

Harnessing Data to Create an Effective Drought Management System

Aibek Musaev

The University of Alabama, Tuscaloosa, USA

Kimberly Stowers

The University of Alabama, Tuscaloosa, USA

Jonghun Kam

The University of Alabama, Tuscaloosa, USA

ABSTRACT

Drought is a billion-dollar weather and climate disaster. It is a complex climate phenomenon that slowly emerges and quickly vanishes, making it hard for local stakeholders and the public to respond in a timely manner. To reduce the vulnerability of our society to future drought, a better understanding of how society responds to drought is critical. Here, we propose a pilot study about social response to a recent California drought through social media. In this study, we identify the most important users in the collected dataset using an extension of PageRank, the algorithm developed for Google Search. We investigate the key drivers of the peak of the Twitter activities in February, 2014, during the California drought. We also create a word cloud from the top 2,000 most retweeted tweets. Lastly, we specify the information sources from those tweets. The findings of this study show that big data can help us to improve current drought response plans through fundamental understanding of social response to drought.

Keywords

Droughts, social response, PageRank.

INTRODUCTION

Drought is a billion-dollar weather and climate disaster. In 2015, the California drought cost the U.S. 2.7 billion dollars and nearly 21,000 jobs (Howitt et al. 2015). In 2016, droughts across the continental U.S. cost 3.6 billion dollars because of the onset of a new drought in the eastern U.S. How can we give society the foresight and wisdom to respond to drought risk before they lead to multi-billion-dollar losses?

Drought is a naturally occurring, but rare, climate phenomenon lasting up to several years through the stages of initiation, onset, and recovery. Drought is initiated by a lack of precipitation over an extended period, is deteriorated by high water demand and persistent water supply deficit during the onset, and is terminated by an extremely intense precipitation event or a series of precipitation events. Because of the slow emergence and quick disappearance of drought, it is hard for local stakeholders and the public to respond in a timely manner (Switzer and Vedlitz 2017). The drought research community has believed that due to its persistency, drought allows society to plan and develop mitigation strategies ahead of time as long as citizens are given actionable information from a reliable drought monitoring and forecasting system. However, despite more reliable drought monitoring and forecasting systems, we have still faced multi-billion-dollar losses from recent droughts. Drought monitoring systems are limited in elevating the level of drought awareness and ultimately managing the social vulnerability to drought. To reduce significantly the vulnerability of our society to future drought, current research must shift to an examination of how society responds to drought. It is critical to stimulate timely social response to drought risk in order to mitigate its adverse effects on our society.

Social behavior and dynamics during the onset of drought have been poorly studied for several reasons. First, it requires an interdisciplinary collaboration approach; and second, there has been a lack of available data for social response. The hydroclimate research community can provide data regarding the stage and severity of drought. The social psychology research community can provide insight to social response to drought through theoretical models examining human reactions to disaster. According to the social-cognitive preparation model (Paton 2003), it has been

suggested that information from behavioral interventions for intention formation can inform education programs designed to facilitate environmental awareness and disaster preparedness in people (Ardalan et al. 2013; Ejeta et al. 2015). However, in order to maximize such facilitation, it is imperative to understand how risk communication and warnings are interpreted by the public prior to intention formation (Paton 2003).

The computer science research community can extract meaningful information for social behavior via big data and machine-learning techniques. In recent decades, significant advancements in these areas allow us to validate and investigate the governing mechanisms of social response to specific events. Integrating state-of-art knowledge and techniques across the disciplines of hydroclimatology, social psychology, and computer science is a key solution to tackle the challenge problem of drought mitigation.

Here, we demonstrate a pilot study about the above-mentioned problem in which we analyze social response to droughts. We downloaded Twitter data related to droughts in California using the date of range from 2010 to 2017. We chose California as it experienced multiple droughts during this period (Howitt et al. 2015). Our overarching research question was “What is a drought related lexicon to which the public is most responsive?” An effective drought management system must utilize such lexicon that makes social response to drought more timely.

In this study, we focused on a subset of data related to the period of the highest public activity on the recent California drought (2011 - 2014). The objectives of this study were to 1) to determine the most popular tweets during this period and the most important terms used in those tweets; and 2) to analyze the most important Twitter users in the collected dataset. Since this was a pilot study, we focused more on developing the methodology as a proof of concept, instead of finding the actual answers to our research question.

The rest of the paper is organized as follows. The next section discusses related work. The methodology section describes the process of creating a lexicon of important terms and how to rank Twitter users based on their importance. The experimental evaluation section presents results using real-world data. The final section concludes the paper and discusses potential implications of this work.

RELATED WORK

Research has shown that the use of Twitter for disseminating and receiving information during disasters such as floods and wildfires has grown over the years (Sutton et al. 2008). In particular, the use of Twitter changes during disaster events to focus on information broadcasting (Hughes and Palen 2009). However, it remains unclear how the use of Twitter changes as a function of silent, slow-growing disasters such as droughts. The general public appears to recognize the threat and need for information concerning floods, or even fires caused by droughts (Glaser 2007; Stelter 2017). But do they also recognize the threat and need for information concerning droughts themselves?

In general, it is known that several factors converge to influence human behavior in the face of impending disasters. According to the social-cognitive preparation model by Douglas Paton (Paton 2003), these factors can be organized into precursors, the formation of intention, and linking intentions to preparedness. It has been suggested that information from behavioral interventions for intention formation can inform education programs designed to facilitate environmental awareness and disaster preparedness in people (Ardalan et al. 2013; Ejeta et al. 2015). However, in order to maximize such facilitation, it is imperative to understand how risk communication and warnings are interpreted by the public prior to intention formation (Paton 2003). We argue that a trend analysis of human communication surrounding environmental hazards and disaster events can fill this gap and offer insights to improve risk communication and pre-disaster planning. Specifically, such a trend analysis should include an examination of social media conversations to uncover the following precursors to drought response, according to Paton’s model: 1) critical awareness of hazards, 2) risk perception, and 3) hazard anxiety.

The computer science research community has made significant advancements in the areas of big data and machine learning, which can allow us to investigate governing mechanisms for social response to specific events. Social media constitute, perhaps, the biggest source of unstructured big data, especially when their use rises during disasters (Imran et al. 2015). The huge success of deep learning in a broad area of applications, including natural language processing, has enabled the analysis of social media data at an unprecedented scale and with high accuracy (Tang et al. 2014; Severyn and Moschitti 2015).

The topic of measuring user influence in social networks has generated a lot of interest (Kwak et al. 2010; Bakshy et al. 2011). An in-depth comparison of three measures of influence has been presented in (Cha et al. 2010) and it was found that the top Twitter users had a disproportionate amount of influence. However, it was also noted that social users have varying amounts of influence in different subjects (Pal and Counts 2011). In addition, existing influence measurement techniques are vulnerable to sybil users, hence sybil-resilient systems have been introduced (Zhang et al. 2016).

Recently, Weng et al. (2010) proposed an extension of the PageRank algorithm called TwitterRank to measure the influence of users in Twitter. TwitterRank measures the influence taking both the topical similarity between users and the link structure into account. The topical similarity is computed using Latent Dirichlet Allocation (LDA) model, which is used to learn a set of latent topics. Each topic in LDA is a multinomial distribution over words. The link structure is based on "following" relationship. We also measure the influence of users in Twitter using an extension of the PageRank algorithm, but the link structure is based on the "reply" relationship. We also deal with a specific topic, "California drought", shared by all users in our dataset.

METHODOLOGY

Towards building a lexicon of important terms

An effective drought management system should not only implement a reliable drought monitoring and forecasting component, but it should also stimulate timely social response. In this study, we determined a lexicon of the most important terms related to droughts based on the public's activity. To achieve this goal, a dataset of tweets related to droughts is needed. Thus, we collected a dataset of tweets discussing California drought during a wide range of dates spanning multiple years.

Given the dataset of relevant tweets, we can compute the most popular tweets based on public's opinion. We used the number of retweets as a measure of tweet's popularity. A tweet can be forwarded to all of the user's followers, in which case it is known as a retweet. Retweets are used to pass along information on Twitter. Retweets indicate the popularity and network centrality of the tweet writer as opposed to the content of the tweet itself (Kwak et al. 2010), leading to potential bias towards messages posted by popular users.

Given a list of the most retweeted tweets, we can determine the most important terms in those tweets. There are different approaches to this problem, for example a frequently used TF-IDF scheme, short for term frequency-inverse document frequency (Salton et al. 1983). This approach builds a term-by-document matrix X whose columns contain the TF-IDF values for each of the documents in the corpus. However, this approach may not be applicable to Twitter as tweets (the documents in our corpus) can only have up to 140 characters.

An alternative approach to analyze texts is a word cloud. Word clouds generated for a body of text can serve as a starting point for a deeper analysis (Heimerl et al. 2014). Thus, we used a word cloud to distill the most popular tweets down to terms that appear with highest frequency. This also served as a visualization method for text.

Ranking users by importance

In this study, we identified the most important users in the collected dataset. Note that tweets do not exist in a vacuum. They are posted by real people who have a wide range of influence and activity in a given topic. We want to find the users whose posts on droughts attract the most amount of attention.

Our dataset consisted of tweets containing "California" and "drought" keywords. Each tweet also contains information whether it is an original tweet or a reply to another tweet. We built a graph based on users and their "reply" relationships.

Given the community of users who post tweets containing the mentioned keywords, we built a graph where such users serve as nodes. The links between nodes are represented by replies, such that the direction of a link is a directed edge from the responder to the original poster. Given the directed graph, we applied the PageRank algorithm (Page et al. 1999) to compute the importance of users in this community.

PageRank is an algorithm used by Google Search to rank websites in their search results. It is a way of measuring the importance of website pages by counting the number of incoming links to those pages to determine a rough estimate of their importance. In our study we treated users as nodes and used the number of replies to their messages as a measure of each user's importance. The more replies a user had, the higher importance score they received. Figure 5 shows the visualization of the proposed user rankings approach in the Experimental Evaluation section.

EXPERIMENTAL WORK

Data collection

We conducted a set of experiments using real-world data collected from Twitter. In these experiments, we focused on the 2011-14 California drought (Howitt et al. 2015). Droughts in California can have a severe economic impact since it is the most populous state in the United States and one of the major agricultural producers.

Table 1. Dataset Overview

	2010	2011	2012	2013	2014	2015	2016	2017	Total
Tweets	1,524	2,277	2,460	3,321	140,249	129,034	23,844	11,786	314,495
Users	1,085	1,618	2,014	2,465	58,592	64,366	14,137	7,835	128,945

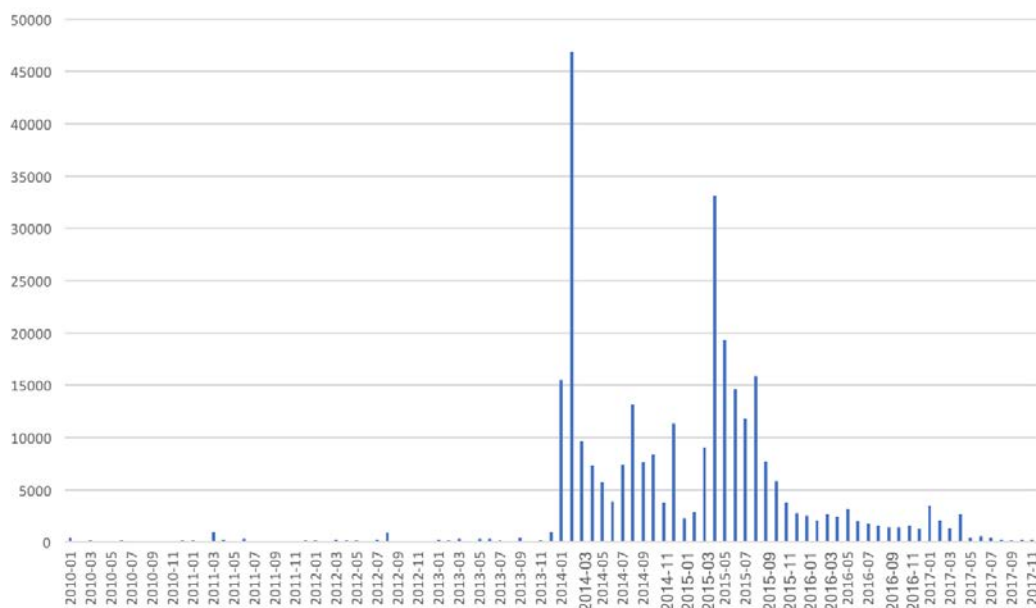
For data collection, we used Twitter’s advanced search feature¹. It allows a logged-in user to specify multiple options for refined search, including specific date ranges. We used “California” and “drought” as search keywords and provided a range of dates from January, 2010, to November, 2017, as droughts can last up to several years. We exported the search results as HTML pages. HTML pages were then parsed, such that tweet IDs were retrieved among other information. Note that Twitter’s advanced search returns only a subset of supported attributes for each tweet. Hence, to collect full information for downloaded tweets, we used Twitter’s statuses/lookup API² that returns detailed information for up to 100 tweets per request.

In total, there were about 310k tweets by 130k unique Twitter users during this period (Table 1). The tweets were not evenly distributed.

We released the collected dataset as a contribution to research community³. It is the first published dataset that contains Twitter data covering a 7 year period and dedicated to a comprehensive coverage of a particular event type. The dataset is provided as a listing of tweet IDs in accordance with the Twitter Developer Agreement and Policy⁴. The actual tweet contents can be retrieved using Twitter’s statuses/lookup API.

Analysis of Twitter activity on California drought

We aggregated the tweets in our dataset by each month from 2010 to 2017 (Figure 1).

**Figure 1. Monthly Twitter activity on droughts from 2010 to 2017**

Results show that there are multiple peaks in this chart, including in years 2014 and 2015. For example, the largest peak occurred in February, 2014. To understand the key drivers of this peak, we analyzed Twitter activity during the months before and after the peak at a higher temporal resolution (the daily time step).

Figure 2 shows daily time series of Twitter activity from January through March of 2014, which represents a 30-time higher granular resolution of Twitter activity on droughts compared to the monthly resolution shown in Figure 1.

¹<https://twitter.com/search-advanced>

²<https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup>

³http://aibek.cs.ua.edu/files/droughts_ids.txt

⁴<https://developer.twitter.com/en/developer-terms/agreement-and-policy>

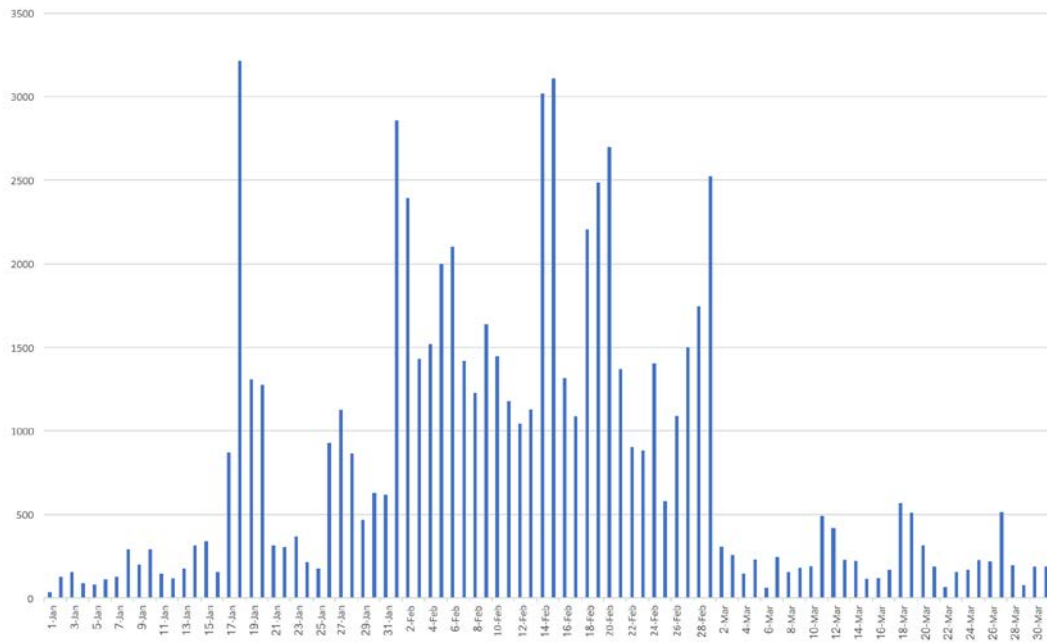


Figure 2. Daily Twitter activity on droughts from January to March in 2014

By February 1, 2014, California officials announced that they would not send any water from the state’s reservoir system to local agencies amid severe drought conditions⁵. Figure 3 shows one of the most retweeted tweets that expresses a sentiment shared by many during this period. This tweet is posted by a regular user, which is why we cover it with an anonymous label to preserve her privacy.



Figure 3. The most retweeted tweet on California droughts

Table 2 shows the text of the top 10 most retweeted tweets with the number of retweets and posted date. Note that we preserve the original spelling and punctuation. These results provide us key words for the lexicon used by the most popular social media messages.

⁵https://www.washingtonpost.com/national/health-science/amid-drought-california-agency-wont-allot-water/2014/02/01/7477aa5c-8b88-11e3-916e-e01534b1e132_story.html

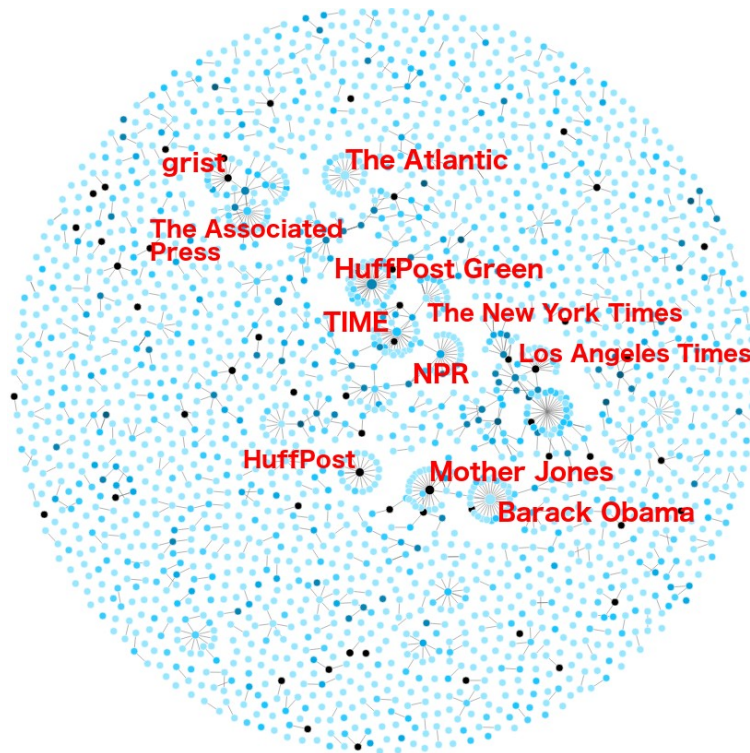


Figure 5. User ranking based on reply relationships

For visualization purposes, we have added labels to the users with the highest ranking scores. The majority of those users are news agencies, such as TIME, The Associated Press, and Los Angeles Times. There are also nonprofit online magazines, such as Mother Jones⁷ and grist⁸. However, no federal or state agency dealing with droughts has a near the top ranking score. Twitter accounts among the top users include Barack Obama's account⁹.

CONCLUSION AND FUTURE WORK

The results of this study show a new insight for how the public talks about drought via social media. Beyond that, we are also able to see triggers of public discussion during a recent California drought. While there was a spike in January, the entire month of February was riddled with high Twitter usage regarding drought. However, this change in use was caused less by the drought itself and more by a temporary change in policy pursued by the California government. After California officials announced they would be halting the transfer of water from state to local agencies, many Twitter users took to social media to discuss this.

This study finds that publicized decisions by the California government influence significantly Twitter activity about droughts, which inform of many potential implications for critical awareness and risk perception of the public during droughts. While the general public may be aware of droughts as they occur, it is possible such awareness increases dramatically as a result of high-profile discussions. This is supported by research that shows that mass communication can be used to alter public behavior (Abroms and Maibach 2008). Furthermore, state government communication strategies may be bolstered by several other variables on social media, such as the presence of high profile figures or news agencies in the discussion, and the presence of highly compelling posts by everyday users. Research has shown that there is a relationship between retweet patterns and the content of tweets themselves (Nagarajan et al. 2010). This study shows an example of a tweet with a powerful image seemingly became popular overnight. More research into this phenomenon could help us create/design specific strategies for adjusting critical awareness and risk perception during droughts.

These findings have several potential implications that need further exploration. It seems that news agencies, rather than government agencies, remain the go-to source for most individuals obtaining information regarding drought-related topics. However, it remains unclear why this is. First, individuals may not realize they can follow various government agencies such as NOAA on Twitter. Additionally, they may not find reports released by such

⁷<https://twitter.com/MotherJones/>

⁸<https://twitter.com/grist>

⁹<https://twitter.com/BarackObama>

agencies very approachable or interesting. Finally, there is the possibility that individuals reading about drought from news agencies simply stumble upon the information, and are not actively seeking it in the first place. They are instead taking a passive approach to information collection, and then responding to that information when they happen to notice it.

More research still needs to be done to better understand the sensitivity of social response to drought. This can be done via interdisciplinary research collaboration across computer science, hydrologic engineering, and social psychology. To that end, we will apply state-of-the-art techniques in computer science to examine big social data, and test the validity of the known theories for social disaster response. This will include the development of an algorithm to estimate society's levels of hazard anxiety, hazard awareness, and risk perception in relation to drought by applying a new model, fastText (Joulin et al. 2016). This algorithm will not only aid in understanding of social response to drought, but will allow us to pinpoint ways to help improve this social response.

REFERENCES

- Abroms, Lorien C and Maibach, Edward W (2008). "The effectiveness of mass communication to change public behavior". In: *Annu. Rev. Public Health* 29, pp. 219–234.
- Ardalan, Ali, Mowafi, Hani, Ardakani, Hossein Malekafzali, Abolhasanai, Farid, Zanganeh, Ali-Mohammad, Safizadeh, Hossein, Salari, Sirous, and Zonoobi, Vahid (2013). "Effectiveness of a primary health care program on urban and rural community disaster preparedness, Islamic Republic of Iran: A community intervention trial". In: *Disaster medicine and public health preparedness* 7.5, pp. 481–490.
- Bakshy, Eytan, Hofman, Jake M, Mason, Winter A, and Watts, Duncan J (2011). "Everyone's an influencer: quantifying influence on twitter". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 65–74.
- Bostock, Michael, Ogievetsky, Vadim, and Heer, Jeffrey (2011). "D³ data-driven documents". In: *IEEE transactions on visualization and computer graphics* 17.12, pp. 2301–2309.
- Cha, Meeyoung, Haddadi, Hamed, Benevenuto, Fabricio, and Gummadi, P Krishna (2010). "Measuring user influence in twitter: The million follower fallacy." In: *Icwsn* 10.10-17, p. 30.
- Ejeta, Luche Tadesse, Ardalan, Ali, and Paton, Douglas (2015). "Application of behavioral theories to disaster and emergency health preparedness: A systematic review". In: *PLoS currents* 7.
- Glaser, Mark (2007). "California wildfire coverage by local media, blogs, Twitter, maps and more". In: *PBS MediaShift*.
- Heimerl, Florian, Lohmann, Steffen, Lange, Simon, and Ertl, Thomas (2014). "Word cloud explorer: Text analytics based on word clouds". In: *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. IEEE, pp. 1833–1842.
- Howitt, Richard E., MacEwan, Duncan, Medellín-Azuara, Josué, Lund, Jay R., and Sumner, Daniel A. (2015). *Economic Analysis of the 2015 Drought for California Agriculture*. Tech. rep. Center for Watershed Sciences, University of California – Davis, Davis, CA.
- Hughes, Amanda Lee and Palen, Leysia (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International Journal of Emergency Management* 6.3-4, pp. 248–260.
- Imran, Muhammad, Castillo, Carlos, Diaz, Fernando, and Vieweg, Sarah (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Computing Surveys (CSUR)* 47.4, p. 67.
- Joulin, Armand, Grave, Edouard, Bojanowski, Piotr, and Mikolov, Tomas (2016). "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759*.
- Kwak, Haewoon, Lee, Changhyun, Park, Hosung, and Moon, Sue (2010). "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 591–600.
- Nagarajan, Meenakshi, Purohit, Hemant, and Sheth, Amit P (2010). "A qualitative examination of topical tweet and retweet practices." In: *ICWSM* 2.010, pp. 295–8.
- Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry (1999). *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford InfoLab.
- Pal, Aditya and Counts, Scott (2011). "Identifying topical authorities in microblogs". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 45–54.
- Paton, Douglas (2003). "Disaster preparedness: a social-cognitive perspective". In: *Disaster Prevention and Management: An International Journal* 12.3, pp. 210–216.

- Salton, Gerard, Fox, Edward A, and Wu, Harry (1983). "Extended Boolean information retrieval". In: *Communications of the ACM* 26.11, pp. 1022–1036.
- Severyn, Aliaksei and Moschitti, Alessandro (2015). "Twitter sentiment analysis with deep convolutional neural networks". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 959–962.
- Stelter, Brian (2017). "How social media is helping Houston deal with Harvey floods". In: *CNN Money*.
- Sutton, J, Palen, L, and Shklovski, I (2008). "Backchannels on the front lines: Emergent use of social media in the 2007 Southern California fire. proceedings of Information Systems for Crisis Response and Management Conference (ISCRAM), Washington DC". In: *Washington DC*.
- Switzer, David and Vedlitz, Arnold (2017). "Investigating the Determinants and Effects of Local Drought Awareness". In: *Weather, Climate, and Society* 9.4, pp. 641–657.
- Tang, Duyu, Wei, Furu, Yang, Nan, Zhou, Ming, Liu, Ting, and Qin, Bing (2014). "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." In: *ACL (1)*, pp. 1555–1565.
- Weng, Jianshu, Lim, Ee-Peng, Jiang, Jing, and He, Qi (2010). "TwitterRank: Finding Topic-sensitive Influential Twitterers". In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 261–270.
- Zhang, Jinxue, Zhang, Rui, Sun, Jingchao, Zhang, Yanchao, and Zhang, Chi (2016). "Truetop: A sybil-resilient system for user influence measurement on twitter". In: *IEEE/ACM Transactions on Networking* 24.5, pp. 2834–2846.