Finding Eyewitness Tweets During Crises

Fred Morstatter[1], Nichola Lubold[1], Heather Pon-Barry[1], Jürgen Pfeffer[2], and Huan Liu[1]
[1]Arizona State University, Tempe, Arizona, USA
[2]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
{fred.morstatter, nlubold, ponbarry, huan.liu}@asu.edu, jpfeffer@cs.cmu.edu

Abstract

Disaster response agencies incorporate social media as a source of fast-breaking information to understand the needs of people affected by the many crises that occur around the world. These agencies look for tweets from within the region affected by the crisis to get the latest updates on the status of the affected region. However only 1% of all tweets are "geotagged" with explicit location information. In this work we seek to identify non-geotagged tweets that originate from within the crisis region. Towards this, we address three questions: (1) is there a difference between the language of tweets originating within a crisis region, (2) what linguistic patterns differentiate within-region and outside-region tweets, and (3) can we automatically identify those originating within the crisis region in real-time?

1 Introduction

Due to Twitter's massive popularity, it has become a tool used by first responders—those who provide first-hand aid in times of crisis—to understand crisis situations and identify the people in the most dire need of assistance (United Nations, 2012). To do this, first responders can survey "geotagged" tweets: those where the user has supplied a geographic location. The advantage of geotagged tweets is that first responders know whether a person is tweeting from within the affected region or is tweeting from afar. Tweets from within this region are more likely to contain emerging topics (Kumar et al., 2013) and tactical, actionable, information that contribute to situational awareness (Verma et al., 2011).

A major limitation of surveying geotagged tweets is that only 1% of all tweets are geotagged (Morstatter et al., 2013). This leaves the first responders unable to tap into the vast majority of the tweets they collect. This limitation leads to the question driving this work: can we discover whether a tweet originates from within a crisis region using *only the language used of the tweet*?

We focus on the language of a tweet as the defining factor of location for three major reasons: (1) the language of Twitter users is dependent on their location (Cheng et al., 2010), (2) the text is readily available in every tweet, and (3) the text allows for real-time analysis. Due to the short time window presented by most crises, first responders need to be able to locate users quickly.

Towards this goal, we examine tweets from two recent crises: the Boston Marathon bombing and Hurricane Sandy. We show that linguistic differences exist between tweets authored inside and outside the affected regions. By analyzing the text of individual tweets we can predict whether the tweet originates from within the crisis region, in real-time. To better understand the characteristics of crisis-time language on Twitter, we conclude with a discussion of the linguistic features that our models find most discriminative.

2 Language Differences in Crises

In order for a language-based approach to be able to distinguish tweets inside of the crisis region, the language used by those in the region during crisis has to be different from those outside. In this section, we verify that there are both regional and temporal differences in the language tweeted. To start, we introduce the data sets we use throughout the rest of this paper. We then measure the difference in language, finding that language changes temporally and regionally at the time of the crisis.

2.1 Twitter Crisis Datasets

The Twitter data used in our experiments comes from two crises: the Boston Marathon bombing and Hurricane Sandy. Both events provoked a significant Twitter response from within and beyond

Table 1: Properties of the Twitter crisis datasets.

Property	Boston	Sandy
Crisis Start	15 Apr 14:48	29 Oct 20:00
Crisis End	16 Apr 00:00	30 Oct 01:00
Epicenter	$42.35, -71.08$	$40.75, -73.99$
Radius	19 km	20 km
$\lvert\mathbf{IR}\rvert$	11,601	5,017
$\lvert\mathbf{OR}\rvert$	541,581	195,957
$\lvert\mathbf{PC\text{-}IR}\rvert$	14,052	N/A
$\lvert\mathbf{PC\text{-}OR}\rvert$	228,766	N/A

the affected regions.

The **Boston Marathon Bombing** occurred at the finish line of the Boston Marathon on April 15th, 2013 at 14:48 Eastern. We collected geotagged tweets from the continental United States from 2013-04-09 00:00 to 2013-04-22 00:00 utilizing Twitter's Filter API.

Hurricane Sandy was a "superstorm" that ravaged the Eastern United States in October, 2012. Utilizing Twitter's Filter API, we collected tweets based on several keywords pertaining to the storm. Filtering by keywords, this dataset contains both geotagged and non-geotagged data beginning from the day the storm made landfall (2012-10-29) to several days after (2012-11-02).

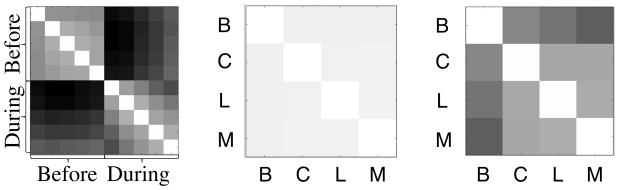
2.2 Data Partitioning

For the Boston Bombing and Hurricane Sandy datasets, we partitioned the tweets published *during the crisis time* into two distinct parts based on location: (1) inside the crisis region (**IR**), and (2) outside the crisis region (**OR**).

For the Boston Bombing dataset, we are able to extract two additional groups: (1) pre-crisis tweets (posted before the time of the crisis) from inside the crisis region (**PC-IR**) and (2) pre-crisis tweets from outside the crisis region (**PC-OR**). We take a time-based sample from 10:00–14:48 Eastern on April 15th, 2013 to obtain **PC-IR** and **PC-OR**. Because the bombing was an abrupt event with no warning, we choose a time period immediately preceding its onset. The number of tweets in each dataset partition is shown in Table 1.

2.3 Pre-Crisis vs. During-Crisis Language

For the Boston dataset, we compare the words used hour by hour between 10:00–19:00 on April 15th. For each pair of hours, we compute the Jensen-Shannon (J-S) divergence (Lin, 1991) of the probability distributions of the words used within those hours. Figure 1(a) shows these J-S divergence values. We see an abrupt change in language in the hours before the bombing (10:00–14:00) and those after the bombing (15:00–19:00). We also note that the tranquil hours are relatively stable. This suggests that language models trained on tweets from tranquil time are less informative for modeling crisis-time langauge.



(a) Temporal language differences.

(b) Geographic language differences: tranquil time.

(c) Geographic language differences: crisis time.

Figure 1: Temporal and geographic differences of language (calculated using Jensen-Shannon divergence); darker shades represent greater difference. To illustrate geographic differences, we compare Boston with three other major U.S. cities.

2.4 IR vs. OR Language

We verify that the tweets authored inside of the crisis use different words from those outside the region. We compare the difference in Boston (**B**) to three other major U.S. cities: Chicago (**C**), Los Angeles (**L**), and Miami (**M**). To obtain a baseline, we compare the cities during tranquil times using **PC-IR** and **PC-OR** datasets. The results are shown in Figure 1. The tranquil time comparison, shown in Figure 1(b), displays a low divergence between all pairs of cities. In contrast, Figure 1(c) shows a wider divergence between the same cities, with Boston displaying the greatest divergence.

3 Linguistic Features

As Twitter is a conversational, real-time, microblogging site, the structure of tweets offers many opportunities for extracting different types of features that represent the different linguistic properties of informal text. Our approach is to compare the utility, in classifying tweets as **IR** or **OR**, of several linguistic features. We preprocess the tweets by extracting tokens using the CMU Twitter NLP tokenizer (Owoputi et al., 2013).

Unigrams and **Bigrams** We extract the raw frequency counts of the word unigrams and bigrams.

POS Tags We extract part-of-speech tags for each word in the tweet using the CMU Twitter NLP POS tagger (Owoputi et al., 2013). We con-

sider CMU ARK POS tags, developed specifically for the dynamic and informal nature of tweets, as well as Penn Treebank (PTB) style POS tags. The ARK POS tags are coarser than the PTB tags and can identify Twitter-specific entities in the data like hashtags. By comparing both tag sets, we can measure the effectiveness of both the fine-grained versus coarse-grained tag sets.

Shallow Parsing In addition to the POS tags, we extract shallow parsing tags along with the headword associated with the tag using the tool provided by Ritter et al. (2011). For example, in the noun phrase "the movie" we would extract the headword "movie" and represent it as [...movie...]$_{NP}$. The underlying motivation is that this class may give more insight into the syntactic differences of **IR** tweets versus **OR** tweets.

Crisis-Sensitive (CS) Features We create a mixed-class of "crisis sensitive" features composed of *word-based*, *part of speech*, and *syntactic constituent* attributes. These are based on our analysis of the Boston Marathon data set. We later apply these features to the Hurricane Sandy data set to validate whether the features are generalizable across crises and discuss this in the results.

• We extract **"in" prepositional phrases** of the form [in ... /N]$_{PP}$. For example, "in Boston." The motivation is this use of "in," such as with a location or a nonspecific time, may be indicative of crisis language.

• We extract verbs in relationship to the **existential *there***. As the existential *there* is usually the grammatical subject and describes an abstraction, it may be indicative of situational awareness messages within the disaster region.

• **Part-of-Speech tag sequences** that are frequent in IR tweets (from our development set) are given special consideration. We find sequences which are used more widely during the time of this disaster. Some of the ARK tag sequences include: ⟨N R⟩, ⟨L A⟩, ⟨N P⟩, ⟨P D N⟩, ⟨L A !⟩, ⟨A N P⟩.

4 Experiments

Here, we assess the effectiveness of our linguistic features at the task of identifying tweets originating from within the crisis region. To do this we use a Naïve Bayes classifier configured with an individual set of feature classes. Each of our features are represented as raw frequency counts of the number of times they occur within the tweet. The output is a prediction of whether the tweet is inside region (**IR**) or outside region (**OR**). We identify the features that can differentiate the two classes of users, and we show that this process can indeed be automated.

Table 2: Top Feature Combinations: Unigrams (Uni), Bigrams (Bi) and Crisis-Sensitive (CS) combinations have the best results.

Top Feature Combos	**Prec.**	**Recall**	**F1**
	Boston Bombing		
Uni + Bi	0.853	0.805	0.828
Uni + Bi + Shallow Parse	0.892	0.771	0.828
Uni + Bi + CS	0.857	0.806	**0.831**
All Features	0.897	0.742	0.812
	Hurricane Sandy		
Uni + Bi	0.942	0.820	0.877
Uni + Bi + Shallow Parse + CS	0.956	0.803	0.873
Uni + Bi + CS	0.947	0.826	**0.882**
All Features	0.960	0.786	0.864

4.1 Experiment Procedure

We ensure a 50/50 split of **IR** and **OR** instances by sampling the **OR** dataset. Using the classifier described above, we perform 3×5-fold cross validation on the data. Because of the 50/50 split, a "select-all" baseline that labels all tweets as **IR** will have an accuracy of 50%, a precision of 50%, and a recall of 100%. All precision and recall values are from the perspective of the **IR** class.

4.2 Feature Class Analysis

We compare all possible combinations of individual feature classes and we report precision, recall, and F1-scores for the best combinations in Table 2.

In both crises all of the top performing feature combinations contain both bigram and unigram feature classes. However, our top performing feature combinations demonstrate that bigrams in combination with unigrams have added utility. We also see that the crisis-sensitive features are present in the top performing combinations for both data sets. The CS feature class was derived from Boston Bombing data, so its presence in the top groups from Hurricane Sandy is an indication that these features are general, and may be useful for finding users in these and future crises.

4.3 Most Informative Linguistic Features

To see which individual features within the classes give the best information, we make a modification to the experiment setup described in Section 4.1: we replace the Naïve Bayes classifier with a Logistic Regression classifier to utilize the coefficients it learns as a metric for feature importance. We report the top three features of each class label from each feature set in Table 3.

The individual unigram and bigram features with the most weight have a clear semantic rela-

Table 3: Top 3 features indicative of each class within each feature set for both crises.

Feature Set (Class)	Boston Marathon Bombing	Hurricane Sandy
Unigram (**IR**)	#prayforboston, boston, explosion	@kiirkobangz, upset, staying
Unigram (**OR**)	money, weather, gone	#tomyfuturechildren, #tomyfutureson, bye
Bigram (**IR**)	⟨in boston⟩, ⟨the marathon⟩, ⟨i'm safe⟩	⟨railroad :⟩, ⟨evacuation zone⟩, ⟨storm warning⟩
Bigram (**OR**)	⟨i'm at⟩, ⟨s/o to⟩, ⟨, fl⟩	⟨you will⟩, ⟨: i've⟩, ⟨hurricane ,⟩
ARK POS (**IR**)	⟨P $ ^ ⟩, ⟨L !⟩, ⟨! R P⟩	⟨P #⟩, ⟨∼ ^ A⟩, ⟨@ @ #⟩
ARK POS (**OR**)	⟨O #⟩, ⟨! N O⟩, ⟨L P R⟩	⟨P V $⟩, ⟨A ^ ^⟩, ⟨N L A⟩
PTB POS (**IR**)	⟨CD NN JJ⟩, ⟨CD VBD⟩, ⟨JJS NN TO⟩	⟨USR DT JJS⟩, ⟨VB TO RB⟩, ⟨IN RB JJ⟩
PTB POS (**OR**)	⟨NNP -RRB-⟩, ⟨. JJ JJ⟩, ⟨JJ NN CD⟩	⟨NNS IN NNS⟩, ⟨PRP JJ PRP⟩, ⟨JJ NNP NNP⟩
Shallow Parse (**IR**)	[...explosion...]$_{NP}$, [...marathon...]$_{NP}$, [...bombs...]$_{NP}$	[...bomb...]$_{NP}$, [...waz...]$_{VP}$, [...evacuation...]$_{NP}$
Shallow Parse (**OR**)	[...school...]$_{NP}$, [...song...]$_{NP}$, [...breakfast...]$_{NP}$	[...school...]$_{NP}$, [...head...]$_{NP}$, [...wit...]$_{PP}$
CS (**IR**)	[in boston/N]$_{PP}$, [for boston/N]$_{PP}$, ⟨i'm/L safe/A⟩	⟨while/P a/D hurricane/N⟩, ⟨of/P my/D house/N⟩, [in http://t.co/UxkKJLoX/N]$_{PP}$
CS (**OR**)	⟨to/P the/D beach/N⟩, [at la/N]$_{PP}$, [in love/N]$_{PP}$	[like water/N]$_{PP}$, ⟨shutdowns/N on/P⟩, ⟨prayer/N for/P⟩

tionship to the crisis. Comparing the two crises, the top features for Hurricane Sandy are more concerned with user-user communication. For example, the heavily-weighted ARK POS trigram ⟨@ @ #⟩ is highly indicative of users spreading information between each other. One explanation is that the concern with communication could be a result of the warning that came from the storm. The bigram ⟨`hurricane ,`⟩ is the 3rd most indicative of a tweet originating from *outside* the region. This is likely because the word occurs in the general discussion outside of the crisis region.

5 Related Work

Geolocation: Eisenstein et al. (2010) first looked at the problem of using latent variables to explain the distribution of text in tweets. This problem was revisited from the perspective of geodesic grids in Wing and Baldridge (2011) and further improved by flexible adaptive grids (Roller et al., 2012). Cheng et al. (2010) employed an approach that looks at a user's tweets and estimates the user's location based on words with a local geographical scope. Han et al. (2013) combines tweet text with metadata to predict a user's location.

Mass Emergencies: De Longueville et al. (2009) study Twitter's use as a sensor for crisis information by studying the geographical properties of users' tweets. In Castillo et al. (2011), the authors analyze the text and social network of tweets to classify their newsworthiness. Kumar et al. (2013) use geotagged tweets to find emerging topics in crisis data. Investigating linguistic features, Verma et al. (2011) show the efficacy of language features at finding crisis-time tweets that contain tactical, actionable information, contributing to *situational awareness*. Using a larger dataset, we automatically discover linguistic features that can help with situational awareness.

6 Conclusion and Future Work

This paper addresses the challenge of finding tweets that originate from a crisis region using only the language of each tweet. We find that the tweets authored from within the crisis region do differ, from both tweets published during tranquil time periods and from tweets published from other geographic regions. We compare the utility of several linguistic feature classes that may help to distinguish the two classes and build a classifier based on these features to automate the process of identifying the **IR** tweets. We find that our classifier performs well and that this approach is suitable for attacking this problem.

Future work includes incorporating the wealth of tweets preceding the disaster for better predictions. Preliminary tests have shown positive results; for example we found early, non-geotagged reports of flooding in the Hoboken train tunnels during Hurricane Sandy[1]. Future work may also consider additional features, such as sentiment.

Acknowledgments

This work is sponsored in part by the Office of Naval Research, grants N000141010091 and N000141110527, and the Ira A. Fulton Schools of Engineering, through fellowships to F. Morstatter and N. Lubold. We thank Alan Ritter and the ARK research group at CMU for sharing their tools.

[1] An extended version of this paper is available at: http://www.public.asu.edu/~fmorstat/paperpdfs/lang_loc.pdf.

References

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684. ACM.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768. ACM.

Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. 2009. "OMG, from here, I can see the flames!": a use case of mining Location Based Social Networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 73–80. ACM.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A Stacking-based Approach to Twitter User Geolocation Prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013): System Demonstrations*, pages 7–12.

Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. 2013. Whom Should I Follow?: Identifying Relevant Users During Crises. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 139–147, New York, NY, USA. ACM.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of The International Conference on Weblogs and Social Media*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP*.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised Text-Based Geolocation using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

United Nations. 2012. *Humanitarianism in the Network Age*. United Nations Office for the Coordination of Humanitarian Affairs.

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*.

Benjamin Wing and Jason Baldridge. 2011. Simple Supervised Document Geolocation with Geodesic Grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 955–964.