EventPeriscope: a Novel Pipeline to Analyze Impact of Real-World Events on Twitter Discussions

Lida Safarnejad[1 [0000−0002−6377−7314]], Yaorong Ge[1 [0000−0002−9576−0293]], Qian Xu[2 [0000−0002−2354−0208]], and Shi Chen[1 [0000−0002−2316−111X]]

[1] University of North Carolina at Charlotte, Charlotte NC 28223, USA
{lsafarne,yge,schen56}@uncc.edu
[2] Elon University, Elon, NC, 27244
qxu@elon.edu

Abstract. Social media, such as Twitter, have become major hubs for people to seek, post, and share information. The rich interactions among Twitter users are valuable resources for researchers to identify and investigate dynamic public opinions regarding trending topics, such as public health emergencies. In this paper, we focus on the direct effect of real-world events on changing the dynamics of discussions on social media. This research topic has not received much attention in the field today. However, we believe it is a significant one because a good understanding of the events that impact social media discussions will allow more effective communication of important messages such as public health warnings that are critical to public safety and population health. This paper presents EventPeriscope, a pipeline to systematically analyze Twitter discussions in order to discover the effect of a particular real-world event on changing the dynamics of discussions. EventPeriscope utilizes and seamlessly combines signal processing, text mining, and natural language processing techniques to quantify the effects of real-world events in four sequential modules: signal constructor, peak detector, content analyzer, and visualizer. Moreover, this pipeline is capable of identifying additional closely-related events that have also invoked the discussions. We provided two case studies of real-world events, the WHO announcement of Zika as a public health emergency of international concern (PHEIC), and Rio 2016 Olympic Games. We have demonstrated how EventPeriscope helps us investigate the dynamic impact of these two real-world events on the Zika discussion on Twitter in 2016. Therefore, EventPeriscope can be adopted in studying other topics on various social media platforms.

Keywords: Events detection · Social media · Twitter · Signal processing · Content analysis.

1 Introduction

Social media, a fast, inexpensive and omnipresent channel, has become an inseparable part of people's daily life. A growing number of people around the world

are joining social media to get information and share their opinions. Among all social media platforms, Twitter, a microblogging service with 100 million active users posting 500 million tweets every day [8], has been one of the main targets of practitioners and decision makers to study public opinion and foster public relation. At the same time, politicians, health agencies, marketers, and celebrities, to name but a few, are considering Twitter as one of the main conduits to communicate with their audiences. To effectively utilize Twitter as a means of communication with the public, one key factor is having awareness of ongoing discussions and debates among users of this platform. In other words, Twitter needs to be monitored to discover users' needs, interests, and concerns, and also stimuli that provoke users' attention and engage them in discussions. To address this problem, various event detection methods have been proposed, such as [14,2], that mainly monitor users' tweeting activities to detect trending topics discussed among users on social media. One important yet overlooked aspect is to investigate whether and how much real-world events are able to stimulate and elevate discussions on social media.

In this paper, we propose an analytical pipeline, EventPeriscope, to reveal and quantify the impact of real-world events on the dynamics of Twitter discussions. It can be used to analyze both planned (with a pre-determined time/location that people have been aware of) and unplanned (without *a priori* notice) events [11]. For example, Olympic Games are planned events that the public is aware of a long time before their occurrences; while an announcement by a government agency is usually unplanned in the sense that people do not anticipate it.

To investigate the impact of a particular real-world event on discussions on Twitter, EventPeriscope models the input tweet stream as a signal to capture the temporal changes in the number of tweets, and uses Wavelet Transform to detect peaks in the constructed signal. If a peak is detected in the close proximity of the event occurrence time, textual contents of the tweets posted around the event time are analyzed using text mining and natural language processing (NLP) techniques. The result of this process is used to generate regular expression (regex) rules that describe the event. The tweets matched with the regex rules are then analyzed and their time series is visualized to quantify the contribution of the event on Twitter discussions.

In the following sections, we first review the most relevant research works to this paper. Section 3 presents our proposed pipeline, EventPeriscope, in details. Section 4 then reports the results of employing EventPeriscope to analyze the effect of two real-world events on the dynamics of Twitter discussions regarding Zika outbreak in 2016. Finally, we conclude the paper in section 5.

2 Related Work

Proposed methods and frameworks in the area of event detection, generally rely on signal processing, Natural Language Processing (NLP), and text mining techniques to analyze social media discussions. [11], [5]. In this section, we briefly review some of these research works.

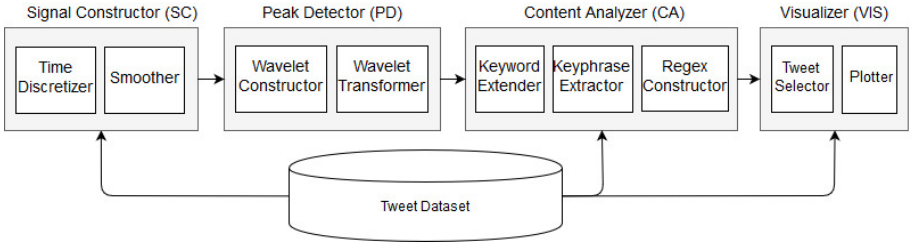

Fig. 1: EventPeriscope Pipeline.

In [14], Weng *et al.* model words occurrence in tweet streams using signals to capture the temporal changes in their appearance. Afterward, they use cross correlation to find similarity between word-specific signals and therefore determine events. In a close work [2], Cordeiro *et al.* consider bursts in hashtags occurrence, specified by wavelet signals, to be events. Next, Latent Dirichlet Allocation (LDA) topic modeling is performed to uncover the topic of them. In [3], a pipeline is proposed to detect and visualize events. They construct a time series from texts related to a topic within a specific time window. Then, the cumulative sum control chart is utilized to spot temporal changes in the topic-specific signals. With a purely NLP approach, Ren *et al.* in [12] propose an LDA based topic modeling and adopt Support Vector Machine for Twitter sentiment classification. In a close work, authors in [6] cluster tweets using Locality Sensitive Hashing technique. As previously discussed, current studies mainly investigate Twitter discussions to uncover events that have sparked or heated the discussions. In comparison, our proposed pipeline has been designed to examine the effect of a real-world event on Twitter discussions.

3 Methodology

In this section, we present an analytical pipeline, EventPeriscope, designed to systematically examine the influence of real-world events on Twitter discussions. Fig. 1 demonstrates the overall architecture of EventPeriscope, which has four main modules: Signal Constructor (SC), Peak Detector (PD), and Content Analyzer (CA), and Visualizer (VIS).

To capture temporal dynamics of discussions of a specific topic on Twitter, SC module constructs a signal from the tweets posted in a particular time window specified by the user. The constructed signal is then passed as an input to the PD module. This module uses the wavelet transform to locate peaks in the input signal. Observing a peak in close proximity to the time point when the hypothesized real-world event happens is necessary but insufficient to conclude that the event has directly caused the discussions. CA module examines tweets posted within a short time interval, specified by the user, around the time when the event of interest occurred to create a set of regular expression (regex) rules to detect tweets discussing the event. Afterward, all tweets in the tweet stream

dataset are tested against the regex rules to find all matched tweets, and subsequently, estimate the percentage of tweets related to this event. In the following sections, each module is explained in detail.

3.1 Signal Constructor

To construct a signal from a tweet stream, first, the time interval between the first tweet and the last tweet in the stream is partitioned into fixed-length time slots or bins, and a signal is created from tweet counts in each bin. We use two attributes, magnitude and width, to characterize a peak in the tweet signal. In other words, we are interested in rises that are above a certain threshold in the signal and also persist for more than a pre-specified time interval. Therefore, the signal is smoothed to filter out small oscillations that last for a short period of time. For Smoothing, we employ Kernel Density Estimation (KDE) [13]. KDE is a smoothing technique which estimates the value of each data point by the average or weighted sum of its value and the values of its neighboring data points. For every data point, its new value $\hat{f}(x)$ is calculated by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{1}$$

where $K(.)$ is a kernel function, and h is the smoothing parameter. h determines smoothness of a curve. Based on the shape of the peaks observed in the constructed tweet signal, we use Gaussian kernel for smoothing:

$$K(U) = \frac{1}{\sqrt{2\pi}} \exp(u^2), \tag{2}$$

The resulted smoothed signal is then passed to PD module to locate peaks.

3.2 Peak Detector

PD module utilizes the wavelet transform to detect peaks in the signal constructed by SC. In the wavelet transform, wavelet coefficients associated with a function $f(x)$, the tweet signal in our case, are calculated by

$$W_f(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(x) \overline{\Psi(\frac{x - b}{a})} dx, \tag{3}$$

where $\overline{\Psi(t)}$ denotes the complex conjugate of the mother wavelet $\Psi(t)$, a is the scale, and b is the time shift. Coefficients at different scales and time shifts are represented by a matrix. When a peak appears in the signal, maximum values of different scales occur at close time shifts, resembling a ridge in the coefficient matrix [4]. In this paper, we adopt Mexican hat wavelet as the mother wavelet:

$$\Psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}} (1 - (\frac{t}{\sigma})^2) \exp(\frac{-t^2}{2\sigma^2}), \tag{4}$$

where σ is the scale.

3.3 Content Analyzer

CA module interactively constructs a set of regular expression (regex) rules describing an event. Subsequently, it utilizes the constructed rules to retrieves all tweets in the stream that are related to the event of interest.
CA module requires two user-provided inputs: 1) a set of keywords, which can be single words, hashtags, or mentions, related to the real-world event of interest, and 2) a time interval around the event that CA module uses for its analysis. Keyword Extender in CA module extends the input keyword set by examining tweets posted within the specified time interval to find other keywords that have the highest correlation with the user-specified keywords. To this end, Keyword Extractor uses Pointwise Mutual Information (PMI) to quantify the association between keywords. PMI is a statistical approach for measuring the level of dependency between two observations [10], in our case two words. PMI between two words w_1 and w_2 is calculated by:

$$PMI\left(w_1,\ w_2\right) = log\ \left(\frac{p\left(w_1, w_2\right)}{p\left(w_1\right) \times \left(w_2\right)}\right), \tag{5}$$

where $p(.)$ is probability function. Two words that have a high PMI value are strongly associated with each other. In other words, they co-occur frequently. Keywords are sorted based on their PMI value and a set of keywords with high PMI value are selected. Keyphrase Extractor extracts keyphrases containing the keywords generated by Keyword Extractor. In this step, we adopt the NLP technique [7]. First, textual contents of tweets are lemmatized and tagged with a Part-of-Speach (POS) tagger. Then, Keyphrase Extractor utilizes a predefined context-free grammar [7], to retrieve all noun phrases. Next, noun phrases containing the event-relevant keywords are extracted and sorted based on their frequencies. The user will be provided with this list to choose keyphrases that describe the event. Matching all tweets in the dataset with selected keyphrases to find relevant tweets is technically challenging as such keyphrases can be rewritten in multiple forms and many of the possible forms may not be presented in the short time interval that is considered to develop event-relevant keyphrases. For instance, spaces between words in a keyphrase might be dropped (e.g. OlympicGames), spaces might be replaced by other characters such as dot or dash (e.g. Olympic.Games), or other words might be inserted between the words (e.g. Olympic Summer Games). To handle such variation and combinations, Regex Generator constructs a set of regular expressions (regex) rules based on the selected keyphrases. Moreover, the user can customize the generated rules.

3.4 Visualizer

VIS module utilizes the resulted regex rules to capture tweets relevant to the event of interest. To be more specific, VIS tests all tweets in the dataset against these regex rules, and constructs time series from the number and percentage of matched tweets. The constructed event-specific time series demonstrate when the event appeared in the discussions and what percentage of discussions were

dedicated to this event. Moreover, peaks in these time series indicate the time points when the event had elevated discussions. Analysts can then carry out more investigation to uncover the underlying reason for the observed rises.

4 Empirical Evaluation

To demonstrate the the feasibility of EventPeriscope, we examine the impact of two distinct types of real-world events on Twitter discussions during Zika outbreak in 2016:

- WHO-PHEIC: World Health Organization (WHO) declared that Zika outbreak was a Public Health Emergency of International Concern (PHEIC) (Feb 1, 2016) [15]
- RIO2016: Rio 2016 Summer Olympics (Aug. 5-21, 2016)

To this end, we retrieved approximately 6 million English tweets (with all associated metadata, such as retweet status, tweet created time, tweet content) regarding Zika during the 2016 outbreak, of which approximately 4 million were original posts, through the Gnip API.
Fig. 2 shows the signal constructed from posted Zika-related tweets. The red dots are the peaks detected by PD module. The vertical dash lines indicate the dates on which the above critical events occurred. As shown in the figure, the tweeter signal peaked almost immediately after WHO-PHEIC event on day 32 (Feb. 1, 2016). For RIO2016, the signal peaked just before the opening ceremony day of Rio 2016 on day 218 (Aug. 5, 2016), and again on Aug. 21 (day 234), the last day of the Olympics.

4.1 WHO-PHEIC Event

On Feb. 1, 2016, Director-general of WHO, Margaret Chan, declared a PHEIC because of the potential Zika pandemic, especially in South America [15]. In this statement, in addition to raising concern over the possible linkage of Zika virus with microcephaly and other neurological disorders, WHO provided advice about

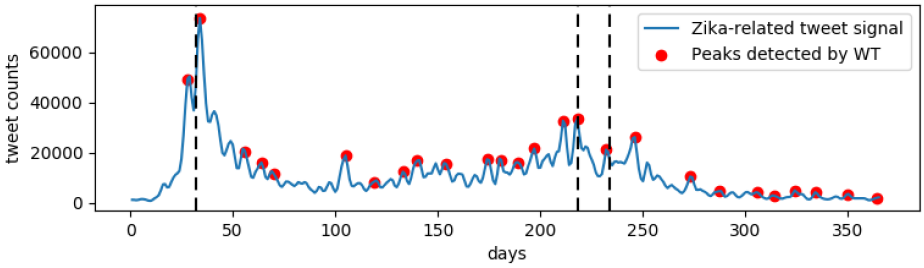

Fig. 2: The blue signal shows Zika tweets posted in 2016. Red points are peaks detected by the wavelet transform. Dash lines depict the two events WHO-PHEIC and RIO2016.

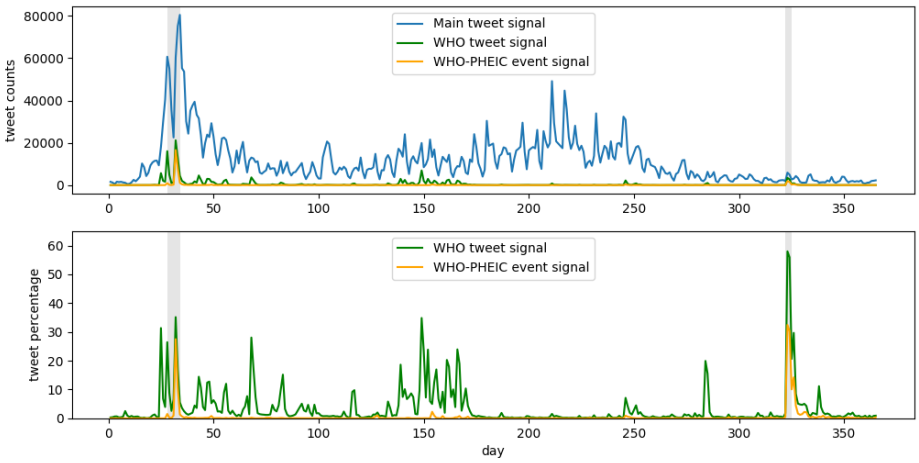

Fig. 3: **Upper Panel:** The blue curve depicts Zika-related tweet counts in 2016. The green and orange curves represent the number of tweets containing WHO and WHO-PHEIC keyphrases, respectively. **Lower Panel:** Green and Orange curves show the percentage of tweets containing WHO and WHO-PHEIC keyphrases, respectively.

Zika virus transmission, and travel measures. WHO-PHEIC announcement was an unplanned event. The public did not have *a priori* knowledge of its occurrence. Thus, we did not expect to see a high volume of tweets related to this event before its occurrence. We initiated CA module with the set of tweets posted in a two-day interval; the day of WHO-PHEIC and one day after. Considering the statement, we let the initial set of keywords describing WHO-PHEIC event be $\{\#who, \#pheic\}$. Keyword Extractor in CA module calculates PMI between each of these keywords and all the keywords extracted from the initial tweet set, and sort them from the largest PMI to the smallest. We selected a few keywords with the highest PMI values and added them to the initial set. The following is the resulted set: $\{\#who, \#pheic, emergency, public, international, global, world, health\}$.

Next, Key Extractor extracted keyphrases, such as *public health emergency*, *global emergency*, and *world health*, containing these keywords. Based on the keyphrases, Regex Constructor crafted two regex rules. One represented the keyword *WHO* and its variants, such as *W.H.O* and *World Health Organization*. The other regex rule described the WHO-PHIC event. Next, Visualizer tested all tweets in the tweet stream dataset with the regex rules to catch those talking about *WHO* or WHO-PHEIC, and created daily time series for each of them.

In the upper panel of Fig. 3, the blue curve shows daily tweet counts in the Zika-related tweet stream. The green and orange curves show tweet counts related to WHO and WHO-PHEIC, respectively. In the blue curve, a sudden rise is clearly observable between day 31 and 32. The number of tweets about Zika increased drastically from 1481 on day 31 to 21171 on day 32, when WHO declared PHEIC

for Zika outbreak.
The green and orange curves in the lower panel of Fig. 3 represents the percentage of tweets containing WHO and WHO-PHEIC keyphrases. In day 32, 35% of Zika-related tweets talked about WHO and 27% of Zika-related tweets were about the PHEIC announcement. Based on the WHO signal, WHO had a strong presence in Zika discussions on Twitter through the first two quarters of 2016. In addition to the peak near day 32, WHO-PHEIC signal had another sharp peak around day 323 (Nov 18). On this day, about 32% of the Zika-related tweets were WHO-PHEIC related. After further investigation, we realized that on Nov 28, WHO declared that Zika was no longer a PHEIC. This can justify the peak observed on this day.
Further investigations revealed that the PHEIC statement also caused cascading announcements by various governmental agencies in countries such as Brazil, Honduras and the United States; this resulted in the highest number of tweets (80000) posted on a single day regarding Zika in 2016. It is noteworthy that the discussion about PHEIC had started on Jan 28, when the director-general of WHO announced that she convened International Health Regulations (IHR) emergency committee and would have a meeting on Feb 1 [16].

4.2 RIO2016 Event

Rio 2016 Olympic Games was held from Aug. 5- Aug. 21, 2016 in Rio de Janeiro, Brazil, amidst global concerns on Zika outbreak. In November 2015, Brazilian authorities declared a national public health emergency due to a high rate of confirmed Zika cases [9].
RIO2016 can be considered as a planned event which users anticipated before it happened. Thus, we expected to see the trace of this event on tweets a few days before its beginning. We initialized CA with tweets posted from August 4

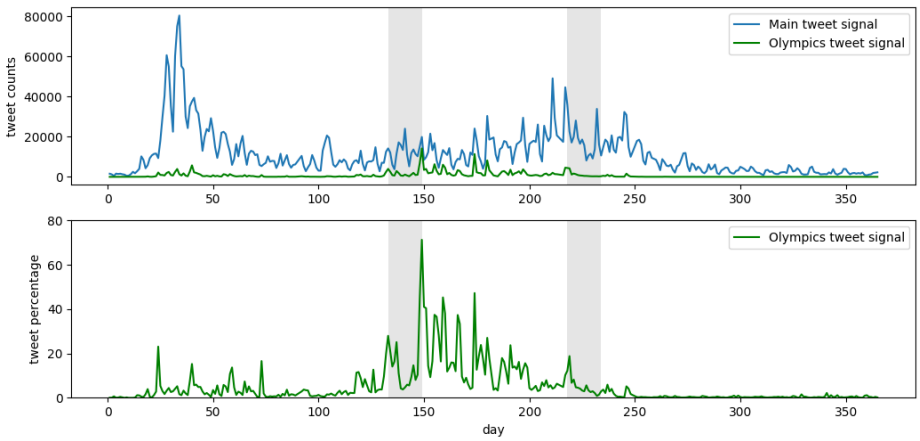

Fig. 4: The green signal in the lower panel represents the percentage of tweets related to Rio 2016 Olympic Games.

to 6 (day 217 to 219). With a similar procedure for the WHO-PHEIC event, a regex rule describing RIO2016 event was crafted. Using the regex rule, Visualizer then extracted all tweets matched with it and created two plots from them; green curves in the upper and lower panel of Fig. 4. The lower one depicts the portion of tweets that were related to the Olympics on each day in 2016. It can be seen that the discussion about Zika and its implications on the Olympics started from the beginning of 2016 until a few days after the Olympics ended. We believe that tweets about the Rio 2016 Olympic Games had been started in 2015; however, we did not have enough data to verify this. On the Olympics opening ceremony day (Aug. 5), 12% of all Zika tweets were Olympics-related, and it increased to 18% the next day. From the lower plot in Fig 4, it is observable that RIO2016 had a noticeable presence in other time points. As an example, on May 28 (day 149), RIO2016-related tweets constituted 71% of all Zika tweets. Our further investigation to uncover the underlying reason showed that On May 12, a letter [1] was published proposing the idea of changing the location of the RIO 2016 Olympic Games, postponing or canceling it. On May 28, WHO released a statement [17] explaining such an action was not necessary. Coincidentally, the WHO statement on May 28 (day 149) also caused the WHO-related tweet signal to reach 34% of the all Zika tweets as shown in Fig. 3.

5 Conclusion

In this paper, we presented a novel analytical pipeline, EventPeriscope, to investigate the effects of real-world events on the dynamics of discussions on Twitter. EventPeriscope characterizes temporal dynamics of Twitter discussions as a signal. The wavelet transform is then applied to detect peaks in the signal. Next, using text mining and NLP techniques, EventPeriscope examines the textual contents of tweets posted around the identified peaks and quantifies the association between the events and discussions on Twitter. Using EventPeriscope, we conducted two case studies to explore the effect of two real-world events, WHO PHEIC announcement and Rio 2016 Olympic Games, on Twitter discussions about Zika outbreak in 2016. Our results show that the proposed pipeline is highly effective in measuring the contribution of these events to the dynamics of Twitter discussions. It also helps us uncover other hidden real-world events closely related to the events under study. In the future, we plan to perform a careful evaluation of EventPeriscope components, such as peak detector, keyphrase extractor, and regex constructor, by studying more ongoing real-world events and measuring their impacts on discussions in various domains such as politics, economics, and sociology.

References

1. Attaran, A., et al.: Off the podium: why public health concerns for global spread of zika virus means that rio de janeiro's 2016 olympic games must not proceed. Harvard Public Health Review (2016)

2. Cordeiro, M.: Twitter event detection: combining wavelet analysis and topic inference summarization. In: Doctoral symposium on informatics engineering. pp. 11–16 (2012)
3. Dou, W., Wang, X., Skau, D., Ribarsky, W., Zhou, M.X.: Leadline: Interactive visual analysis of text data through event identification and exploration. In: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. pp. 93–102. IEEE (2012)
4. Du, P., Kibbe, W.A., Lin, S.M.: Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics **22**(17), 2059–2065 (2006)
5. Hasan, M., Orgun, M.A., Schwitter, R.: A survey on real-time event detection from the twitter data stream. Journal of Information Science p. 0165551517698564 (2017)
6. Kaleel, S.B., Abhari, A.: Cluster-discovery of twitter messages for event detection and trending. Journal of Computational Science **6**, 47–57 (2015)
7. Kim, S.N., Baldwin, T., Kan, M.Y.: Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In: Proceedings of the 23rd international conference on computational linguistics. pp. 572–580. Association for Computational Linguistics (2010)
8. Liew, J.K.S., Wang, G.Z.: Twitter sentiment and ipo performance: a cross-sectional examination. Journal of Portfolio Management **42**(4), 129 (2016)
9. Lowe, R., Barcellos, C., Brasil, P., Cruz, O., Honório, N., Kuper, H., Carvalho, M.: The zika virus epidemic in brazil: from discovery to future implications. International journal of environmental research and public health **15**(1), 96 (2018)
10. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. Association for Computational Linguistics (2010)
11. Panagiotou, N., Katakis, I., Gunopulos, D.: Detecting events in online social networks: Definitions, trends and challenges. In: Solving Large Scale Learning Tasks. Challenges and Algorithms, pp. 42–84. Springer (2016)
12. Ren, Y., Wang, R., Ji, D.: A topic-enhanced word embedding for twitter sentiment classification. Information Sciences **369**, 188–198 (2016)
13. Silverman, B.W.: Density estimation for statistics and data analysis. Routledge (2018)
14. Weng, J., Lee, B.S.: Event detection in twitter. ICWSM **11**, 401–408 (2011)
15. (WHO), W.H.O.: Who statement on the first meeting of the international health regulations (2005) (ihr 2005) emergency committee on zika virus and observed increase in neurological disorders and neonatal malformations. https://bit.ly/2t7Imzr (2 2016), accessed January 2019
16. (WHO), W.H.O.: Who statement on the first meeting of the international health regulations (2005) (ihr 2005) emergency committee on zika virus and observed increase in neurological disorders and neonatal malformations. https://bit.ly/2WJjE6b (2 2016), accessed January 2019
17. (WHO), W.H.O.: Who statement on the first meeting of the international health regulations (2005) (ihr 2005) emergency committee on zika virus and observed increase in neurological disorders and neonatal malformations. https://www.who.int/en/news-room/detail/28-05-2016-who-public-health-advice-regarding-the-olympics-and-zika-virus (2 2016), accessed January 2019