


# Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters

Bairong Wang<sup>1</sup> · Jun Zhuang<sup>1</sup> 

Received: 27 November 2017 / Accepted: 1 May 2018 / Published online: 11 May 2018  
© Springer Science+Business Media B.V., part of Springer Nature 2018

**Abstract** The rapid spread of rumors occurring on social media is a critical problem that poses a great risk to emergency situation navigation, especially during disasters. Many research questions, such as how misinformed users judge potential rumors or how they respond to them, are crucial issues for crisis communication, but have not been extensively studied. This paper fills this gap by originally documenting and studying Twitter users' rumor and debunking response behaviors during disasters, such as Hurricane Sandy in 2012 and the Boston Marathon bombings in 2013. To this end, two rumors from each disaster and their related tweets are documented for analysis. Users who were misinformed and involved in the rumor topic by posting tweet(s), could respond to a rumor by: (1) spreading (85.86–91.40%), (2) confirmation-seeking (5.39–9.37%), or (3) doubting (0.71–8.75%). However, if the rumor-spreading users were debunked, they would respond by: (1) deleting rumor tweet(s) (2.94–10.00%), (2) clarifying rumor information with a new tweet (0–19.75%), or (3) neither deleting nor clarifying (78.13–97.06%). We conclude that Twitter users perform poorly in rumor detection and rush to spread rumors. The majority of users who spread rumors do not take further action on their Twitter accounts to fix their rumor-spreading behaviors.

**Keywords** Crisis information · Twitter · Rumor response · Debunking response · Decision analysis

## 1 Introduction

Social media has been widely used for crisis communication during disasters due to the advanced development of mobile technologies (Gupta et al. 2013; Kang et al. 2015; Zubiaga et al. 2016). Due to this, it has become common for event updates to be available

---

✉ Jun Zhuang  
jzhuang@buffalo.edu

<sup>1</sup> Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, New York, USA

on social media first and then introduced in the mainstream media thereafter (Oh et al. 2011). Information from social media is also commonly used for evacuation decision makings during disasters (Sadri et al. 2017). Despite the advantages of social media, it has also been criticized and deemed a rumor mill for propagating misinformation and gossips during crises (Leberecht 2010), which can cause large-scale panics (Richards and Lewis 2011) and even economic loss (Liu et al. 2015). To make matters worse, social media users are notorious for their poor rumor detection (Rubin 2017) and rushing to share early unverified information (Zubiaga et al. 2016) during disasters. However, there are also individuals who combat rumors based on their bounded knowledge (Rosnow and Fine 1976) and try to turn to other sources for confirmation to verify the claims (Lewandowsky et al. 2012). With those rumor combating users, social media is able to recognize and even correct rumors (Zhao et al. 2016; Tripathy et al. 2010; Jong and Dückers 2016). By recognizing this duality of social media users' behaviors when facing rumor information during disasters, this paper explores both the rumor and debunking response behaviors of social media users based on their different information statuses. This paper addresses these two research questions: (1) How do social media users respond to rumors when they first become misinformed? (2) Will social media users combat rumors when they are debunked by accurate information? and if so, what will they do? Analysis of real tweets data is used in this study to address these research questions. We first choose one recent natural disaster (Hurricane Sandy in 2012) and one recent man-made disaster (Boston Marathon Bombings in 2013). For each of the two disasters, we read news and related articles to identify the top two widely spread rumors, fake news, and hoaxes (Hill 2012; Holt 2012; Sager 2013; Montopoli 2013). With the four rumor case studies, the main contributions of this research include:

- Analysis of rumor awareness and response behaviors of Twitter users during disasters.
- Investigation of the debunked status of rumor-spreading users and their debunking response behaviors. To the best of our knowledge, this work is the first one to analyze both rumor and debunking response behaviors of Twitter users.

The rest of this paper is organized as follows: Sect. 2 introduces the literature review; Sect. 3 presents the research methods and data used in this study; Sect. 4 presents an analysis of our results; Sect. 5 proposes a decision tree model of misinformed social media users; Sect. 6 summarizes and concludes.

## 2 Literature review

### 2.1 Literature on rumor response behaviors

Rumors, defined as “the informally improvised news” (Shibutani 1969), often come along with disasters and other crises (Rosnow and Fine 1976). Social media, which is an important tool for crisis communication during disasters, is susceptible to rumors and other malicious applications due to its accessibility to everyone (Tripathy et al. 2013). The advanced development of mobile devices has made it a norm for an incident to be first reported by a local eyewitness, and then covered by the mainstream media during large-scale crises (Oh et al. 2011). As a result, rumors or misinformation may have propagated widely by the time anti-rumor information from authoritative sources is available (Ozturk et al. 2015). Therefore, social media users are left most of the time to evaluate the accuracy of the information on their own and decide whether to disseminate or combat the potential

rumor (Ozturk et al. 2015). By reviewing this dilemma that occurs to social media users, their behaviors have been examined when facing potential rumor information. Results from Zubiaga et al. (2016), Rubin (2017), Morris et al. (2012) and Starbird et al. (2014) show that social media users are notorious for poor rumor detection during disasters due to people's truth-biased characters (Rubin 2017), which means that people tend to ascribe truth to messages that they have received than deceit (Levine et al. 1999). Research from Zubiaga et al. (2016) shows that Twitter users rush to share early unverified information, rather than the later accurate rumor-debunking information. Traditional socio-psychology research from Rosnow and Fine (1976) has shown that not all of the users will become rumor spreaders, as some individuals are able to make basic judgments based on their bounded knowledge (Knopf 1975; Rosnow 1991). People may turn to other sources for information confirmation based on their prior beliefs on the information that they have received (Lewandowsky et al. 2012). Reviewing the literature, it is clear that even if individuals are misinformed by the same rumor information, the response behaviors of social media users may vary from individual to individual. However, most of the current work focuses on exploring rumor awareness and rumor-spreading behaviors of social media users (Zubiaga et al. 2016; Rubin 2017; Morris et al. 2012; Starbird et al. 2014), and little research has been done to explore different response behaviors of social media users except for rumor spreading. An exception is the work of Dang et al. (2016), which identified three types of response behaviors: (1) support, (2) refute, and (3) joke. However, based on Lewandowsky et al. (2012) and Knopf (1975), social media users could also seek confirmation, or make assessments of the information that they receive during disasters. Therefore, we are motivated to explore more rumor response behaviors of social media users during disasters, which have not been studied extensively in the literature.

## 2.2 Literature on rumor-debunking behaviors

Although social media has been criticized by many researchers for accelerating the spread of rumors, it can still be helpful in rumor control in times of disasters. The self-correction function of social media, identified by Gayo-Avello et al. (2013), Zubiaga et al. (2016) and Castillo et al. (2011), can be explained by the broad mass participation and persistent discussions inherent in social media, with which rumors and misinformation can be identified and corrected by knowledgeable users at an early stage of their spread (Friggeri et al. 2014; Alexander 2014). Traditionally, rumors could be debunked either by an individual's common sense or by actual professional verifications (Rubin 2017). Rumors spread on social media are best quelled in the same word-of-mouth way that they spread, which is via messages spread from one user to another (Tripathy et al. 2013). This indicates that the best rumor quelling method depends heavily on rumor combating behaviors of social media users during disasters. However, research from Zhao et al. (2016) has identified a "higher awareness but lower behavior gap" of social media users in terms of rumor combating. Similar results are also found in Zubiaga et al. (2016), which indicates that social media users are less likely to share confirmation or rumor-debunking information compared with unverified reports in the early stage of rumor spreading. In summary, not all social media users will combat rumors even if they have already been debunked by accurate information. Reviewing the literature, a research gap is the investigation of the rumor-debunking behaviors of social media users when they are debunked by accurate information. To the best of our knowledge, this paper is the first study to investigate rumor-debunking response behaviors of social media users during disasters.

### 3 Research method

#### 3.1 Disasters and rumor cases

*Hurricane Sandy* was the largest hurricane of the Atlantic hurricane season in 2012 and the second most costly natural disaster in the U.S. history (Blake et al. 2013a). The hurricane formed on October 22, 2012 and made its landfall at Brigantine, New Jersey on October 29, 2012. There were 24 states affected by this hurricane, with particularly severe damages in New Jersey and New York. The disaster dissipated on November 2 and caused 147 casualties and \$75 billion in damages (Blake et al. 2013a,b).

The first rumor case studied is the New York Stock Exchange (NYSE) flooding rumor, which was originated by the Twitter user @ComfortablySmug. The user mentioned that the NYSE was flooded and under more than 3 feet of water (Hill 2012; Holt 2012), as shown in Table 1. The rumor spread all the way to CNN, where meteorologist Chad Myers of Piers Morgan's show announced that the NYSE was under 3 feet of water and mentioned that the news came from the National Weather Service (NWS) (Wemple 2012). Another rumor case that we studied pertains to the Coney Island Hospital fire, as shown in Table 1, which originated as a report on a police scanner, indicating that the Fire Department City of New York (FDNY) firefighters had troubles entering the Coney Island hospital to respond to a fire in the building (Inside Breaking News 2012). The rumor ended up spreading widely on Twitter, especially by users who thought the scanner to be a reliable source (Oremus 2012). However, it was revealed finally that the FDNY was actually responding to a car fire in the parking lot rather than a fire in the hospital itself (Inside Breaking News 2012).

*The Boston Marathon bombings in 2013* occurred on April 15, 2013 at 2:49 p.m. EDT, and left 3 people dead and 264 injured (Kotz 2013). Three days later, a subsequent shooting occurred on the MIT campus, leaving one suspect dead, while the other suspect escaped. The two suspects identified by the FBI were Tamerlan and Dzhokhar Tsarnaev. The brother who survived was found and arrested on April 19, 2013 at 9:00 p.m. EDT (Starbird et al. 2014).

**Table 1** Original rumor tweets of four rumor cases

Rumor case	Time stamp	Re-tweets	Likes	Message
NYSE flooding	10/29/2012 18:04	580	35	BREAKING: Confirmed flooding on NYSE. The trading floor is flooded under more than 3 feet of water
Hospital fire	10/29/2012 19:44	281	5	Brooklyn:FDNY is enroute to Coney Island Hospital for a reported fire on the third floor with a heavy smoke condition
RT donation	04/15/2014 11:29	52,173	855	For every re-tweet we receive we will donate \$ 1.00 to the #BsotonMarathon victims #PrayForBoston
Sandy hook girl	04/15/2012 19:41	816	151	R.I.P to the 8 year-old girl who died in Boston explosions while running for the Sandy Hook kids. #prayforboston pic.twitter.com/yiTw4WUcbZ

The first rumor case is about the re-tweet (RT) donation scam, shown in Table 1. It originated from a fake account @\_BostonMarathon claiming that “For every re-tweet we receive we will donate \$1.00 to the #BostonMarathon victims #PrayForBoston.” The tweet received at least 52,173 re-tweets before the account was suspended by Twitter (Sager 2013). Another rumor case is about the killing of an 8-year-old girl who had survived the Sandy Hook massacre. The original rumor tweet, as shown in Table 1, claimed that an 8-year-old girl who survived the Sandy Hook mass shootings was killed in the bombings (Sager 2013). A photograph of the 8-year-old girl was even presented to support the information. However, the child killed in the bombings was actually an 8-year-old boy who was waiting at the finish line of the marathon and was rushing out to hug his dad who had just completed the race (Montopoli 2013). The bombs went off when he returned to his mother and sister, which resulted in his death, severe injuries to his mother and his sister losing one leg (Hollywood Life Staff 2013).

### 3.2 Data collection

Twitter REST API <https://dev.twitter.com/rest/public> is used for data collection. First, we apply keywords to search rumor related tweets. Second, we eliminate reply tweets, and keep only the normal original tweets for analysis. Third, we collect re-tweets and descendant replying tweets to the normal original tweets. We choose 1 month as the data searching period to get relatively complete and comprehensive data given the fact that the number of tweets related with a disaster decreases significantly 15 days after the occurrence of that disaster (Wang and Zhuang 2017). Specifically, the keywords, hashtags, and search period for each rumor case are “NYSE flood” and “New York Stock Exchange flood;” “Coney Island Hospital fire,” “Coney Island fire,” and “Coney Island Hospital” from Oct 29 to Nov 29, 2012; “Sandy Hook child,” “Sandy Hook girl,” “eight year boy,” and “eight year girl” from April 15 to May 15, 2013, and “#BostonMarathon,” “#PrayForBoston,” “re-tweet donate,” “\$1.00 donation,” and “\$1.00 donate” from April 15 to May 15, 2013. Removing all unrelated tweets, results of the data collection are listed in Table 2, within which “Normal,” “Re-tweets,” “Descendants,” and “Deleted” refer to the counts of normal tweets, re-tweets, descendant tweets, and deleted tweets, respectively. There are two types of re-tweets in Twitter: the first one is made by clicking re-tweet button under each original tweet and the re-tweet(s) (without modifications) will appear in the re-tweeter’s timeline. The second type is made by copying the original message and adding “RT @” or “MT @” before a user name to show the citation source(s). We examine only the first type in this study. More than 20,000 tweets were collected for analysis in this study.

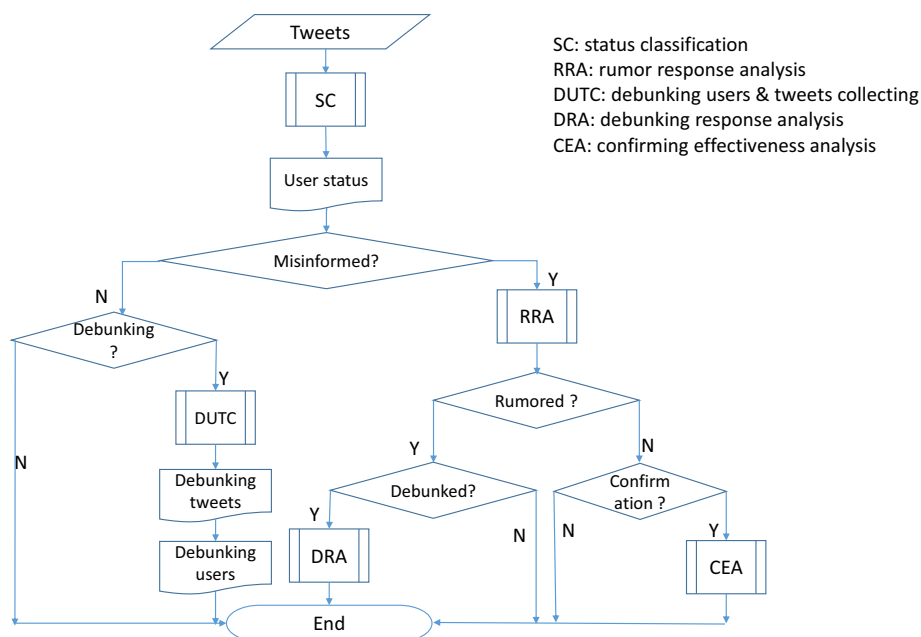
**Table 2** Data collection results for each rumor case

Rumor case	Date	Normal	Re-tweets	Descendant	Deleted	Total
NYSE flooding	10/29-11/29	1259	3719	868	16	5862
Hospital fire	10/29-11/29	878	2655	533	17	4083
RT donation	04/15-05/15	651	1466	403	31	2551
Sandy Hook girl	04/15-05/15	2078	4862	1183	23	8146
Total		4866	12,702	2987	87	20,642

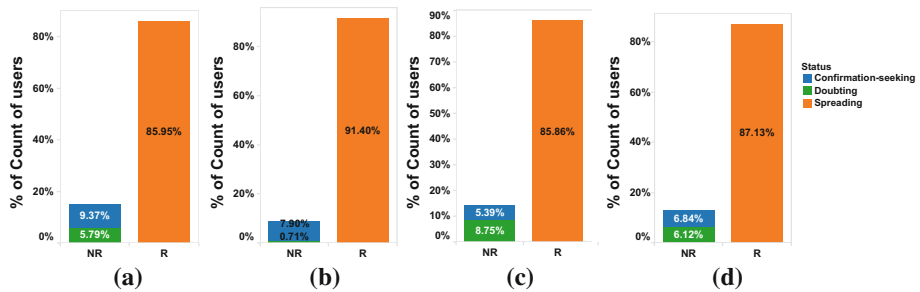
### 3.3 Content analysis

Originating from communication and media research, content analysis (CA) is widely used to derive quantitative information from a body of open-ended qualitative data by analyzing the frequency or occurrence of ideas, or codes (Berg et al. 2004; Neuendorf 2002; Patton 1990). This paper uses CA to categorize the Twitter users' beliefs and responses to rumor information in the four cases. Reliability of CA results is ensured by the coding scheme that requires at least two coders for the coding process. In this study, manual coding is used to classify Twitter users' beliefs and responses to potential rumors.

In this paper, we analyze social media users who were misinformed by rumor information and involved themselves in the rumor topic on Twitter by posting tweet(s). The analysis consists of status classification (SC), rumor response analysis (RRA), debunking users and tweets collecting (DUTC), debunking response analysis (DRA), and confirmation-seeking effectiveness analysis (CEA), as shown in Fig. 1. We first analyze the tweets content to identify the information status of the tweet's author and generate a user status file. If the tweet's content shows that the user is debunking the rumor, then this original debunking tweet, the tweet's author, re-tweets of this rumor-debunking tweet, and users of all these re-tweets will be documented in the DUTC process. Two data sets are generated from the DUTC process, including a debunking tweets data set and a debunking users data set. If the tweet's content shows that the user is misinformed, then RRA will be performed, including spreading (S), confirmation-seeking (C), and doubting (T). If a tweet message shows that the user has spread the rumor ("rumored" in Fig. 2), DRA will be performed to learn how they would respond if they were debunked by accurate information. Confirmation-seeking users refer to users who post tweets to seek confirmation on Twitter; e.g., "I have seen the news that..., is that true?" While doubting users post tweets to show their



**Fig. 1** Flow chart for tweet analysis



**Fig. 2** Statistics of rumor awareness for each case. **a** NYSE flooding. **b** Hospital fire. **c** RT donation. **d** Sandy Hook girl

doubt on the messages based on their knowledge; e.g., “I do not think this message is true since only adults are allowed for marathon.” For confirmation-seeking users, CEA will be performed to learn the effectiveness of their confirmation-seeking behaviors.

## 4 Results

### 4.1 Status classification (SC)

Results of SC for all rumor cases are shown in Table 3. Three information statuses identified are: (1) debunking (D), (2) misinformed (M), and (3) others (O). Debunking refers to a status in which a user already knows it is a rumor and debunks the rumor with a tweet. Misinformed refers to a status in which a user has received the rumor information, but is not clear regarding the information’s accuracy. Others status refers to a status where a user is neither misinformed nor debunking, they may post for commenting; e.g., “I don’t care whether the NYSE is flooding. I care about the thousands of poor families who can’t afford to rebuild from scratch.” We filter these tweets out and tag them as “others” status. Due to the fact that a user can post several tweets with different statuses, the number of users involved in each rumor case is not necessarily the sum of users in each information status. For example, we have 788 total users in the Hospital fire rumor, but the sum of users from each type of information status is 823 ( $= 647 + 150 + 26$ ), as shown in Table 3.

**Table 3** Number of tweets and users with misinformed (M), debunking (D), and other (O) status

Rumor case	Tweets				Users			
	M	D	O	Total	M	D	O	Total
NYSE flooding	363	570	331	1264	363	545	321	1165
Hospital fire	709	160	26	895	647	150	26	788
RT donation	297	203	182	682	296	199	182	675
Sandy Hook girl	1118	784	199	2101	1111	778	199	2081
Total	2487	1717	738	4942	2417	1672	728	4709

Note that the number of “M” users is larger than that of “D” users in the NYSE flooding case, potentially because the NYSE flooding rumor is easier to debunk than others.

## 4.2 Rumor response analysis (RRA)

Response behaviors of misinformed users are summarized in this section, including: spreading (S), doubting (T), and confirmation-seeking (C), which are identified when the tweets indicate that the users are spreading rumors, doubting rumors, and seeking confirmation of the rumors, respectively. We define the status of users who doubt or seek confirmation as non-rumored (NR), and we define the status of rumor-spreading users as rumored (R). Figure 2 shows that of Twitter users who were misinformed and involved in the rumor topic on Twitter, more than 86% of them would be rumored and rush to spread false information, which is consistent with the conclusions from Morris et al. (2012) and Starbird et al. (2014) that Twitter users perform poorly with respect to rumor detection. Additionally, our results reveal that less than 9% would doubt the rumor. Information ambiguity or abnormal features of rumors can increase users’ rumor awareness. For instance, 8.75% doubting users were identified in the RT donation rumor case, which is higher than the percentages in other rumor cases, with 5.79% in the NYSE flooding case, 0.71% in the hospital fire case, and 6.12% in the Sandy Hook girl rumor case. A possible explanation for this higher doubting rate are the abnormal features of the rumor-spreading user @\_BostonMarathon, who registered only several hours before the rumor tweet and had only several prior tweets. All of these features could impair the credibility of this information (Abbasi and Liu 2013). Less than 10% of misinformed Twitter users would respond by seeking rumor confirmation with a new tweet.

## 4.3 Debunking users and tweets collecting (DUTC)

Both debunking tweets and the users posting debunking tweets are collected in the DUTC process. Debunking tweets include original normal debunking tweets and their re-tweets, which share the same doubting status since no modification is made when a re-tweet occurs. With the data set of debunking users, we are able to test friendship between a debunking user and a rumor-spreading user. If a rumor-spreading user follows a rumor-debunking user, then this rumor-spreading user is able to receive the rumor-debunking tweet(s) posted by the rumor-debunking user. Therefore, we are able to figure out whether a rumor-spreading user received anti-rumor tweet(s) from their friends. By testing the friendships of a rumor-spreading user with all rumor-debunking users, we are able to get the frequency of similar rumor-debunking tweet(s) that a rumor-spreading user has received. Results of the DUTC process for each rumor case are summarized in Table 4.

**Table 4** Number of debunking (D) tweets and users collected in each rumor case

Rumor case	Normal (D)	Re-tweets (D)	Total tweets (D)	Total users (D)
NYSE flooding	568	2065	2633	2462
Hospital fire	160	757	917	853
RT donation	203	975	1178	1283
Sandy Hook girl	784	467	1251	1244



About 4000 more debunking users from four rumor cases were identified by collecting re-tweets of original rumor-debunking tweets, which is helpful in obtaining a better estimate of the debunked status of a rumor-spreading user.

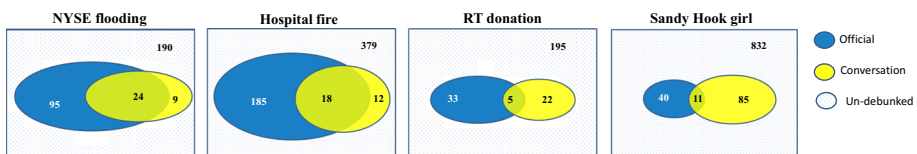
#### 4.4 Debunking response analysis (DRA)

Before we investigate the debunking response behaviors of rumor-spreading users, an analysis of their debunked status has to be made. Within this analyzing process, debunking methods, coverage, and frequency for rumor-spreading users are analyzed as follows.

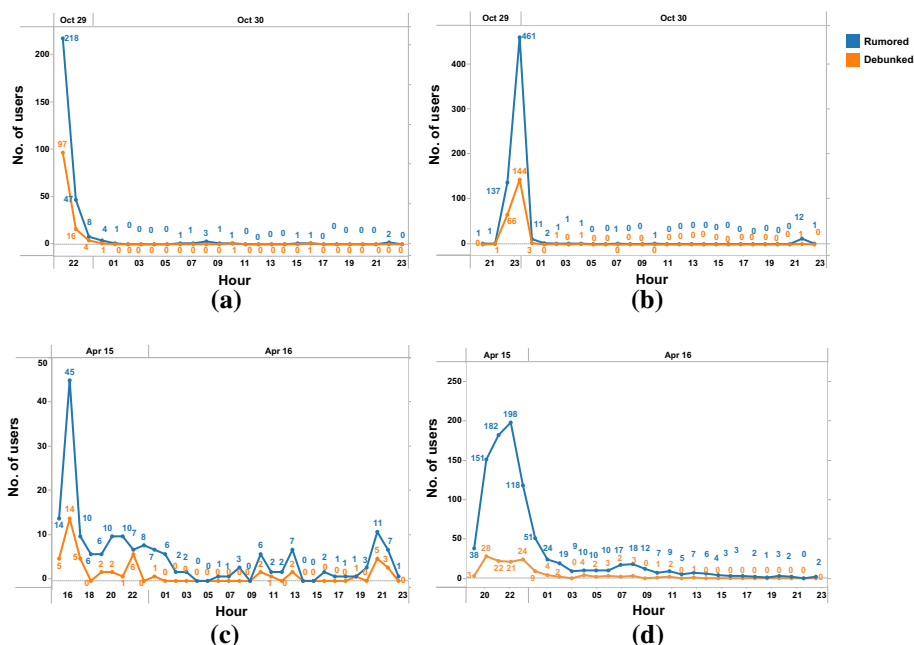
*Debunking methods* include official debunking (receiving rumor-debunking tweets from friends) and conversational debunking (getting rumor-debunking information from the descendant replies under their original rumor tweets). For example, an original tweet user A has received the rumor-debunking information via a descendant reply from user B saying the rumor information is false. However, the user A may also receive another descendant reply from user C confirming the rumor information is true even before the rumor-debunking reply. For all of the rumor-spreading users, we first investigate whether they were debunked or not. If they were debunked, we want to determine by which method they were debunked. Venn diagrams in Fig. 3 illustrate debunking results for each rumor case. The majority of debunked-rumor-spreading users were debunked in the official way, except for the Sandy Hook girl rumor case. The percentage of officially debunked users reaches a very high level of 94% ( $= (185 + 18)/(185 + 18 + 12)$ ) in the Hospital fire rumor case. However, the results in the Sandy Hook girl rumor were reversed with more than 70% of debunked-rumor-spreading users being debunked in the conversational way. There were also rumor-spreading users debunked by both methods, which took about 8% in the Hospital fire, RT donation, and Sandy Hook girl rumor cases. Moreover, the percentage for the NYSE flooding rumor case was 19%.

*Debunking coverage* refers to the fraction of debunked users of all rumor-spreading users. For instance, Fig. 3 shows that 128 ( $= 95 + 24 + 9$ ) out of 318 ( $= 95 + 24 + 9 + 190$ ) rumor-spreading users were debunked in the NYSE rumor case, with a debunking coverage of 40%. The debunking coverage in the Hospital fire rumor, RT donation rumor, and Sandy Hook girl rumor were 36, 23, and 14%, respectively. All were less than 40% among four rumor cases, which is not satisfying. However, we found that the number of rumor-spreading users decreased significantly several hours after the rumor release, as shown in Fig. 4, indicating that the majority of the users were already clear of the truth value of the rumor information and stopped posting rumor tweets despite the large number of un-debunked users in earlier hours. These rumor-spreading users were either debunked by other Twitter users, or by accurate anti-rumor information from other platforms, such as TV news.

*Debunking frequency* is used to measure how much Twitter users are overloaded with similar rumor-debunking information during disasters. Analysis of debunking frequency was only applied to the official debunking method to ensure all of the rumor-spreading



**Fig. 3** Number of official, conversational and un-debunked users in all rumor cases



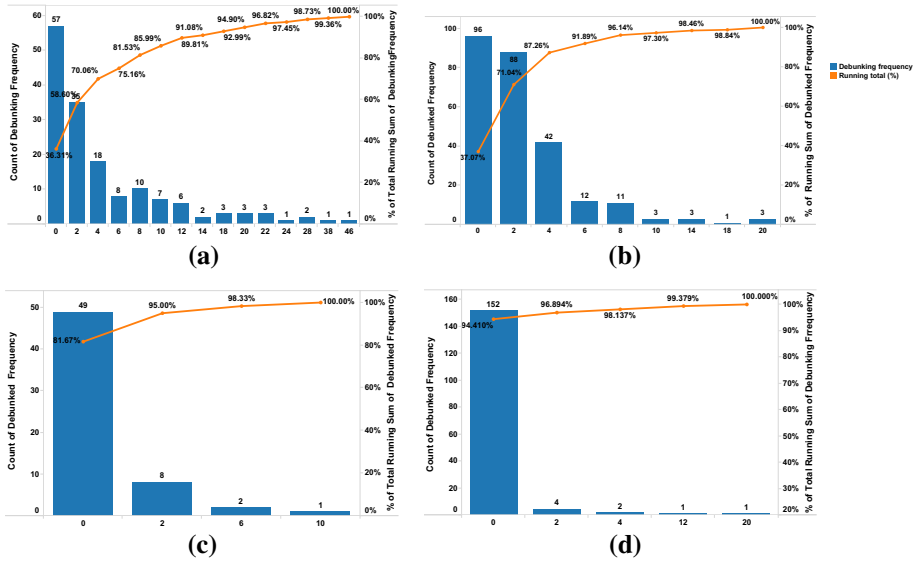
**Fig. 4** Number of rumored and debunked users in first two days of each rumor. **a** NYSE flooding. **b** Hospital fire. **c** RT donation. **d** Sandy Hook girl

**Table 5** Debunking frequency of debunked users in each rumor case

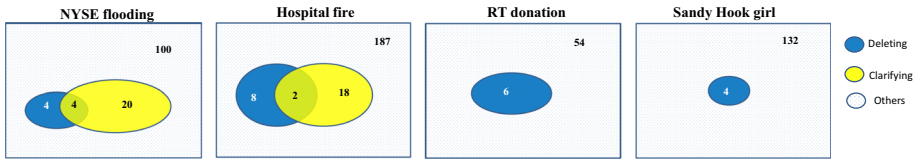
Rumor case	Debunked user	Debunking user	Max	Min	Median	Mean
NYSE flooding	119	2462	47	1	2	4.99
Hospital fire	215	853	20	1	2	2.65
RT donation	60	1283	10	1	1	1.76
Sandy Hook girl	136	1244	21	1	1	1.82

users were tested for friendship by the same debunking users set. Table 5 shows that a debunked-rumor-spreading user had received 1.76–4.99 similar debunking tweets. In the two Hurricane Sandy rumor cases, 50% of debunked-rumor-spreading users had received at least two similar rumor-debunking tweets (median = 2). Debunked-rumor-spreading users in NYSE flooding rumor received an average of five similar rumor-debunking tweets. In an extreme case, they could receive forty seven similar debunking tweets, as shown in Fig. 5a. Figure 5 illustrates that approximately 30% and 13% of officially debunked-rumor-spreading users received more than four similar rumor-debunking tweets in the NYSE flooding and Hospital fire rumor cases. On the other hand, for the RT donation and Sandy Hook girl rumor cases, the majority of the officially debunked-rumor-spreading users received one rumor-debunking tweet, as shown in Fig. 5.

The debunking responses identified were deleting (previous rumor tweets) and clarifying (by posting a new rumor clarification tweet) when a rumor-spreading user was



**Fig. 5** Histograms of debunking frequency of debunked users in each rumor case. **a** NYSE flooding. **b** Hospital fire. **c** RT donation. **d** Sandy Hook girl



**Fig. 6** Number of users responding by deleting, clarifying and others after being debunked in each rumor case

debunked by accurate information. Results for each rumor case are illustrated in Fig. 6, indicating that the majority of debunked-rumor-spreading users would neither respond by deleting their previous rumor tweet(s), nor by posting a new tweet to clarify their rumor-spreading behavior. The aforementioned users took 78% ( $= 100/(100 + 4 + 4 + 20)$ ), 87% ( $= 187/(187 + 8 + 2 + 18)$ ), 90% ( $= 54/(54 + 6)$ ), and 97% ( $= 132/(132 + 4)$ ) in the NYSE flooding, Hospital fire, RT donation, and Sandy Hook girl rumor cases, respectively. Less than 10% of debunked-rumor-spreading users would delete their rumor tweets in all rumor cases and that percentage of users who post clarifying tweets was zero in the two Boston Marathon Bombing rumor cases. Only about 14 and 7% of debunked-rumor-spreading users would both delete and make clarifying statements after being debunked in the two Hurricane Sandy rumor cases.

#### 4.5 Confirmation-seeking effectiveness analysis (CEA)

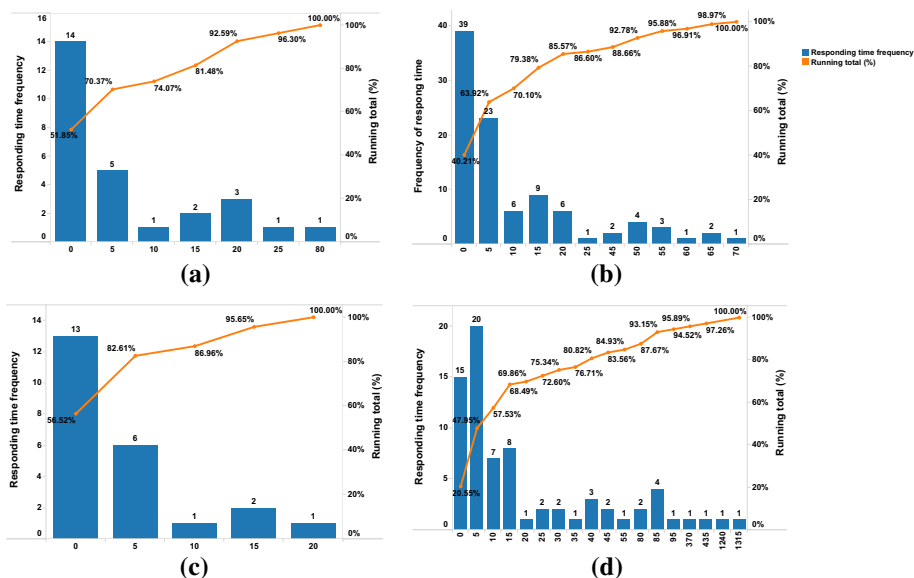
Reply frequency, efficiency, and results are analyzed for the confirmation-seeking effectiveness analysis. Replied frequency is investigated to study how many confirmation-seeking tweets were replied to, and if replied to, how many replies that these tweets obtained. Replied efficiency is also studied to investigate how fast these confirmation-

seeking tweets were replied to. Motivated by the fact that descendant reply tweets of an original confirmation seeking tweet may also contain rumor and debunking information, or both.

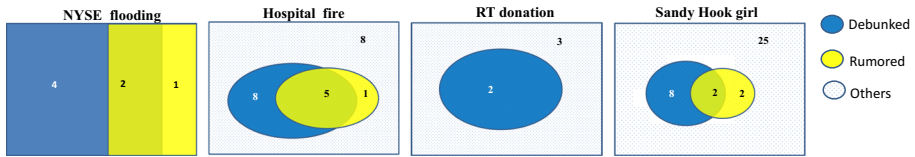
*Replied frequency* results shown that more than 26% of the confirmation-seeking tweets were replied to in each rumor case and the replied rate for non-confirmation-seeking tweets is less than 20% in each rumor case. In addition, the average number of replies to confirmation-seeking tweets ranges from 26.47 to 36.36, which is also larger than that of non-confirmation-seeking tweets, ranges from 17.02 to 19.84 in each rumor case.

*Replied efficiency* is measured by both first efficiency (FE) and average efficiency (AE). FE refers to the response speed of the first reply to a confirmation-seeking tweet. AF refers to the average response speed of all the replies to a confirmation-seeking tweet. Results show that the average response time when the first reply occurred to confirmation-seeking tweets is less than 7 min in each rumor case and that more than 50% of these first replies occurred within 4 min. The average replying time for confirmation-seeking tweets ranges from 5.87 to 65.84 min among four rumor cases. Histograms of replying time (min) in each rumor case are illustrated in Fig. 7, showing that 70.37, 63.92, 82.61, and 47.95% of the replies occurred within 5 min in the NYSE flooding, Hospital fire, RT donation, and Sandy Hook girl rumor cases, respectively. Responses to confirmation-seeking tweets are efficient on Twitter during disasters.

*Replied results* indicate that not all of the descendant replies to confirmation-seeking tweets are accurate. As shown in Fig. 8, except for the RT donation rumor case, all confirmation-seeking tweets from three other cases had rumor descendant replies under their original tweets. However, 67% ( $= 2/(2 + 1)$ ), 83% ( $= 5/(5 + 1)$ ), and 50% ( $= 2/(2 + 2)$ ) of the aforementioned confirmation-seeking tweets had rumor-debunking descendant replies coexisting with the rumor-spreading descendant replies in the NYSE flooding, Hospital fire, and Sandy Hook girl rumor cases, respectively.



**Fig. 7** Reply time histograms and running totals to confirmation-seeking tweets in each rumor. **a** NYSE flooding. **b** Hospital fire. **c** RT donation. **d** Sandy Hook girl



**Fig. 8** Number of debunked and rumored confirmation-seeking tweets in each rumor case

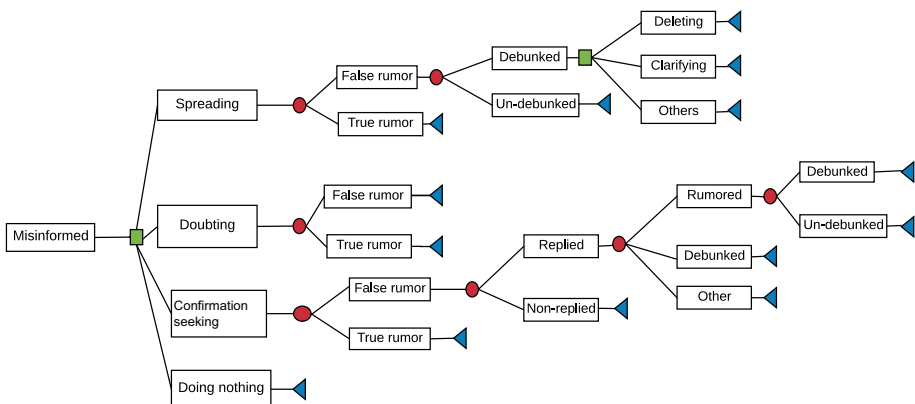
## 5 A decision tree model for misinformed social media users

Based on the response behaviors of Twitter users identified in Sects. 4.2 and 4.4, we propose a decision model of misinformed social media users in the context of rumors, as shown in Fig. 9. Specifically, a false rumor in this model refers to a claim whose truthfulness is in doubt during propagation (Harsin 2006) and turns out to be false eventually. Reading possible rumor information, a social media user may: (1) spread this information to his/her followers, (2) doubt the information and post a tweet with doubting reasons, (3) seek confirmation from other Twitter users, or (4) do nothing.

Responding by spreading, a social media user may face a risk that the information could be a false rumor. Given that the information is a false rumor, a user could be debunked either by the conversational or by the official debunking method. Once the rumor-spreading user is debunked, another decision has to be made on whether to help in rumor-debunking by deleting previous rumor tweets, posting a new rumor clarifying tweet, or by other behaviors.

Responding by doubting, a user can avoid the risk of spreading rumors to his/her followers, regardless the truth value of the information. If the information is a false rumor, this decision contributes to rumor combating during disasters and therefore generates higher utilities. If the information is a true rumor, the user may be worse off by incorrectly doubting accurate news and missing a good chance to update his/her followers with the latest news.

Responding by seeking confirmation, a social media user may face the risk of not being replied to. Given that the user is replied to, however, another risk is posed by the quality of information from the descendant replies to this confirmation seeking. Three possible results



**Fig. 9** A decision making model for a misinformed social media user

will be generated, including: (1) rumored, if replies to this confirmation-seeking tweet contain rumor information; (2) debunked, if replies contain accurate rumor-debunking information; (3) others, if all of the replies to this confirmation-seeking tweet contain nothing pertinent to the information truth value. Fortunately, rumored confirmation-seeking users may also eventually be debunked if there is also accurate rumor-debunking information available in the descendant replies. Finally, a user may also choose to do nothing when reading possible rumor information.

Misinformed social media users may make multiple decisions in different stages. For example, a user may first spread rumor(s), and then debunk the rumor(s) or delete previous rumor tweet(s), which have been identified by this study. Based on our data set, we did not find the case where the users both spread rumors, and seek confirmation at the same time. This may be due to the fact that a user who spreads the rumor may believe that the message is true and may not likely to seek confirmation at the same time. While seeking confirmation shows that a user is not sure whether the information is true or not, they are not likely spread the information before he/she gets confirmed.

In addition, not all of the possible responding decisions of misinformed social media users were identified. Future research could further identifying more rumor responding behaviors, such as spreading rumor(s) by replying to other tweets.

Results of the decision model of misinformed social media users could be used for motivating future research on decision makings of social media users when facing potential false rumors, including identifying and qualifying the factors impacting decisions of social media users.

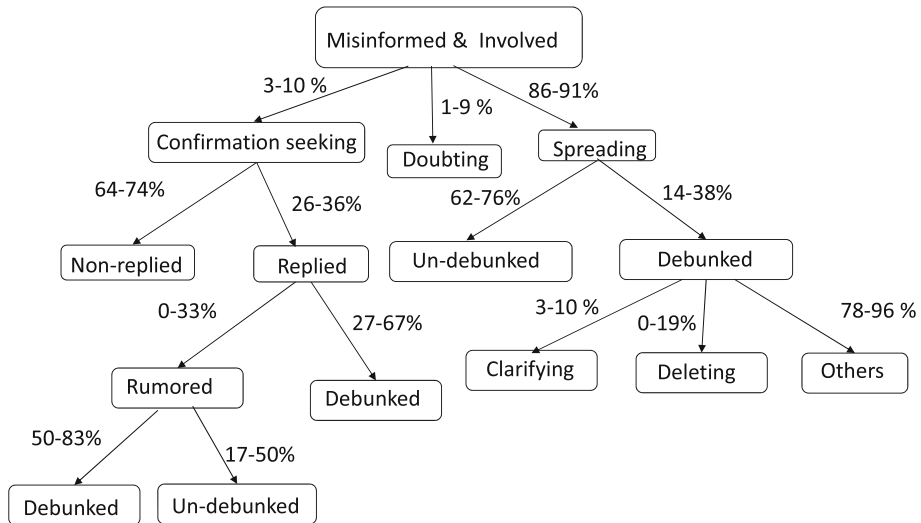
## 6 Conclusions and future research

### 6.1 Results summary and conclusions

A tree with one root node, as shown in Fig. 10, was designed to summarize the results of rumor response analysis, debunking response analysis, and confirmation-seeking effectiveness analysis. Starting from a misinformed and involved status, 3 to 10% of these Twitter users would post a new tweet seeking confirmation of the information they read, among whom 26 to 36% received replies. First replies to confirmation-seeking tweets occurred within 7 min on average, and more than 50% of all the replies occurred within 4 min. However, despite the efficient responses to confirmation-seeking tweets, the confirmation-seeking behavior is still a double-edged sword since not all of the confirmation-seeking tweets were replied to with accurate information. Results show that 0 to 33% of the confirmation-seeking users could be rumored by their descendant tweets. Fortunately, 50 to 83% of those rumored confirmation-seeking users could also be replied to with accurate rumor-debunking information.

Misinformed and involved users could also doubt the rumor information based on some abnormal features of the rumor tweet(s). Doubting users increased from 1% in the Hospital fire rumor to 9% in the RT donation rumor, which may rise from the fact that the initial rumor spreader of the RT donation rumor was a fake account with only several newly posted tweets.

Most of the misinformed and involved users spread the rumors, and the percentage ranges from 86% to 91% among the four cases. Less than 40% of these rumor-spreading users were debunked. However, we also find that the number of rumor-spreading users



**Fig. 10** Responses of misinformed and involved Twitter users identified

decreased dramatically several hours after the release of the original rumor, since rumor-spreading users could be debunked by information from other platforms (e.g., TV news) or by other debunking Twitter users. Once debunked, only 3 to 10% of these rumor-spreading users would post a new tweet to clarify the rumor tweets they had posted before. Less than 19% of the debunked-rumor-spreading users would delete their original rumor tweets. Less than 14% of the debunked-rumor-spreading users would both delete their original rumor tweets and post a clarification tweet. The majority (78–96%) of the debunked-rumor-spreading users would neither delete nor clarify.

Four major conclusions of this paper are as follows. First, three types of rumor response behaviors are identified for misinformed and involved Twitter users. Our results indicate that Twitter users could respond to rumor information by spreading it, doubting, or seeking confirmation based on the condition that they were misinformed and involved in the rumor topic on Twitter through their posting of tweets. However, the majority ( $\geq 86\%$ ) of these misinformed and involved Twitter users rushed to spread the rumor information, indicating the poor rumor detection ability of Twitter users. This finding aligns with the results from Morris et al. (2012) and Starbird et al. (2014).

Second, two types of debunking response behaviors were identified for rumor-spreading Twitter users when they were debunked. Our results show that a rumor-spreading Twitter user can respond by either deleting previous rumor tweet(s) or clarifying rumor tweet(s) they had posted with a new tweet. Both response behaviors are necessary to stop further rumor-spreading during disasters. By deleting previous rumor tweet(s), a user can avoid further re-tweets of that rumor tweet. By posting a rumor clarifying tweet, a user can debunk at least his/her misinformed followers. Unfortunately, the majority ( $\geq 78\%$ ) of rumor-spreading users would not delete previous rumor tweets or post a clarifying tweet after they were debunked.

Thirdly, posting a confirmation-seeking tweet is an effective way for information navigation during disasters on Twitter. Response to confirmation-seeking tweets was efficient since average response time of the first reply and all of the replies are short.

Despite the fact that some confirmation-seeking users could be rumored at first, at least 50% of them would be debunked eventually.

Fourth, rumor-spreading users are debunked quickly on Twitter during disasters. Our study shows that rumor-spreading Twitter users decreased significantly only several hours after the release of original rumor. Although results based on our debunking users set show that less than 40% of rumor-spreading users were debunked, they could be debunked by information from other platforms or by other Twitter users. As a result, rumor-debunking on social media is good in terms of efficiency with the assistance from other platforms.

## 6.2 Future research directions

Some further research directions include: (1) discovering how the information statuses of Twitter users change during disasters, including the possibility that a misinformed Twitter user would first seek confirmation, but would post a misleading or rumor-spreading tweet after receiving inaccurate information as a reply; (2) learning factors that may affect rumor and debunking response behaviors of social media users, which would help in predicting social media users' decisions when facing rumors; (3) designing and optimize a decision model of misinformed social media users during disasters with the theory of decision analysis.

**Acknowledgements** This research was partially supported by the United States National Science Foundation (NSF) under award numbers 1730503, 1760586, 1762807. This research was also partially supported by China Scholarship Council (CSC) to support Bairong Wang. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the NSF, or CSC. We also thank two referees for providing constructive comments.

## References

- Abbasi M-A, Liu H (2013) Measuring user credibility in social media. Springer, Berlin, pp 441–448
- Alexander DE (2014) Social media in disaster risk reduction and crisis management. *Sci Eng Eth* 20(3):717–733
- Berg BL, Lune H, Lune H (2004) Qualitative research methods for the social sciences, 5th edn. Pearson, Boston
- Blake ES, Kimberlain TB, Berg RJ, Cangialosi JP, Beven II, John L (2013a) Tropical cyclone report: Hurricane Sandy (AL182012), Technical report, National Hurricane Center, Miami, FL, USA, 22–29 October 2012
- Blake ES, Kimberlain TB, Berg RJ, John PC, Beven II, John L (2013b) Hurricane Sandy: October 22–29, 2012. Tropical Cyclone Rep
- Castillo C, Mendoza M, Poblete B (2011) Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, WWW '11, pp 675–684, Hyderabad, India. ACM
- Dang A, Smit M, Moh'd A, Minghim R, Milios E (2016) Toward understanding how users respond to rumours in social media. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 777–784
- Friggeri A, Adamic LA, Eckles D, Cheng J (2014) Rumor cascades. In: Proceedings of the 8th International Conference on Weblogs and Social Media, pp 101–110, Ann Arbor, Michigan
- Gayo-Avello EMM, Strohmaier HS, Metaxas PT, Gloor DP, Castillo C, Ma Mendoza, Poblete B (2013) Predicting information credibility in time-sensitive social media. *Internet Res* 23(5):560–588
- Gupta A, Lamba H, Kumaraguru P (2013) \$1.00 per RT # bostonmarathon # prayforboston: analyzing faking content on Twitter. In: eCrime Researchers Summit, pp 1–12, San Francisco. IEEE
- Harsin J (2006) The rumor bomb: a convergence theory of contemporary mediated American politics. Southern Review: Politics, Communication, Culture, Spring



- Hill K (2012) Hurricane Sandy, @comfortablysmug, and the flood of social media misinformation. [EB/OL], <http://www.forbes.com/sites/kashmirhill/2012/10/30/hurricane-sandy-and-the-flood-of-social-media-misinformation/#63cd322cd967>. Accessed 30 Oct 2012
- Hollywood Life Staff (2013) Martin Richard's death: How I desperately tried to save him. [EB/OL], <http://hollywoodlife.com/2013/04/19/martin-richard-death-firefighter-saves-boston-marathon-bombing/>. Accessed 19 April 2013
- Holt K (2012) How false Sandy news spread on Twitter. [EB/OL], <http://www.dailydot.com/news/false-sandy-reports-spread-twitter/>. Accessed 30 Oct 2012
- Inside Breaking News (2012) Three Sandy rumors that circulated on social media. [EB/OL], <http://blog.breakingnews.com/post/34652885735/three-sandy-rumors-that-circulated-on-social>. Accessed 30 Oct 2012
- Jong W, Dückers MLA (2016) Self-correcting mechanisms and echo-effects in social media: an analysis of the “gunman in the newsroom” crisis. *Comput Hum Behav* 59:334–341
- Kang B, Höllerer T, O'Donovan J (2015) Believe it or not? Analyzing information credibility in microblogs. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15, pp 611–616, Paris, ACM
- Knopf TA (1975) Rumors, Race, and Riots. Transaction Publishers, Abingdon
- Kotz D (2013) Injury toll from Marathon bombs reduced to 264. [EB/OL], <http://www.bostonglobe.com/lifestyle/health-wellness/2013/04/23/number-injured-marathon-bombing-revised-downward/NRpaZ5mmvGquP7KMA6XsIK/story.html>. Accessed 24 April 2013
- Leberecht T (2010) Twitter grows up in aftermath of Haiti earthquake. [EB/OL], <https://www.cnet.com/news/twitter-grows-up-in-aftermath-of-haiti-earthquake/>. Accessed 19 Jan 2010
- Levine TR, Park HS, McCormack SA (1999) Accuracy in detecting truths and lies: documenting the “veracity effect”. *Commun Monogr* 66(2):125–144
- Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction continued influence and successful debiasing. *Psychol Sci Public Interest* 13(3):106–131
- Liu X, Nourbakhsh A, Li Q, Fang R, Shah S (2015) Real-time rumor debunking on Twitter. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM '15, pp 1867–1870, Melbourne, Australia. ACM
- Mohaimin SA, Ukkusuri SV, Hugh G (2017) The role of social networks and information sources on hurricane evacuation decision making. *Nat Hazards Rev*. [https://doi.org/10.1061/\(ASCE\)NH.1527-6996.0000244](https://doi.org/10.1061/(ASCE)NH.1527-6996.0000244)
- Montopoli B (2013) Eight-year-old Martin Richard killed in Boston bombings. [EB/OL], <http://www.cbsnews.com/news/eight-year-old-martin-richard-killed-in-boston-bombings/>. Accessed 16 April 2013
- Morris MR, Counts S, Roseway A, Hoff A, Schwarz J (2012) Tweeting is believing?: understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12, pp 441–450, Seattle, Washington. ACM
- Neuendorf KA (2002) The content analysis guidebook. Sage, Thousand Oaks
- Oh O, Agrawal M, Rao HR (2011) Information control and terrorism: tracking the Mumbai terrorist attack through Twitter. *Inf Syst Front* 13(1):33–43
- Oremus W (2012) Dear Twitter, don't believe everything you hear on a police scanner. [EB/OL], [http://www.slate.com/blogs/future\\_tense/2012/10/30/false\\_hurricane\\_sandy\\_rumors\\_police\\_scanner\\_fools\\_twitter\\_into\\_spreading.html](http://www.slate.com/blogs/future_tense/2012/10/30/false_hurricane_sandy_rumors_police_scanner_fools_twitter_into_spreading.html). Accessed 30 Oct 2012
- Ozturk P, Li H, Sakamoto Y (2015) Combating rumor spread on social media: the effectiveness of refutation and warning. In: 2015 48th Hawaii International Conference on System Sciences (HICSS), pp 2406–2414, Hawaii, USA
- Patton MQ (1990) Qualitative evaluation and research methods, 2nd edn. SAGE, Newbury Park
- Richards J, Lewis P (2011) How Twitter was used to spread—and knock down—rumours during the riots. [EB/OL], <https://www.theguardian.com/uk/2011/dec/07/how-twitter-spread-rumours-riots>. Accessed 07 Dec 2011
- Rosnow RL (1991) Inside rumor: a personal journey. *Am Psychol* 46(5):484
- Rosnow RL, Fine GA (1976) Rumor and gossip: the social psychology of hearsay. Elsevier, Amsterdam
- Rubin VL (2017) Deception detection and rumor debunking for social media. In: Sloan L, Quan-Haase A (eds) The SAGE handbook of social media research methods. Sage, Beverly Hills
- Sager J (2013) 10 Boston Marathon bombing rumors that need to be stopped immediately. [EB/OL], [http://thetir.cafemom.com/crime/154242/10\\_boston\\_marathon\\_bombing\\_rumors](http://thetir.cafemom.com/crime/154242/10_boston_marathon_bombing_rumors). Accessed 17 April 2013
- Shibutani T (1969) Improvised news: a sociological study of rumor. *Soc Res* 36(1)
- Starbird K, Maddock J, Orand M, Achterman P, Mason RM (2014) Rumors, false flags, and digital vigilantes: misinformation on Twitter after the 2013 Boston Marathon bombing. In: iConference 2014 Proceedings, pp 654–662, Berlin, Germany. iSchools

- Tripathy RM, Bagchi A, Mehta S (2013) Towards combating rumors in social networks: models and metrics. *Intell Data Anal* 17(1):149–175
- Tripathy RM, Bagchi A, Mehta S (2010) A study of rumor control strategies on social networks. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pp 1817–1820, Toronto, ON, Canada. ACM
- Twitter Developer Documentation. REST APIs. [EB/OL]. <https://dev.twitter.com/rest/public>
- Wang B, Zhuang J (2017) Crisis information distribution on Twitter: a content analysis of tweets during hurricane sandy. *Nat Hazards* 89(1):161–181
- Wemple E (2012) Hurricane Sandy: NYSE not flooded! [EB/OL]. [https://www.washingtonpost.com/blogs/erik-wemple/post/hurricane-sandy-nyse-not-flooded/2012/10/30/37532512-223d-11e2-ac85-e669876c6a24\\_blog.html?utm\\_term=.c6ec07080562](https://www.washingtonpost.com/blogs/erik-wemple/post/hurricane-sandy-nyse-not-flooded/2012/10/30/37532512-223d-11e2-ac85-e669876c6a24_blog.html?utm_term=.c6ec07080562). Accessed 30 OCT 2012
- Zhao L, Yin J, Song Y (2016) An exploration of rumor combating behavior on social media in the context of social crises. *Comput Hum Behav* 58:25–36
- Zubiaga A, Liakata M, Procter R, Hoi GWS, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. *Plos One* 11(3):e0150989