# A Demographic Analysis of Online Sentiment during Hurricane Irene

**Benjamin Mandel***, **Aron Culotta***, **John Boulahanis**+,
**Danielle Stark**+, **Bonnie Lewis**+, **Jeremy Rodrigue**+
*Department of Computer Science and Industrial Technology
+Department of Sociology and Criminal Justice
Southeastern Louisiana University
Hammond, LA 70402

## Abstract

We examine the response to the recent natural disaster Hurricane Irene on Twitter.com. We collect over 65,000 Twitter messages relating to Hurricane Irene from August 18th to August 31st, 2011, and group them by location and gender. We train a sentiment classifier to categorize messages based on level of concern, and then use this classifier to investigate demographic differences. We report three principal findings: (1) the number of Twitter messages related to Hurricane Irene in directly affected regions peaks around the time the hurricane hits that region; (2) the level of concern in the days leading up to the hurricane's arrival is dependent on region; and (3) the level of concern is dependent on gender, with females being more likely to express concern than males. Qualitative linguistic variations further support these differences. We conclude that social media analysis provides a viable, real-time complement to traditional survey methods for understanding public perception towards an impending disaster.

## Introduction

In 2011, natural disasters cost the United States more than 1,000 lives and $52 billion. The number of disasters costing over $1 billion in 2011 (twelve) is more than in the entire decade of the 1980s.[1] As the number of people living in disaster-prone areas grows, it becomes increasingly important to have reliable, up-to-the-minute assessments of emergency preparedness during impending disasters. Understanding issues such as personal risk perception, preparedness, and evacuation plans helps public agencies better tailor emergency warnings, preparations, and response.

Social scientists typically investigate these issues using polling data. The research shows significant demographic differences in response to government warnings, personal risk assessment, and evacuation decisions (Perry and Mushkatel, 1986; Perry and Lindell, 1991; Goltz et al., 1992; Fothergill et al., 1999; West and Orr, 2007; Enarson, 1998). For example, Fothergill et al. (1999) find that minorities differ in their risk perception and in their response to emergency warnings, with some groups having fatalistic sentiments that lead to greater fear and less preparedness. Goltz et al. (1992) find that people with lower income and education, Hispanics, and women all expressed greater fear of earthquakes.

This past research suggests governments could benefit by tailoring their messaging and response to address the variability between groups. While survey data have advanced our knowledge of these issues, they have two major drawbacks for use in disaster research. First, most surveys rely on responses to hypothetical scenarios, for example by asking subjects if they would evacuate under certain scenarios. This *hypothetical bias* is well-known (Murphy et al., 2005). Second, surveys are often impractical in disaster scenarios. In a rapidly-changing environment, governments cannot wait for a time-consuming survey to be conducted and the results analyzed before making warning and response decisions. Additionally, survey response rates shortly before or after a disaster are likely to be quite low, as citizens are either without power or are busy preparing or rebuilding. Thus, it is difficult to collect data

[1] "Record year for billion-dollar disasters", CBS News, December 7, 2011.

*Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pages 27–36,
Montréal, Canada, June 7, 2012.

during the critical times immediately before and after the disaster.

In this paper, we investigate the feasibility of assessing public risk perception using social media analysis. Social media analysis has recently been used to estimate trends of interest such as stock prices (Gilbert and Karahalios, 2010), movie sales (Asur and Huberman, 2010), political mood (O'Connor et al., 2010a), and influenza rates (Lampos and Cristianini, 2010; Culotta, 2010; Culotta, 2012). We apply a similar methodology here to assess the public's level of concern toward an impending natural disaster.

As a case study, we examine attitudes toward Hurricane Irene expressed on Twitter.com. We collect over 65,000 Twitter messages referencing Hurricane Irene between August 18th and August 31st, 2011; and we train a sentiment classifier to annotate messages by level of concern. We specifically look at how message volume and sentiment varies over time, location, and gender.

Our findings indicate that message volume increases over the days leading up to the hurricane, and then sharply decreases following its dispersal. The timing of the increase and subsequent decrease in messages differs based on the location relative to the storm. There is also an increasing proportion of concerned messages leading up to Hurricane Irene's arrival, which then decreases after Irene dissipation. A demographic analysis of the proportion of concerned messages shows significant differences both by region and gender. The gender differences in particular are supported by previous survey results from the social science literature (West and Orr, 2007). These results suggest that social media analysis is a viable technology for understanding public perception during a hurricane.

The remainder of the paper is organized as follows: First, we describe the data collection methodology, including how messages are annotated with location and gender. Next, we present sentiment classification experiments comparing various classifiers, tokenization procedures, and feature sets. Finally, we apply this classifier to the entire message set and analyze demographic variation in levels of concern.

## Data Collection

Irene became a tropical storm on August 20th, 2011, and hit the east coast of the United States between August 26th and 28th. This hurricane provides a compelling case to investigate for several reasons. First, Irene affected many people in many states, meaning that regional differences in responses can be investigated. Second, there was considerable media and political attention surrounding Hurricane Irene, leading to it being a popular topic on social network sites. Third, the fact that there was forewarning of the hurricane means that responses to it can be evaluated over time.

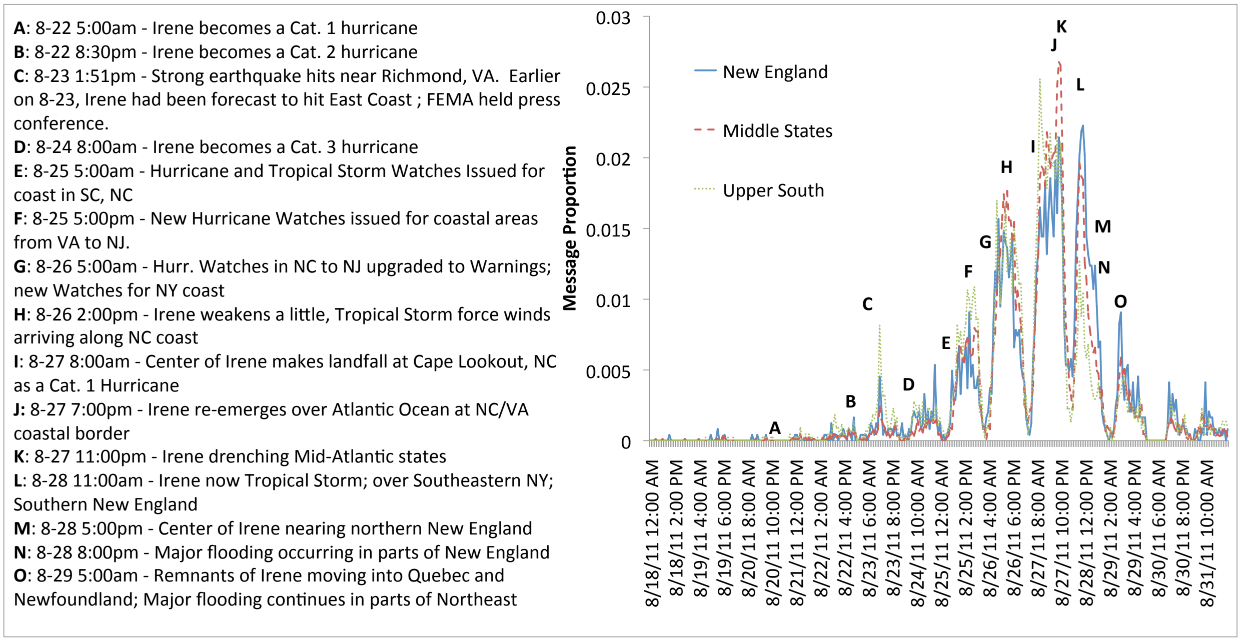
Twitter is a social networking site that allows users to post brief, public messages to their followers. Using Twitter's API[2], we can sample many messages as well as their meta-data, such as time, location, and user name. Also, since Twitter can be used on smart phones with batteries, power outages due to natural disasters will presumably have less of an effect on the volume of messages.

Using Twitter's sampling API ("spritzer"), we sample approximately uniformly from all messages between August 18 and August 31. We then perform keyword filtering to collect messages containing the words "Irene" or "Hurricane", or the hashtag "#Irene". During the period of August 18th to August 31st, messages containing these keywords are overwhelmingly related to Hurricane Irene and not some other event. This results in 65,062 messages.

### Inferring Location

In order to determine the location of the message sender, we process the user-reported location data from that user's profile. Since not all users enter accurate location data, we search for specific keywords in order to classify the messages by state. For example, if the location data contains a token "VT" or "Vermont," it is labeled as coming from Vermont. (See Appendix A for more details.) The locations we consider are the 13 states directly affected by Hurricane Irene, plus Washington DC. These locations are then grouped into 3 regions. First, the New England region consists of the states of Connecticut, Massachusetts, Rhode Island, New Hampshire, Vermont, and Maine. Second, the Middle States region

[2] http://dev.twitter.com



**A**: 8-22 5:00am - Irene becomes a Cat. 1 hurricane
**B**: 8-22 8:30pm - Irene becomes a Cat. 2 hurricane
**C**: 8-23 1:51pm - Strong earthquake hits near Richmond, VA. Earlier on 8-23, Irene had been forecast to hit East Coast ; FEMA held press conference.
**D**: 8-24 8:00am - Irene becomes a Cat. 3 hurricane
**E**: 8-25 5:00am - Hurricane and Tropical Storm Watches Issued for coast in SC, NC
**F**: 8-25 5:00pm - New Hurricane Watches issued for coastal areas from VA to NJ.
**G**: 8-26 5:00am - Hurr. Watches in NC to NJ upgraded to Warnings; new Watches for NY coast
**H**: 8-26 2:00pm - Irene weakens a little, Tropical Storm force winds arriving along NC coast
**I**: 8-27 8:00am - Center of Irene makes landfall at Cape Lookout, NC as a Cat. 1 Hurricane
**J:** 8-27 7:00pm - Irene re-emerges over Atlantic Ocean at NC/VA coastal border
**K**: 8-27 11:00pm - Irene drenching Mid-Atlantic states
**L**: 8-28 11:00am - Irene now Tropical Storm; over Southeastern NY; Southern New England
**M**: 8-28 5:00pm - Center of Irene nearing northern New England
**N**: 8-28 8:00pm - Major flooding occurring in parts of New England
**O**: 8-29 5:00am - Remnants of Irene moving into Quebec and Newfoundland; Major flooding continues in parts of Northeast

Figure 1: Results from Hurricane Irene Twitter data showing the influence of disaster-related events on the number of messages from each region. The y-axis is the proportion of all Irene-related messages from each region that were posted during each hour.

consists of New York, New Jersey, and Pennsylvania. Third, the Upper South region consists of North Carolina, Virginia, Maryland, Delaware, and Washington DC.

Of the messages that we collect between August 18th and August 31st, 15,721 are identified as belonging to one of the directly affected areas. Grouped into regions, we find that 2,424 are from New England, 8,665 are from the Middle-States region, and 4,632 are from the Upper South region.

Figure 1 displays the messages per hour from each of the three regions. The y-axis is normalized over all messages from that region — e.g., a value of 0.02 for New England means that 2% of all messages from New England over the 10 day span were posted in that hour. This allows us to see which time periods were the most active for that region. Indeed, we see that the spikes occur in geographical order of the hurricane's path, from the South, to the Mid-Atlantic region, and finally New England. Additionally, Figure 1 is marked with data points indicating which events were occurring at that time.

There are several obvious limitations of this approach (as explored in Hecht et al. (2011)). For example, users may enter false location information, have an outdated profile, or may be posting messages from a different location. Assuming these issues introduce no systemic bias, aggregate analyses should not be significantly impacted (as supported by the observed trends in Figure 1).

## Inferring Gender

To determine the gender of the message sender, we process the name field from the user's profile obtained from the Twitter API. The U.S. Census Bureau provides a list of the most popular male and female names in the United States. The lists contain over 1,000 of the most common male names and over 4,000 of the most common female names. After removing names that can be either male or female (for example, Chris or Dana), we match the first name of the user to the list of names obtained from the census. Users that cannot be classified in such a manner are labeled as unsure. The data contains a total of 60,808 distinct users, of which 46% are assigned a gender (of those, 55% are female, 45% male). We find that many of the unlabeled users are news agencies. A similar methodology is used by Mislove et al. (2011). As with geographic inference,

| **Total Sample 8/18/2011-8/31/2011** | 25,253,444 |
|---:|:---:|
| **Matching Irene Keywords** | 65,062 |
| **Female-indicative names** | 16,326 |
| **Male-indicative names** | 13,597 |
| **Mid-Atlantic states** | 8,665 |
| **Upper-South states** | 4,632 |
| **New England states** | 2,424 |

Table 1: Number of messages in sample for each filter.

we make no attempt to model any errors introduced by this process (e.g., users providing false names). Table 1 displays statistics of the overall dataset. A sample of 100 messages revealed no misattributed location or gender information.

## Sentiment Classification

In this section, we describe experiments applying sentiment classification to assess the level of concern of each message. Our goal is not to investigate new sentiment classification techniques, but instead to determine whether existing, well-known methods are applicable to this domain. While there is an extensive literature in sentiment classification technology (Pang and Lee, 2008), binary classification using a bag-of-words assumption has been shown to provide a strong baseline, so that is the approach we use here. We also evaluate the impact of lexicons and tokenization strategies.

We define "concerned" messages to be those showing some degree of apprehension, fear, or general concern regarding Hurricane Irene. Examples of unconcerned messages include links to news reports or messages expressing an explicit lack of concern. The idea is to assess how seriously a particular group is reacting to an impeding disaster.

To train the classifier, we sample 408 messages from the 66,205 message collection and manually annotate them as concerned or unconcerned. The final training set contains 170 concerned messages. Examples are shown in Table 2. To estimate inter-annotator agreement, we had a second annotator sample 100 labeled messages (50 concerned, 50 unconcerned) for re-annotation. The inter-annotator agreement is 93% (Cohen's kappa $\kappa = .86$).

| **Examples of concerned messages** |
|---|
| wonderful, praying tht this hurricane goes back out to sea. |
| Im actually scared for this hurricane... |
| This hurricane is freaking me out. |
| hope everyone is #safe during #irene |
| **Examples of unconcerned messages** |
| for the very latest on hurricane irene like our fb page ... |
| am i the only one who doesn't give a shit about this hurricane?? |
| tropical storm irene's track threatens south florida - miamiherald.com |

Table 2: Examples of concerned and unconcerned messages from the training set.

### Tokenization and features

We train a simple bag-of-words classifier, where the basic feature set is the list of word frequencies in each message. Given the brevity and informality of Twitter messages, tokenization choices can have a significant impact on classification accuracy. We consider two alternatives:

- **Tokenizer0:** The tokenizer of O'Connor et al. (2010b), which does very little normalization. Punctuation is preserved (for the purpose of identifying semantics such as emoticons), URLs remain intact, and text is lower-cased.
- **Tokenizer1:** A simple tokenizer that removes all punctuation and converts to lowercase.

We also consider two feature pruning options:

- **Stop Words:** Remove words matching a list of 524 common English words.
- **Frequency Pruning:** Remove words occurring fewer than 2 times in the labeled data.

We also consider the following features:

- **Worry lexicon:** We heuristically create a small lexicon containing words expressing worry of some kind, based on a brief review of the data.[3] We replace all such tokens with a WORRIED feature.

[3] The words are afraid, anxiety, cautious, die, died, nervous, pray, prayers, prayin, praying, safe, safety, scared, scary, terrified, thoughts, worried, worry, worrying

| **Classifier** | **Acc** | **Pr** | **Re** | **F1** |
|---|---|---|---|---|
| MaxEnt | **84.27 ± 2.0** | 90.15 | 70.00 | 78.81 |
| Dec. Tree | 81.35 ± 1.8 | 79.72 | 67.06 | 72.84 |
| Naive Bayes | 78.63 ± 2.2 | 75.78 | 71.76 | 73.72 |
| Worry Lex. | 79.41 | 95.74 | 52.94 | 68.18 |

Table 3: Average accuracy (with standard error) and micro-averaged precision, recall, and F1 for the three sentiment classifiers, using their best configurations. The difference in accuracy between MaxEnt and the other classifiers is statistically significant (paired t-test, $p < 0.01$).

| **System Configuration** | **Avg Acc** | **Max Acc** |
|---|---|---|
| **Tokenizer0** | 77.78 | 81.10 |
| **Tokenizer1** | 80.59 | 84.27 |
| **Keep Stop Words** | 77.99 | 81.34 |
| **Remove Stop Words** | 80.38 | 84.27 |
| **No Freq. Pruning** | 79.67 | 83.29 |
| **Freq. Pruning** | 78.71 | 84.27 |
| **No Worry lexicon** | 77.62 | 81.82 |
| **Worry lexicon** | 80.76 | 84.27 |
| **No Humor Lexicon** | 79.15 | 83.78 |
| **Humor Lexicon** | 79.23 | 84.27 |
| **No Emoticons** | 79.26 | 84.27 |
| **Emoticons** | 79.11 | 84.27 |

Table 4: Summary of the impact of various tokenization and feature choices. The second and third columns list the average and maximum accuracy over all possible system configurations with that setting. All results use the MaxEnt classifier and 10-fold cross-validation. **Tokenizer1**, **Remove Stop Words**, and **Worry Lexicon** result in the largest improvements in accuracy.

- **Humor lexicon:** Similarly, we create a small lexicon containing words expressing humor.[4] We replace all such tokens with a HUMOR feature.
- **Emoticon:** Two common emoticons “:)” and “:(“ are detected (prior to tokenization in the case of Tokenizer 1).

Finally, we consider three classifiers: MaxEnt (i.e., logistic regression), Naive Bayes, and a Decision Tree (ID3) classifier, as implemented in the MALLET machine learning toolkit (McCallum, 2002). We use all the default settings, except we set the maximum decision tree depth to 50 (after preliminary results suggested that the default size of 4 was too small).

Enumerating the possible tokenization, features, and classifier choices results in 192 possible system configurations. For each configuration, 10-fold cross-validation is performed on the labeled training data. Table 3 reports the results for each classifier using its best configuration. The configuration Tokenizer1/Remove Stop Words/Freq. Pruning/Worry lexicon/Humor lexicon/Emoticons was the best configuration for both MaxEnt and Naive Bayes. Decision Tree differed only in that its best configuration did not use Frequency Pruning. Table 3 also compares to a simple baseline that classifies messages as concerned if they contain any of the words in the worry lexicon (while accuracy is competitive, recall is quite low).

MaxEnt exhibits the best accuracy, precision, and F1; Naive Bayes has slightly better recall. Table 4 provides a summary of the numerical impact each configuration choice has. Using MaxEnt, we compute the accuracy over every possible system configuration, then average the accuracies to obtain each row. Thus, the Tokenizer1 row reports the average accuracy over all configurations that use Tokenizer1. Additionally, we report the highest accuracy of any configuration using that setting. These results indicate that Tokenizer1, Remove Stop Words, and Worry Lexicon result in the largest accuracy gains. Thus, while some unsupervised learning research has suggested that only light normalization should be used for social media text analysis (O’Connor et al., 2010b), for this supervised learning task it appears that more aggressive normalization and feature pruning can improve accuracy.

We select the best performing MaxEnt classifier for use in subsequent experiments. First we retrain the classifier on all the labeled data, then use it to label all of the unlabeled data from the original 65,062 messages. To estimate performance on this new data, we sample 200 additional documents of this testing data and manually label them (35 positive, 165 negative). We find that the automated classifications are accurate in 86% of these documents. Many of the remaining errors appear to be difficult cases. For example, consider the message: “1st an earthquake, now a hurricane? Damn NY do you

[4] The words are lol, lmao, rofl, rotfl, ha, haha.

miss me that bad?” The classifier labels this as concerned, but the message is likely intended to be humorous. In another message (“#PrayForNYC and everyone that will experience Hurricane Irene”), a hashtag #PrayForNYC complicates tokenization, so the word “pray” (highly correlated with concern) is not detected, resulting in a false negative.

## Demographic Analysis

We next apply this classifier to assess the demographic determinants of concerned messages. By classifying all remaining messages, we can analyze trends in sentiment over time by gender and region.

Figure 2 displays the total number of messages by day as well as the subset (and percentage) that are classified as concerned. Consulting the timeline in Figure 1, we see that the peak volume occurs on August 27th, the day the eye of the hurricane makes landfall. The percentage of messages labeled as concerned actually peaks a day earlier, on August 26th.

### Geographic Analysis

We first make several observations concerning Figure 1, which does not use the sentiment classifier, but only displays message volume. There appears to be a regional difference in when message volume peaks. Data point C in the figure, which marks the time around 2pm on August 23rd, represents the first noticeable spike in message count, particularly in the Upper South region. Two important events were occurring around this time period. First, the strongest earthquake to hit the Eastern United States since WWII (measured as 5.8 on the Richter scale) occurs near Richmond, Virginia. Also on August 23rd, a few hours prior to the earthquake, FEMA holds a press conference regarding the impeding threat that Hurricane Irene will pose to East Coast states. It appears likely that the combination of these events leads to the increase in messages on August 23rd as revealed in the figure. In fact, in examining some of the messages posted on Twitter during that time period, we notice some people commenting on the unlikeliness that two natural disasters would hit the region in such a narrow time frame.

Also in Figure 1, we see that the frequency of Twitter messages relating to Hurricane Irene for each region increases greatly over roughly the period of August 25th to August 28th, before decreasing later on August 28th and beyond. The increase and decrease roughly parallel the approach of Hurricane Irene toward and then beyond each region. Data point I represents the time (August 27th at 8am) when the center of Hurricane Irene makes landfall on the North Carolina coast. This point represents the highest message count for the Upper South region. Later on August 27th, as the hurricane moves north toward New Jersey and then New York, we see the peak message count for the Middle States region (Data point K). Finally, on August 28th in the late morning, as Hurricane Irene moves into the New England region, we see that the New England regions peak message count occurs (Data Point L).

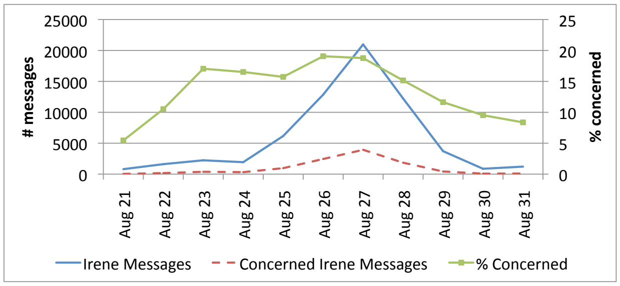


Figure 2: Total number of Twitter messages related to Hurricane Irene, as well as the count and percentage classified as *concerned* by the sentiment classifier.

With the sentiment classifier from the previous section, we can perform a more detailed analysis of the regional differences than can be performed using message volume alone. Figure 3 applies the sentiment classifier to assess the proportion of messages from each region that express concern. Figure 3 (top) shows the raw percentage of messages from each region by day, while the bottom figure shows the proportion of messages from each region that express concern. While the New England region has the lowest volume of messages, on many days it has the highest proportion of concerned messages.

Comparing regional differences in aggregate across all 10 days would be misleading – after the hurricane passes a region, it is expected that the level of concern should decrease. Indeed, these aggregate regional differences are not statistically significant (NE=15.59%, MID=15.4%, SOUTH=15.69%). Instead, for each day we compare the levels of concern

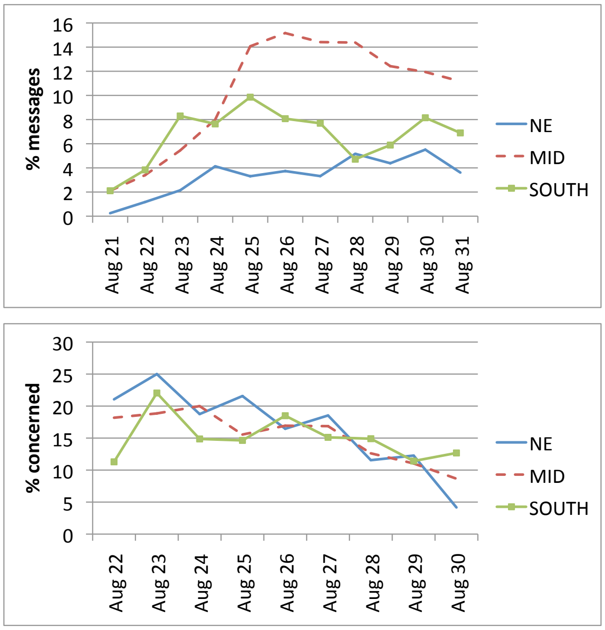


Figure 3: Message proportion and percent classified as concerned by the sentiment classifier, by region.

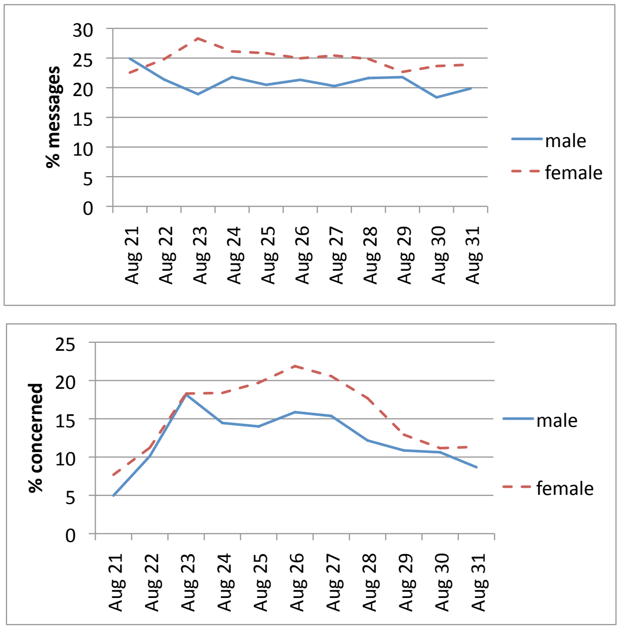


Figure 4: Message proportion and percent classified as concerned by the sentiment classifier, by gender.

for each region, testing for significance using a Chi-squared test. Two days show significant differences: August 25 and August 27. On both days, the proportion of concerned messages in New England is significantly higher ($p < 0.05$) than that of the Southern region (August 25: NE=21.6%, SOUTH=14.6%; August 26: NE=18.5%, SOUTH=15.1%). It is difficult to directly attribute causes to these differences, although on August 25, a Hurricane Watch was issued for the New England area, and on August 27 that Watch was upgraded to a Warning. It is also possible that states that experience hurricanes more frequently express lower levels of concern. Further sociological research is necessary to fully address these differences.

### Gender Analysis

We apply a similar analysis to assess the differences in levels of concern by gender. Figure 4 shows that for roughly the period between August 24th and August 29th, messages written by females are more likely to express concern than those written by males. Over the entire period, 18.7% of female-authored messages are labeled as concerned, while over the same period 13.9% of male-authored messages are labeled as concerned. We perform a Chi-squared test over the entire period, and find that gender differences in concern are significant (p $<$ .01). We conclude that messages attributed to female authors are significantly more likely to be classified as concerned than messages authored by males.

In order to assess a possible gender bias in our classifier, we examine the proportion of concern for males and females in the labeled training set. We find that of the original 408 labeled messages, 69 are from males, 112 are from females, and 227 cannot be determined. 24 male messages, or 34.8%, are marked as concerned. In contrast, 57 female messages, or 50.9%, are marked as concerned. 88 of the undetermined gender messages, or 38.9%, are concerned. We therefore down-sample the female messages from our labeled training set until the proportion of female-concerned messages matches that of male-concerned messages. Repeating our classification experiments shows no significant difference in the relative proportions of messages labeled as concerned by gender. We therefore conclude that the training set is not injecting a gender bias in the classifier.

| |
|---|
| **Female:** i my safe praying this everyone died jada butistillloveu brenda who love t me thank school pets retweet respects all please here so stay neverapologizefor wine sleep rainbow prayers lord |
| **Male:** http co de en el hurac media breaking la rooftoproofing track obama jimnorton gay ron blames smem change seattle orkaan becomes disaster zona zan lean vivo por es location dolphin |
| **New England:** boston MAirene ct vt ri england sunday connecticut malloy ma vermont tropical maine wtnh massachusetts haven rhode VTirene va power CThurricane cambridge mass lls gilsimmons mbta gunna storm slut NHirene |
| **Middle States:** nyc ny nj nycmayorsoffice york jersey mta brooklyn zone nytmetro va ryan philly shut dc mayor city manhattan lls new subways con team longisland bloomberg evacuation evacuate yorkers catskills queens |
| **South:** nc dc va lls earthquake raleigh maryland dmv ncwx virginia ncirene richmond isabelle perdue isabel mdhurricane bout carolina capitalweather sniper rva norfolk goin feeds nycmayorsoffice baltimore ilm mema tho aint |

Table 5: Top 30 words for each demographic ranked by Information Gain.

## Qualitative Analysis

In Table 5 we provide a brief qualitative analysis by displaying the top 30 words for each demographic obtained using Information Gain (Manning and Schtze, 1999), a method of detecting features that discriminate between document classes. To provide some of the missing context: “jada” refers to the divorce of celebrities Will Smith and Jada Pinkett; “hurac” refers to the Spanish word *Huracán*; “smem” stands for Social Media for Emergency Management; “dolphin” refers to a joke that was circulated referencing the hurricane; “lls” is an abbreviation for “laughing like shit”.

Some broad trends appear: male users tend to reference news, politics, or jokes; the Middle States reference the evacuation of New York City, and the South refers back to other disasters (the earthquake, the sniper attacks of 2002, Hurricane Isabel).

## Related Work

Recent research has investigated the effectiveness of social media for crisis communication (Savelyev et al., 2011) — indeed, the U.S. Federal Emergency Management Agency now uses Twitter to disseminate information during natural disasters (Kalish, 2011). Other work has examined the spread of false rumors during earthquakes (Mendoza et al., 2010) and tsunamis (Acar and Muraki, 2011) and characterized social network dynamics during floods (Cheong and Cheong, 2011), fires (Vieweg et al., 2010), and violence (Heverin and Zach, 2010). While some of this past research organizes messages by topic, to our knowledge no work has analyzed disaster sentiment or its demographic determinants.

Survey research by West and Orr (2007) concluded that women may feel more vulnerable during hurricanes because they are more likely to have children and belong to a lower socio-economic class. Richer people, they find, tend to have an easier time dealing with natural disasters like hurricanes. These reasons might explain our finding that women are more likely on Twitter to show concern than men about Hurricane Irene. West and Orr also find differences in regional perceptions of vulnerability between coastal areas and non-coastal areas. Our location annotation must be more precise before we can perform a similar analysis.

More generally, our approach can be considered a type of *computational social science*, an emerging area of study applying computer science algorithms to social science research (Lazer et al., 2009; Hopkins and King, 2010).

## Conclusion and Future Work

Our results show that analyzing Twitter messages relating to Hurricane Irene reveals differences in sentiment depending on a person’s gender or location. We conclude that social media analysis is a viable complement to existing survey methodologies, providing real-time insight into public perceptions of a disaster. Future directions include investigating how to account for classifier error in hypothesis testing (Fuller, 1987), adjusting classification proportions using quantification methods (Forman, 2007), as well as applying the approach to different disasters and identifying additional sentiment classes of interest. Finally, it will be important to infer a greater variety of demographic attributes and also to adjust for the demographic bias inherent in social media.

## References

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

S. Asur and B. A. Huberman. 2010. Predicting the future with social media. In *Proceedings of the ACM International Conference on Web Intelligence*.

France Cheong and Christopher Cheong. 2011. Social media data mining: A social network analysis of tweets during the 2010–2011 Australian floods. In *PACIS 2011 Proceedings*.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Workshop on Social Media Analytics at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Aron Culotta. 2012. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation, Special Issue on Analysis of Short Texts on the Web*. to appear.

Elaine Enarson. 1998. Through women's eyes: A gendered research agenda for disaster social science. *Disasters*, 22(2):157–73.

George Forman. 2007. Quantifying counts, costs, and trends accurately via machine learning. Technical report, HP Laboratories, Palo Alto, CA.

A. Fothergill, E.G. Maestas, and J.D. Darlington. 1999. Race, ethnicity and disasters in the united states: A review of the literature. *Disasters*, 23(2):156–73, Jun.

W.A. Fuller. 1987. *Measurement error models*. Wiley, New York.

Eric Gilbert and Karrie Karahalios. 2010. Widespread worry and the stock market. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, D.C., May.

J.D. Goltz, L.A. Russell, and L.B. Bourque. 1992. Initial behavioral response to a rapid onset disaster: A case study. *International Journal of Mass Emergencies and Disasters*, 10(1):43–69.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 237–246, New York, NY, USA.

T. Heverin and L. Zach. 2010. Microblogging for crisis communication: Examination of Twitter use in response to a 2009 violent crisis in Seattle-Tacoma, Washington area. In *Proceedings of the Seventh International Information Systems for Crisis Response and Management Conference*, Seattle, WA.

Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Brian Kalish. 2011. FEMA will use social media through all stages of a disaster. Next Gov, February.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, pages 411–416.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

Chris Manning and Hinrich Schtze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, May.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA '10)*, July.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, , and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain.

James Murphy, P. Allen, Thomas Stevens, and Darryl Weatherhead. 2005. A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, 30(3):313–325.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010a. From Tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media*, Washington, D.C.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010b. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

R.W. Perry and M.K. Lindell. 1991. The effects of ethnicity on decision-making. *International journal of mass emergencies and disasters*, 9(1):47–68.

R.W. Perry and A.H. Mushkatel. 1986. *Minority citizens in disasters*. University of Georgia Press, Athens, GA.

Alexander Savelyev, Justine Blanford, and Prasenjit Mitra. 2011. Geo-twitter analytics: Applications in crisis management. In *25th International Cartographic Conference*, pages 1–8.

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 *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088, New York, NY, USA.

Darrell M. West and Marion Orr. 2007. Race, gender, and communications in natural disasters. *The Policy Studies Journal*, 35(4).

## Appendix A: Location String Matching

The following strings were matched against the user location field of each message to determine the location of the message. Matches were case insensitive, except for abbreviations (e.g., VT must be capitalized to match).

Vermont, VT, Maine, ME, New Hampshire, Rhode Island, RI, Delaware, DE, Connecticut, CT, Maryland, MD, Baltimore, North Carolina, NC, Massachusetts, MA, Boston, Mass, W Virginia, West Virginia, Virginia, VA, RVA, DC, D.C., PA, Philadelphia, Pittsburgh, Philly, New Jersey, Atlantic City, New York, NY, NYC, Long Island, Manhattan, Brooklyn, Staten Island, The Bronx, Queens, NY, N.Y.