A System Analytics Framework for Detecting Infrastructure-Related Topics in Disasters Using Social Sensing

Chao Fan, Ali Mostafavi(✉), Aayush Gupta, and Cheng Zhang

Texas A&M University, College Station, TX 77840, USA
{chfan,ayu2224,czhang}@tamu.edu, amostafavi@civil.tamu.edu

Abstract. The objective of this paper is to propose and test a system analytics framework based on social sensing and text mining to detect topic evolution associated with the performance of infrastructure systems in disasters. Social media, like Twitter, as active channels of communication and information dissemination, provide insights into real-time information and first-hand experience from affected areas in mass emergencies. While the existing studies show the importance of social sensing in improving situational awareness and emergency response in disasters, the use of social sensing for detection and analysis of infrastructure systems and their resilience performance has been rather limited. This limitation is due to the lack of frameworks to model the events and topics (e.g., grid interruption and road closure) evolution associated with infrastructure systems (e.g., power, highway, airport, and oil) in times of disasters. The proposed framework detects infrastructure-related topics of the tweets posted in disasters and their evolutions by integrating searching relevant keywords, text lemmatization, Part-of-Speech (POS) tagging, TF-IDF vectorization, topic modeling by using Latent Dirichlet Allocation (LDA), and K-Means clustering. The application of the proposed framework was demonstrated in a study of infrastructure systems in Houston during Hurricane Harvey. In this case study, more than sixty thousand tweets were retrieved from 150-mile radius in Houston over 39 days. The analysis of topic detection and evolution from user-generated data were conducted, and the clusters of tweets pertaining to certain topics were mapped in networks over time. The results show that the proposed framework enables to summarize topics and track the movement of situations in different disaster phases. The analytics elements of the proposed framework can improve the recognition of infrastructure performance through text-based representation and provide evidence for decision-makers to take actionable measurements.

Keywords: System analytics framework · Social sensing
Infrastructure-related topics · Disaster resilience · Text mining

1 Introduction

Infrastructure systems such as water networks, highways, and power grid are critical components to human lives and community functions [1]. The performance of

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10864, pp. 74–91, 2018.
https://doi.org/10.1007/978-3-319-91638-5_4

infrastructure systems can affect other systems due to their interdependencies [2, 3]. Timely mapping of infrastructure disruptions and damages is essential for response and restoration of infrastructure services. However, the limited resources and accessibility inhibit situational awareness and effective detection of infrastructure disruptions. Social sensing and crowdsourced data collection (such as social media), processes in which social posts deliver the users' observations and emotions regarding their physical environment [4], have been shown to be potentially useful in improving the situational awareness of agencies involved in disaster response. The advantages of using social sensing data are twofold: (1) sheer volume of messages and users; and (2) high velocity of generating posts. In existing studies, for example, Olteanu et al. [5] showed that more than 2.7 million tweets relative to hurricane Sandy were generated in three days, while 3.3 million tweets relative to Boston Bombings were generated in five days. Lu and Brelsford [6] collected 14.2 million messages in the 2011 Japanese Earthquake and Tsunami, where over 2000 tweets per second were generated following this disaster [7]. As such, social media impart a great opportunity for obtaining important information in disasters.

In order to use social sensing in assessing infrastructure service disruptions and resilience, there are two technical issues that should be addressed: event detection and tracking. Infrastructure-related events exhibit the severity and duration of failures. Detecting those events is essential for disaster responders to recognize the situation and distribute the resources for relief. On the other hand, the situations associated with infrastructure systems were changing over time since the intervention from humans or other interdependent systems. Tracking the evolution of the events is necessary for evaluating the performance of infrastructure and making response decisions. Some research [8–10] involving text mining approaches in detecting and tracking domain-specific events has been conducted on social media. For example, Ashktorab et al. [11] classified tweets, extracted tokens, and phrases that reported damages and casualties based on predefined terms by using a Twitter-mining tool, Tweedr. Yin et al. [12] investigated natural language processing and data mining techniques to conduct some burst detection, tweet classification, and geotagging by employing empirical search terms. Bala et al. [13] applied regression analysis to text data for finding the course of the disaster by counting the frequencies of sample words. Tien et al. [1] detected damages and failures in transportation and energy systems by using given search terms. However, all outputs in the literature were binary: can or cannot be detected, since the events were predefined by keywords. Another stream of research concentrates on the tentative analysis of filtered relevant social data. For example, Olteanu et al. [5] employed crowdsourcing and supervised learning to create a lexicon of crisis-related terms. Prasetyo et al. [14] conducted content analysis, emotion analysis, activity analysis and network analysis on social posts that were retrieved by selected keywords in the case of Singapore Haze. However, the existing studies still cannot be applied to understanding the extent of detecting and tracking infrastructure-related events on social media.

To address this gap, this study proposes a framework integrating social sensing and text mining for building a detailed text-based representation of infrastructure-related events and tracking the evolution of those major events. It enables to precisely filter infrastructure-related social posts, classify tweets into different clusters, summarize

relevant events, and analyze the changes of these events during the development of disasters. This paper is organized into two parts: system analytics framework and a case study. In section two, the architecture of our novel framework including innovative algorithms is presented and explained. In section three, a case study of hurricane Harvey in Houston area is conducted, relevant events are summarized in four disaster phases (e.g., before Harvey landed, hurricane period, flooding period, and after flooding receded), and the changes of the events over time are explored. Section four presents the limitation of current results and discusses the future work that is potentially capable of improving the outputs. Section five concludes the study in this paper and the contributions of this work towards situational awareness of infrastructure disruptions.

2 System Analytics Framework

The proposed system analytics framework supports a series of social sensing methods and text-mining approaches, such as text classification, vectorization, topic modeling, data clustering, and text-based representation. The overall framework of this study is shown in Fig. 1. This framework includes four components: (1) context recognition; (2) data collection and preprocessing; (3) text classification; (4) topic modeling and representation. First, context recognition is to learn basic information about disasters (e.g., spatial distribution, severity, measures, and prominent events) for figuring out the inputs (e.g., geolocations and keywords) to the algorithms. The novelties of this framework are two important iterations which contribute to: (1) stop words and keywords updates; (2) dynamic changes of the number of classes in text classification. The update of keywords in the first iteration is to obtain a comprehensive list of keywords for retrieving relevant social posts, while the update of stop words is to delete some common words that also appear in other irrelevant documents for reducing the size of datasets and improving the accuracy and efficiency of following procedures. The process of updates is manual but adopts POS (Part-of-Speech) tagging to filter out new relevant keywords and stop words. The iteration of text classification is to determine the number of major topics in each dataset without predefined numbers. Meanwhile, text classification integrates text-mining elements such as TF-IDF (term frequency-inverse document frequency) vectorization, K-means clustering, LDA (Latent Dirichlet Allocation) topic modeling for precisely extracting specific topics. These two iterations contributed to making the algorithm be more adaptive to the evolving streaming data than previous studies. The merits of these iterations will be validated by their applications. The last component, topic representation, is to make a detailed summary of each detected topic so that the results can be used directly by decision makers for developing strategies. In the following subsections, we present the functions and merits of each component.

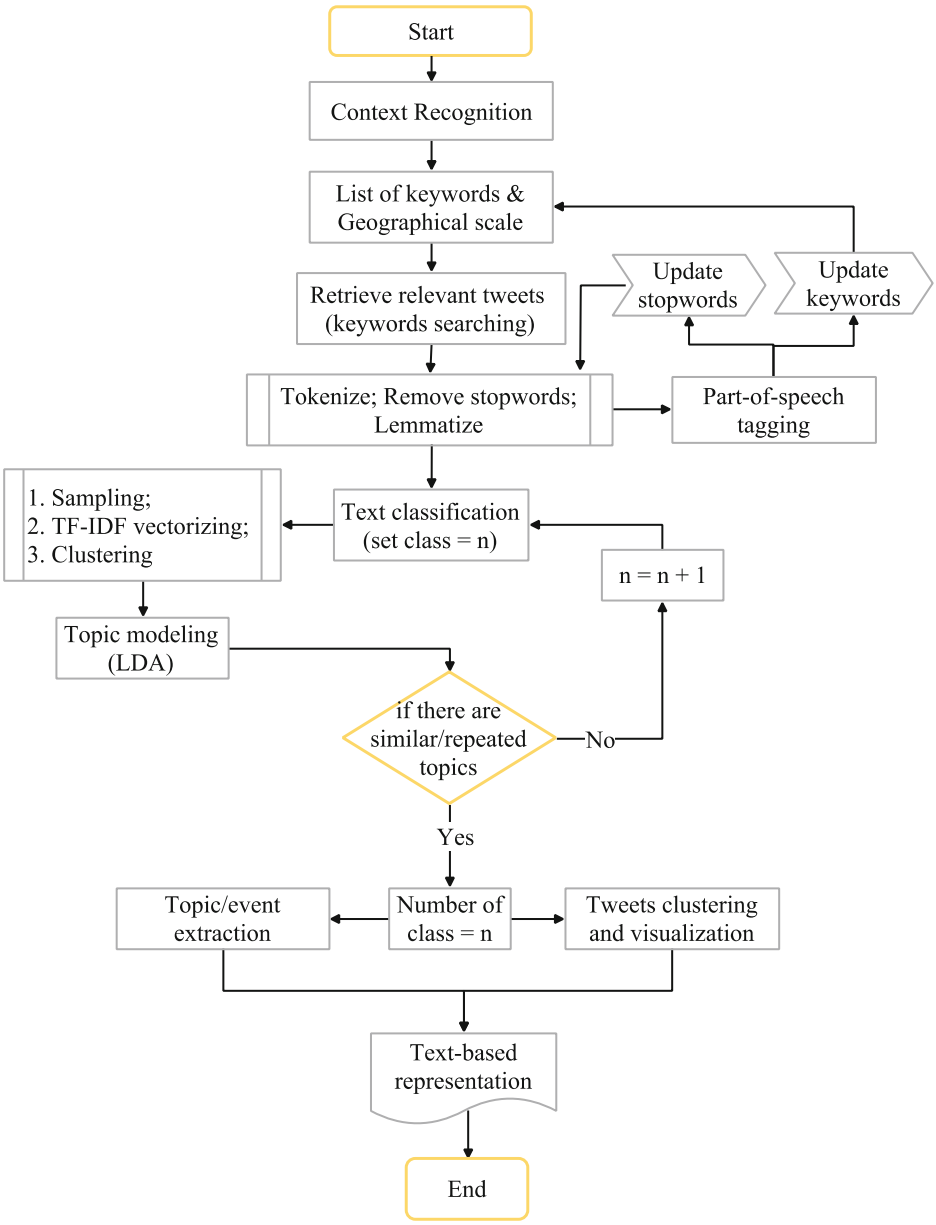


Fig. 1. The architecture of the system analytics framework.

2.1 Context Recognition

In the first component, context recognition and awareness of users are fundamental requirements of conducting social sensing and text mining, because the context of disasters provides the basis for data collection and processing. As the architecture of this framework shows, data retrieval includes keyword-based and geolocation-based retrieval on social media. The understanding of context is helpful to preliminarily identify relevant keywords and define the scale of affected areas for data collection. Further, defining phases based on the context is essential for dividing the life cycle of disasters into several time periods, in order to analyze the changes of detected events over time. In the proposed framework, disaster phases are defined by the dominant events and

threats so that the derived damages and disruption of infrastructure systems by those dominant events can be summarized and assessed. In the end, the extracted topics in each disaster phase will be validated by the context. As such, context recognition and disaster phases definition are significant for users to conduct this framework. For details of context recognition, there are a vast number of sources (e.g., news articles, web portals, governmental reports, photos as well as interviews) relative to disasters can be accessed. Thus, deep understanding of the development of disasters is easy to be achieved.

2.2 Data Collection and Preprocessing

After getting to know the context of the disasters, we can determine a list of keywords describing the disaster and geolocation of affected areas for identifying and gathering relevant tweets from Twitter. Due to the rate limitation of Twitter's public API (i.e., the constraints of login and scraping data), both keyword-based and location-based collection are commonly used in the current social sensing literature [15]. Keywords (e.g., disruption, damage, sad, emergency, and survivor) are constructed by human reaction to the effects of disasters and are representative of the sentences in messages. In this framework, the list of keywords is the combination of two parts: infrastructure-related keywords and disaster event-related keywords, for collecting domain-specific data. These search terms contribute to obtaining precise topics relative to infrastructure performance in the following components of this framework. However, keyword-based collection always leads to a lot of noise in datasets since similar keywords or events may appear in other areas around the world. Thus, a combination of the location-based and keyword-based collection is particularly helpful for reducing noise in our datasets and improving the accuracy of collected data in our specific research domain. Then, the lexicon and geographical scale can be applied to retrieve microblogged communications and messages.

The next step in processing social data before text classification is data cleaning to get focused content which includes tokenization, removing stop words as well as lemmatization. Tokenization is a process of splitting the text into meaningful tokens [16]. As the microblogged communications are always no more than 140 words, so the complexity of tokenization is much lower than dealing with fictions and scripts. However, because the users are used to using a lot of symbols and punctuations (e.g., "(: - :)", "???", and "\^ - ^/") to express their emotions and feelings, it is necessary to remove these symbols and punctuations by employing regular expressions. After getting the clean datasets, lemmatization including multiple rules of converting all tokens to their root words is utilized. For example, "affect" and "affects" have the same meaning in out topics. Thus, "affects" should be converted to "affect" so that these tokens can be filtered into the same category. It is essential for updating keywords, stop words as well as text classification since almost every derived word can be identified. Furthermore, in order to get more representative content, removing the words that occur commonly across all the other documents, called stop words, are used as a preprocess in natural language processing (NLP). In our infrastructure-specific research domain, the stop words and relevant keywords are not very discriminative and cannot be determined

previously and completely. Therefore, an iteration with POS tagging and identification is designed for updating stop words and relevant keywords from retrieved datasets. It categorizes tokens into different part-of-speech categories so that the users can filter keywords in certain categories. For example, the word with symbol (SYM) tag is straightforward to be identified and assigned to the stop words category, while the word with NN (noun, singular or mass) tag can be filtered to recognize whether it is relevant or not.

2.3 Text Classification

Text classification is a machine learning algorithm for reducing the dimensionality of social media messages and employing NLP applications (e.g., vectorization, counting the frequency of words, clustering and topic modeling). The content of texts from social media covers several different disaster topics, such as power outage, heavy rain, strong wind, and governmental relief. Therefore, clustering the tweets with the same topic into the same class is a preliminary step to understand the content of texts. First, the overall dataset should be divided into two parts: training set and test set. It can be split randomly for simplicity, or based on well-researched approaches (e.g., probability distribution, Bayesian networks, and empty network) to improve accuracy. Then, as all machine learning methods are developed for numeric features, the text corpus of social communications should be converted in a format with numeric features as well. In our framework, the texts are transformed to vectors because computers are good at handling vectors in an efficient way. As such, different words in the dataset can be the elements of a matrix, while the frequencies of the words in each tweet are the values. In this way, each tweet can be represented by a vector. In this framework, the transformation is conducted by employing a refinement of term frequency method to downscale weights for words that are commonly used in other documents. This approach is called term frequency-inverse document frequency (TF-IDF) [17]. In the context of infrastructure resilience domain, some words such as “nervous”, “strong”, “washer”, and “repose” are not highly relevant but may appear with infrastructure-related keywords in the same tweet. The weights of those words should be very low even if their frequencies are a little bit high. Thus, TF-IDF plays a critical role to address these conditions. Subsequently, a term-document matrix is obtained after vectorization of the texts. The next step is to train a text classifiers such as Naïve Bayes classifier, decision trees, stochastic gradient descent, and support vector machines which are widely used for this purpose. However, these classification algorithms are supervised algorithms which require historic pre-labeled training data. Also, the disruptions and bursts of infrastructure services are changing over time since disruptions and bursts are sensitive to the intervention of human activities and interdependent system. For example, a large-scale power outage can be solved by utility companies rather than continuing after flooding receded. Thus, for simplicity and flexibility for dynamic data streams, the proposed framework adopts an unsupervised learning, called K-means clustering, which does not need any labeled data but can identify important hidden patterns in unlabeled data. However, K-means clustering method requires the number of clusters to be predefined. To obtaining the exact number of major clusters and topics from social media datasets, an iteration

involving topic modeling and repeatability detecting is implemented for determining a precise number of clusters (see Fig. 1). As discussed earlier, when the topics are repeated, the number of clusters should be reduced to maintain their uniqueness. For example, if two clusters of tweets are both talking about the road closure, it should be combined and get a complete a dataset for this topic. Thereby, the extracted topics from clusters are representative of the majority of texts and precise enough to be used for topic evolvement analysis.

2.4 Topic Modeling and Representation

In the infrastructure-specific domain, the detailed representation and development of events on social media communications can provide evidence for performance understanding and assessment (e.g., how severe is the interruption, how many people are affected by this disruption, and what further effects by this disruption should be taken into account). For example, the highway, Interstate-10 from Texas to Louisiana, closed when it was covered by floodwaters during Harvey, but it reopened after six days because the water receded. Thus, people who need to use this road should be presented with information in accordance to the situation of this road including which section of this road was closed, when it was closed, when it would reopen, and what damage of this road was caused by floodwaters. Therefore, not only the detailed representation, but also the evolvement of events should be carefully analyzed on the social data. In this framework, the events are detected based on the summarization of topics. By definition, topics are the summary and representation of certain clusters of tweets and are practical for tracking the development of disasters and resilience performance of infrastructure. For example, the airport was closed, and all flights were canceled when hurricane Harvey landed, while it reopened and played an important role in disaster relief when Harvey passed. Based on the extracted topics about the airport, disaster responders can further detect potential issues (e.g., deficient in the capacity of transporting and distributing resources) in this cluster of tweets. To get the initial insights regarding the unlabeled social posts, topic modeling is an effective approach to deal with large volumes of texts through Latent Dirichlet Allocation (LDA) and reduce the dimensionality of the dataset for further studies (e.g., specific entities detection, geo-tagging, and relation extraction). Intuitively, the topic modeling technique is based on the probability of the words in each document and the latent semantic structures of the text. For example, if a document is 80% talking about “flooding” and 20% talking about “birthday party”, the topic about “flooding” will definitely be extracted in priority. As such, the clusters of text can be represented by semantic coherent keywords, while the trend of topics can be tracked over different disaster phases.

3 A Case Study of Hurricane Harvey

The presented system analytics framework provides chances to identify developments of events related to the performance of infrastructure systems in disaster situations based on social media. In this paper, the process on a set of tweets relative to hurricane Harvey

shows the possibilities of the application. Using this system analytics framework, we filtered out relevant tweets, cleaned data sets, classified texts, and modeled topics with the context of Harvey in Texas. The findings in this specific case study show the potential of the proposed framework for detecting and tracking the disasters using tweets.

3.1 Context Recognition

The proposed framework is examined in a case study of hurricane Harvey, which made a landfall in Texas on 25th August 2017 [3]. The rainfall level during hurricane Harvey in Texas is shown in Fig. 2 [18]. Harvey caused severe disruptions in infrastructure systems. Based on a Texas Department of Transportation report [19], more than 200 highway locations were closed or flooded, and all flights were suspended in the Houston Airport System. Customers served by 166 water systems received boil-water orders and another 50 water systems were shut down completely due to storm impact [20].

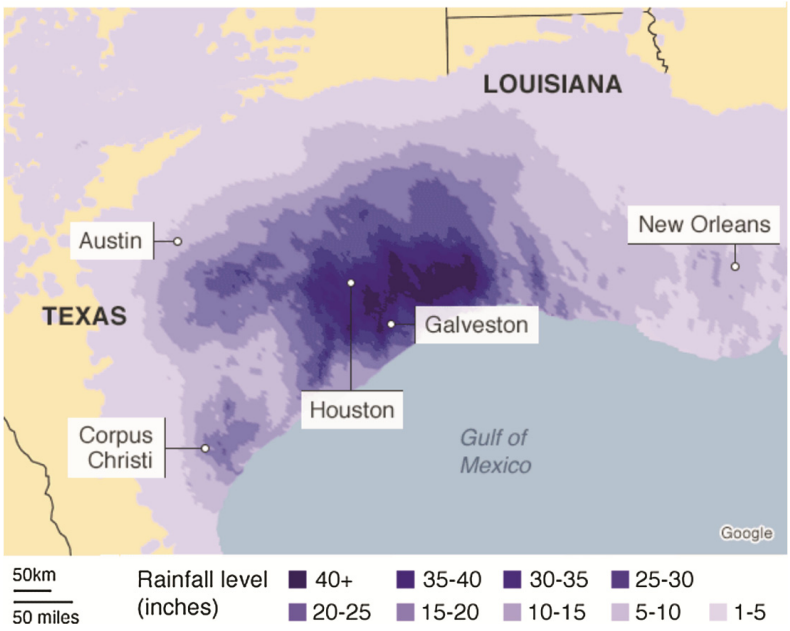


Fig. 2. The rainfall in Texas by hurricane Harvey [18].

In the case study, we identified relevant keywords including infrastructure-related, and disaster event-related keywords (see Table 1) from multiple online sources such as news articles, government reports, and social media. We applied the proposed framework to retrieve relevant tweets through the process of keyword-based collection and geography-based collection, process texts, update stop words and keywords, and iterate retrieving tweets. This process was repeated until a complete dataset is collected, at which point no stop words and keyword needed to be updated and no more tweets would be retrieved. In this paper, 63263 tweets were gathered from Aug. 23rd to Sept. 30th in the 150-mile radius around Houston.

Table 1. Keywords used for collecting relevant tweets.

Infrastructure-related keywords	Disaster event-related keywords
Shelter, power outage, electricity, road, closure, floodwater, water, watersheds, drainage, infrastructure, damage, restoration, transport, shortage, devastation, flight, airport, bridge, car, neighborhood, river, gas, build, rebuild, debris, sewage, electric, energy, utilities, roadway, Addicks, Barker, bayou, sh6, I10, reservoir	Harvey, flood, flooding, Hurricane, flooding, recovery, storm, surge, warning, evacuation, emergency, remnants, survivor, disaster, FEMA, medicine, rescue, rainfall, relief, victims, food, safety, health, donate, insurance, wind, response, rainy, refuge, resources, aid, demand, volunteer, landfall, needs, restore

The authors defined specific time periods of disaster phases to map the temporal distribution of infrastructure-related events. In this study, the disaster phases were defined based on time of duration (see Table 2): before the hurricane landed (preparedness phase), after the hurricane landed (response phase). From Aug. 23rd to Aug. 24th, people began to be aware of the threats of hurricane Harvey and prepare for the response (e.g., stocked up on food and water, and reinforced roofs). August 25th was the date that hurricane Harvey landed on the coast of Texas and moved towards Houston. After Aug. 29th, Harvey weakened as it drifted inland [3]. But, the heavy rainfall continued and caused flooding even after Harvey passed Houston. Besides, some areas in West Houston were flooded due to water release from two flood control reservoirs, Barker and Addicks. Thus, neighborhoods in West Houston continued to be affected by flooding until surcharge release ended on Sept. 14th. Hence, the flooding phase was defined from Aug. 30th to Sept. 14th. After flooding, residents, responders, government agencies, and infrastructure agencies took emergency response measures for disaster rescue and infrastructure restoration. Although the recovery was continued, our study did not retrieve tweets after Sept. 30th. Meanwhile, considering the limitation of the size of the dataset, the last disaster phase is from Sept. 15th to Sept. 30th. Accordingly, the collected tweets were separated into four subsets based on their predefined phase. It should be noted that the tweets processed in this case study were original tweets without retweeting. These tweets can reflect the actual condition at the time when they were posted.

Table 2. Time periods of disaster phases.

Disaster phases	Time of duration
Before Harvey landed	Aug. 23rd–Aug. 24th
Hurricane period	Aug. 25th–Aug. 29th
Flooding period	Aug. 30th–Sept. 14th
After the flooding receded	Sept. 15th–Sept. 30th

3.2 Text Classification and Topic Modeling

In the next component, text sampling, vectorization, clustering and topic modeling were conducted. The output of this component is listed in Table 3 and mapped in Fig. 3. Because the duration of time in different phases was various and the physical

environment was dynamic in disasters, the number of clusters in each disaster phase was different. This process was controlled by the iteration that increased the number of classes until repeated topics appear.

As shown in Table 3, before the Harvey landed, three major topics were dominant: (1) George Bush Intercontinental airport was affected and needed to take rapid response to the hurricane; (2) Pearland is humid so that residents and infrastructure agencies should prepare; (3) City of Houston encourage citizens to get prepared. The relevant tweets were clustered into three categories (see Fig. 3). Topic 1 which is associated with the airport and topic 3 which is associated with the early warning in Houston area are more frequent than topic 2 which is associated with the early warning in Pearland since the size of tweets cluster of topic 1 and topic 3 are distinctively larger than the size of the cluster of topic 2. In addition, topic 2 showed its significance in early warning before the disaster since the water sewage treatment system in Pearland was sensitive to storms and flooding [21].

When Harvey landed, topics were more about the devastation and severity of the disaster. As displayed in Table 2, five topics were detected in the hurricane phase: (1) Because Houston was attacked by Harvey at that time, the public was conducting relief and donating for damaged properties; (2) the intensity of rainfall in Pasadena was heavy so that residents and responders were encouraged to be careful; (3) freeway and tollway were both closed because of flooding and extreme weather; (4) the floodwaters level and status of freeway were reported for the public and responders; (5) downtown and Eastside were highly affected by the hurricane. Most of the topics were talking about the damage in infrastructure systems. For example, the freeway was the most sensitive infrastructure for hurricane due to rainfall, and it was essential for evacuation, search and rescue, and refuge. As shown in Table 3, a technical challenge remained in which some noisy words (e.g., “bitlyhsijn”, “bublyusahrz”) were scratched inevitably should be mentioned here. The reason for that is there are some technical issues (e.g., decode error) of parsing texts from the Twitter platform. The problem can be addressed when cleaning up the data, but it is possible to delete some critical and relevant information at the same time. Therefore, in this case study, some noisy words were left to keep the semantic integrity of the texts.

After the hurricane passed through Houston, flooding continued, not only because of the rainstorm but also due to water release from reservoirs that was recognized in context recognition. In addition, the public, agencies, and organizations were joining in the relief efforts. Thus, the topics in this phase were: (1) Bush intercontinental airport started operation for relief; (2) reducing mph when driving in Houston was important since some roads were still humid; (3) freeway and tollway were still closed in the Westside; (4) The residents in Pasadena should still be in response to heavy wind; (5) fund was raised for aid and shops were open. The severe flooding after hurricane appeared on the Westside of Houston. Thus, the freeway and tollway in the west of Houston were still closed for security. In topic 3, a damage on the highway, Sam, in west Houston was detected and the Sam tollway reopened when the 14 feet of water drained from it [22]. Meanwhile, other parts of Houston gradually restored from the disasters. A dominate event was that the airport resumed its operation. It helped stranded passengers leave Houston and transport resources from other states. Thus, a large

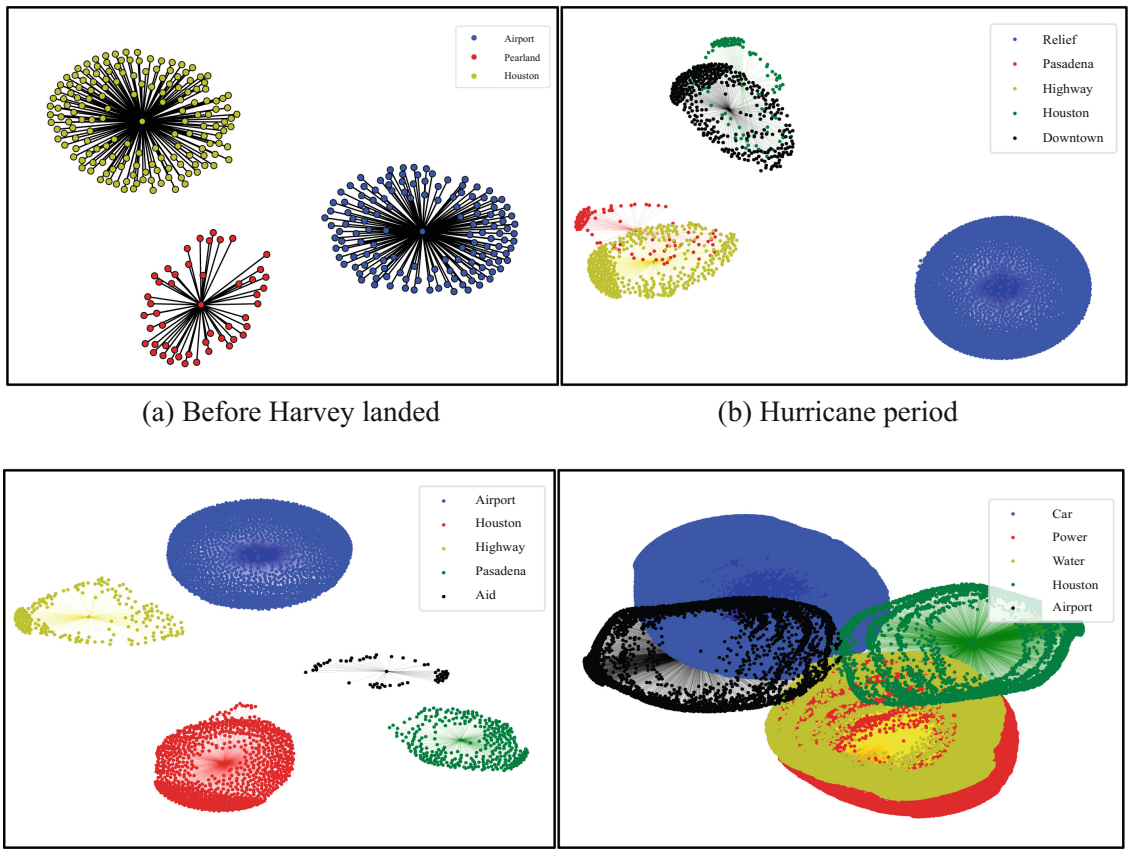
Table 3. Extracted topic in four disaster phases.

Phases	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Before the Harvey landed	mph	mph	hurricane		
	Houston	pressure	Houston		
	Harvey	current	Texas		
	airport	weather	Harvey		
	wind	humidity	mph		
	hurricane	wind	ready		
	bush	Pearland	wind		
	intercontinental	sky	tropical		
	George	clear	Hurricane Harvey		
Hurricane period	Houston	humidity	fwy	Houston	main
	Harvey	pressure	closed	Harvey	affecting
	hurricane	weather	Sam	repost	high
	Texas	intensity	flooding	bitlyhsijn	downtown
	water	heavy	lane	help	eastside
	storm	Pasadena	traffic	Texas	bublyusahrz
	relief	rain	tollway	fwy	lane
	donate	wind	frontage	water	fwy
Flooding period	make	hurricane	temperature	weather	raise
	money	Houston	Sam	pressure	shop
	travel	humidity	traffic	cloud	aid
	airport	relief	flooding	wind	unique
	bush	wind	closed	Pasadena	sale
	lifestyle	mph	fwy	sky	help
	intercontinental	Harvey	tollway	current	Obama
	start	Texas	Westside	humidity	fund
After the flooding receded	car	need	energy	Harvey	buy
	need	food	need	benefit	coming
	Houston	warning	water	relief	hopefully
	water	power	service	concert	does
	recovery	info	health	Houston	flight
	fund	ppb	power	strong	airport
	relief	Houston	get	car	know

percentage of tweets were associated with this topic. However, some topics such as power outage were not among the topics based on the modeling on current dataset. This was because the power outage just appeared in some parts of Houston areas so that it was not the commonly concerned issues among the public. Except for the infrastructure-related topics, human behaviors about raising funds, opening shops, and relief were discussed as well.

When flooding receded, the damaged infrastructure systems were in need of recovery. Thereby, the topics in this phase were: (1) many cars in Houston were destroyed by flooding so the victims were seeking funds for recovery; (2) food and power were needed in some parts of Houston; (3) the public were concerned about the health impacts of contaminated flood water; (4) stakeholders were conducting relief during this period; (5) airport performed very well for transporting resources by flights. Because of severe flooding, 300,000 to 500,000 vehicles in Houston were destroyed by Harvey [23]. Thus, victims and companies were seeking funds for making up for a loss. After the disaster, resources including power, oil, and boiled water were the most urgent needs for all victims. Thus, topic 2 and 3 account for a large proportion of the collected tweets in Fig. 3 (e.g., green and yellow nodes). A serious issue detected in the results is the contaminated floodwaters with Nasty and dangerous bacteria that people with underlying illness, the elderly and children were susceptible to [24]. Thus, this cluster of tweets was warning for the public to be careful of the floodwaters and for the agencies to develop strategies for reducing the health effects.

The summarized topics were used to analyze the infrastructure conditions in disasters from relevant tweets. Figure 3 mapped the tweets based on spring layout, a force-directed layout in graph theory. The lengths of the edges and the locations of nodes were based on Hooke's law which was used to attract pairs of endpoints of the edges towards each other [25]. It should be mentioned that the clusters in each figure were mapped randomly



(a) Before Harvey landed (b) Hurricane period

(c) Flooding period (d) After the flooding receded

Fig. 3. Clusters of tweets in four disaster phases. (Color figure online)

and automatically by computers. Thus, as the limitation of the size of the figures, some clusters were partially overlapping. In addition, different colors represent different clusters. For example, in the first phase, three topics were represented by three colors: red, yellow, and blue. As Fig. 3 shows, the tweets were connected to their centroid topics. Although the number of major topics in last three phases was equal, the topics are quite different, and the trend of these topics will be discussed in the following section.

3.3 Topic Evolution and Resilience Analysis

Analyzing the evolution of topics over the development of disasters is essential for understanding the effects of disasters on infrastructure systems. For example, Harris County Tolls started up all toll roads on September 7th except Sam Houston Tollway because a section of Sam was still closed [22]. This situation can be an indicator of the performance of Sam Houston Tollway during flooding. Also, the performance of the tollway can be the evidence for decision makers to develop mitigation plans including repairing or replacing road drainage systems for improving the flood resistance. Because of the importance of infrastructure performance in disasters, we summarized the detected topics, and computed the percentage of relevant tweets in each topic among the dataset in certain disaster phase in Fig. 4. Figure 4 shows the trend of the major topics as well as the importance of the topic in each disaster phase. For example, the topic about airport appeared before the Harvey landed, in flooding period, and after the flooding receded. When the Harvey began to affect the airport, all flights were canceled for the purpose of security, and passengers were trapped in the airport, waiting for rescue. Thus, the public was worried about their journey and tweeted their situations for help. Hence, the tweets about airport topic account for 40% of the total number of tweets in first disaster phase. During Harvey, the airport was closed and only allowed a few flights for relief such as transporting rescue staffs and resources. Therefore, the topic about the situation of the airport did not become a core topic in this disaster phase. In flooding period, the airport restored and operated for transporting relief supplies and passengers, playing a critical role in disaster relief. Thereby, the tweets about airport conditions account for 58% of the dataset in this phase. After flooding receded, affected communities started recovery measures, such as cleaning up the ruins, restoring transportation systems, and repairing power grid. The topic about highway appeared and increased significantly during hurricane and flooding period since sections of important roads such as Interstate-10 were closed by water. But, after the hurricane, the percentage of tweets about highway dropped from 11% to 6% since flooding receded in most parts of Houston but still affected the Westside. For example, in flooding period, outbound interstate 45 North to and from Dallas was accessible, and US-290 was open, even though the State Highway 225 at Richey in and out of Pasadena remains covered in water and still closed [26]. Meanwhile, in the last disaster phase, three new topics about car and power were detected by the framework. It is because floodwaters destroyed more than 300,000 cars and power grid in some communities.

To support the results from topic modeling in this framework, the daily frequencies of infrastructure-related keywords were computed and displayed in Fig. 5. Before the Harvey landed, social communications were hardly related to infrastructure service and

their damages. Thus, the frequencies of the keywords were lower than the frequencies of these keywords in other disaster phases. During Harvey, however, the keywords including "roadway", "lanes", "freeway (fwy)", and "blocked" were obviously increased and accounted for a large proportion. It should be mentioned that "power" and "outage" were not frequent in the collected datasets even though the power outage appeared in some areas during hurricane Harvey. After the Harvey passed, the size of tweets about infrastructure damages was decreased since the stakeholders were conducting recovery measures in full swing at that time. It was evidenced by the high frequency of "relief" in the phase of flooding period. Due to the limitation of the dataset which will be discussed in next section, the size of the dataset in this disaster phase was larger than the sizes of datasets in other phases. Thus, the word frequencies were higher than the corresponding frequencies in other phases. Nevertheless, comparing the word frequencies in this dataset, "car", "electric", "power", "road", "energy", "tollway", and "gas" reached high frequencies in this disaster phase. The results proved the conclusions made from topic modeling. Thus, the summarized topics in each disaster phase were precise enough for representing the situations of some infrastructure systems.

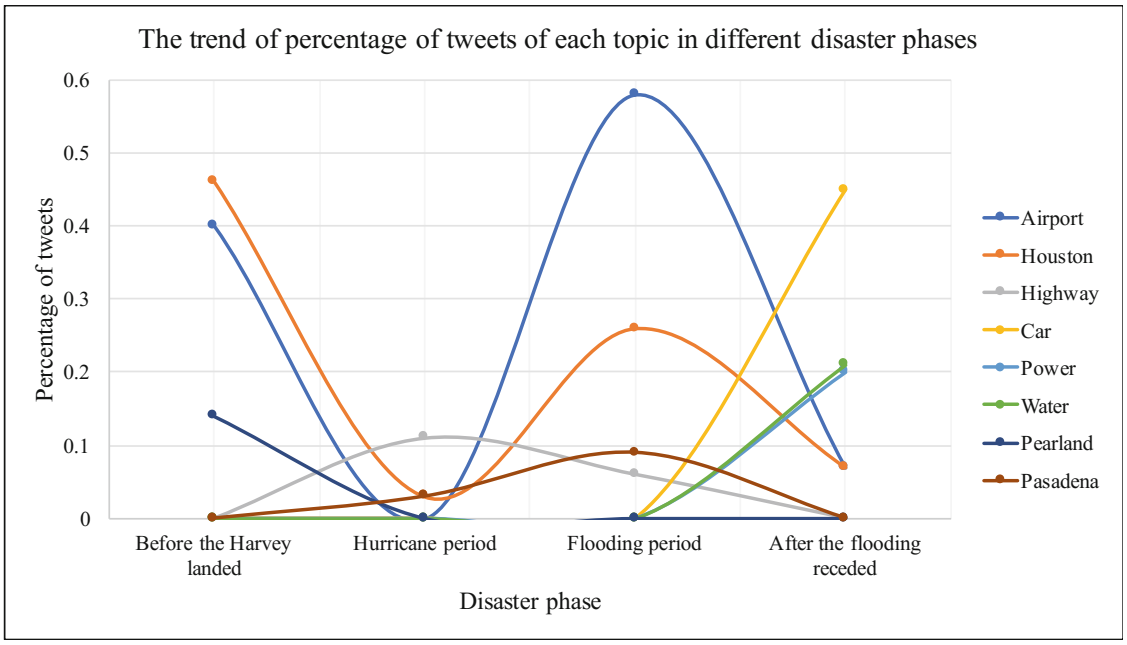


Fig. 4. The trend of percentage of tweets of each topic in different disaster phases

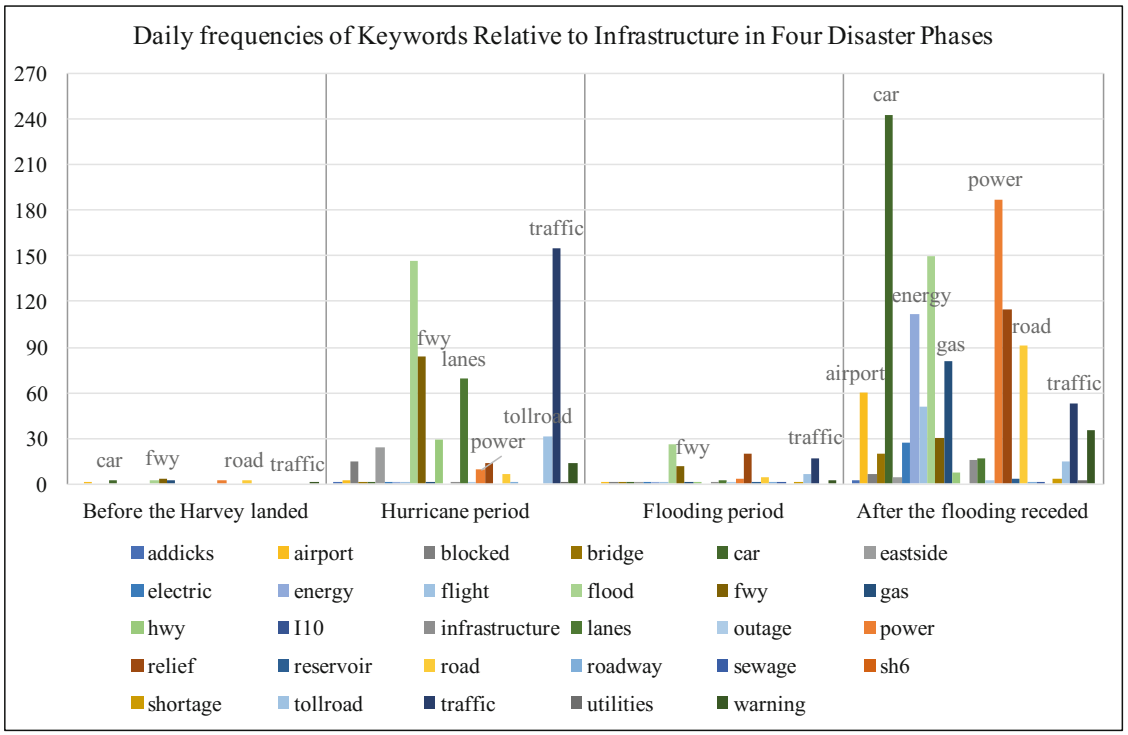


Fig. 5. Daily frequencies of keywords relative to infrastructure in four disaster phases

4 Limitations and Future Work

This paper answered the research question: to what extent can infrastructure-related events be detected and tracked in disasters on social media, through a system analytics framework and a case study of hurricane Harvey. However, limitations exist in our current work. First, the dataset of tweets used in this case study is incomplete. This dataset was not collected by Twitter Rest API or Streaming API during disasters, instead, it was collected by a web crawler through Twitter search platform. The limitation of using Twitter search platform was that part of relevant tweets in two weeks ago cannot be scrapped. Thus, the size of tweets in first three disaster phases was relatively smaller than the size of tweets in last disaster phase. In further studies, we will employ a complete dataset with 21 million tweets over Houston area in Harvey and flooding period.

The proposed framework can incorporate more text mining approaches (e.g., Named Entity Recognition, and event detection) [27, 28] for detecting detailed information in social communications. For example, among the blocked roads was Interstate 45 from Dallas to Houston [29]. This event did not appear in the major topics, but it is essential for the people who want to evacuate themselves or providing helps to victims through this road. Therefore, detecting detailed entities and related events are needed for situational awareness and relief operation. However, there is no well-trained model to conduct detection and summarization of infrastructure-specific entities and events. So, supervised learning algorithms can be employed to train a model which can identify the domain-specific entities and events automatically. The procedure for the supervised learning approach can be: (1) cleaning, vectoring, and clustering the data; (2) sampling

the corpus into training set and test set, and labeling the data in training set with infrastructure-specific labels (may need crowdsourcing); (3) adopting an appropriate machine learning algorithm to train the model; (4) applying the model to the test set. The larger the training set is, the better the output will be. Furthermore, the well-trained model can be enveloped into a system which can provide evidence of damage and support for decision making automatically.

5 Conclusion

This paper presented and examined a system analytics framework involving social sensing and text mining for text-based representation of infrastructure resilience performance in disasters. The framework creatively developed an iteration for building a complete infrastructure-specific lexicon and a stop-words list to retrieve and clean relevant tweets. In addition, this framework integrated text classification (e.g., K-means clustering) and topic modeling (e.g., LDA) for summarizing the major topics in clusters of tweets. Therefore, the topics that the public and agencies were highly concerned were detected and analyzed for understanding the extent that the events were discussed on social media. The application of the proposed framework was shown in a preliminary study of infrastructure-related event detection and tracking in 2017 hurricane Harvey and derived flooding around Houston area. The case study investigated 63263 tweets in four disaster phases which were defined by the duration of disasters. The results showed that the computational algorithm exhibits capabilities of gathering complete and relevant datasets, classifying tweets into several clusters, modeling major topics in each cluster, and analyzing the changes of infrastructure-related topics along with the development of disasters. Thus, the results well addressed the gap of applying social sensing to assessing infrastructure service disruptions and highlighted the extent of the public and agencies discussing infrastructure on social media. Based on the output of the framework and the results of the analysis, the disaster responders and residents can develop their measures for enhancing the resilience of infrastructure systems, including improving the drainage capacity of the highways and early warning for airports. Their implementation can eventually contribute to the improvement of infrastructure resilience performance and property safety in times of disasters.

The framework proposed in case study can integrate the machine learning algorithms and disaster informatics on social media for data collection, cleaning, clustering and topic modeling relative to infrastructure systems. This system analytics framework can be further developed into several aspects: (1) apply it to other domain-specific research (e.g., sports event, finance & stock, and business marketing), and adjust some components to make it more adaptive to varieties of different domains; (2) incorporate other data processing and analyzing methods (e.g., dependency parsing, chunking, and word sense disambiguation) to explore new features of the collected data; (3) extend this framework to conduct some other analysis such as dynamic network analysis from systematic operation or civil engineering perspectives.

Acknowledgement. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1759537. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Tien, I., Musaev, A., Benas, D., Pu, C.: Detection of damage and failure events of critical public infrastructure using social sensor big data. In: Proceedings of International Conference on Internet of Things and Big Data, April 2016
2. Eusgeld, I., Nan, C., Dietz, S.: "System-of-systems" approach for interdependent critical infrastructures. Reliabil. Eng. Syst. Saf. **96**(6), 679–686 (2011)
3. Wikipedia Contributors: Hurricane Harvey. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Hurricane_Harvey. Accessed 10 Dec 2017
4. Wang, D., Abdelzaher, T., Kaplan, L.: Social Sensing: Building Reliable Systems on Unreliable Data. Morgan Kaufmann, Burlington (2015)
5. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In: ICWSM, June 2014
6. Lu, X., Brelsford, C.: Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami. Sci. Rep. **4**, 6773 (2014)
7. iRevolutions Contributors: What is Big (Crisis) Data? iResolutions. https://irevolutions.org/2013/06/27/what-is-big-crisis-data/. Accessed 10 Dec 2017
8. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. ACM Comput. Surv. (CSUR) **47**(4), 67 (2015)
9. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Extracting information nuggets from disaster-related messages in social media. In: ISCRAM, May 2013
10. Acar, A., Muraki, Y.: Twitter for crisis communication: lessons learned from Japan's tsunami disaster. Int. J. Web Based Commun. **7**(3), 392–402 (2011)
11. Ashktorab, Z., Brown, C., Nandi, M., Culotta, A.: Tweedr: mining twitter to inform disaster response. In: ISCRAM, May 2014
12. Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R.: Using social media to enhance emergency situation awareness. IEEE Intell. Syst. **27**(6), 52–59 (2012)
13. Bala, M.M., Navya, K., Shruthilaya, P.: Text mining on real time Twitter data for disaster response. Int. J. Civ. Eng. Technol. **8**(8), 20–29 (2017)
14. Prasetyo, P.K., Gao, M., Lim, E.P., Scollon, C.N.: Social sensing for urban crisis management: the case of Singapore haze. In: SocInfo, pp. 478–491, November 2013
15. Bruns, A., Liang, Y.E.: Tools and methods for capturing Twitter data during natural disasters. First Monday **17**(4), 1–8 (2012)
16. Hardeniya, N.: NLTK Essentials. Packt Publishing Ltd., Birmingham (2015)
17. Sarkar, D.: Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data. A Press, New York (2016)
18. BBC NEWS Contributor: In maps: Houston and Texas flooding. http://www.bbc.com/news/world-us-canada-41094872. Accessed 31 Dec
19. ENR Editors: How badly has hurricane Harvey damaged Texas infrastructure? In: Engineering News-Record, August 2017
20. Hernandez, A.R., Zezima, K., Achenbach, J.: Texas faces environmental concerns as wastewater, drinking water systems compromised. In: Washington Post, September 2017

21. Nix, K.: Hurricane Harvey brings record rains to Pearland, September 2017. http://www.chron.com/neighborhood/pearland/news/article/Hurricane-Harvey-brings-record-rains-to-Pearland-12169450.php. Accessed 14 Jan
22. KHOW Contributor: Sam Houston Tollway fully reopens after Harvey flooding, damage. http://www.khou.com/news/hctra-sam-houston-tollway-southbound-to-reopen-sunday-night/473114404. Accessed 19 Jan
23. Fortune Contributor: Hurricane Irma and Harvey damaged 1 million cars. What happens now? http://fortune.com/2017/09/20/hurricane-irma-harvey-damaged-cars/. Accessed 14 Jan
24. Scutti, S., CNN Contributor: Sewage, fecal bacteria in hurricane Harvey floodwaters, September 2017. http://www.cnn.com/2017/09/01/health/houston-flood-water-contamination/index.html. Accessed 19 Jan
25. Kobourov, S.G.: Spring embedders and force-directed graph drawing algorithms (2012). Freely accessible: arXiv:1201.3011
26. Whaley, K., Eyewitness NEWS Contributor: Check out Houston freeway conditions post-Harvey. http://abc13.com/traffic/check-out-houston-freeway-conditions-post-harvey/2358197/. Accessed 19 Jan
27. Gao, X., Yu, W., Rong, Y., Zhang, S.: Ontology-based social media analysis for urban planning. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 1, pp. 888–896. IEEE, July 2017
28. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M.: Rapid assessment of disaster damage using social media activity. Sci. Adv. **2**(3), e1500779 (2016)
29. QUARTZ Contributor: One map shows just why the Texas flooding is so disastrous: it's preventing people from escaping, August 2017. https://qz.com/1066033/hurricane-harvey-highways-closed-by-record-rainfall-are-trapping-texans-in-flooded-communities/. Accessed 31 Dec