Urban Crisis Detection Technique: A Spatial and Data Driven Approach based on Latent Dirichlet Allocation (LDA) Topic Modeling

Yan WANG[1] and John E. TAYLOR[2]

[1] Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech, 121 Patton Hall, Blacksburg, VA 24061, U.S.A.; email: wangyan@vt.edu

[2] School of Civil and Environmental Engineering, Georgia Institute of Technology, Mason 4140c, Atlanta, GA 30332, U.S.A.; email: jet@gatech.edu.

ABSTRACT

Social networking platforms have been widely employed to detect, track, and visualize physical events in population-dense urban areas. They can be effective tools to understand when, where, and what happens retrospectively or in real time. Correspondingly, a variety of approaches have been proposed for detecting either targeted or general events. However, neither type of event detection technique has been developed to detect urban disasters in specific geographic locations and with unpredictable characteristics. Therefore, we propose a spatial and data-driven technique for detecting urban disasters. The method addresses both geographical and semantical dimensions of events (geo-topic detection module) and evaluates their crisis levels based on the intensity of negative sentiment (ranking module). Our approach was designed specifically for georeferenced tweets. To demonstrate the system, we conducted an experiment with four-hours of geotagged tweets in London. Our urban crisis detection technique successfully identified the Grenfell Tower fire among all the candidate geo-topics. Our future work focuses on enabling online-mode detection with high scalability in large-volumes of streaming data. The completed research will contribute to efficient disaster informatics and urban resilience regarding crisis detection and tracking, situation awareness, and information diffusion.

INTRODUCTION

Newly available and massive sets of data have played an increasingly important role at different stages of crisis/disaster management (i.e. early warning, monitoring and evaluation), especially in forming bottom-up perspectives in understanding the evolving process of events (Ford et al. 2016). Currently, diverse sources of digital data have brought enormous opportunities in the research area of urban resilience. These sources mainly include cell phone (Lu et al. 2012), and a few social networking platforms (Wang and Taylor 2017). Among them, Twitter is suitable for emergency environments in terms of its open design, wide usage, geo-enabled function and limited message lengths environments (Kryvasheyeu et al. 2016). Geo-referenced tweets can document geographical locations and collective reactions to crises unfolding at both spatial and temporal scales. Specifically, crisis detection is an emerging topic, where disaster managers can take advantage of the crowdsourced data from social networking platforms to enhance situation awareness. Early detection is crucial because it enables immediate responses and helps to reduce potential casualties and damage (Li et al. 2017). Early detection also contributes to characterize an event in terms of spatiotemporal scale, collective emotions, semantic topics, and its dynamic evolving process over time.

An increasing number of studies concerning event detection techniques in the context of Twitter have been recently published. Some studies focus on targeted events with supervised methods (Sakaki et al. 2010; Sun et al. 2016) while others intend to identify general events which burst in contents, time and space (Maurya et al. 2016; Xie et al. 2016; Yu et al. 2017; Zhang et al. 2016). These proposed detection techniques can be built based on clustering, supervised classification, Latent Dirichlet Allocation (LDA) or hybrid ways. However, few have explored the context of urban disasters (e.g. infrastructure failure, building fires, city bombing, natural disasters, etc.). Compared to other events, urban disasters can be regarded as more "targeted", but also unpredictable in types and forms. It is difficult to employ current supervised techniques for targeted events to identify an un-characterized disaster. Moreover, detection approaches for general events have not stressed the distinct characteristics of disasters in terms of their geographical and thematic impact, and high-intensity of negative sentiment. Therefore, to address this methodological gap, we propose a data-driven technique to detect urban disasters with a focus on geotagged tweets from a Twitter Streaming API. We describe the system as the Urban Crisis Detection technique to highlight its specification in detecting crises occuring in the confined physical locations of cities.

BACKGROUND

Event detection from Twitter streams has witnessed a mounting number of publications in the literature. We classified the most cutting-edge approaches based on their detection objectives into two types: targeted event detection and general event detection. Existing techniques for general event detection were either retrospective or real-time. We exclusively discussed the latter since most real-time methods were built based on retrospective methods, and our final goal is to enable real-time detection.

Targeted event detection

Targeted event detection requires pre-defined keywords and mainly adopts supervised detection techniques. For example, Sakaki et al. (2010) proposed a targeted event detection system that monitored tweets and delivered prompt notifications. Their system was specifically applied in reporting *earthquakes* with Japanese tweets. They firstly devised classifiers to classify event-related tweets and unrelated tweets. Then the related tweets were used to develop a probabilistic spatiotemporal model for event detection and location/trajectory estimation. Sun et al. (2016) designed a novel method to detect and locate power outages from Twitter. The system was based on a heterogeneous information network, which includes time, locations, and texts. Supervised LDA was then used to compute the probability of the topics of tweets that were related with a power outage. Gu et al. (2016) proposed a real-time traffic incident (TI) detection approach based on tweet texts. Each imported tweet was mapped into binary vector of a dictionary and classified as TI-related or not. The TI-related tweets were further geo-coded and classified into different incident categories.

Detection techniques for distinct targeted events are effective in identifying specific events with pre-envisioned and pre-defined characteristics. However, due to the diverse types of urban disasters, it may require a large volume of keywords to describe different types of potential events, not to mention unexpected types of events. Therefore, it is impractical to employ a supervised approach to detect any general and unknown disaster without pre-defining its specific characteristics.

General event detection

Clustering-based approaches. Cluster-based approaches include threshold-based online approaches, graph-based clustering algorithms and other new approaches. For example, Yu et al. (2017) proposed a real-time emerging anomaly monitoring system over microblog text streams, named RING. The system was based on a graph stream model. It was able to detect events at an early stage, to conduct correlation analysis between emerging events, and to track evolution of events over time. Specifically, the graph regarded keywords as nodes, their co-occurrence in each tweet as edges, and an accumulated frequency as weights of edges. A k-clique percolation method was then employed to identify communities (events) in the built graph. SigniTrend is a scalable detection technique developed by Schubert et al. (2014), which measured significance of terms to detect trending words based on their co-occurrences, and used hashing technique to track all the keyword pairs. The final stage of this approach was to cluster the detected keywords into larger topics. This method was used to detect emerging topics early before they become "hot tags". Some techniques included geolocations as a main dimension to capture real-world occurrences. EvenTweet (Abdelhaq, et al., 2013) identified localized events using geotagged tweets. It extracted keywords based on the burstiness degree of words, and then computed the spatial density distribution (spatial signature) over a keyword in a spatial grid. The event keywords were further partitioned based on the cosine similarity of their spatial signatures. Finally, the clusters were scored to uncover the real-world local events. GeoBurst (Zhang et al. 2016) was also designed to extract local events from streams of geotagged tweets in real time. It identified candidate events based on both geographical and semantic impact between each pair of tweets, and ranked the candidates according to their spatial and temporal burstiness. However, most of the clustering based approaches used co-occurrences of keywords to measure the semantic relationship between documents and, as such, they cannot reveal the latent structure of topics underlying the text corpora.

LDA-topic-model-based method. Latent Dirichlet Allocation (LDA) is a basic probabilistic topic model, which analyzes the words of the original texts to reveal the underlying themes and their connections (Blei 2012). More recently, researchers have explored the advantages of LDA in allowing for the examination of multiple topics within a document and generating a probabilistic distribution of words under a topic. This has been employed using LDA as a basis to extract thematic content from social networks for event detection. For example, Semantic Scan (Maurya et al. 2016) used a contrastive topic modeling based on LDA to identify new topics in text stream, it then used statistical scanning to find the spatially localized events. The proposed technique has been tested on Yelp and Emergency Department datasets, and the moving window size is three days, which is too long for detecting emergencies from Twitter. Moreover, the method requires a pre-defined number of topics for both background corpus and foreground topics. Topic Sketch (Xie et al. 2016) was designed to detect bursty topics from Twitter, with the assumption that each tweet is only related to one latent topic. Topics were generated based on sketch-based topic modeling using Singular Value Decomposition of word pair frequency matrices or tensor decomposition of word triple frequency matrices. It also employed a hashing-based dimension reduction technique, and conducted an effective sketch maintenance based on acceleration of words.

These methods only included time and semantic dimensions in the detection process, without considering the geographic dimension. However, in terms of urban disasters, the physical locations and spatial pattern of an event are as important as text contents and time. Current LDA-based event detection methods have not been tailored to detect disasters, which are different from other events, such as celebrations, football games, and marathons, in terms of the intensity of negative emotions. Therefore, we propose a technique to specifically detect urban emergencies.

The technique takes both geolocations and texts into consideration, and the semantic correlation is measured based on LDA. To separate urban crises from other events, we employ intensity of negative sentiment to filter the events. Details regarding the modules and designs are described in the following section.

URBAN CRISIS DETECTION TECHNIQUE

We defined urban crises in our studies as having the following properties: (i) they are geographically proximal and semantically relevant; and (ii) they influence people's emotion and trigger a high intensity of negative sentiment over certain time windows. The properties also form the basic assumption of the detection pipeline. The main modules of our proposed pipeline technique are illustrated in Figure 1. The first module for collecting geo-referenced tweets from Twitter Streaming API is a built system in our lab (Wang and Taylor 2015). On the basis of the collection module, we developed the "Geo-Topic Detection" module to identify distinct events. This module combines both geographic closeness and LDA-based semantic similarity to extract topics. To further evaluate the "crisis level" of the candidate topics, we employed intensity of negative sentiment to rank the events in the "Geo-Topic Ranking" module. Events with higher average intensity over certain consecutive time windows are identified as Urban Crises.

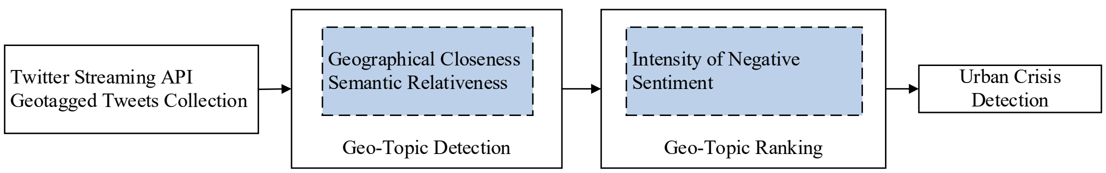

Figure 1. Urban Crisis Detection Technique

Data preprocessing

Twitter allows its users to share short 140-character messages. The texts can include words, URLs, mentions, emotions, abbreviations, etc. To analyze this data, we clean the texts by removing URL links and user mentions (@), which are not relevant with the core meaning of the text in a large amount of tweets. We also remove the special characters (i.e. '#$%^&*', punctuations and numeric numbers) that are unnecessary for further analysis. We tokenize the tweets to unigrams, convert all of them to lowercase, and remove the stop words and words with less than two characters. We also delete the words that appear less than a certain number of times (depending on text frequencies in different cases). Notably, these procedures are for topic modeling, and normalization steps for sentiment analysis have some differences: we try to keep as much information as possible for sentiment analysis and only URL links and user mentions (@) are removed from the original tweets.

Candidate geo-topics generation

LDA-based topic similarity. We employ Latent Dirichlet Allocation (LDA) to identify latent topic information in the corpora of tweets. A tweet is treated as the "bag of words", disregarding the word order and grammar. The basic idea of this generative probabilistic model is that each tweet is represented as a probability distribution over pre-defined number of topics, and each topic is represented as a probability distribution over words (Blei et al. 2003). Compared to other semantic measurements, results of LDA-based cosine similarity is much closer to human perception of document similarity (Towne et al. 2016). We then use cosine similarity of tweets' topics for similarity measure (see Eq. 1).

$$C\left(t \to t'\right) = \cos\left(topic_t, topic_{t'}\right) \qquad (1)$$

t and t' are two distinct tweets. $topic_t$ represents the topic vector of a tweet, and the topic vector is a probability distribution over the pre-defined number of topics.

Geographical closeness. We adopt Epanechnikov kernel to measure the geographical closeness of two tweets due to its high efficiency and simplicity (Zhang et al. 2016) (See Eq. 2).

$$K(t \to t') = 0.75 \times \left(1 - c(\left\|loc_t - loc_{t'}\right\|)^2\right) \qquad (2)$$

$\left\|loc_t - loc_{t'}\right\|$ is the Haversine Distance between locations of two distinct tweets; c is the scaling function that transforms the distance to the range $(0,1)$.

Geo-Topic clustering. We build an undirected graph named "Tweet Geo-Topic Graph": $G = (V, E)$ where V is the set of tweets and E is the set of edges. The weight of an edge is the product of semantic similarity and geographical closeness. If the weight is less than the pre-defined threshold, the edge will not be built. The node with the highest sum weight is regarded as a hub. Each hub with each connected neighbors will be regarded as a candidate geo-topic.

Geo-topics ranking and crisis detection

We rank the candidate geo-topics based on their intensity of negative sentiment: Geo-topics with higher average intensity of negative sentiment over certain time windows have a higher rank. This module is based on the basic property of geo-crisis.

Sentiment analysis. We adopt an unsupervised lexicon-based method to measure the sentiment. The method is based on an affective word list AFNN to assign sentiment scores to words in tweets (Nielsen 2011). The latest version of the word list includes 2,477 words. The valence of a word ranges from -5 (very negative) to +5 (very positive) as an integer. The sum of valence without normalization of words represents the combined sentiment strength for a tweet. A Python Package "afinn" was used to compute the sentiment scores. The AFNN is a Twitter-based sentiment lexicon including Internet slangs and obscene words. It has been tested in different types of tweets corpora and performs consistently at a satisfactory level of accuracy for both two classes (positive and negative) and three classes (positive, negative and neutral), compared with other unsupervised methods for sentence-level sentiment analysis (Ribeiro et al. 2015). Moreover, the AFNN word list has shown its advantages in analyzing tweets for disaster and crisis sentiment detection, e.g. Nagy and Stamberger (2012) and Walther and Kaisser (2013). We therefore selected the AFNN lexicon to evaluate the sentiment polarity of our collected tweets.

Intensity of negative sentiment. Since we are mainly concerned with negative tweets (in a crisis situation), we convert the numeric scores of sentiment to binary score. Specifically, we use one to represent negative sentiment, and zero for both positive and neutral sentiment. We employ intensity of negative sentiment under each candidate geo-topic to quantify their crisis level (see Eq. 3).

$$S_t = \frac{n_{tweet^-}}{n_{tweet}} \qquad (3)$$

n_{tweet^-} is the number of tweets with negative sentiment in each cluster; and n_{tweet} is the total number of tweets in each cluster. To include the impact duration of each crisis, we calculate the average values of the intensity over consecutive time windows (see Eq. 4).

$$\overline{S} = (\sum_{1}^{i} S_t) \Big/ i \qquad (4)$$

i is the number of consecutive time windows and $i \geq 1$.

VALIDATION

We applied our proposed technique on a dataset collected from a Twitter Streaming API with only the filter "geotagging" to assess and validate its functionality. The dataset was filtered in a spatial bounding box of Greater London (longitude: from -0.489 to 0.236; latitude: from 51.28 to 51.686) during 5:57AM to 9:56AM (UTC) on June 14, 2017. During the time duration, the Grenfell Tower in London was experiencing a severe building fire. We sliced the duration into four-equal-length time windows to analyze the average of intensity of negative sentiment for geo-topics. Our experiment was conducted with Python 2.7 in the Anacondor environment.

Data pre-processing. The normalization of the tweets followed the aforementioned procedures. In addition, we solely focused on tweets in English, which occupied 92% of our filtered tweets. The final data volume of the English tweets over the four hours in London is 1,309. We also removed common words in the local area which do not contribute to the topics to avoid their disturbance in topic modeling, including: "London", "United", "Kindgdom", "unitedkingdom", "trending", "trndnl", "trend", "trends", "hashtag", "tweet", "tweets", "twitter", "today", "photo", "photos", "Instagram", "posted". Words that appeared less than two times were also removed.

Candidate event generations. We applied a batch mode to model the topics of our cleaned corpora, and set the number of topics as 15. One of generated 15 topics with its top words and distributions is demonstrated in Table 1.

Table 1. Topic-most frequented words.

Topic#	Distribution of Words Over Topics
4	0.061*"tower" + 0.057*"grenfell" + 0.034*"fire" + 0.020*"city" + 0.019*"food" + 0.017*"white" + 0.015*"wednesdaywisdom" + 0.015*"plfixtures" + 0.014*"bst" + 0.011*"new" + 0.010*"building" + 0.010*"block" + 0.009*"smoke" + 0.008*"best" + 0.008*"mwuk" + 0.008*"clothes" + 0.008*"project" + 0.008*"yeah" + 0.008*"blaze" + 0.006*"love" + 0.006*"help" + 0.006*"days" + 0.006*"boxing" + 0.006*"look" + 0.006*"closed" + 0.006*"donate" + 0.006*"campaign" + 0.006*"year" + 0.006*"meet" + 0.006*"really"

We further used the topic model to generate the topic distribution for each tweet. Table 2 shows an example of the topic distributions for a tweet. The semantic similarity between distinct pairs of tweets were calculated based on the cosine similarity of their distributions over the 15 topics.

Table 2. A distribution over topics for a sample tweet.

Topic #	Topic 1	Topic 2	Topic 3	…	Topic 15
Distribution	0.013	0.013	0.013		0.81

We also computed the Epanechnikov Kernel distance between each pair of tweets. To generate geo-topics based on both semantic similarity and geographical closeness, we set the threshold of semantic similarity as 0.5 and geographical closeness as 0 to generate neighbors of each tweet. Finally, we obtained 24 unique geo-topics. Table 3 shows the different geo-topics, the number of tweets under each geo-topic, and the hub tweet of each geo-topic (tweet with highest sum of weights).

Table 3. Candidate Geo-topics

Geo-topic #	**#Tweets**	**Hub Tweet**
1	**197**	**West London fire: 45 fire engines sent to #grenfell tower, 200 firefighters still tackling the…<url>**
2	179	Welcome back #boroughmarket @ username <url>
3	270	I first met Suzanne when she came to a Moonclub last year. That night she said she wanted to… <url>
4	256	Morning! @ username<url>
5	216	Back in heels for the first time since foot surgery. Feels weird - bringing shoes to change… <url>
6	214	Good morning #London #travel #cats #morningmotivation @ London, United Kingdom <url>
7	228	Chocolate frangipane birthday cake.<url>
8	200	Geraniums are looking @ Stratford, London <url>
9	245	Arches @ London Bridge Station <url>

10	231	#GoesCampingOnce<url>
11	249	Join us Thursday 15th at the Glasshouse in Camden st 6:30pm for a glass of bubbly. RSVP and… <url>
12	223	It's not where is Wally anymore but where is Sandy @ username<url>
13	253	prophetsofrage so good! #killinginthenameof #takethepowerback @ O2… <url>
14	231	Trend Alert: 'Burnley'. More trends at <url>#trndnl
15	258	Percentage I’m going to the park today: 100
16	222	Open again :) (at @BoroughMarket in London, Greater London) <url>
17	199	When your friend has guitar pic and you have only your tongue to show off such a legend… <url>
18	287	Reshape, resurfaced and renovated bmx track is done! @ Brockwell Park <url>
19	277	@username @username Come back to Newcastle and have some more, when you next have a break!
20	194	Trend Alert: 'Carrington'. More trends at <url> #trndnl
21	**194**	**"I'm not with the government". #grenfell tower #London #WestLondon @ London, United… <url>**
22	273	I'm at London City Airport - @username in London, Greater London <url>
23	196	prophetsofrage in @o2academybrix last night! #prophetsofrage #rageagainstthemachine #cypresshill… <url>
24	231	#food #boroughmarket #london#londonbridge what shall I have for brunch <url>

Geo-crisis ranking. We further ranked the 24 candidate geo-topics based on their average intensity of negative sentiment over the four time windows (See Figure 2). The threshold for selecting crisis was set to be 0.2 in this case. We identified two events as geo-crisis: Event #1 and

Event #21 with average intensity of negative sentiment at 0.206 and 0.243, respectively. The hub tweet for both events are related to the Grenfell Tower fire.

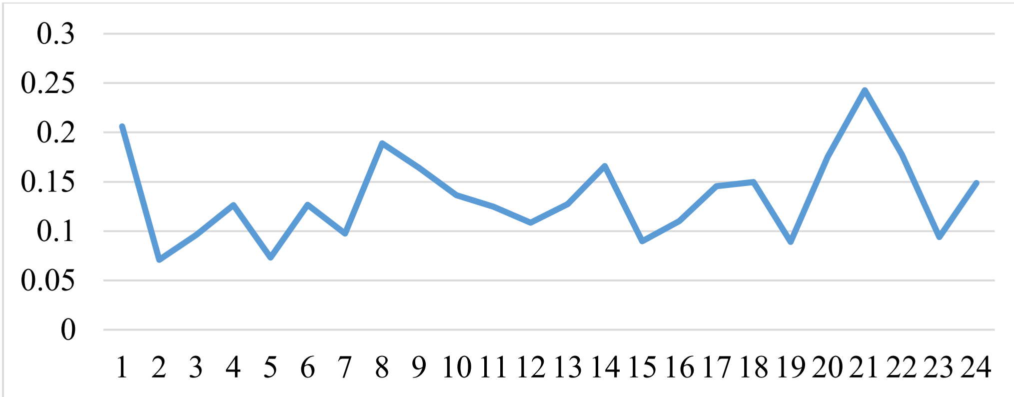

Figure 2. Average intensity of negative sentiment over geo-topics

CONCLUSION

In this paper, we developed and validated a novel technique for detecting urban crises. The technique effectively detected the Grenfell Tower fire in London with short-term geo-referenced tweets. This new technique for detecting urban disasters leverages both semantic and geographic similarity in generating events, and evaluates the crisis level based on the intensity of negative sentiment. The next step of this study is to expand the technique for scalable geo-crisis detection in high-volume tweets. We will also incorporate an online mode to identify emergencies in real time, and add modules for disaster tracking and visualization. We will build an open visualization platform to visualize the evolving crises in a geographical map. Different emergencies will be represented by different keywords and symbols in the interactive map. The open map will provice information regarding the type, content, location and sentiment level of an emergency. After completing the system, we will employ it in detecting urban disasters of different types and forms (e.g. infrastructure failure, earthquake, etc), and compare it with current methods in terms of efficiency and accuracy. The completed platform will also provide first responder and emergency management agencies in disaster-affected areas with an updated understanding of the role geo-referenced social media can play in increasing the effectiveness of disaster response efforts. In doing so, this work is intended to improve disaster and crisis management, to enhance situation awareness, and to take steps toward achieving urban resilience.

ACKNOWLEDGEMENTS

This study is supported by the National Science Foundation under Grant No. 1142379 and a Virginia Tech BioBuild IGEP grant. 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 or Virginia Tech.

REFERENCES

Blei, D. M. (2012). "Probabilistic topic models." *Communications of the ACM*, 55(4), 77-84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent dirichlet allocation." *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Ford, J. D., Tilleard, S. E., Berrang-Ford, L., Araos, M., Biesbroek, R., Lesnikowski, A. C., MacDonald, G. K., Hsu, A., Chen, C., and Bizikova, L. (2016). "Opinion: Big data has big potential for applications to climate change adaptation." *Proceedings of the National Academy of Sciences*, 113(39), 10729-10732.

Gu, Y., Qian, Z. S., and Chen, F. (2016). "From Twitter to detector: real-time traffic incident detection using social media data." *Transportation Research Part C: Emerging Technologies*, 67, 321-342.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., and Cebrian, M. (2016). "Rapid assessment of disaster damage using social media activity." *Science Advances*, 2(3), e1500779.

Li, T., Xie, N., Zeng, C., Zhou, W., Zheng, L., Jiang, Y., Yang, Y., Ha, H.-Y., Xue, W., and Huang, Y. (2017). "Data-driven techniques in disaster information management." *ACM Computing Surveys (CSUR)*, 50(1), 1.

Lu, X., Bengtsson, L., and Holme, P. (2012). "Predictability of population displacement after the 2010 Haiti earthquake." *Proceedings of the National Academy of Sciences*, 109(29), 11576-11581.

Maurya, A., Murray, K., Liu, Y., Dyer, C., Cohen, W. W., and Neill, D. B. (2016). "Semantic scan: detecting subtle, spatially localized events in text streams." *arXiv preprint arXiv:1602.04393*.

Nagy, A., and Stamberger, J. "Crowd sentiment detection during disasters and crises." *Proc., Proceedings of the 9th International ISCRAM Conference*, 1-9.

Nielsen, F. Å. (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." *arXiv preprint arXiv:1103.2903*.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Benevenuto, F., and Gonçalves, M. A. (2015). "SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods." *arXiv preprint arXiv:1512.01818*.

Sakaki, T., Okazaki, M., and Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proc., Proceedings of the 19th International Conference on World Wide Web*, ACM, 851-860.

Schubert, E., Weiler, M., and Kriegel, H.-P. "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds." *Proc., Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 871-880.

Sun, H., Wang, Z., Wang, J., Huang, Z., Carrington, N., and Liao, J. (2016). "Data-driven power outage detection by social sensors." *IEEE Transactions on Smart Grid*, 7(5), 2516-2524.

Towne, W. B., Rosé, C. P., and Herbsleb, J. D. (2016). "Measuring Similarity Similarly: LDA and Human Perception." *ACM TIST*, 8(1), 7:1-7:28.

Walther, M., and Kaisser, M. "Geo-spatial event detection in the twitter stream." *Proc., European Conference on Information Retrieval*, Springer, 356-367.

Wang, Q., and Taylor, J. E. (2015). "Process Map for Urban-Human Mobility and Civil Infrastructure Data Collection Using Geosocial Networking Platforms." *Journal of Computing in Civil Engineering*, 30(2), 04015004.

Wang, Y., and Taylor, J. E. (2017). "Tracking urban resilience to disasters: a mobility network-based approach." *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management* I. e. A. M. M. L. C. H. F. B. T. C. (Ed.), ed. Albi, France., pp. 97–109.

Xie, W., Zhu, F., Jiang, J., Lim, E.-P., and Wang, K. (2016). "Topicsketch: Real-time bursty topic detection from twitter." *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2216-2229.

Yu, W., Li, J., Bhuiyan, M. Z. A., Zhang, R., and Huai, J. (2017). "Ring: Real-Time Emerging Anomaly Monitoring System over Text Streams." *IEEE Transactions on Big Data*.

Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., Wang, S., and Han, J. "Geoburst: Real-time local event detection in geo-tagged tweet streams." *Proc., Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 513-522.