	503,841	8,186,000
North Carolina	483,899	9,752,000

Table 5: The number of tweets for the 5 most common states.

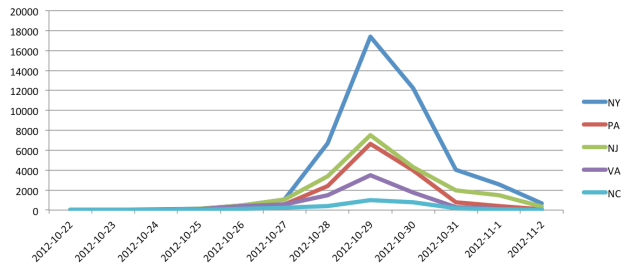

Figure 3: The number of tweets per day that contain the word "sandy" by state.

Figure 6 shows the top words for the topics produced when run on all the data and only those produced when run only on tweets that contained the word "sandy." The first topic found in the whole corpus is related to Hurricane Sandy, while other topics are more general. The topics inferred for the "sandy" tweets are focused on aspects of the hurricane, such as flooding, weather and school closings.

Related Work

Social media has emerged as a popular resource for providing new public health data sources (Ayers, Althouse, and Dredze 2014; Dredze 2012). For disaster response, this has included work on analysis tools (Kumar et al. 2011), crowdsourcing efforts (Goodchild and Glennon 2010; Rogstadius et al. 2011; Gao et al. 2011), developing situational awareness (Yin et al. 2012; Dufty 2012), and finding critical information in a disaster (Imran et al. 2013; Neubig et al. 2011). There are several case studies regarding the use of social media in disasters, such as the 2010 Haitian earthquake (Yates and Paquette 2011), the 2010 Yushu earthquake (Qu et al. 2011) and the 2011 Japanese tsunami (Acar and Muraki 2011). A key problem in disaster management is event detection from social media. This has included first story detection (Petrović, Osborne, and Lavrenko 2010b), using Wikipedia to improve detection (Osborne et al. 2012), open domain event extraction (Ritter et al. 2012), and structured event retrieval (Metzler, Cai, and Hovy 2012).

Data Release

Under the permission for this distribution from twitter, The corpus is only released as a file of tweet IDs. The actual tweets can then be downloaded using the Twitter API. For each ID, we indicate the date of the tweet and indicate which tweets contained the word "sandy".

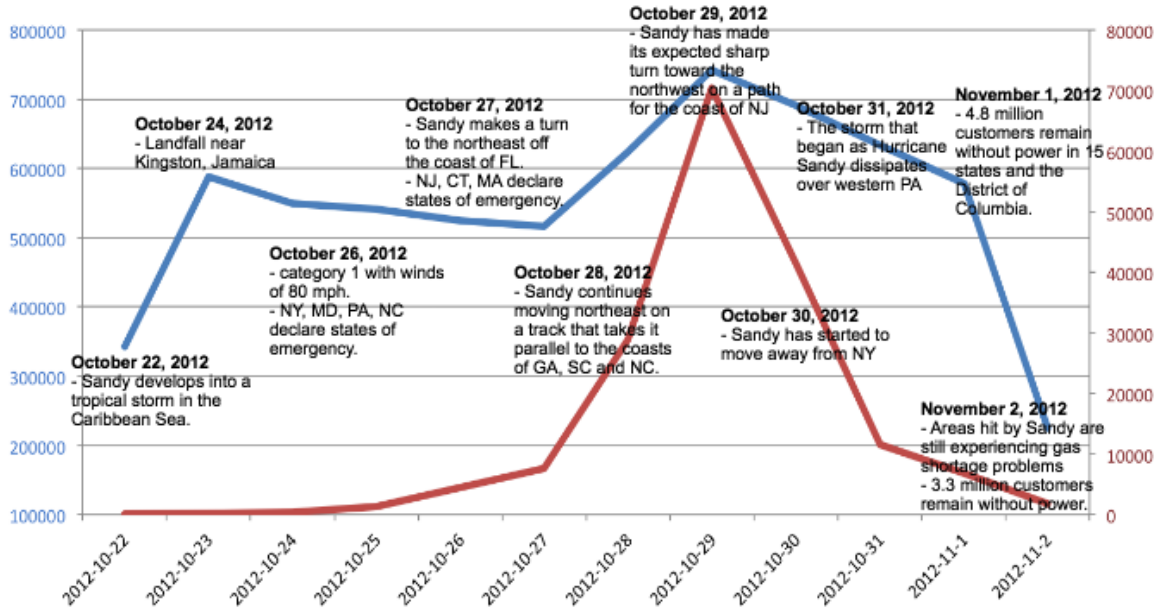

Figure 2: The number of tweets per day and significant events regarding Hurricane Sandy. The blue line shows the number of tweets while the red line shows the number of tweets containing the word "sandy".

All					"Sandy"				
new	school	game	happy	power	tomorrow	wind	ready	will	safe
sandy	tomorrow	god	too	from	school	rain	food	news	everyone
others	morning	watching	haha	more	work	down	water	about	stay
york	work	new	thanks	phone	day	going	gas	your	coast
hurricane	off	tonight	miss	one	thanks	weather	some	how	east
park	today	play	halloween	has	today	outside	wine	obama	hope
center	class	thank	birthday	obama	back	crazy	getting	has	who
house	fuck	watch	follow	our	will	into	bring	weather	jersey
city	going	team	yes	romney	days	got	got	state	those
from	why	show	back	got	but	will	phone	emergency	god

Table 6: The top words for selected topics found for the entire corpus (left) and tweets containing the word "sandy" (right).

References

Acar, A., and Muraki, Y. 2011. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *International Journal of Web Based Communities* 7(3):392–402.

Ayers, J. W.; Althouse, B. M.; and Dredze, M. 2014. Could behavioral medicine lead the web data revolution? *JAMA* 311(14):1399–1400.

Blake, E. S.; Kimberlain, T. B.; Berg, R. J.; Cangialosi, J.; and Beven II, J. L. 2013. Tropical cyclone report: Hurricane sandy. *National Hurricane Center* 12.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Dredze, M. 2012. How social media will change public health. *Intelligent Systems, IEEE* 27(4):81–84.

Dufty, N. 2012. Using social media to build community disaster resilience. *The Australian Journal of Emergency Management* 27(1):40–45.

Gao, H.; Barbier, G.; Goolsby, R.; and Zeng, D. 2011. Harnessing the crowdsourcing power of social media for disaster relief. Technical report, DTIC Document.

Goodchild, M. F., and Glennon, J. A. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *Int. Journal of Digital Earth* 3(3):231–241.

Imran, M.; Elbassuoni, S. M.; Castillo, C.; Diaz, F.; and Meier, P. 2013. Extracting information nuggets from disaster-related messages in social media.

Kumar, S.; Barbier, G.; Abbasi, M. A.; and Liu, H. 2011. Tweettracker: An analysis tool for humanitarian and disaster relief. In *ICWSM*.

Merchant, R. M.; Elmer, S.; and Lurie, N. 2011. Integrating social media into emergency-preparedness efforts. *New England Journal of Medicine* 365(4):289–291.

Metzler, D.; Cai, C.; and Hovy, E. 2012. Structured event retrieval over microblog archives. In *NAACL*.

Neubig, G.; Matsubayashi, Y.; Hagiwara, M.; and Murakami, K. 2011. Safety information mining-what can nlp do in a disaster-. In *IJCNLP*, 965–973.

Osborne, M.; Petrovic, S.; McCreadie, R.; Macdonald, C.; and Ounis, I. 2012. Bieber no more: First story detection using twitter and wikipedia. In *Workshop on Time-aware Information Access*, volume 12.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2010a. The edinburgh twitter corpus. In *NAACL Workshop on Computational Linguistics in a World of Social Media*.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010b. Streaming first story detection with application to twitter. In *NAACL*.

Qu, Y.; Huang, C.; Zhang, P.; and Zhang, J. 2011. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *CSCW*, 25–34. ACM.

Ritter, A.; Etzioni, O.; Clark, S.; et al. 2012. Open domain event extraction from twitter. In *KDD*.

Rogstadius, J.; Kostakos, V.; Laredo, J.; and Vukovic, M. 2011. Towards real-time emergency response using crowd supported analysis of social media. In *CHI workshop on crowdsourcing and human computation, systems, studies and platforms*.

Wikipedia. 2014. Hurricane Sandy — Wikipedia, the free encyclopedia. [Online; accessed 12-September-2014].

Yates, D., and Paquette, S. 2011. Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake. *International Journal of Information Management* 31(1):6–13.

Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; and Power, R. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems* 27(6):52–59.