Rumour Detection on Social Media for Crisis Management

Sooji Han
Department of Computer Science,
University of Sheffield,
Sheffield, S1 4DP, UK
sooji.han@sheffield.ac.uk

Fabio Ciravegna
Department of Computer Science,
University of Sheffield,
Sheffield, S1 4DP, UK
f.ciravegna@sheffield.ac.uk

ABSTRACT

We address the problem of making sense of rumour evolution during crises and emergencies. We study how understanding and capturing emerging rumours can benefit decision makers during such event. To this end, we propose a two-step framework for detecting rumours during crises. In the first step, we introduce an algorithm to identify noteworthy sub-events in real time. In the second step, we introduce a graph-based text ranking method for summarising newsworthy sub-events while events unfold. We use temporal and content-based features to achieve the effective and real-time response and management of crises situations. These features can improve efficiency in the detection of key rumours in the context of a real-world application. The effectiveness of our method is evaluated over large-scale Twitter data related to real-world crises. The results show that our framework can efficiently and effectively capture key rumours circulated during natural and human-made disasters.

Keywords

Rumours, large-scale data, event summarisation, sub-event detection, social media analysis

INTRODUCTION

Nowadays social media platforms are main sources of a variety of information with rapidly growing rates of user engagement and global mobile social media usage. According to the Global Digital Report 2018, 3.196 billion users are using social media. Along with the growth of social media, its role during emergencies and crises has become prominent (Andrews et al. 2016; Arif et al. 2017; Castillo 2016; Imran et al. 2015; Rudra et al. 2018; Zeng et al. 2016). For instance, emergency services can remotely identify areas affected by crises situations based on social media users' posts reporting what they are seeing and hearing (Yin et al. 2012) or find victims seeking help (Zubiaga et al. 2018a). Emergency responders can then make adequate decisions such as the allocation of resources and police. The enormous and continuing growth of user activities on social media is accompanied by the real-time acquisition of a substantial amount of user-generated data with an unprecedented volume and variety and different levels of veracity. As the speed of event evolution is very fast, the existence of such information in the real-time stream of social media content poses a challenge to decision makers, in particular, in time-sensitive situations such as natural disasters and terrorist attacks. It is impossible for humans to manually inspect rapidly changing and noisy messages in a timely manner. In addition, citizens can be confused by unfounded or conflicting rumours regardless of whether they were generated purposely or unintentionally.

Research areas using social media that have been in the limelight recently are *automated rumour and fake news detection* (Helmstetter and Paulheim 2018; Kwon et al. 2017; Shu et al. 2017, 2018; Wong et al. 2018) and *fact-checking* (Boididou et al. 2018; Vosoughi et al. 2017). Social media are origins of rumors and where they spread among a large number of people (Ma et al. 2018). Despite the growing interest in rumours and fake news on social media and several attempts to make sense of them in both academia and industry, the utilisation of rumours for emergency response poses several challenges (Andrews et al. 2016; Starbird et al. 2016). In this study, we aim at partially filling this gap by studying how to identify rumours appearing during crises without manually examining a large number of messages. Our framework comprises of real-time key sub-event detection and summarisation. Summaries obtained by our method can include both rumours and non-rumours. According to our experiments, most summaries contain informative details related to emergencies and reflect what draws the public's attention.

Our system can help practitioners identify not only newsworthy events but rumours to debunk or verify without examining an entire stream of messages.

In this paper, we address the problem of the early and semi-automatic identification of *rumours* on social media during crises. To this end, we first analyse rumours during emergency situations. We adopt the definition of rumour proposed by (DiFonzo and Bordia 2007): "an unverified and instrumentally relevant information statement in circulation that arises in contexts of ambiguity, danger or potential threat, and that functions to help people make sense and manage risk". We refer to (Olteanu et al. 2015) to categorise rumours by type. We claim that some information types such as 'donations and volunteer', 'caution and advice', and 'sympathy and emotional support' can contribute to situational awareness but cannot be considered as rumours. Our scope of rumour during crises include a) affected individuals, infrastructure, and utilities, b) the location of the event, c) other useful information (e.g. details about suspects and weapons). We use Twitter as a representative social media.

To help decision makers and journalists identify breaking news or understand unfolding situations as early as possible, we pose the following research question: "Which moments in event evolution are worth being paid attention to from the perspective of emergency responders?" To address the question, we introduce "noteworthy moment detection" into the framework of rumour studies. Several studies have studied bursts in the popularity of events on social media (Hu et al. 2017; Keneshloo et al. 2016; Kong et al. 2015; Matsubara et al. 2012; Wang et al. 2016). The main focus of such studies is often a 'peak'. However, a peak indicates that a rumour has already become popular (i.e. a large amount of information regarding the rumour is available). This phenomenon makes it difficult for professional analysts to take action to deal with the rumour in a timely manner. We emphasize the importance of capturing bursts of user activity along with peaks, and collectively refer to both types of events as *noteworthy moments.* The identification of such moments in rumour evolution without domain knowledge is often challenging and tricky as a boundary between noteworthy and normal moments may not be obvious to a lay person. Noteworthy sub-events are not necessarily equivalent to novelties concerning their content (i.e. topics) as considered in the area of first story detection, but they are considered to be novel in temporal context.

Having identified noteworthy sub-events, our proposed framework generates summaries of a high volume of messages. There are commonly two approaches for event summarisation. One is generating a whole new text which contains important information. The other is extracting representative messages that have high scores. In this work, we choose the latter approach. Several studies on crisis management via social media have proposed sub-event summarisation methods (Kedzie et al. 2015; Nguyen et al. 2015; Rudra et al. 2015, 2016, 2018). The state-of-the-art work on event summarisation usually focuses on achieving higher Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores (Lin 2004). ROUGE is a metric which evaluates the quality of a summary by comparing overlapping n-grams between the summary and gold standard references. Our framework also consists of two steps: noteworthy moment detection and sub-event summarisation. What differentiates our study from the state-of-the-art work is that we undertake an extensive analysis of rumours that emerge and evolve during crises. We believe that understanding rumour evolution during crises can benefit both the improvement of situational awareness during crises and diverse realms of rumour studies. Attempts to understand differences between true information and false rumours have been extensively studied in the community of rumour studies. Crisis management via social media can benefit from findings of rumour studies. Meanwhile, examples of crises and emergencies can be useful resources for rumour studies.

The main contributions of our work are summarised as follows:

1. We propose a semi-automatic framework for the early detection and summarisation of key sub-events. Our framework can identify potential rumours that draw people's attention before their popularity reaches a peak. Our framework makes it possible for decision makers to understand developing situations quickly and to take action proactively. This can lead to the minimization of the extent of the damage.
2. We demonstrate that our framework can work with events with different types of language and different sizes. We conduct experiments over real-world datasets that mainly consist of English, Italian and Russian tweets.
3. We evaluate our framework's effectiveness quantitatively and qualitatively by comparing different methods for key sub-event detection.
4. Our framework generates readable and meaningful summaries that cover a wide range of information so that end users who do not have background knowledge on models can easily understand unfolding events.

REQUIREMENTS AND SYSTEM DESIGN

- **Early detection**: A system must be able to capture informative reports in the early stages of event evolution. This makes it possible for decision makers to be prepared for potential hazards in the near future.

- **Real-time detection**: The speed of the dissemination of information during crisis is very fast. A system must be able to process a stream of posts in real time so that decision makers can plan urgent and planned remedial actions in a timely manner.
- **Scalability**: Rumours during crises and emergencies are usually unexpected, sudden, and unusual. One major drawback of using a threshold to identify key moments such as emerging patterns and spikes in event evolution is that system performance is highly dependent on data sizes.
- **Generalisation**: A system must be transferrable to events from different domains and with different characteristics (e.g., language, burstiness, and geolocation).

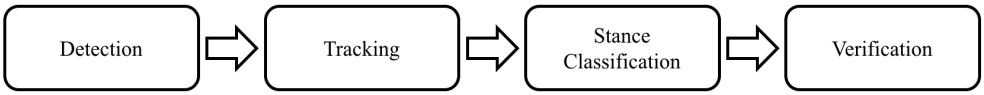


Figure 1. Architecture for rumour studies.

Figure 1 shows a commonly used architecture in rumour studies (Zubiaga et al. 2018b). Systems for rumour detection perform binary classification to decide whether a piece of information (e.g. a single tweet) is a rumour or non-rumour. In general, tweets during crises can be collected using a set of keywords related to an event or geolocation. For example, relevant keywords such as 'bostonexplosion', 'bostonbombing', 'bostonblast', and 'prayforboston' can be identified at the early stage of the Boston marathon bombings. In particular, the hashtag starting with 'prayfor' is very popular worldwide and usually appears when tragic events take place. Due to this data collection procedure, there can be lots of irrelevant or uninformative posts in a stream. Therefore, we argue that it is impractical to apply a rumour detection system to a raw stream of posts. Our framework will bridge the gap between real-time data collection and the automatic detection of breaking news and rumours.

We further discuss why 'peaks' are not appropriate signals for key moments detection. Let us think about the task of determining whether the time window marked in square in Figure 2a is noteworthy or not. This task may become less challenging if we observe the next time window. For example, we can confirm that the time window marked in square is a key moment without doubt in Figure 2b because it is a 'peak'. On the other hand, it is still challenging to decide whether the target window is noteworthy or not in Figure 2c. The target lies on the increasing line of the graph. However, the amount of increase is not significant compared to that between the time windows marked in square and star. By looking at this simple example, we can draw two conclusions. Firstly, it is impossible to identify a peak without observing future instances. This means the detection of peaks in event evolution cannot be done in real time. This is why several studies on key moment detection employ retrospective approaches for detecting peaks. Secondly, the term 'noteworthy moment' can be considered an anomaly of interest to analysts or domain experts. As several studies on anomaly analysis point out, it is subjective to judge whether a point in the evolution of an event can be considered to be noteworthy or not (Aggarwal 2015). The distinction between noteworthy and non-noteworthy moments could be regulated based on the interest of analysts.

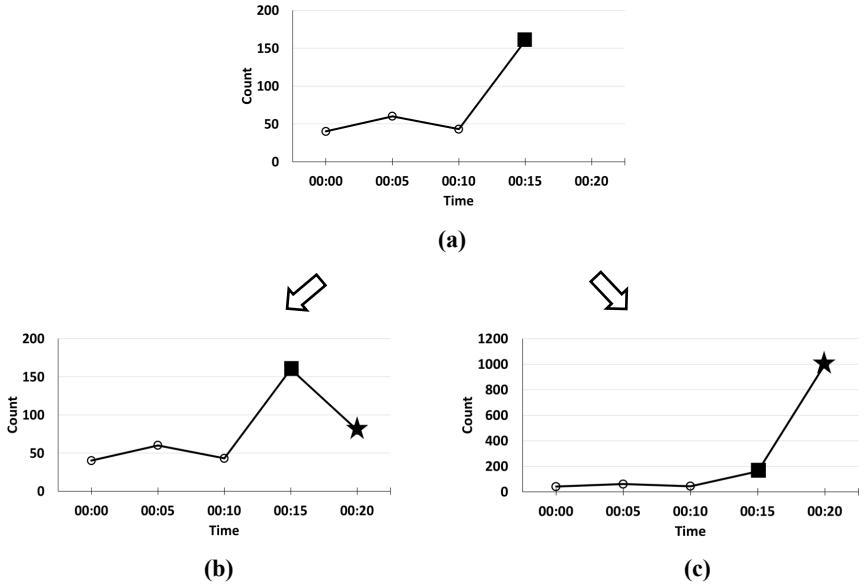


Figure 2. Examples of event evolution.

Previous studies on key sub-event detection utilize thresholds. Specifically, their methods recognize sub-events the volume of messages of which exceeds a certain threshold. A drawback of most existing methods based on thresholds is that they tend to capture a large number of instances in periods of persistent decline. However, we observe that these instances are generally spurious anomalies in the context of content novelty. We are interested in detecting newly emerging or developing stories rather than already seen sub-events with a high volume of messages. Our observation shows that the same topic can be actively discussed even after its popularity reaches a peak for several reasons such as the time difference between locations of users. Instances located in a persistently falling line are not of interest because they are likely to be duplicates of key sub-events in the past. We will further investigate this argument using real-world datasets. We design our architecture for *event-level* studies of crises in the real world. We aim to aid tweet-level rumour studies by reducing data granularity. We propose our method as a pre-processing step for rumour detection in crises and emergencies. The output of our framework can be used as *candidates* for the construction of annotated data for rumour detection. The state-of-the-art work on rumour detection annotated only highly retweeted tweets (Ma et al. 2017; Zubiaga et al. 2016b; a). This data sampling may lead to biased results for rumour detection as some minor but important sub-events do not get a high number of reactions (Meladianos et al. 2015). The fine-grained detection, tracking, stance classification, and verification of rumours are beyond the scope of this paper.

NOTEWORTHY MOMENT DETECTION

We propose a rule-based algorithm for detecting noteworthy moments in quasi real-time. Our method identifies key moments by solely using the left-hand side of any instances (i.e. time window) of a time series. Table 1 describes features associated with the number of tweets. We use the Boston marathon bombings to empirically choose our parameters. Specifically, we visualise key moments detected using different combinations of parameter values and select a set of parameters which satisfies best our hypothesis: noteworthy moments lie in increasing lines in even evolution graphs. We set α for EWM_MEAN to 0.3, S for DIFF_RMEAN to 3, and λ for ψ to 0.01. Recall that it is subjective to judge whether a point in the evolution of an event can be considered to be noteworthy or not. Therefore, the used parameter values are rules of thumb. Even so, our experiments show that they are scalable and generalize well to datasets with different sizes and from different types of crises.

Table 1. Features extracted from the tweets within the i^{th} time window.

Name	Description	Mathematical Equation
COUNT	The number of tweets	c_i
EWM_MEAN	Exponentially weighted average of COUNT	$s_i = \begin{cases} c_i & i = 1 \\ (1-\alpha) * s_{i-1} + \alpha * c_i & i = 2, \dots, N \end{cases}$, where $0 < \alpha \leq 1$
DIFF	Difference between s_i and s_{i-1}	$d_i = \begin{cases} 0 & i = 1 \\ s_i - s_{i-1} & i = 2, \dots, N \end{cases}$
DIFF_RMEAN	Rolling mean of d with a window size of S	$z_i = \begin{cases} 0 & i = 1 \\ \frac{\sum_{k=1}^{\min\{S, i-1\}} d_{i-k}}{\min\{S, i-1\}} & i = 2, \dots, N \end{cases}$
STDEV	Sample standard deviation of 4 time windows	$std_i = \begin{cases} 0 & i = 1 \\ \sqrt{\frac{\sum_{k=0}^{\min\{3, i-1\}} (s_{i-k} - \bar{s})^2}{\min\{i, 4\} - 1}} & i = 2, \dots, N \end{cases}$, where $\bar{s}$ is the mean of s_i.
θ	A combination of EWM_MEAN and STDEV	$\theta_i = \begin{cases} 0 & i = 1 \\ s_{i-1} + std_i & i = 2, \dots, N \end{cases}$
ψ	A combination of DIFF and DIFF_RMEAN	$\psi_i = \begin{cases} 0 & i = 1 \\ \lambda * z_{i-1} + (1-\lambda) * d_i & i = 2, \dots, N \end{cases}$

We formulate noteworthy sub-event detection into a binary classification task. Algorithm 1 describes the rule-based procedure of identifying noteworthy moments using the above features. In order to simulate a real-time scenario, we assume that a time series consists of N time windows. Given the N instances with the features, the algorithm assigns binary labels to time windows based on a set of conditions. A label is 1 if a window is noteworthy and is 0 otherwise. Some existing studies define thresholds for the absolute number of messages of each time window (Gillani et al. 2017; Shamma et al. 2009; Zubiaga et al. 2012). Others define thresholds for differences in the number of messages over several time windows (Nichols et al. 2012; Peng et al. 2018). Our algorithm overcomes several limitations that existing methods for key sub-event detection have. Firstly, results of the state-of-the-art approaches for sub-event detection are highly dependent on the total number of tweets posted during event evolution. This poses a challenge because the performance of a detection method can be poor if an inappropriate threshold is chosen. To overcome this, we adjust thresholds in the early stages of event evolution (LINE 3-7). As rules of thumb, we set H to 24 for instantaneous events and to 36 for progressive events. In other words, for instantaneous events, thresholds are adjusted during the first 24 time windows regardless of resampling window size. Using a larger value of H indicates our algorithm needs a longer time to learn event evolution. In general, it takes longer to draw people's attention in the case of progressive events as interesting sub-events develop slower than they do in instantaneous events. Secondly, our algorithm considers both the absolute volume (LINE 10-14) and differences in the number of messages (LINE 15-17) to detect noteworthy sub-events. Consequently, detected sub-events have a significant number of messages to be considered anomalous. Data instances lying in decreasing lines in time series plots are filtered out. We implement our method via Python.

Algorithm 1. Algorithm for the real-time and automatic detection of noteworthy moments.

Algorithm 1. Noteworthy sub-event detection

Input: N time series with features
Output: Binary labels

```
1: for i=1 to N do  // Iterate N time windows
2:     // Calculate thresholds
3:     H = constant
4:     while i < H do
5:         AVG = mean(COUNT)  // the mean of COUNT up to the i^th time window
6:         THRESHOLD = min(AVG, 100)
7:         DIFF_THRESHOLD = max(θ)  // the maximum θ up to the i^th time window
8:     if i == 0 then  LABEL ← 0
9:     else
10:        if (COUNT > THRESHOLD) and (COUNT >10) then
11:            if PREVIOUS LABEL == 1 then
12:                if EWM_MEAN ≥ θ then  LABEL ← 1
13:                elif PREVIOUS COUNT ≤ COUNT then  LABEL ← 1
14:                else  LABEL ← 0
15:            elif (PREVIOUS LABEL == 0) and (EWM_MEAN > θ) then
16:                if ψ > DIFF_THRESHOLD then  LABEL ← 1
17:                else  LABEL ← 0
18:            else  LABEL ← 0
19:        else  LABEL ← 0
```

EVENT SUMMARISATION

Given detected noteworthy moments, we extract representative tweets that can summarise unfolding events within each key moment. One simple approach to extract summaries are to assign scores to words in each tweet using a frequency or term frequency-inverse document frequency. However, this approach generally produces poor results. We use an unsupervised graph-based ranking algorithm for text called TextRank (Mihalcea and Tarau 2004). We decide to use TextRank for two reasons. One reason is that it is fully unsupervised. The performance of supervised text ranking models is highly dependent on training data. On the other hand, it is capable of extracting summaries solely based on the text itself. This is particularly important for crises as they are usually unseen and evolve unexpectedly. Another reason is that TextRank can be adapted to short text, which is suitable for tweets. We implement the TextRank using an open source software in Python (Barrios et al. 2016). However, the representation of the graphs of words in their implementation favours long tweets. To avoid this limitation, we follow the representation proposed by (Meladianos et al. 2015). Let a 'document' refer to a set of tweets

published within a time window. Given a document, we firstly generate a graph G whose vertices (V) are words and edges (E) are co-occurrence relation. If two words w_i and w_j co-occur in a tweet, an edge between two vertices V_i and V_j is created. Next, a weight $w_{ij} = \frac{1}{p-1}$ is added to the edge, where p is the number of unique terms in the tweet. If vertices already exist in a graph, the edge weight is accumulated. By repeating this procedure for the document, the algorithm returns scores for words in the document. Given the word scores, the method calculates scores for each tweet in the document by adding up the scores of words appearing in each tweet. Finally, we sort tweets in reverse order and use the top-K tweets as extractive summaries of each key sub-event. We use 10 for K in our experiments. Using a higher number for K may help users discover more sub-events. However, it is not viable for humans to read and annotate all available tweets due to the large volume of tweets during breaking events and time constraints (Zubiaga et al. 2016b; a).

EXPERIMENTS

Datasets

We use two publicly available datasets. A summary is shown in Table 2.

Twitter Event 2012-2016 (Zubiaga 2018) : This data consists of over 147 million tweets about 30 real-world events that took place between 2012 and 2016. The tweets were collected using Twitter's streaming API with keywords and hashtags related to each event. We select three human-made crises: Boston marathon bombings (2013), Ferguson unrest (2014), and Sydney siege (2014). We refer to this data as 'Event1216' hereinafter.

CrisisLexT26 (Olteanu et al. 2014): This data includes tweets regarding 26 hazardous events that unfolded between 2012 and 2013. It is designed to enable decision makers and citizens to understand unfolding disaster situations and obtain information of interest from social media. We select two natural disasters: Northern Italy earthquakes (2012) and Russian meteor (2013). The data size is small, but we decide to use this to show the scalability and generalization of our framework. We refer to this data as 'CrisisLex' hereinafter.

In Table 2, we define the development type of the Ferguson unrest as 'Mixed'. The initial shooting was instantaneous, but the developments were progressive. Table 3 shows the keywords used by the authors of the datasets when retrieving tweets. Some keywords such as 'ferguson' are too general. It is time-consuming to examine the entire set of collected messages as it contains a large number of non-rumours. Therefore, we can appreciate the importance of the detection of noteworthy moments in studying rumours during crises and emergencies. We carry out *event-level* analysis of tweets posted during crises in the real world. Specifically, we are interested in discovering what stories attract the public's attention and candidates for rumours to be debunked and verified by human experts while events unfold. There exists a trade-off between identifying key sub-events in real-time and manually examining candidates for key sub-events. It is not viable to identify key sub-events for every second as large-scale and noisy messages are generated at a rapid pace in social media (Atefeh and Khreich 2015). To tackle the trade-off, we employ a sliding window as previous studies on sub-event detection proposed. We use a 1-minute window for the Boston marathon bombings and a 5-minute window for the others. We use different window sizes for two reasons. One is to show that our method is versatile as it can work for different window sizes. The other is to embody different characteristics of event evolution. For example, sub-events such as affected individuals and details regarding suspects change very rapidly and situations tend to be chaotic and dynamic during the Boston marathon bombings. It is crucial to minimise delays in key sub-event detection by using a shorter window. During the Sydney siege, situations are less likely to require instant emergency responses as a gunman was holding hostages at a café for several hours. It is reasonable to use a longer window.

Table 2. Description of datasets.

Dataset	Event	Year	Country	Start	End	Domain	Development	Tweet Count
Event1216	Boston marathon bombings	2013	US	Apr 15	Apr 15	Bombings	Instantaneous	1,753,836
	Ferguson unrest	2014	US	Aug 9	Aug 25	Shooting/Riots	Mixed	4,183,421
	Sydney siege	2014	US	Dec 14	Dec 17	Hostage taking	Instantaneous	1,292,948
CrisisLex	Northern Italy earthquakes	2012	Italy	May 18	May 31	Earthquake	Instantaneous	4,231
	Russian meteor	2013	Russia	Feb 14	Feb 15	Meteorite	Instantaneous	4,498

Table 3. Keywords used to collect tweets (Olteanu et al. 2014; Zubiaga 2018)

Dataset	Event	Keywords
Event1216	Boston marathon bombings	boston, marathon, #prayforboston
	Ferguson unrest	ferguson
	Sydney siege	#sydneysiege, sydney, gunman, lindt, martin place
CrisisLex	Northern Italy earthquakes	earthquake italy, quake italy, #modena, #sanfelice, San Felice, modena terremoto, #terremoto, #norditalia, modena earthquake, modena quake, terremoto italia
	Russian meteor	#RussianMeteor, #Chelyabinsk, #челябинск, #метеорит, #meteor, #meteorite, russia meteor, russian meteor

Baselines

We compare our method with two baselines which exploit thresholds to identify key sub-events. We reproduce the baselines via Python.

Moving-threshold burst detection (Hsieh et al. 2012): Given the resampled time series, it computes the mean μ and standard deviation σ of the number of tweets observed up to the current time t_i. Then, the threshold at the t_i is defined by $\alpha * (\mu + \beta * \sigma)$. The parameters α and β are set between 0.7 and 1.0, and between 1.5 and 2.0, respectively. However, the authors did not specify default values for their parameters. If the number of tweets at t_i is greater than the threshold, the method annotates the instance at t_i as a key moment.

Outlier-based sub-event detection (Zubiaga et al. 2012): Given the resampled time series, it identifies the instance at t_i as an event of interest if the tweeting rate at t_i is higher than 90% of previously seen tweeting rates.

RESULTS AND EVALUATION

Quantitative Evaluation

Table 4. The results of noteworthy moment detection.

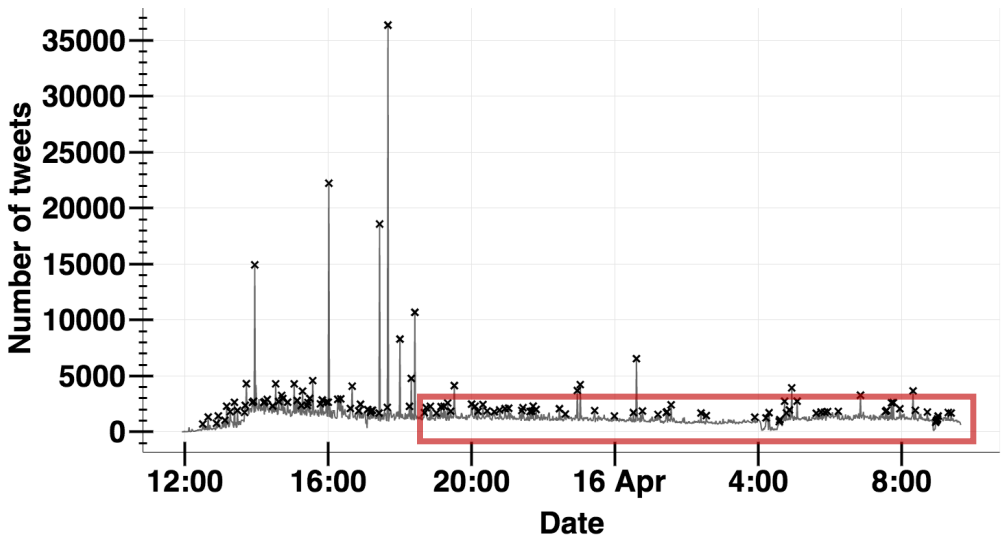
		Our method		Zubiaga		Hsieh	
Event	**# of windows**	**# of key moments**	**% of key moments**	**# of key moments**	**% of key moments**	**# of key moments**	**% of key moments**
Boston bombings	1,305	125	9.6	1,299	99.5	40 - 95	3.1 – 7.3
Ferguson unrest	4,651	603	13.0	4,093	88.0	341 - 646	7.3 – 13.9
Sydney siege	708	11	1.6	688	97.2	13 - 42	1.8 – 5.9
Italy earthquakes	3,585	40	1.1	13	0.4	126 -180	3.5 – 5.0
Russian meteor	337	29	8.6	191	56.7	20 - 50	5.9 – 14.8

Table 4 compares the number and percentage of detected noteworthy moments. (Hsieh et al. 2012) can detect different numbers of sub-events by adjusting parameters. The authors did not propose default values for their parameters. For a fair comparison, we randomly and uniformly draw parameter values from the ranges proposed in the original work. We perform the random sampling 100 times for each event. In Table 4, we show the minimum and maximum number of detected key moments. We find that the higher α and β are, the smaller number of windows are detected. For our method, we set H in Algorithm 1 to 36 the Ferguson unrest and to 24 for the others. We find that the number of detected sub-events depends not only on the total number of tweets but on burstiness. For example, the Sydney siege exhibits huge bursts (i.e. spikes) separated by long periods of inactivity. Small fluctuations between large spikes tend to be considered less noteworthy. On the other hand, the evolution of the Boston marathon bombings displays more bursty behaviour, and therefore relatively small bursts are identified as noteworthy moments. As can be seen in Table 4, Zubiaga's method performs poorly. In particular, their method cannot generalise to datasets with different characteristics. It identifies most of the time windows as key moments for the Event1216. It is not beneficial to perform sub-event detection if an algorithm recognises most time windows as key sub-events. In addition, previous study on sub-detection also showed that Zubiaga's method detect a large number of spurious outliers (Meladianos et al. 2015). On the contrary, it captures about 0.4% of the

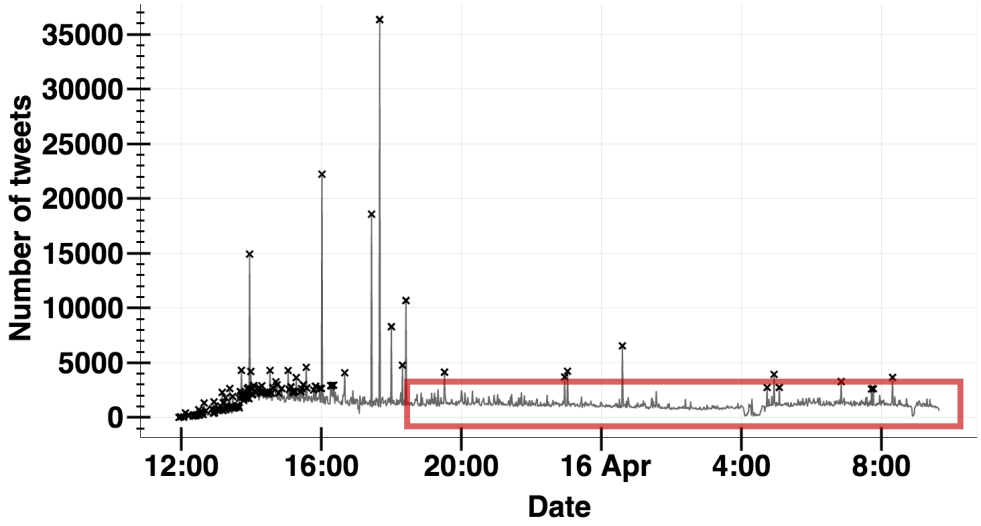
time windows for the Northern Italy earthquakes. Hsieh's method detects a reasonable number of windows. Therefore, we further compare the results in the qualitative evaluation.

Qualitative Evaluation

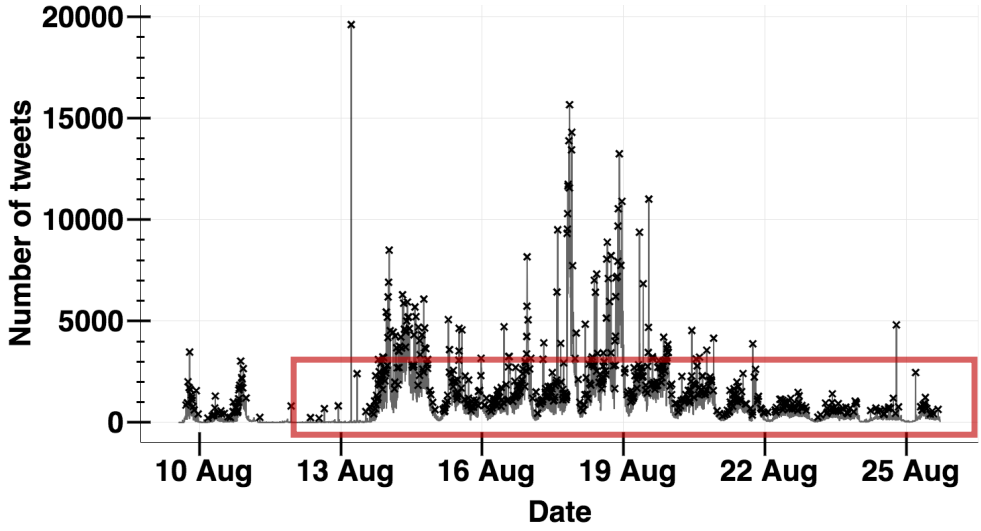
Not only the number of tweets but the distribution of key moments does matter. Here we compare the results obtained by our method and Hsieh's method. The major limitation of their method is their threshold is not adaptive to abrupt and extreme changes in event evolution. Recall that their threshold is based on the mean and standard deviation of the number of messages of preceding time windows. If there exist time windows the volume of messages of which is high in the past, their threshold tends to stay high afterwards. The evolution patterns of events on social media exhibit bursts of intense activity separated by long periods of inactivity or low-frequency periods (Barabási 2005). Taken all together, their method becomes less capable of detecting key sub-events with the relatively low number of messages as an event unfolds. We explain this phenomenon in detail with two real-world examples. Figure 3a and 3b show detected key sub-events for the Boston marathon bombings. For a fair comparison, we experimented with Hsieh's parameters outside the proposed ranges so that the number of sub-events detected by our method is similar to that detected by their method. Figure 3a shows detected key moments when $\alpha = 0.661$ and $\beta = 1.45$. One observation we can make is the results obtained by our method more evenly spread across the entire event evolution than those obtained by their method does. This means that our method is capable of detecting key moments in various phases of event evolution. Hsieh's method tends to detect more instances in the early stages of event evolution than our method does. However, their method only identifies spikes after 16:15 on 15 April. As we have already addressed previously, focusing on peaks in event evolution can prohibit the early detection of rumours. Figure 3c and 3d show detected key moments for the Ferguson unrest. It is clearly shown that a large number of instances are not detected in the red box in Figure 3d. This indicates that our method is better at early detection than Hsieh's method is. Figure 4a and 4b illustrate this observation more clearly. Hsieh's method detects not only instances on increasing lines but also those on decreasing lines. Figure 5 shows a part of Figure 3d. We can observe that all time windows, the number of messages of which is above a certain value are detected as key sub-events. We further investigate why these data examples are likely to be false anomalies in the following section.



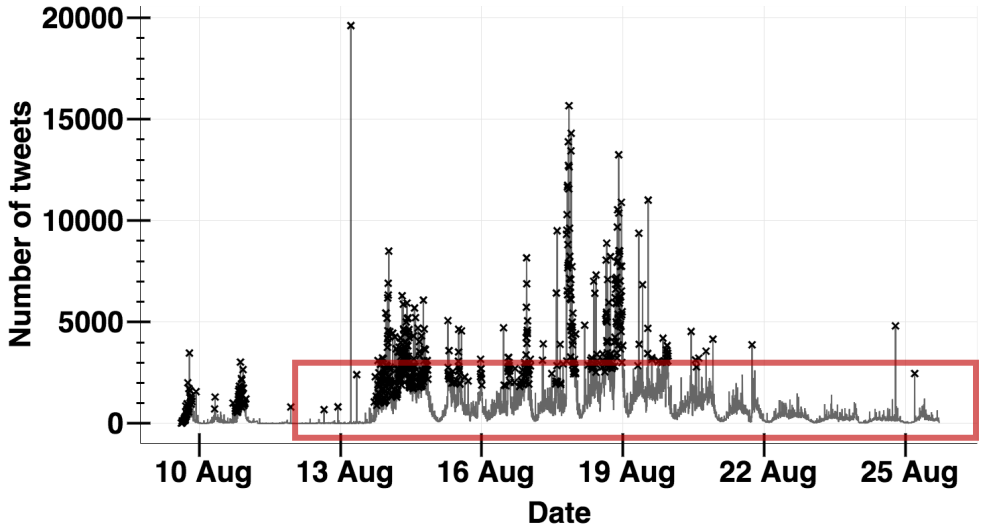
(a) 125 time windows are detected via our method.



(b) 129 time windows are detected via Hsieh's method



(c) 603 time windows are detected via our method



(d) 599 time windows are detected via Hsieh et al.'s method

Figure 3. Visual comparison of the results obtained by our method (a, c) and Hsieh's method (b, d). Detected key moments are marked with ×. (a) and (b) show the Boston marathon bombings. (c) and (d) show the Ferguson unrest.

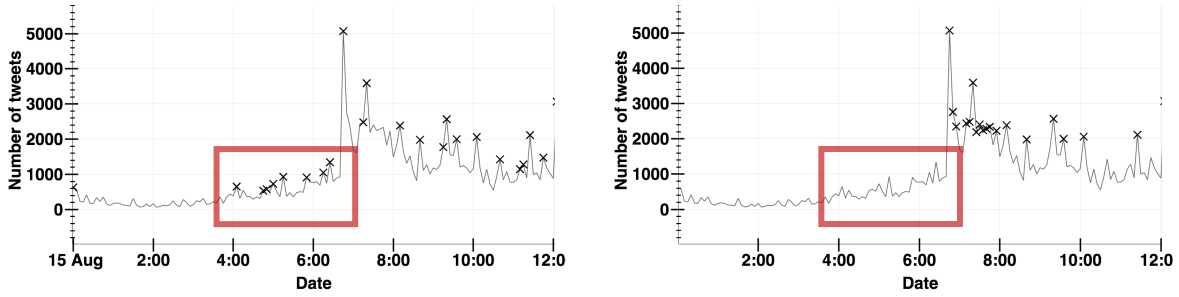


Figure 4. Partial plots of the Ferguson unrest. Detected noteworthy moments are marked with ×. (a) the results of our method (b) the results of Hsieh's method

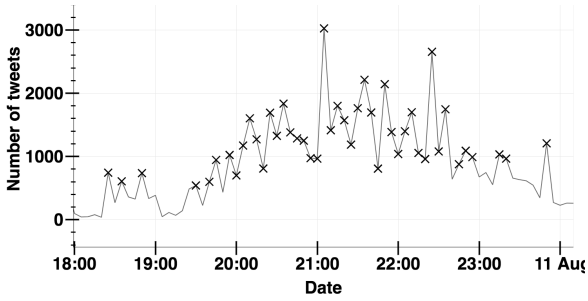


Figure 5. An example showing that Hsieh's method detects both increasing and fading patterns in event evolution.

Rumours During Crises

We apply the event summarisation method to the detected noteworthy moments. We then manually investigate the top 10 representative tweets of each key sub-event. Here, we show some examples of rumours appeared in the noteworthy moments detected by our method and their types. Note that there are more than 10 rumours as the given examples are found across an entire event rather than from one specific key moment. Rumours are classified into four categories: a) affected individuals, infrastructure, and utilities, b) the location of an event, c) other useful information (e.g. details about suspects and weapons). We present some rumours found in the detected noteworthy moments for two different datasets. As can be seen in the following tables, several informative and useful sub-events can be captured by focusing on noteworthy moments rather than the whole sequence. We find that several rumours spread during the Boston marathon bombings and Sydney siege. In particular, several posts related to a certain sub-event but with different entities (e.g. numbers) circulated. For the Boston marathon bombings, for instance, people reported different numbers and locations of explosive devices and whether a victim is an 8-year-old boy or girl. For the Sydney siege, different numbers of hostages and injured people were reported. There were also lots of details related to suspects.

- Boston marathon bombings

Type	Rumour	Type	Rumour
b	Two explosions near finish line	a	8-year-old girl or boy?
b	The 1st explosion reported on Boylston Street	a	An 8-year old girl is not one of the victims, an 8-year-old boy is dead
c	2 explosive devices are found	a	An 8-year-old girl died while running for her best friend
c	2 more explosive devices are found	a	An 8-year old girl/boy died while running for the Sandy Hook kids
c	Boston bombings are a "False Flag" operation	a	An 8-year-old boy killed has been identified as Martin Richard
a	8-year-old child died	a	The killed boy was waiting for his dad to finish his marathon
a	8-year-old boy/girl died	a	Mom and sister of the 8-year-old boy who was killed are injured
a	8-year-old girl died	b	The third explosion at the JFK library (unknown connection)
c	4 explosive devices	c	A fire broke out at the JFK library. Not an explosion
c	No one is in custody	c	Pressure cooker bombs were placed in black duffel bags
c	The third explosion has been reported	c	The third device is found

- Sydney siege

Type	Rumour	Type	Rumour
b, c	A gunman is taking hostages at a café in Martin Place.	c	The hostage taker is named as Man Haron Monis
b, c	An ISIS flag is being displayed on the window of a café	c	Man Haron Monis fled from Iran to Australia in 1996
a	2 killed and 3 seriously injured	a	6 more escaped and 11 in total have escaped
c	Hostages are being forced to hold an ISIS flag	c	Gunman is reportedly killed
b, c	Two gunmen inside the café under siege	a	2 dead (no information regarding the gunman's death)
a	Up to 20 hostages are being held	a, c	3 dead including the gunman (2 hostages and gunman dead)
a, b	Opera house has been evacuated	c	Prime Minister Abbott says Man Monis held a gun license
c	Sydney airspace closed	c	Shots fired
c	Gunman wants to talk to the Prime Minister	a, c	An Islamic extremist took 17 people hostages today
c	5 possible bombs in the city	c	Gunman wants ISIS flags
a	40-50 hostages being held	c	The gunman is carrying 2 bombs
b, c	Gunman says he has 4 bombs/devices around Sydney.	a	2 more hostages escape, 5 now have escaped.
a, c	2 dead including the gunman	c	The gunman is a lone wolf/self-styled.
c	Man Haron Monis is an Iranian cleric on bail for 40 sexual offences and accessory to the murder of his ex-wife.	a	Tori Johnson was the Lindt Cafe manager and tried to take the gun away from the man
a	Hostages who died are named as Tori Johnson, 34 & Katrina Dawson, 38	c	The flag at the window of the café is not an ISIS flag.

A hypothesis of our study is that newsworthy stories and rumours can be detected from time windows lying in increasing lines in even evolution graphs. We further study when rumours are detected for the first time using summaries obtained via our framework. The experimental results confirm our hypothesis. Table 5 shows whether some key rumours appear in noteworthy moments (i.e. increasing lines and peaks) or decaying lines for the first time. Most of the key rumours are detected at noteworthy moments, and therefore, we prove that our hypothesis is acceptable.

Table 5. Table shows whether key rumours detected from our summaries appear in noteworthy moments or time windows lying in decreasing lines in event evolution.

Timestamp	Rumour	Time window type
11:59 Apr 15	Two explosions near finish line	Noteworthy
12:24 Apr 15	The 1st explosion reported on Boylston Street	Decaying
13:28 Apr 15	Boston bombings are a “False Flag" operation	Noteworthy
22:24 Apr 15	An 8-year-old boy killed has been identified as Martin Richard	Noteworthy
19:49 Apr 15	The killed boy was waiting for his dad to finish his marathon.	Noteworthy
12:44 Apr 15	The third device has been found.	Noteworthy
12:42 Apr 15	Police tell people to stay away from the JFK library	Noteworthy
14:27 Apr 15	A fire broke out at the JFK library. Not an explosion.	Noteworthy
14:12 Apr 15	A 20-year-old Saudi Arabian man is in custody.	Noteworthy
09:03 Apr 16	Bombs are made from pressure cookers.	Noteworthy
09:23 Apr 16	Pressure cooker bombs were placed in black duffel bags.	Noteworthy

CONCLUSION AND FUTURE RESEARCH

In this paper, we have proposed a two-step framework to aid rumour detection during crises in real time. There have been several attempts to verify rumours in various domains in the community of rumour studies. However, rumours during crises are new, and therefore there is no *evidence* to debunk or verify claims when crises are unfolding. It is very challenging to develop an automatic system which is capable of verifying such rumours. Rumour detection during crises, however, can be automated and provide useful information for managers and the public. Therefore, our research aims at helping emergency responders and citizens detect emerging and developing rumours without examining an enormous number of messages. The first part of our framework detects *noteworthy moments* that draw the public’s attention during crises. We view noteworthy moments or key sub-events as outliers of interest to decision makers and other stakeholders. The performance of sub-event detection methods based on

thresholds is highly dependent on characteristics of data such as the total number of tweets and burstiness. To overcome this limitation, we incorporate several rules which take into account the absolute volume of messages and differences in the number of messages. Our rule-based sub-event detection system can be adaptive to abrupt and extreme changes in event evolution on social media. Our system can work in real time. This will enable decision makers to plan appropriate actions as early as possible. The second part extracts summaries that represent detected key moments. We employ an unsupervised graph-based text ranking algorithm called TextRank to assign scores to messages. We conducted experiments over large-scale and real-world crises datasets. We show the generalization and scalability of our method by using the datasets with different languages and characteristics such as bursty patterns and the total number of messages. The experimental results show that our method can perform the early detection of informative reports and rumours spreading during different types of crises in a real-time scenario. A common role of social media in crisis management is to improve situational awareness by analysing early reports of events. We have found that there exist several rumours during crises and their contents evolve over time. This behaviour of rumours has been extensively studied in the community of rumour studies. The application of different realms of rumour studies such as rumour tracking and stance classification can benefit emergency responders in real time. For example, decision makers will be able to distinguish rumours that are likely to be false from facts in the initial stages of rumour evolution. They can then take action to debunk and verify rumours in a timely manner. This can lead to the minimization of disaster-related losses. In conclusion, this paper aims to benefit emergency management by mining rumours evolving during crises. We have shown what rumours evolve during crises. Our experiments with real-world examples show that key rumours and informative reports can be detected from *noteworthy moments*. Our research can bridge the gap between the state-of-the-art work on rumours on social media and real-world needs for crisis management.

Future research will delve into how to fully automate rumour detection during crises. Our current system is semi-automatic. It produces summaries of detected noteworthy sub-events. However, these summaries can include rumours and non-rumours. Humans still need to judge whether each summary tweet is a rumour or non-rumour. We will further investigate characteristics of rumours during crises so that a system can distinguish rumours from non-rumours without a manual inspection.

ACKNOWLEDGEMENT

We thank Vitaveska Lanfranchi for helpful comments and discussions on an early draft of this paper. This project was partially funded by the European Commission as part of the Seta European Project (contract no. 688082), Horizon 2020 framework.

REFERENCES

Aggarwal, C. C. (2015). "Outlier Analysis." *Data Mining: The Textbook*, C. C. Aggarwal, ed., Springer International Publishing, Cham, 237–263.

Andrews, C. A., Fichet, E. S., Ding, Y., Spiro, E. S., and Starbird, K. (2016). "Keeping Up with the Tweet-dashians: The Impact of `Official- Accounts on Online Rumoring." *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, ACM Press, San Francisco, California, USA, 451–464.

Arif, A., Robinson, J. J., Stanek, S. A., Fichet, E. S., Townsend, P., Worku, Z., and Starbird, K. (2017). "A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors." *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, ACM, New York, NY, USA, 155–168.

Atefeh, F., and Khreich, W. (2015). "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence*, 31(1), 132–164.

Barabási, A.-L. (2005). "The origin of bursts and heavy tails in human dynamics." *Nature*, 435(7039), 207–211.

Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. (2016). "Variations of the Similarity Function of TextRank for Automated Summarization." *arXiv:1602.03606 [cs]*.

Boididou, C., Middleton, S. E., Jin, Z., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., and Kompatsiaris, Y. (2018). "Verifying information with multimedia content on twitter." *Multimedia Tools and Applications*, 77(12), 15545–15571.

Castillo, C. (2016). *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

DiFonzo, N., and Bordia, P. (2007). *Rumor psychology: social and organizational approaches*. American Psychological Association, Washington, DC.

Gillani, M., Ilyas, M. U., Saleh, S., Alowibdi, J. S., Aljohani, N., and Alotaibi, F. S. (2017). "Post Summarization of Microblogs of Sporting Events." *Proceedings of the 26th International Conference on World Wide*

Web Companion - WWW '17 Companion, ACM Press, Perth, Australia, 59–68.

Helmstetter, S., and Paulheim, H. (2018). "Weakly Supervised Learning for Fake News Detection on Twitter." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 274–277.

Hsieh, L. C., Lee, C. W., Chiu, T. H., and Hsu, W. (2012). "Live Semantic Sport Highlight Detection Based on Analyzing Tweets of Twitter." *2012 IEEE International Conference on Multimedia and Expo*, 949–954.

Hu, Y., Hu, C., Fu, S., Fang, M., and Xu, W. (2017). "Predicting Key Events in the Popularity Evolution of Online Information." *PLOS ONE*, (W.-B. Du, ed.), 12(1), e0168749.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing Social Media Messages in Mass Emergency: A Survey." *ACM Comput. Surv.*, 47(4), 67:1–67:38.

Kedzie, C., McKeown, K., and Diaz, F. (2015). "Predicting Salient Updates for Disaster Summarization." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Beijing, China, 1608–1617.

Kong, S., Ye, F., Feng, L., and Zhao, Z. (2015). "Towards the prediction problems of bursting hashtags on Twitter." *Journal of the Association for Information Science and Technology*, 66(12), 2566–2579.

Kwon, S., Cha, M., and Jung, K. (2017). "Rumor Detection over Varying Time Windows." *PLOS ONE*, 12(1), 1–19.

Ma, J., Gao, W., and Wong, K.-F. (2017). "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 708–717.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2012). "Rise and Fall Patterns of Information Diffusion: Model and Implications." *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, ACM, New York, NY, USA, 6–14.

Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., and Vazirgiannis, M. (2015). "Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream." *undefined*, </paper/Degeneracy-Based-Real-Time-Sub-Event-Detection-in-Meladianos-Nikolentzos/20a29b5af88d739ef2e7b524c144104cd5cfbf54> (Aug. 13, 2018).

Mihalcea, R., and Tarau, P. (2004). "TextRank: Bringing Order into Texts." *Proceedings of EMNLP 2004*, D. Lin and D. Wu, eds., Association for Computational Linguistics, Barcelona, Spain, 404–411.

Nguyen, M.-T., Kitamoto, A., and Nguyen, T.-T. (2015). "TSum4act: A Framework for Retrieving and Summarizing Actionable Tweets During a Disaster for Reaction." *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, and H. Motoda, eds., Springer International Publishing, 64–75.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises." *Eighth International AAAI Conference on Weblogs and Social Media*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises." *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, ACM Press, Vancouver, BC, Canada, 994–1009.

Peng, S., Tseng, V. S., Liang, C.-W., and Shan, M.-K. (2018). "Emerging Product Topics Prediction in Social Media Without Social Structure Information." *Companion Proceedings of the The Web Conference 2018*, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1661–1668.

Rudra, K., Banerjee, S., Ganguly, N., Goyal, P., Imran, M., and Mitra, P. (2016). "Summarizing Situational Tweets in Crisis Scenario." *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, HT '16, ACM, New York, NY, USA, 137–147.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., and Ghosh, S. (2015). "Extracting Situational Information from Microblogs During Disaster Events: A Classification-Summarization Approach." *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, ACM, New York, NY, USA, 583–592.

Rudra, K., Goyal, P., Ganguly, N., Mitra, P., and Imran, M. (2018). "Identifying Sub-events and Summarizing Disaster-Related Information from Microblogs." *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, ACM, New York, NY, USA, 265–274.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). "Fake News Detection on Social Media: A Data Mining Perspective." *SIGKDD Explor. Newsl.*, 19(1), 22–36.

Starbird, K., Spiro, E., Edwards, I., Zhou, K., Maddock, J., and Narasimhan, S. (2016). "Could This Be True?: I Think So! Expressed Uncertainty in Online Rumoring." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, ACM Press, Santa Clara, California, USA, 360–371.

Vosoughi, S., Mohsenvand, M. `Neo', and Roy, D. (2017). "Rumor Gauge: Predicting the Veracity of Rumors on Twitter." *ACM Trans. Knowl. Discov. Data*, 11(4), 50:1–50:36.

Wang, S., Yan, Z., Hu, X., Yu, P. S., Li, Z., and Wang, B. (2016). "CPB: a classification-based approach for burst time prediction in cascades." *Knowledge and Information Systems*, 49(1), 243–271.

Wong, K.-F., Gao, W., and Ma, J. (2018). "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks." *ACL*.

Zeng, L., Starbird, K., and Spiro, E. S. (2016). "Rumors at the Speed of Light? Modeling the Rate of Rumor Transmission During Crisis." *2016 49th Hawaii International Conference on System Sciences (HICSS)*, IEEE, Koloa, HI, USA, 1969–1978.

Zubiaga, A. (2018). "A longitudinal assessment of the persistence of twitter datasets." *Journal of the Association for Information Science and Technology*, 69(8), 974–984.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018a). "Detection and Resolution of Rumours in Social Media: A Survey." *ACM Comput. Surv.*, 51(2), 32:1–32:36.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., Cohn, T., and Augenstein, I. (2018b). "Discourse-aware rumour stance classification in social media using sequential classifiers." *Information Processing & Management*, 54(2), 273–290.

Zubiaga, A., Liakata, M., and Procter, R. (2016a). "Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media." *arXiv:1610.07363 [cs]*.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016b). "Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads." *PLOS ONE*, 11(3), e0150989.

Zubiaga, A., Spina, D., Amigó, E., and Gonzalo, J. (2012). "Towards real-time summarization of scheduled events from twitter streams." *Proceedings of the 23rd ACM conference on Hypertext and social media - HT '12*, ACM Press, Milwaukee, Wisconsin, USA, 319.