Understanding Patterns and Mood Changes through Tweets about Disasters

Liuqing Li and Edward A. Fox
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061, USA
[liuqing, fox]@vt.edu

ABSTRACT

We analyzed a sample of large tweet collections gathered since 2011, to expand understanding about tweeting patterns and emotional responses of different types of tweeters regarding disasters. We selected three examples for each of four disaster types: *school shooting*, *bombing*, *earthquake*, and *hurricane*. For each collection, we deployed our novel model TwiRole for user classification, and an existing deep learning model for mood detection. We found differences in the daily tweet count patterns, between the different types of events. Likewise, there were different average scores and patterns of moods (fear, sadness, surprise), both between types of events, and between events of the same type. Further, regarding surprise and fear, there were differences among roles of tweeters. These results suggest the value of further exploration as well as hypothesis testing with our hundreds of event and trend related tweet collections, considering indications in those that reflect emotional responses to disasters.

Keywords

Disaster, Pattern, User Classification, Mood Detection, Twitter

INTRODUCTION

Disasters can appear in any area of the world, and include both natural disasters like earthquakes, hurricanes, floods, and forest fires, as well as man-made disasters such as shootings, bombings, and train crashes. Different types of disasters have distinct patterns. For example, natural disasters have dramatic effects and can bring huge loss, while man-made disasters often proceed rapidly, with long-lasting impact. Further, natural disasters have particular spatial distribution patterns (Shen et al. 2018). Since Twitter users tend to tweet about the latest developments, and convey emotional feelings in a time of disaster, the patterns found in large tweet collections can reflect public concern about disasters, and help researchers dive deeper into understanding them.

Twitter users express different moods during disasters. For instance, some users may be frightened by the aftershocks during an earthquake, some are angered by a bombing disaster, and others verbalize their sadness for the victims in a school shooting event. Their changes in mood, synchronized with the unfolding of disasters, are interesting to explore. In addition, Twitter users assuming different roles (e.g., brand/organization, female, male) behave differently. For example, preliminary work has shown that male users prefer technology and sports (Bamman et al. 2014), female users are more emotional (Rao et al. 2010), and brand users act as broadcasters by spreading information or delivering messages. The gender value can be set by a user or predicted by Twitter on the profile page, but it is unreadable to others and does not include the brand class. Since moods might vary with time among different roles of users, analyzing such changes will be helpful both academically and practically. Social scientists can conduct user-related research, while disaster assistance can help victims recover, through targeted services.

In this paper, we report on a comprehensive analysis of 12 large tweet collections, to understand tweeting patterns and mood changes of different types of tweeters about disasters. We prepared three tweet collections for each of four disaster types: *school shooting*, *bombing*, *earthquake*, and *hurricane*. We developed TwiRole, a novel hybrid model to identify role-related users, and reused an existing recurrent neural network (RNN) model to detect the moods expressed in tweets. Considering both user roles and mood scores, we studied patterns and mood changes at both the event and role level. To help investigate patterns and changes associated with disasters, we developed the following research questions: 1) Is there any tweeting pattern behind disasters of each type? Are these patterns similar between different types? 2) How do tweeters demonstrating different roles relate to these disasters? 3) What is the major mood? Are there exceptions? 4) How does mood change along with the daily tweet count pattern in each disaster? 5) Is there any difference in mood change among roles of tweeters?

The findings of our study indicate that different types of disasters lead to different daily tweet count patterns. Regarding user participation, brand users played an essential role in tweeting, while male users were more active than female users. There were different average scores and patterns of moods (fear, sadness, surprise), both between types of events, and between events of the same type. Fear is the major feeling in most disasters, while sadness is dominant in two of the school shootings analyzed. Mood delays exist in most disasters, where sadness is more time-sensitive. Regarding surprise and fear, mood variation appears among users with different roles.

We next present related work, and then describe our methodology for data selection, user classification, and mood detection. After that, findings, followed by evaluation and discussion, are presented in detail. Our conclusions are drawn in the final section.

RELATED WORK

User Classification on Twitter

Several researchers have worked on Twitter user classification according to role. Multiple features have been considered and extracted for gender classification. (Liu and Ruths 2013) investigated the relationship between first name and gender, and took the first name as an important feature for gender prediction. Some researchers have found that female and male users may apply templates in different colors (Alowibdi et al. 2013a; Alowibdi et al. 2013b; Fortmann-Roe 2013), and so utilized color-based features in gender inference. Information in user profiles, especially the descriptions of users, may also contain gender-based terms or phrases (e.g., man, woman, actor, mother) that are good indicators (Daas et al. 2016; Vicente et al. 2015). Besides these features, a user's network and behavior (Ciot et al. 2013; Lasorsa 2012; Nilizadeh et al. 2016), external sources (e.g., personal website, Facebook) (Burger et al. 2011), tweets (Artwick 2014; Bamman et al. 2014; Li et al. 2014; Pennacchiotti and Popescu 2011a; Geng et al. 2017; Rao et al. 2010), and profile images (Levi and Hassner 2015; Wang et al. 2016), have been advocated for gender classification.

In addition to the traditional bi-classification, (Purohit and Chan 2017) categorized Twitter users into organization, organization-affiliated, and non-affiliated, while (Pennacchiotti and Popescu 2011b) classified Twitter users as either Democrat or Republican. However, there are methodological concerns regarding the evaluation of such studies, because most approaches were evaluated on self-labeled datasets, and performance is still unclear on other datasets. We developed TwiRole, that applies multiple simple features as well as some novel features for 3-way (i.e., brand, female, and male) classification of tweeters. After training and testing TwiRole on an open third-party dataset, we also evaluated its performance on our disaster collections.

Sentiment / Emotion Classification on Twitter

There has been growing interest in sentiment and emotion classification of tweets. (Schulz et al. 2013) took unigram, part-of-speech, syntactic, and sentiment information as features for sentiment analysis. (Nagy and Stamberger 2012) detected tweet sentiments during crisis situations by using sentiment words, emoticons, and a list of out-of-vocabulary words. (Caragea et al. 2014) performed sentiment classification of tweets during Hurricane Sandy, and visualized these sentiments on a geographical map. Many studies have been published on basic sentiment analysis (i.e., positive, negative, and neutral) by applying traditional machine learning techniques (e.g., Naive Bayes, support vector machine, logistic regression). (Wendland et al. 2018) applied an appraisal system to account for the interpersonal assessment of speakers and associated attitudes, which evaluated the sentiments of tweets from several aspects such as appreciation, affect, and judgment. A recent study reported in (Colneriĉ and Demsar 2018) improved the classification performance using RNN deep learning. Instead of detecting the three basic sentiments, their model predicted tweet emotions, which can better describe the feelings of tweeters. The default task is for predicting Ekman's six basic emotions (Ekman 1992), while the other two aim to identify the eight moods developed by Plutchik (Plutchik 1990), and the six mood states from the profile of Mood States (POMS) (Norcross et al. 1984). In our study, we deployed the innovative RNN deep learning model to predict the moods of users through tweets.

Empirical Study in Disasters

Regarding social media studies in crises, some researchers are focusing on single disasters, such as the 2011 Egyptian uprising (Kavanaugh et al. 2013), 2011 Japan Earthquake (Wilensky 2014), 2012 Hurricane Sandy (Caragea et al. 2014; Neppalli et al. 2016), 2013 River Elbe Flood (Herfort et al. 2014), 2016 Ghana Election (Moreno et al. 2017), or 2016 Berlin Terrorist Attack (Fischer et al. 2018). (Yang et al. 2013) designed PhaseVis, a multi-view integrated visualization tool, and applied it to Hurricane Isaac. (Lee et al. 2012) built a prototype of a digital library to detect and visualize water main breaks in disaster and other areas. (Alam et al. 2018) analyzed the multimedia content from three hurricane collections. Novel theories, models, and techniques have been developed and applied, and researchers can make further discoveries or tell exciting stories about these disasters. However, the generality of these approaches is not clear. It is uncertain how well they can work on other disasters. To the best of our knowledge, this is the first investigation of groups of disasters, carried out in a systematic empirical study, combined with both user classification and mood detection.

METHODOLOGY

Figure 1 shows the data flow of our analysis across disaster types. First, we created three tweet collections for each of four types of disasters. Second, we identified users from the cleaned tweets after preprocessing. Third, we designed and implemented TwiRole [1], a hybrid model for user classification on Twitter, which detects brand-, female-, and male-related users. Then, we applied a pre-trained RNN model to predict the moods of tweets. Finally, we carried out a comprehensive analysis of disaster patterns and mood changes, considering both user roles and mood scores.

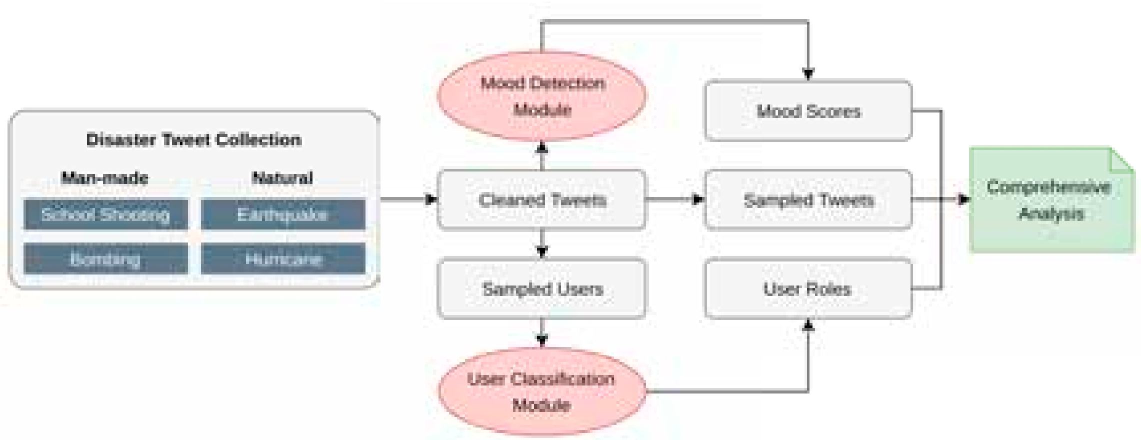


Figure 1. A data flow diagram: selection, classification, and analysis

Data Selection and Cleaning

Though we already had hundreds of tweet collections dating from 2012, we applied GetOldTweets3 (Mottl 2018) to create a dozen larger, more complete, tweet collections with four different types of disasters (i.e., school shooting, bombing, earthquake, and hurricane), three event collections per type. Each search query includes a disaster name and bound dates. The time range of each collection covers up to about one month after the corresponding disaster. For example, regarding hurricanes, we found the date they first formed, according to Wikipedia, and obtained tweets for four weeks starting on that date. Later, retweets (RTs) and non-English tweets were filtered out during cleaning to avoid muddling of results and to meet the input requirements of the mood detection module, respectively. Table 1 shows the details of our collections.

Role-related User Classification

We designed and implemented TwiRole to classify users by role. This hybrid classifier has a basic feature (BF) multi-classifier, an advanced feature (AF) multi-classifier, a convolutional neural network (CNN) model, and a final multi-classifier, as shown in Figure 2. Features for a given user include name, description, tweet, relationship, and profile image. Each classifier/model takes one or more features as input and predicts the probability of each role (i.e., brand, female, and male). Features and their details in each classifier/model are shown in Table 2. We utilized

[1]GitHub link: https://github.com/liuqingli/TwiRole

Table 1. An overview of the twelve tweet collections, across four disaster types

Type	Collection	Starting (UTC)	# of tweets	# of cleaned tweets
School Shooting	Sandy Hook Elementary School shooting	12/14/2012 14:40:00	100,833	97,283
	Stoneman Douglas High School shooting	02/14/2018 19:27:00	22,441	22,321
	Umpqua Community College shooting	10/01/2015 17:48:00	17,930	17,821
Bombing	Boston Marathon bombing	04/15/2013 18:49:00	220,221	211,142
	San Bernardino attack	12/02/2015 18:58:00	44,359	44,224
	Manchester Arena bombing	05/22/2017 21:31:00	6,119	6,105
Earthquake	Nepal earthquake	04/25/2015 06:11:00	677,067	671,323
	Japan earthquake	03/11/2011 05:46:00	644,814	566,627
	Taiwan earthquake	02/05/2016 19:57:00	49,052	48,894
Hurricane	Hurricane Sandy	10/22/2012 00:00:00	2,528,523	2,399,334
	Hurricane Matthew	09/28/2016 00:00:00	1,092,684	1,088,212
	Hurricane Florence	08/31/2018 00:00:00	639,888	636,281

an existing third-party dataset (Kaggle 2016) to train and test TwiRole, and also evaluated it on the users extracted from our disaster collections. TwiRole performs better than other existing models on the Kaggle dataset, and its overall accuracy is about 90%.

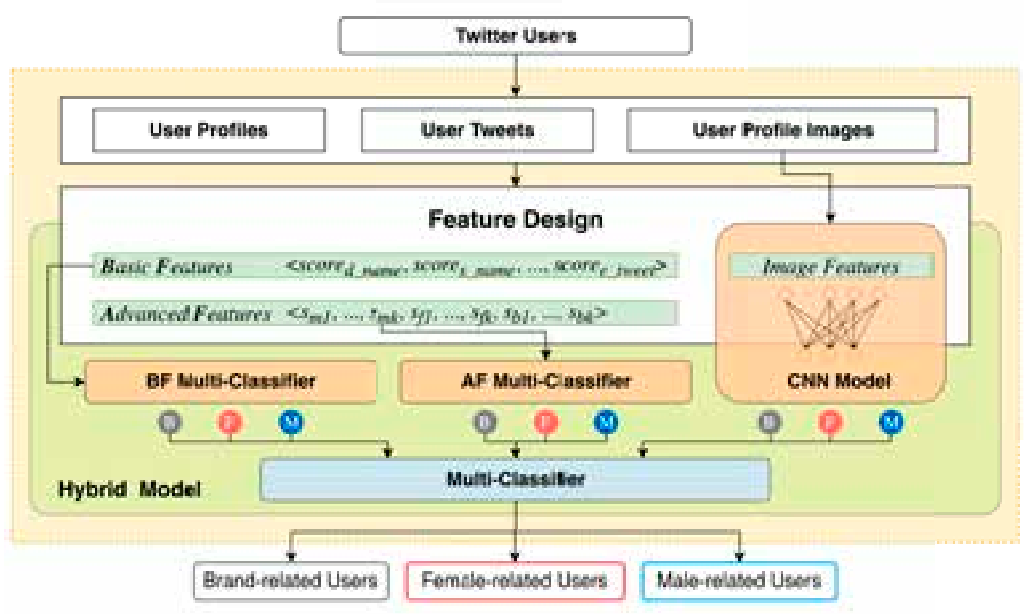


Figure 2. TwiRole is a hybrid model consisting of four components. The BF multi-classifier takes the basic features from the user profiles and tweets as input. The AF multi-classifier focuses on the k-top words in user tweets. The CNN works on the user profile images. The final multi-classifier takes the output of the above three modules as input during training and testing.

Mood Detection

We used a pre-trained RNN model (Colneriĉ and Demsar 2018) for predicting Ekman's emotions (Ekman 1992) from tweets, which outperforms several traditional machine learning methods (e.g, SVM, Naive Bayes, Random Forest). The F1 score mentioned in their paper is about 70%, leading to an approximate emotion analysis. The model takes each tweet as input and predicts the score for each of six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise). In our study, as a simple start, we focus on three moods: fear, sadness, and surprise. We assume these are common feelings during disasters.

FINDINGS

Disaster Patterns

For each disaster collection, we calculated the number of tweets posted per hour during the time window of the event. Then we applied min-max normalization to convert the number of hourly tweets to a scale between 0 and 1

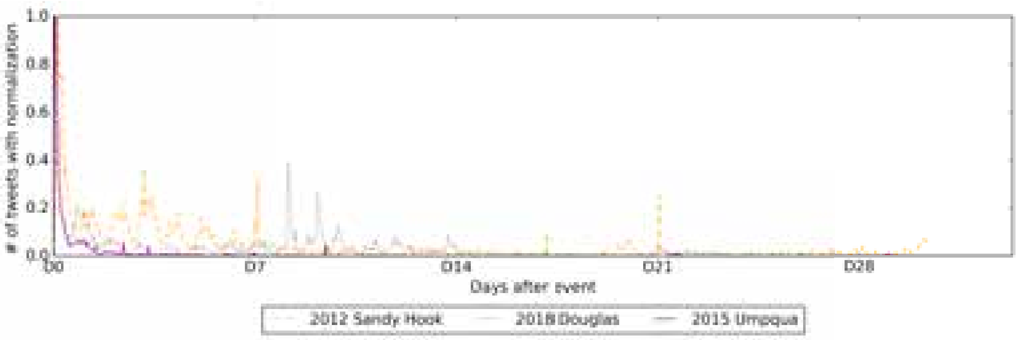
Table 2. Feature Types and Details in TwiRole

No.	Feature Type	Feature Detail		No.	Feature Type	Feature Detail
BF1	*name*	display name screen name		BF3	*relationship*	Twitter Follower-Friend score
				BF4	*profile image*	brightness
BF2	*description*	first-person score term frequency		BF5	*tweet*	first-person score interjection score emotion score
AF	*tweet*	k-top words		CNN	*profile image*	hidden in image

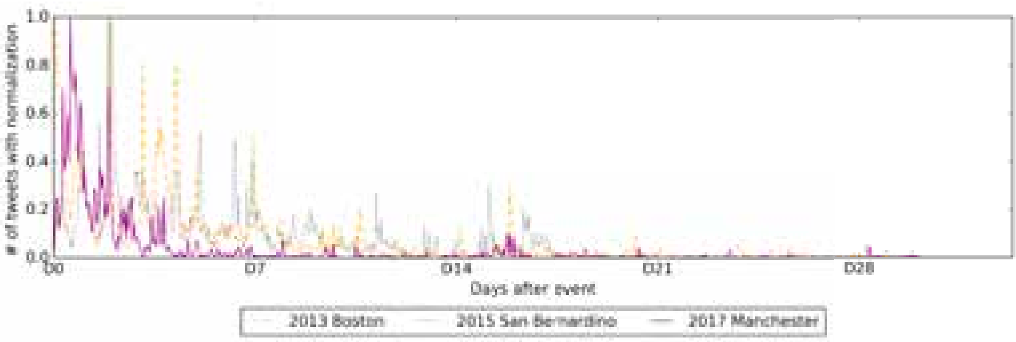
for inter-comparison. Different colors represent different collections in each disaster type. Figures 3a to 3d show the collection timelines across different disaster types.

Figure 3 illustrates that each type of disaster has a pattern that differs from the others. These patterns can also help us better understand user mood changes.

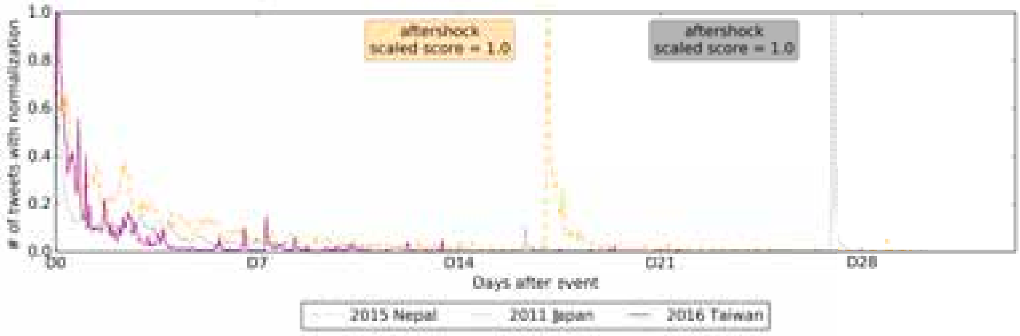
- *School Shooting*: School shootings are not as complicated, among man-made disasters, as bombings. The number of hourly tweets dropped to a low level just one or two days after a shooting. However, further along in the timeline, there are still some smaller peaks, e.g., corresponding to subsequent information releases or discoveries.
- *Bombing*: Each bombing timeline demonstrates highly variable numbers of tweets during the subsequent week. Tweets in significant numbers were posted even two or three days after the incident, which indicates that users paid close attention to information about the disaster.
- *Earthquake*: Twitter users posted many tweets, shown by a peak, when an earthquake occurred, and then the number of tweets gradually reduced over time. Later, when a massive aftershock occurred, another greater peak appeared, followed by a rapid reduction different from that observed after the original event.
- *Hurricane*: As a progressive natural disaster, each hurricane has a well-recognized timeline that approximates a normal distribution. We can roughly distinguish its stages and estimate the time when it hit land and caused severe damage. Moreover, the day-night periodicity in tweet number variations is more significant than for other types of disasters.



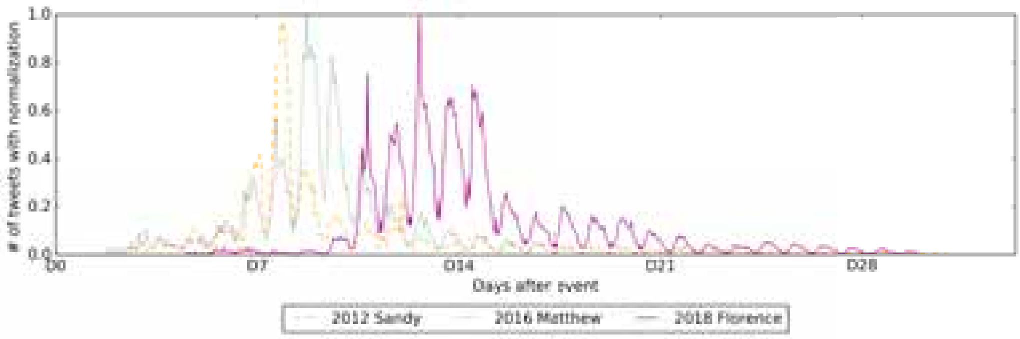
(a) School Shooting



(b) Bombing



(c) Earthquake



(d) Hurricane

Figure 3. Number of tweets (scaled) per hour for different types of disasters

User Distribution

We randomly selected at most 10,000 users from each disaster collection; the total across all collections was 107,323 users. Some accounts were deactivated or protected. Finally we utilized TwiRole to detect a user's role for 102,597 users (95.6% of 107,323 users). Afterward, we calculated the fraction of each role in each collection. The results (Brand: 0.4737 ± 0.0093; Female: 0.2124 ± 0.0098; Male: 0.3139 ± 0.0085) indicate that the distribution of users is relatively consistent among all disasters, as shown in Figure 4. Brand users are the primary group who might publish information or deliver messages, while male users participated more actively than female users in our selected disasters.

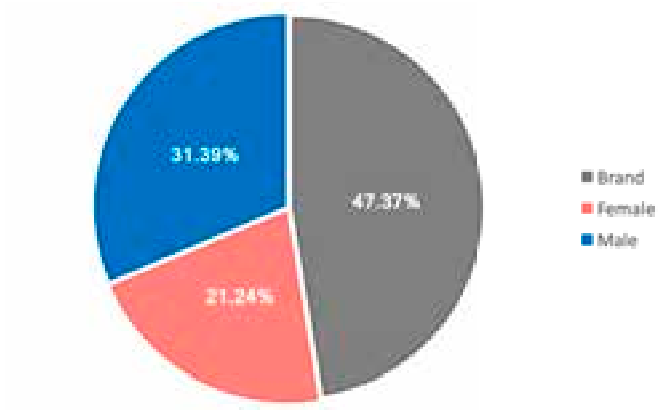


Figure 4. User distribution totals across all disasters

Bots are worth exploring and we applied Botometer for detection. Due to the rate limit, we detected 150 sampled users for each collection and calculated the percentage of bot users. The result (0.0694 ± 0.0270) indicates that only a small number of users are bots, so they have little impact on our results.

Mood Changes in All Users

We randomly sampled at most 100,000 tweets from each disaster collection and predicted the mood scores (i.e., fear, sadness, and surprise) of each tweet with the pre-trained RNN classifier. The scores of each mood were accumulated and divided by the total number of tweets in each collection. Figure 5 shows the average scores of the three moods in our collections.

Fear is the dominant feeling in eight out of the twelve disasters, which is consistent with our expectation. After the Boston bombing in 2013, users posted more fear tweets and expressed their feelings with words: *fear*, *afraid*, *terror*, *deadly*, or *scared*. During Hurricane Matthew in 2016, the fear words include *fear*, *scary*, *terrifying*, *frightening*, and *threatening*.

Twitter users show more fear and surprise than sadness in ten collections. The two counter-examples are school shootings, where sadness is the common mood. Users felt great sadness for the children and students who died or were injured in those massacres, and showed their feelings with sad words or phrases like *R.I.P*, *heart goes out*, *condolence*, *heart is broken*, and *depressed*.

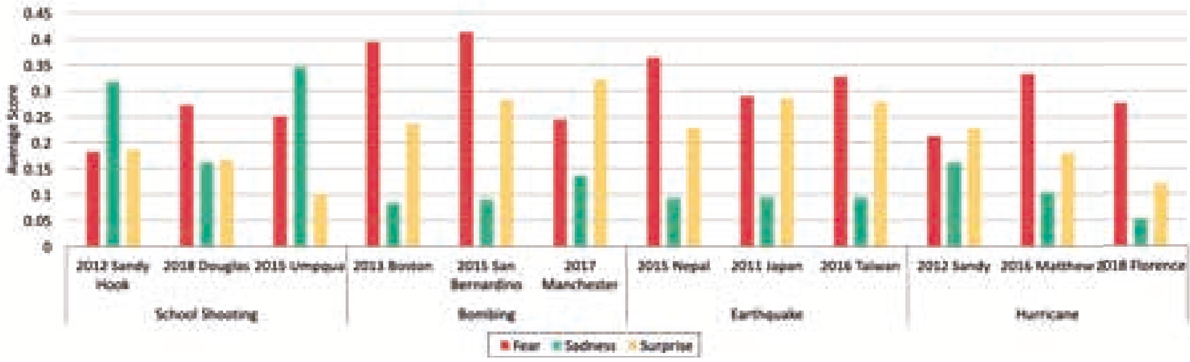


Figure 5. Average scores of fear, sadness, and surprise in different types of disasters

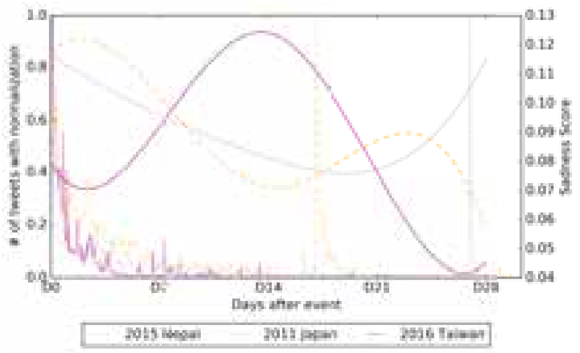
To explore further, we calculated the average score of each mood on the first day of each disaster and in every week after that. We selected two moods and two types of disasters for each mood, and showed changes over time with the corresponding tweet timelines, in Figures 6a through 6d. Figure 6a illustrates the sadness change in the earthquake disasters. The sadness score peaked two weeks after the main shock in Taiwan, while the aftershock played an important role in the Japan and Nepal earthquakes, leading to an increase in sadness. Figure 6b displays the sadness change in the selected three hurricanes. The sadness scores peaked about one week after the tweet count peaks of the three tweet timelines. Figure 6b also supports a quantitative comparison of the impacts of the three hurricanes. Users felt sadder in Hurricane Sandy, since it was the fourth-costliest hurricane in U.S. history, while Hurricane Florence made landfall as a weakened Category 1 hurricane, accompanied by a low sadness score.

The peaks in surprise were delayed by two or three weeks, for the bombing disasters shown in Figure 6c; we discuss more about mood delays below. As Figure 6d indicates, for school shootings, the scores of surprise were still increasing one month after the disasters, especially for the Sandy Hook Elementary School shooting (2012) and the Douglas High School shooting (2018). After browsing through users' tweets, we noted that users posted tweets like "*The Sandy Hook shooting was a hoax?!*" and "*I find it very odd two of the sandy hook funds were created before the shooting*" after the former shooting event, while they tweeted "*Who's Behind the Real Scandal*" about the latter event.

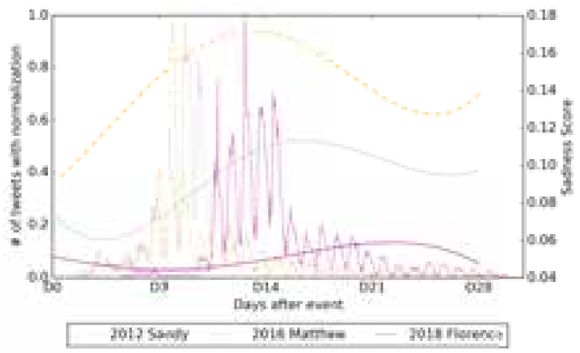
Mood Changes for Different Roles of Users

Based on the user and mood classification results, we enriched each tweet collection with user roles and mood scores, analyzed the mood changes among different role-related users, and carried out two case studies. Each sub-figure in Figure 7 illustrates the mood change across brand, female, and male users in one specific disaster, with its corresponding tweet timeline. Then, for each case study, we sorted the tweets according to mood score, and presented the surprise value for the k-th (k=1, 10, 20, 30) ranking tweets, for all three roles. Finally, we give the texts of some sampled tweets in Tables 3 and 4.

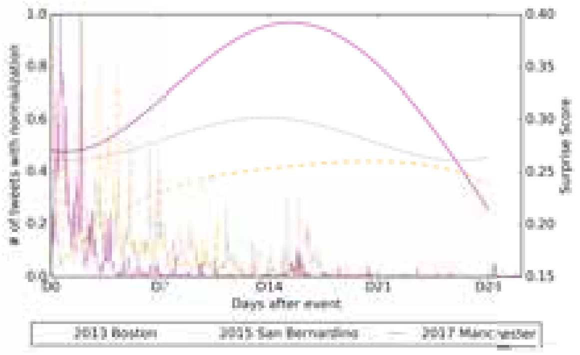
- *Case Study 1*: Surprise change in 2011 Japan Earthquake
  Figure 7a shows that the surprise score of male users had a significant increase after the aftershock, compared with the scores of brand and female users. The ranked k-th surprise scores also show male users were more surprised than brand users, while female users were relatively calm during that period.



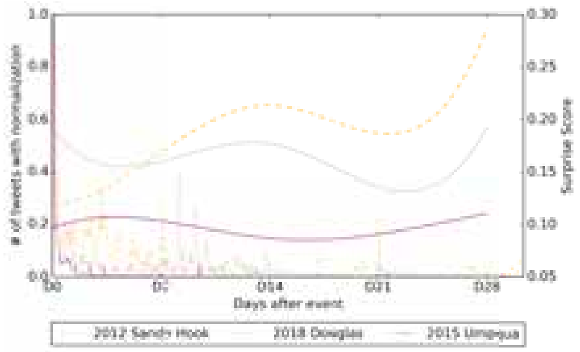
(a) Sadness + Earthquake



(b) Sadness + Hurricane



(c) Surprise + Bombing



(d) Surprise + School Shooting

Figure 6. Mood changes (smoothed) for different types of disasters

Table 3. Ranked k-th surprise score comparison and sample tweets from male users in 2011 Japan Earthquake

	Surprise Score		
k	**Brand**	**Female**	**Male**
1	0.993	0.938	0.996
10	0.899	0.775	0.959
20	0.866	0.664	0.925
30	0.818	0.570	0.880

Sample tweets from male users with high surprise scores
japan had another earthquake? #idontbelieveyou
Japan just had another earthquake. WHHAAAAAAAAA?!?
Mother nature #idontbelieveyou another earthquake Japan?
Another earthquake in Japan! #wow
There was another earthquake in Japan?!?! OMG #PRAYFORJAPAN

- *Case Study 2*: Fear change in 2012 Hurricane Sandy
  Figure 7b shows the divergence of fear scores among different roles that appeared during the last week of the one-month time window. Here, brand users expressed more fear, and the fear scores of female and male users decreased at the same time. From the sample tweets, we noticed that brand users posted fear-related tweets regarding the aftermath of the destructive hurricane.

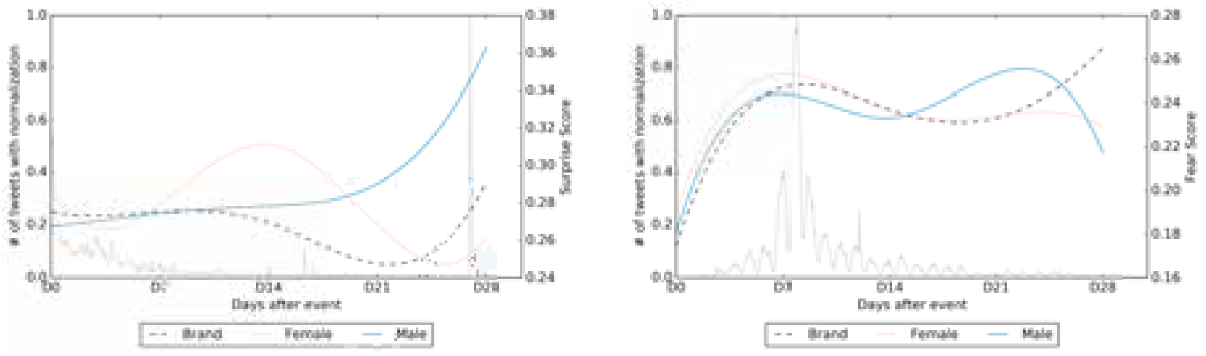
Table 4. Ranked k-th fear score comparison and sample tweets from brand users in 2012 Hurricane Sandy

	Fear Score		
k	**Brand**	**Female**	**Male**
1	0.999	0.991	0.996
10	0.875	0.720	0.848
20	0.789	0.537	0.733
30	0.728	0.441	0.635

Sample tweets from brand users with high fear scores
Haiti fears food crisis in Hurricane #Sandy's aftermath
Hurricane Sandy 2012: What a nightmare!! Lack of power, gas rationing...
Gas Shortage In New York After Hurricane Sandy Caused By Poor Policy
Terrifying Note Left Behind By a New Jersey Man In Hurricane #Sandy...
#Forbes Obviously the destruction caused by Hurricane Sandy posed a risk...

EVALUATION OF TWIROLE

Our new tool TwiRole has been used to predict the role of a Twitter user. To assuage concerns regarding the accuracy of the user classification predictions, we report on an evaluation of that tool. We utilized 10-fold cross-validation to evaluate TwiRole on an existing third-party dataset with gold standard labels. Further, we randomly selected 150 users from our disaster collections, and labeled user roles by hand, for a further evaluation of the prediction results of TwiRole.



(a) Surprise + 2011 Japan Earthquake **(b) Fear + 2012 Hurricane Sandy**

Figure 7. Mood changes (smoothed) among different roles of users in specific disasters

Metrics

We use a confusion matrix to calculate the recall (R), precision (P), and F1 score for each role. For a certain role r, the three values are computed as:

$$\begin{aligned} Recall_r &= \frac{\text{\# of users correctly identified as } r}{\text{\# of users labeled as } r}, \\ Precision_r &= \frac{\text{\# of users correctly identified as } r}{\text{\# of users predicted as } r}, \\ F1_r &= \frac{2 * Recall_r * Precision_r}{Recall_r + Precision_r} \end{aligned} \tag{1}$$

The performance of TwiRole is reflected in the overall accuracy; see Equation 2.

$$Accuracy = \frac{\sum_r (\text{\# of users correctly identified as } r)}{\sum_r (\text{\# of users labeled as } r)} \tag{2}$$

Results

First, we compare TwiRole with Ferrari et al.'s work (Ferrari et al. 2017) on the same third party dataset, since they also categorized Twitter users into the same three classes. The classification results are shown in Table 5. Their model has an advantage in identifying the male-related users, where their F1 score is 0.947 while ours is 0.903. But TwiRole performs better in detecting both female-related ($F1_{female}$ = 0.908) and brand-related users ($F1_{brand}$ = 0.885), and the overall accuracy (Acc = 0.899) is higher than with Ferrari et al.'s approach (Acc = 0.865). In addition, the prediction results of our model are more balanced across different roles, because the difference in F1 score is only 0.023 in TwiRole while it is 0.136 for Ferrari et al.

Table 5. Classification results of TwiRole and Ferrari et al.'s work

Tool	Brand			Female			Male			Acc
	R	P	F1	R	P	F1	R	P	F1	
TwiRole	**0.891**	**0.879**	**0.885**	**0.920**	**0.897**	**0.908**	0.885	0.922	0.903	**0.899**
Ferrari et al., 2017	0.837	0.786	0.811	0.806	0.857	0.831	**0.948**	**0.946**	**0.947**	0.865

Table 6. Classification results of TwiRole on disaster collections

Tool	True Label	# of users	Predict			Accuracy
			Brand	Female	Male	
TwiRole	Brand	50	41	6	3	82%
	Female	50	6	38	6	76%
	Male	50	10	3	37	74%

Second, using our disaster collections, we randomly chose 50 users labeled by TwiRole in each class and checked their roles manually by browsing their Twitter pages. Table 6 lists the manual checking results and the accuracy of prediction in each role. The overall accuracy of TwiRole is about 77%, which is satisfactory in our situation, where it is used to describe mood changes among tweeters.

DISCUSSION

As is mentioned in the Findings section, mood delays appear in some disasters. For instance, sadness maximum values were delayed one week after the peak counts in tweet timelines for hurricanes, while surprise score has a delay of two or three weeks for the bombing disasters. Accordingly, we examined mood changes over four weeks in all of our twelve collections, to check whether mood delay is a widespread phenomenon in disasters and is different among moods. For each mood, we compared its score on the first day with the score in each of the following four weeks. Table 7 lists the occurrences of the peaks of moods in different periods.

Table 7. Mood peak occurrences in different periods of time among all twelve disasters

Moods	Number of peaks occurred on / in					
	Day 0	Week 1	Week 2	Week 3	Week 4	Total
fear	3	2	4	0	3	12
sadness	7	0	2	1	2	12
surprise	1	3	2	3	3	12

The peaks of fear occurred on the first day in three out of the twelve collections (i.e., 2017 Manchester Arena bombing, 2015 Nepal earthquake, and 2012 Hurricane Sandy), while others can be found in the following weeks. Sadness peaks could be spotted on the first day in seven disasters, while others appeared in the next few weeks. For the surprise mood, the first-day peak only existed in one disaster (i.e., 2016 Hurricane Matthew); other peaks were distributed evenly over the weeks. Based on the mood peak distribution, we found that the sadness delay is shorter than the other two moods, which indicates that people tend to show their sadness feelings while facing a disaster right after its occurrence. In contrast, fear and surprise moods are likely to be delayed, as disasters and their reporting develop or unfold.

CONCLUSION

Focusing on patterns and mood changes during disasters on Twitter, we built twelve tweet collections with four disaster types: *school shooting*, *bombing*, *earthquake*, and *hurricane*. Then, we designed and implemented TwiRole to detect different roles of users, and used a pre-trained RNN model to predict moods (i.e., fear, sadness, and surprise) of tweets. Both user roles and mood scores were added to the data in the raw tweet collections to enrich the tweet representation.

Regarding analysis, we first investigated different types of disasters and found that each type has its own pattern. Next, based on the percentage of each role in our disasters, we concluded that brand users played an important role in disasters, while male users seemed more active than female users. Then, we worked on the mood changes at the event level and discovered that fear is the major feeling in most disasters, while sadness is dominant in two of the school shootings analyzed. By exploring each collection at the week level, we carried out a temporal analysis and discussed mood delays in disasters, where sadness is more time-sensitive than fear and surprise. Finally, two case studies illustrated that mood variation appears among users with different roles.

There are some limitations to our study that should be noted. The performance of the two classifiers might not be the best, in identifying users and moods, resulting in an approximate analysis. Three disasters in each type might not be enough for pattern analysis. We accumulated the mood scores of tweets, and it might not be the best approach to describe users' moods. In future work, we plan to add more disaster collections and improve the performance of TwiRole. We also will examine the analysis results on both local and distant populations, since the mood patterns might be different. Another improvement would be to explore the text information (e.g., frequent words, favorite topics) of tweets of users, which can assist in victim recovery in real-world emergencies.

ACKNOWLEDGMENT

Thanks go to the US NSF for its support of the Global Event and Trend Archive Research (GETAR) project, through grant IIS-1619028. We would also thank Ziqian Song and Xuan Zhang for their early work on TwiRole.

REFERENCES

Alam, F., Ofli, F., and Imran, M. (2018). "A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria". In: *ISCRAM*.

Alowibdi, J. S., Buy, U. A., and Yu, P. (2013a). "Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter". In: *Machine Learning and Applications (ICMLA), 12th International Conference on*. Vol. 1. IEEE, pp. 365–369.

Alowibdi, J. S., Buy, U. A., and Yu, P. (2013b). "Language Independent Gender Classification on Twitter". In: *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, pp. 739–743.

Artwick, C. G. (2014). "News sourcing and gender on Twitter". In: *Journalism* 15.8, pp. 1111–1127.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). "Gender identity and lexical variation in social media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). "Discriminating Gender on Twitter". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 1301–1309.

Caragea, C., Squicciarini, A. C., Stehle, S., Neppalli, K., and Tapia, A. H. (2014). "Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy". In: *ISCRAM*.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). "Gender Inference of Twitter Users in Non-English Contexts". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1136–1145.

Colneriĉ, N. and Demsar, J. (2018). "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model". In: *IEEE Transactions on Affective Computing*.

Daas, P. J., Burger, J., Le, Q., Bosch, O. ten, and Puts, M. J. (2016). *Profiling of Twitter users: a big data selectivity study*. Tech. rep. Discussion paper 201606, Statistics Netherlands.

Ekman, P. (1992). "An argument for basic emotions". In: *Cognition & emotion* 6.3-4, pp. 169–200.

Ferrari, F., Jasuja, A., Seth, M., and Sharma, R. D. (2017). *Gender Inference Based On Twitter Profiles*. https://csewe-b.ucsd.edu/classes/wi17/cse258-a/reports/a119.pdf.

Fischer, D., Schwemmer, C., and Fischbach, K. (2018). "Terror Management and Twitter: The Case of the 2016 Berlin Terrorist Attack". In: *ISCRAM*.

Fortmann-Roe, S. (2013). "Effects of hue, saturation, and brightness on color preference in social networks: Gender-based color preference on the social networking site Twitter". In: *Color research & application* 38.3, pp. 196–202.

Geng, L., Zhang, K., Wei, X., and Feng, X. (2017). "Soft Biometrics in Online Social Networks: A Case Study on Twitter User Gender Recognition". In: *Applications of Computer Vision Workshops (WACVW), 2017 IEEE Winter*. IEEE, pp. 1–8.

Herfort, B., Schelhorn, S.-J., Albuquerque, J. P. d., Zipf, A., et al. (2014). "Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the River Elbe Flood in June 2013". In: *ISCRAM*.

Kaggle (2016). *Twitter User Gender Classification*. https://www.kaggle.com/crowdflower/twitter-user-gender-classification.

Kavanaugh, A., Sheetz, S. D., Hassan, R., Yang, S., Elmongui, H. G., Fox, E. A., Magdy, M., and Shoemaker, D. J. (2013). "Between a Rock and a Cell Phone: Communication and Information Technology Use during the 2011 Uprisings in Tunisia and Egypt". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 5.1, pp. 1–21.

Lasorsa, D. (2012). "Transparency and Other Journalistic Norms on Twitter". In: *Journalism Studies* 13.3, pp. 402–417.

Lee, S., Elsherbiny, N., and Fox, E. A. (2012). "A Digital Library for Water Main Break Identification and Visualization". In: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. ACM, pp. 335–336.

Levi, G. and Hassner, T. (2015). "Age and Gender Classification Using Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–42.

Li, J., Ritter, A., and Hovy, E. (2014). "Weakly Supervised User Profile Extraction from Twitter". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 165–174.

Liu, W. and Ruths, D. (2013). "What's in a Name? Using First Names as Features for Gender Inference in Twitter." In: *AAAI Spring Symposium: Analyzing Microtext*. Vol. SS-13-01. AAAI Technical Report. AAAI.

Moreno, A., Garrison, P., and Bhat, K. (2017). "WhatsApp for Monitoring and Response during Critical Events: Aggie in the Ghana 2016 Election". In: *ISCRAM*.

Mottl, D. (2018). *GetOldTweets3*. https://github.com/Mottl/GetOldTweets3.

Nagy, A. and Stamberger, J. (2012). "Crowd Sentiment Detection during Disasters and Crises". In: *ISCRAM*, pp. 1–9.

Neppalli, V. K., Medeiros, M. C., Caragea, C., Caragea, D., Tapia, A. H., and Halse, S. E. (2016). "Retweetability Analysis and Prediction during Hurricane Sandy". In: *ISCRAM*.

Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.-Y., Kapadia, A., and Rojas, F. (2016). "Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility". In: *ICWSM*, pp. 289–298.

Norcross, J. C., Guadagnoli, E., and Prochaska, J. O. (1984). "Factor structure of the Profile of Mood States (POMS): two partial replications". In: *Journal of Clinical Psychology* 40.5, pp. 1270–1277.

Pennacchiotti, M. and Popescu, A.-M. (2011a). "A Machine Learning Approach to Twitter User Classification". In: *ICWSM*, pp. 281–288.

Pennacchiotti, M. and Popescu, A.-M. (2011b). "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 430–438.

Plutchik, R. (1990). "Emotions and psychotherapy: A psychoevolutionary perspective". In: *Emotion, psychopathology, and psychotherapy*. Elsevier, pp. 3–41.

Purohit, H. and Chan, J. (2017). "Classifying User Types on Social Media to inform Who-What-Where Coordination during Crisis Response". In: *ISCRAM*.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). "Classifying Latent User Attributes in Twitter". In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, pp. 37–44.

Schulz, A., Thanh, T. D., Paulheim, H., and Schweizer, I. (2013). "A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts". In: *ISCRAM*.

Shen, S., Cheng, C., Song, C., Yang, J., Yang, S., Su, K., Yuan, L., and Chen, X. (2018). "Spatial distribution patterns of global natural disasters based on biclustering". In: *Natural Hazards* 92.3, pp. 1809–1820.

Vicente, M., Batista, F., and Carvalho, J. P. (2015). "Twitter gender classification using user unstructured information". In: *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. IEEE, pp. 1–7.

Wang, Y., Feng, Y., Luo, J., and Zhang, X. (2016). "Voting with Feet: Who are Leaving Hillary Clinton and Donald Trump". In: *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, pp. 71–76.

Wendland, J., Ehnis, C., Clarke, R., and Bunker, D. (2018). "Sydney Siege, December 2014: A Visualisation of a Semantic Social Media Sentiment Analysis". In: *ISCRAM*.

Wilensky, H. (2014). "Twitter as a Navigator for Stranded Commuters during the Great East Japan Earthquake". In: *ISCRAM*.

Yang, S., Chung, H., Lin, X., Lee, S., Chen, L., Wood, A., Kavanaugh, A. L., Sheetz, S. D., Shoemaker, D. J., and Fox, E. A. (2013). "PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media". In: *ISCRAM*.