Can We Predict a Riot? Disruptive Event Detection Using Twitter

NASSER ALSAEDI, PETE BURNAP, and OMER RANA, Cardiff University, UK

In recent years, there has been increased interest in real-world event detection using publicly accessible data made available through Internet technology such as Twitter, Facebook, and YouTube. In these highly interactive systems, the general public are able to post real-time reactions to "real world" events, thereby acting as social sensors of terrestrial activity. Automatically detecting and categorizing events, particularly small-scale incidents, using streamed data is a non-trivial task but would be of high value to public safety organisations such as local police, who need to respond accordingly. To address this challenge, we present an end-to-end integrated event detection framework that comprises five main components: data collection, pre-processing, classification, online clustering, and summarization. The integration between classification and clustering enables events to be detected, as well as related smaller-scale "disruptive events," smaller incidents that threaten social safety and security or could disrupt social order. We present an evaluation of the effectiveness of detecting events using a variety of features derived from Twitter posts, namely temporal, spatial, and textual content. We evaluate our framework on a large-scale, real-world dataset from Twitter. Furthermore, we apply our event detection system to a large corpus of tweets posted during the August 2011 riots in England. We use ground-truth data based on intelligence gathered by the London Metropolitan Police Service, which provides a record of actual terrestrial events and incidents during the riots, and show that our system can perform as well as terrestrial sources, and even better in some cases.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; **Robotics**; • **Networks** → *Network reliability;*

Additional Key Words and Phrases: Social media, event detection, classification, clustering, feature selection, evaluation

ACM Reference Format:
Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. Can we predict a riot? Disruptive event detection using twitter. ACM Trans. Internet Technol. 17, 2, Article 18 (March 2017), 26 pages.
DOI: http://dx.doi.org/10.1145/2996183

1. INTRODUCTION

The rapid growth of Internet-enabled communication technology in the form of social networking services (often collectively referred to as social media) and associated smartphone apps has enabled billions of global citizens to broadcast news and "on-the-ground" information during "real-world" events as they unfold. Twitter, for example, has been studied as an emerging news reporting platform [Phuvipadawat and Murata 2010; Weng and Lee 2011; Osborne et al. 2013] and has been widely used to disseminate information about the Arab Spring [Starbird and Palen 2012; Alsaedi and Burnap 2015] and other disaster-related incidents [Imran et al. 2015; Shamma. et al. 2010; Thelwall et al. 2011; Burnap et al. 2014; Williams and Burnap 2015]. The interaction between people, events, and Internet-enabled technology presents both an opportunity and a challenge to social computing scholars, public sector organisations (e.g.,

Authors' addresses: N. Alsaedi, P. Burnap, and O. Rana, School of Computer Science and Informatics, Cardiff University, UK; emails: {AlsaediNM, burnapp, ranaof}@cardiff.ac.uk.

ACM 1533-5399/2017/03-ART18 $15.00
DOI: http://dx.doi.org/10.1145/2996183

governments and policing agencies), and the private sector, all of whom aim to understand how events are reported using social media and how millions of online posts can be reduced to accurate but meaningful information that can support decision making and lead to productive action.

Research in recent years has uncovered the increasingly important role of utilising data from social networking sites in disaster situations and shown that information broadcast via social media can enhance situational awareness during a crisis situation [Vieweg et al. 2010; Alsaedi et al. 2015; Vieweg et al. 2014]. In particular, members of the public, formal response agencies, and local, national, and international aid organizations are all aware of the ability to use social media to gather and disperse timely information in the aftermath of disaster [Imran et al. 2014; Chowdhury et al. 2013; Iyengar et al. 2011]. However, many existing approaches to event detection are limited to global or large-scale event detection (e.g., natural disasters and terror attacks), while detecting small-scale incidents such as fires, car accidents, and public order events remains an ongoing research topic due to several key challenges.

One challenge is that online posts are often constrained in length (referred to as microblogs), which means that only a small amount of text is available to be analysed to gain insights. Within the text, there are other challenges, such as frequent use of informal, irregular, and abbreviated words; a large number of spelling and grammatical errors; and the use of improper sentence structure and mixed languages [Becker et al. 2011a; Farzindar and Wael 2015; Imran et al. 2015]. Some languages are more challenging than others, for example, Arabic users use dialects heavily as well as a mixture of Latin and Arabic characters (Arabizi) [Alsaedi and Burnap 2015]. These dialects may differ in vocabulary, morphology, and spelling from the standard Arabic, and most do not have standard spellings. Additionally, social networking services' popularity have attracted spammers and other content polluters to spread advertisements, pornography, viruses, phishing, and other malicious material that cloud the information analysis [Farzindar and Wael 2015; Burnap et al. 2015].

Despite these challenges, it has been noted that detecting small-scale events is essential to improving situational awareness of both citizens and decision makers [Schulz et al. 2015; Walther and Kaisser 2013; Li et al. 2012] and thus remains a well-motivated research topic for the social computing community. In this article, we propose a novel approach to event detection that aims to overcome many of the challenges to provide a system to detect large-scale events and related small-scale events. The approach is based on the integration of supervised machine-learning algorithms to detect larger scale events and unsupervised approaches to cluster, disambiguate, and summarize smaller sub-events, with a goal of improving situational awareness in emergency situations through automatic methods. Our contributions can be summarized as follows:

—Using temporal, spatial, and textual features, our approach is able to detect small-scale events in a given place and time better than existing algorithms, to which we compare our performance results;
—While other related work focuses on large *or* small scale events, our approach can identify large and related small scale events. Thus, our approach retains the context of smaller events (e.g., distinguishing between public disorder related to an event, and general disorder);
—Much of the related event detection work is dependent on utilising event-specific terms and phrases, but we propose a novel approach to summarizing microblog posts corresponding to events without the need for prior knowledge of the entire data set, that is, in real time and not post event. Our approach is based on modifying a term frequency algorithm to include a dynamic temporal aspect;

—We demonstrate that our proposed approach can identify the relationship between content posted via social media and "real-world" events by using time-stamped social media data and actual crime reports to accurately flag events prior to their known reporting time throughout a study period, using human annotated Twitter data as an example data source.
—We present a case study of our approach by evaluating it against other leading approaches using Twitter posts from the UK riots in 2011 and a publicly accessible account of *actual reported* intelligence obtained and reports received by the Metropolitan Police Service during this event. Smaller scale events include localized looting, violence, and criminal damage. Results show that our system can perform as well as terrestrial sources at detecting events related to the riots; in some cases, we detect the event *before* intelligence reports were recorded.

The rest of this article is organized as follows: Section 2 reviews related work. Sections 3 and 4 define the problem of event detection using data from social networking services and discuss the technical architecture and algorithms developed as part of our proposed system. In Section 5 we present and analyze several features, namely temporal, spatial, and textual features. Section 6 presents our experiments and discusses the results. In Section 7, we conclude and highlight some directions for future research.

2. RELATED WORK

The general topic of detecting real-world events from social media has received considerable research interest. Research efforts have focused on real-time event detection and tracking, social media analysis, micro-blog summarization, and information visualisation. We describe relevant related work in three areas: large-scale (global) event detection, small-scale (local) event detection, and systems used to extract crisis relevant information from social media.

For large-scale events, Petrović et al. [2010] presented an approach to detect breaking stories from a stream of tweets using locality-sensitive hashing (LSH). Becker et al. [2011a] proposed an online clustering framework to identify different types of real-world events. Then, they use different machine-learning models to predict whether a pair of documents belong to real-world events or not. These approaches are limited to widely discussed events and fail to report rare and potentially disruptive small-scale incidents.

Large-scale event detection has also been explored through clustering of discrete wavelet signals built from individual words generated by Twitter [Weng and Lee 2011]. Auto-correlation then filters away the trivial words (noise) and cross correlation groups together words that relate to an event by modularity-based graph partitioning. Similarly, Cordeiro [2012] proposed a continuous wavelet transformation based on hashtag occurrences combined with a topic model inference using Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. In fact, LDA and its variants are widely used statistical modelling approach implemented in event detection tasks [Vavliakis et al. 2013; Pan and Mitra 2011; Cordeiro 2012; Vieweg et al. 2014]. However, these methods have the main drawback of requiring *a priori* specification of the number of total topics, which leads to problems when the total number of events exceeds this number.

Other approaches have focused on structural networks and graph models to discover events in social media feeds. Benson et al. [2011] presented a structured graphical model that simultaneously analyzes individual messages, clusters them according to event and induces a canonical value for each event property. Using a different graph analytical approach, Sayyadi and Raschid [2013] used a KeyGraph algorithm [Ohsawa et al. 1998] to convert text data into a term graph based on co-occurrence relations between terms. Then they employed a community detection approach to partition the

graph. Eventually, each community is regarded as a topic, and terms within the community are considered as the topic's features. Moreover, Schinas et al. [2012] used the Structural Clustering Algorithm for Networks (SCAN) for detecting "communities" of documents. These candidate social events were further processed by splitting the events that exceeded a predefined time range into shorter events. Then they used a classification approach based on median geolocations and accumulated TF-IDF vectors for each cluster to separate relevant and irrelevant candidate events. Nevertheless, these graph partitioning algorithms are not ideal for social media event detection problems because of their complexity [Agarwal et al. 2012] and limitation that they do not capture the highly skewed event distribution of social media event data due to their bias towards balanced partitioning [Karypis et al. 1997]. In addition, the multiple events and sub-events discovery becomes computationally expensive using graph partitioning algorithms due to velocity and scale of updates in a highly dynamic real-time situation [Agarwal et al. 2012].

Various methods have been proposed to identify small-scale events from social media streams such as fire incidents, traffic jams, and so on. Walther and Kaisser [2013] developed spatiotemporal clustering methods where they monitor specific locations of high tweeting activity and cluster tweets that are geographically and temporally close to each other. A machine-learning module is then used to evaluate whether a cluster of tweets refer to an event based on 41 features including the tweet content. Another clustering approach is presented in Schulz et al. [2015], with a small-scale incident detection pipeline based on the clustering of incident-related micro-posts using three properties that define an incident: (1) incident type, (2) location, and (3) time period. Various techniques are adopted to increase the quality of their clustering approach: (a) the incident type determination using supervised machine learning (Semantic Abstraction), (b) geotagging of tweets based on tweets geolocalization, and (c) the extraction of time period of the incident. Yet, both methods are very specific without giving aspects of the general context, and it is critical that the system can provide insight into ongoing sub-events arising amid the protest to better inform how to react accordingly and to improve both event reasoning and system performance. That could explain the low recall/precision of the Schulz et al. [2015] and Walther and Kaisser [2013] approaches when validated using real-world official reports, 32.14% and 4.75%, respectively.

Another event detection system, Twitcident [Abel et al. 2012], presents a Web-based application for searching, filtering, and aggregating information about known events reported by emergency broadcasting services in the Netherlands. In addition, Watanabe et al. [2011] proposed a system called Jasmine for detecting local events in the real-world using geolocation information from microblog documents. They obtain the name list of locations from geotagged tweets and add positional information to tweets by matching the location name. A similar work is Boettcher and Lee [2012], which introduces a statistical method for detecting local events using a temporal and spatial analysis by considering 7-day historic data. The main contribution of EventRadar is that it detects local events without keeping a list of locations by finding clusters of Tweets that contain the same subset of words. Another related system is proposed by Li et al. [2012] to detect crime- and disaster-related events (CDE) from tweets. They use spatial and temporal information of tweets to detect new events with a number of text-mining techniques to extract the meta information (e.g., geo-location names, temporal phrase, and keywords) for event interpretation. Most of these small-scale event detection approaches are novel and automatic; however, the performance and detection reliability of these systems are highly dependent on the incident type so they are limited to certain specific types of event content that they can handle.

Regarding the use of social media data during disasters, researchers have proposed several visual analytics approaches aiming at real-time microblog analysis that often

facilitate interactive means for exploration and anomaly indication. TwitterMonitor [Mathioudakis and Koudas 2010] performs trend detection in two steps and analyzes trends in a third step. During the first phase, it identifies bursty keywords that are then grouped based on their co-occurrences. Once a trend is identified, additional information from the tweets is extracted to analyze and describe the trend. Artificial Intelligence for Disaster Response (AIDR) [Imran et al. 2014] is a platform for filtering and classifying messages posted to social media during humanitarian crises in real time. AIDR uses human-assigned labels (crowdsourcing messages) and pre-existing classification techniques to classify Twitter messages into a set of user-defined situational awareness categories in real time. Vieweg et al. [2010] analyze the Twitter logs for a pair of concurrent emergency events: the Oklahoma Grassfires (April 2009) and the Red River Floods (March and April 2009). Their automated framework is based on the relative frequency of geo-location and location-referencing information from users' posts.

In a related work, Olteanu et al. [2014] created a lexicon of crisis-related terms (380 single-word terms) that frequently appear in relevant messages posted during six crisis events. Then they demonstrated how we use the lexicon to automatically identify new terms by employing pseudo-relevance feedback mechanisms to extract crisis-related messages during emergency events. Vieweg et al. [2014] enable filtering, searching, and analyzing of Twitter during another natural disaster (the 2013 Typhoon Yolanda). They used supervised classification algorithm to automatically classify tweets into three categories: Informative, Not informative, and Not related to this crisis. Then they employed topic modelling using the LDA [Blei et al. 2003] model to further classify the informative tweets into 10 clusters according to the Humanitarian Clusters Framework. Similarly, Twitinfo [Marcus et al. 2011] automatically detects and labels unusual bursts in real-time Twitter streams. However, they used different approach as TwitInfo adapts signal processing and streaming techniques to extract peaks and label them meaningfully using text from the tweets.

In research that is mostly analytical, Shamma. et al. [2010] presented Tweetgeist for identifying structure and semantics in Twitter about media events and providing that information back to the microbloggers to enhance their experience. Most recently, Thapen et al. [2015] built a situational awareness system that uses frequency statistics and cosine similarity based measures to produce terms characterising localized events (the detection of illness outbreak) and then retrieve relevant news and representative tweets. However, most of the current disaster identification approaches are limited to detect certain events such as earthquake, tornado, and so on, and cannot be generalized to detect other disaster-related events. Another presumption of these approaches is that users have to know the event in advance to represent keyword queries to be detected.

It is worth reporting some studies that have been proposed to identify event phases and the temporal boundaries of mass disruption events. For instance, Chowdhury et al. [2013] introduced a system called Tweet4act to automatically determine different phases of an event by extracting content features from each message. They applied the popular k-means clustering algorithm to classify messages for three crisis events (the Joplin Tornado in the US, the Nesat Typhoon in the Phillipines, and the Haiti Earthquake in Haiti). Similarly, but with a broader perspective of events, Iyengar et al. [2011] described an approach to automatically determine when an anticipated event started and ended by analyzing the content of tweets using a support vector machines (SVM) classifier and hidden Markov model with various textual features such as bag of words, part-of-speech tags, and so on. Both studies aim to automatically classify tweets into three phases of an event: before, during, and after. Additionally, Yin et al. [2015] investigated several approaches that have been shown useful when analyzing Twitter messages generated during humanitarian crises even to local levels.

They evaluate these key relevant methods for burst detection, tweet filtering and classification, online clustering, and geotagging.

Several recent efforts proposed techniques for automatic microblog event summarization from social media. The centroid-based method is one of the most popular extractive summarization methods, such as MEAD by Radev et al. [2001] and that by Becker et al. [2011b], who presented and evaluated three centrality-based approaches to select the high-quality messages from clusters. Another approach is the graph-based LexRank, which was introduced by Erkan and Radev [2004]. The TextRank algorithm [Mihalcea and Tarau 2004] is another graph-based approach that implements two unsupervised approaches for keyword and sentence extraction in order to find the most highly ranked sentences in a document using the PageRank algorithm [Brin and Page 1998]. Recently, Xu et al. [2013] extended the PageRank ranking algorithm and investigate a graph-based approach that leverages named entities, event phrases, and their connections across tweets to create summaries of variable length for different topics. Moreover, Olariu [2014] proposed a graph-based abstractive summarization scheme where bigrams extracted from the tweets are considered as the graph nodes.

Feature-based approaches are statistical and linguistic features that have been extensively investigated; for example, Sharifi et al. [2010] proposed a phrase reinforcement algorithm to summarize the Twitter topic in one sentence. Nichols et al. [2012] extended this idea and generated journalistic summary for events in World Cup games. More fine-grained summarization was proposed by considering sub-events detection and combining the summaries extracted from each sub-topic (tweet selection, tweet ranking) [Shen et al. 2013; Zubiaga et al. 2012; Yajuan et al. 2012]. Other researchers have proposed various models including the use of Non-negative Matrix Factorization [Yang et al. 2012], a structured retrieval approach [Metzler et al. 2012], Structured Probabilistic Latent Semantic Analysis [Lu et al. 2009], and many more [Chua and Asur 2013]. However, some of these algorithms can only be applied to periodic events such as sports events and not on longer-term events or aperiodic events and others do not perform particularly well on large real-world multilingual corpora.

Previously, we have focused on online real-world events identification for large-scale events such as sport events [Alsaedi et al. 2014]. In Alsaedi and Burnap [2015], we presented an approach to detect stories from a stream of Arabic tweets with the main focus on motivation and challenges of identifying events in Arabic language. In Alsaedi et al. [2015], we described an improved model for feature selection based on the popular Mitra et al. [2002] algorithm and made it suitable for microblog data such as Twitter. In this article, we further explored features that could be useful to enhance situational awareness. We validate the effectiveness of our framework using a large noisy dataset of over 40 million Twitter messages. We also show that our framework yields better performance than many leading approaches in the real-time event detection.

In summary, although many approaches exist for the event detection task, they are generally either used for large-scale events and cannot capture important small-scale events or are very specific and are limited to detect certain events only—thus missing the context of larger events. In contrast to the above approaches, our system automatically identifies as many real-world events in a given region as possible. Then, using an online clustering algorithm with a sliding window timeframe, it can be utilised to detect large and small-scale events from social media streams—with particular attention to filtering from large to small-scale events. Employing supervised classification of each tweet before clustering (large-scale event detection) reduces the computational overhead at the clustering stage as the number of tweets is significantly reduced (containing only event-related tweets). Thus clustering (small-scale event detection), feature selection, and summarization are much faster and suitable for real-time analysis. The

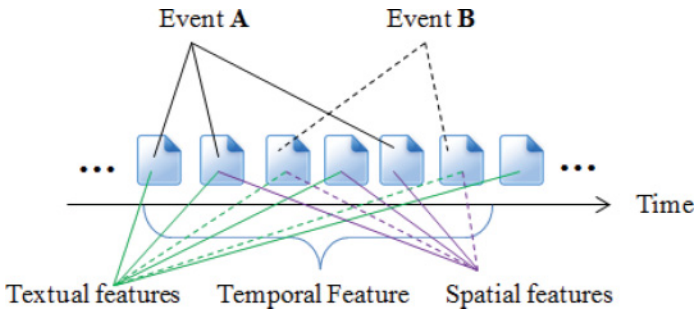

Fig. 1. Document clustering using different sets of features.

presented case study of our approach by evaluating it against other leading approaches using Twitter posts from the UK riots in 2011 itself can be considered a contribution.

3. PROBLEM DEFINITION

"Events," as captured via social media, are real-world happenings that are reflected by change in the volume of text data that discusses the associated topic at a specific time [Dong et al. 2015]. Hence, an event can be characterized by one or more of the following attributes: Topic, Time, People, and Location [Imran et al. 2015]. A "disruptive event" in the context of social media can be defined as follows.

Definition. A disruptive event is an event that interferes with (disrupts) the achieving of the objective of an event or interrupts ordinary event routine. It may occur over the course of one or several days, causing disorder and destabilizing security and may results in a displacement or discontinuity.

Different events have different context-specific features but are generally conveyed on social media using verbs (actions), nouns (names, places, topics), adjectives (descriptive), and prepositional phrases (proximity and location descriptive). Our aim is to represent the data extracted from social media as a timeline of events (clusters), where each cluster contains sufficient data to discriminate between events and summarize them as actionable information for use by public safety officials and policy makers.

We assume the task of event detection includes summarizing one or more of the following scenarios: (a) different events occurring in the same location for a time period, where we assume each event can be characterized using different *textual features*; (b) similar multiple events in different locations, where we assume the most appropriate features will be *temporal* and the *spatial features*; and (c) similar events in the same location at almost the same time, and in this case we assume they are the same event and will group them together, where new documents are updates of earlier ones.

One of the main empirical foci of this work is an exploration of the most effective features within social media data for event detection. Consider a text stream $D = (D_1, D_2, \ldots, D_n)$, where D_i is a document and the length of D is $|D|$. A document d_i consists of a set of features, $(F_1, F_2, \ldots, F_k)$ and is reported at time t_i. In the text stream D, $t_i \leq t_j$ if $i < j$. Dividing the text stream, D, into time windows, W_i, of the same length, for example, per day, per 12h, per minute. The problem of real-time event detection is to find an optimal set of features to detect events in each unit time, where all known events are identified and correctly summarised. Figure 1 illustrates the relationship between events and the different sets of features.

4. METHODS AND SYSTEM DESIGN

Social media generally produces a large number of posts per hour with a wide variety of topics, rendering human monitoring impractical. Our proposed framework is based on collecting data over time windows for a given location that supports the automatic detection and summarization of events from social media. In this section, we describe each step in more detail.

4.1. Data Collection

We collect user-generated updates directly from social media using streaming Application Programming Interface (API), as it allows subscription to a continuous live stream of data. Our goal is to detect events in a given location without prior knowledge of these events. The most open and widely used social media is Twitter. Facebook and Google+ are much more closed by design to their users. Terrestrial events always occur in space and thus we collect tweets based on a set of keywords that describe a region (e.g., Iraq, Syria, Egypt) using different languages. We also collect tweets from users who selectively add the required region as their location in their profile metadata or turn on GPS on the smartphones. Finally, we make use of geographic Hashtags (e.g., #Ramadi, #Aleppo, #Cairo, #Dubai). Data are stored using a MongoDB database that is suitable for storing short texts and supports different indices with a standardised querying interface [Alsaedi et al. 2014].

4.2. Pre-Processing

We perform basic text processing techniques to optimise the text for use as features. This included stop-word elimination and stemming (Khoja stemmer [Diab et al. 2004]) for Arabic text and a Porter stemmer [Porter 1997] for English and other Latin documents. Also, posts that were fewer than three words long were removed, as were messages where over half the total words were the same word, since these posts were less likely to have useful information.

4.3. Classification

This step aims to distinguish events from noise or irrelevant posts. The classification step identifies large-scale events and subsequently reduces the number of posts to be processed in the following steps (small-scale event identification and summarization), because these steps will process only event-related tweets. Words from each status are considered as features, and a Naive Bayes classifier [Lewis 1998] was chosen for the classification task over a number of leading methods, such as SVMs [Joachims 1998] and Logistic Regression [Friedman et al. 1998] following empirical baseline testing where Naive Bayes outperformed other algorithms when detecting large-scale events (see Section 6.2.1). Naive Bayes is relatively fast to compute and easy to construct, with no need for any complex iterative parameter estimation schemes. Unlike SVMs or Logistic Regression, the Naive Bayes classifier treats each feature independently. Naive Bayes also tends to overfit less than Logistic Regression [Petrović et al. 2010].

To train and test the classifier, we use human annotators to manually label 5,000 randomly selected tweets into two classes, “Event” and “Non-Event,” which were collected at five different hours of the first and second weeks of October 2015 (first dataset, Section 6.1.1). These five hours were sampled uniformly at random from five bins partitioned according to the volume of messages per hour over these two weeks. To ease the annotation process, examples were shown to the annotators along with their respective classes. The details of the collection and annotation process are expanded in Section 6.2. Training data (tweets) where transformed into feature vectors (see Section 5) and their corresponding category (event or non-event) were provided to the classifier,

constituting the training set. From the training data the likelihood of each post belonging to either class was derived on the basis of feature occurrence in the training data. When a new example is presented, the class likelihood for the unseen data is predicted on the basis of the training instances.

4.4. Online-Clustering

The classification step separates event-related documents from non-event posts (such as chats, personal updates, spam, incomprehensible messages). Consequently, non-event posts are filtered. To identify the topic of an event, including determining potentially disruptive events, we define a temporal, spatial, and textual set of features, which are detailed in the next section. We then apply an online clustering algorithm, which is outlined in Algorithm 1.

Using a set of features $(F_1, \ldots, F_k)$ for each document $(D_1, \ldots, D_n)$, we compute the cosine similarity measure between the document and each cluster $(C_1, \ldots, C_k)$, where the similarity function is computed against each cluster c_j in turn for $j = 1, \ldots, m$ and m is the number of clusters (initially $m = 0$). We use *the average* weight of each term across all documents in the cluster to calculate the centroid similarity function $E(D_i, c_j)$ of a cluster. The threshold parameters are determined empirically in the training phase (as was shown in Section 6.2.2).

ALGORITHM 1: Online Clustering Algorithm

Input: n set of documents $(D_1, \ldots, D_n)$
Threshold τ
Output: k clusters $(C_1, \ldots, C_k)$
while τ ;
is ;
given **do**
  compute the centroid similarity function $E(D_i, c_j)$ of each cluster c_j;
  if *centroid similarity* $E(D_i, c_j) \geq \tau$ **then**
    1) A new cluster is formed containing D_i;
    2) The new centroid value $= D_i$.
  else
    1) Assign it to the cluster which gives the maximum value of $E(D_i, c_j)$;
    2) Add D_i to cluster j and recalculate the new centroid value c_j.
  end
end

The decision to use an online clustering algorithm was taken for three main reasons: (i) it supports high-dimensional data as it effectively handles the large volume of social media data produced around events; (ii) many clustering algorithms such as kf-means require previous knowledge of the number of clusters. Because we do not know the number of events *a priori*, online clustering is suitable in that it does not require such input; (iii) partitioning algorithms are ineffective in this case because of the high and constant sheer scale of the user contributed messages (as discussed in Section 2).

4.5. Summarization

After clustering the documents, the next natural step is to automatically summarize and represent the topics being discussed within the clusters. Each cluster may contain hundreds of posts, images, or videos, and the task of finding the most representative update or extracting top terms (topics) is crucial to making the output useful and interpretable for policy and decision makers.

Our approach is inspired by the fact that users tend to use similar words when describing a particular event, as well as observations obtained from Reed et al. [2006]:

(1) High-frequency words like stop-words occur in approximately the same percentage of documents no matter whether the document set is small or large and, similarly, low-frequency words like "murder" occur very rarely across small and large datasets.
(2) The document frequency distribution of one corpus can be used to approximate another.

We propose a novel temporal Term Frequency–Inverse Document Frequency (TF-IDF) that generates a summary of top terms without the need of prior knowledge of the entire dataset, unlike the existing TF-IDF approach [Salton and Buckley 1988] and its variants. Temporal TF-IDF is based on the assumption that words that occur more frequently across documents over a particular interval (timeframe) have a higher probability of being selected for human created multi-document summaries than words that occur less frequently [Vanderwende et al. 2007].

Typically, the TF-IDF approach requires knowing the frequency of a term in a document (TF) as well as the number of documents in which a term occurred at least once (DF). The need for *a priori* knowledge of the entire data set introduces a significant challenge of using this approach where continuous data streams must be summarized in real time as an event unfolds. In addition, the adopted scheme must be flexible to update frequently (every minute, 10 mins, hourly, 3h depending on the time-frame size). Hence, the iterative calculation of term weights should be taken into consideration.

To overcome these limitations, we introduce the temporal TF-IDF where we consider a set of posts in a cluster to be represented as a document. The total number of clusters equals the total number of documents that is a subset of the entire dataset or corpus. This reduces the overall computational complexity and overcomes the limitations of the TF-IDF-based approaches in which the document set to be clustered must be known in advance. After the first cluster timeframe, we use clusters from the previous timeframe with the documents in the recent one to add more relevance and usefulness to our results such as emerging keyword. Consequently, we use the document frequency distribution of two timeframes instead of one, taking into account the changing event dynamic and narrative. We define the TF-IDF weighting scheme of a new document d for a collection C (from two clusters) as

$$w_{ji} = \frac{1}{norm(d_i)} f_{ji} \times log\left(1 + \frac{N}{N_j}\right),$$

where f_{ji} is the term frequency of word in document d_i, N_j is document frequency of word in a collection, and N is the total number of documents in the collection. In order to avoid the bias caused by different document lengths, the length of each document vector is normalized so it is of unit length $norm(d_i)$. This summarizer selects the most weighted post as summary as determined by the Temporal TF-IDF weighting.

5. FEATURE SELECTION

Feature selection is a fundamental problem in mining large data sets. The problem is not limited to the total processing time but involves dimensionality reduction to achieve better generalization. In Alsaedi et al. [2015], we analysed in-depth three types of features, namely temporal, spatial, and textual features. We used an improved version of the unsupervised feature selection proposed by Mitra et al. [2002] to optimize the textual features. In this article, we use the standard metric of Normalized Discounted

Table I. Overview of Features Used by Our Framework

	Description
Temporal features	We retain the most frequently occurring terms in a cluster in hourly time frames and compare the number of posts published during an hour that contain term t to the total number of posts during that hour. The 1h time window leads to the best performance, as it requires much less computational time producing the second best accuracy compared to the other settings as shown in Alsaedi et al. [2015].
Spatial features	We use three statistical location approaches to extract geographic content from clusters: (1) The source latitude and longitude coordinates are extracted (if provided by the user). (2) The use of shared media (photos and videos) GPS coordination of the capture device. (3) OpenNLP (http://opennlp.sourceforge.net) and Named-Entity Recognition (NER) are used for geotagging the tweet content to identify places, street names, landmarks, and so on.
Textual features	*Near-Duplicate measure*: The average content similarity over all pairs of messages posted in a (1h time slot) cluster. If the two posts have a very high similarity (the cosine similarity is above 0.9), then we assume that one of them is a near duplicate of the other.
	Share ratio: We calculate this attribute by normalizing the number of times a post (photo or video) appears in a timeframe to the total number of messages in that timeframe.
	Mention ratio: Number of mentions (@) relative to the number of posts in the cluster.
	Hashtag ratio: Number of hashtags (#) relative to the number of posts in the cluster.
	Url ratio: Number of posts that contain links relative to the number of posts in the cluster.
	Text sentiment: For each post, we use the SentiStrength [Thelwall et al. 2011] algorithm to compute a positive, neutral, or negative sentiment score. Then we compute the average cluster-level sentiment in order to study the effect of average positive or negative sentiment with respect to events.
	Dictionary-based feature: This bag-of-words model uses a dictionary of trigger words to detect and characterize events; these are manually labelled by experts and decision makers. We use a subset of verbs, nouns, and adjectives from (events and actions) category from WordNet (http://globalwordnet.org) to create a dictionary model. We have created nine lexicons regarding events from the clustering scheme, one for each popular topic, including weather, communication, energy, transportation, health, crime, terrorism, politics, and others. The total number of terms is 1,538. Table II shows our lexicons with topics and examples in each category.

Cumulative Gain (NDCG) [Croft et al. 2009] for the feature selection task. Table I gives a brief description of these features.

6. EVALUATION

We evaluate our identification framework using two real-world datasets through a set of carefully designed experiments. This section details the datasets used and our approach to evaluating the proposed system. We analyse three sets of features associated with tweets that define an event (temporal, spatial, and textual features) to determine the best feature combinations.

6.1. Experimental Settings

6.1.1. Datasets. We use two real-world datasets.

Middle East 2015 Our first dataset consists of 40 million tweets and was collected from October 1, 2015, until November 30, 2015, using Twitter's Streaming API. This is a general collection of tweets used to show that our event detection system is useful for extracting information from sociallygenerated content on a broad range of topics. Our aim is to monitor and analyze events and disruptive events in a particular

Table II. Topics and Sub-Topics with Examples Taken from the Corresponding Lexicons

Topics	Sub-Topics	Examples	Total
Weather	Heavy rain, Wind, Fog, Storm, High waves, Flooding, Heat waves, Cold.	Verb: rain, suffer, Noun: fog, visibility, Adjective: heavy, cold, hot,	155
Energy	Blackout, Power lost, Fire, Electricity cut, Water supply, Gas leak.	Verb: lose, leak, continue, Noun: power, signal, authority, Adjective: long, delay,	82
Communication	Signal, Communication lost, Breakdown.	Verb: communicate, restore, Noun: signal, company, Adjective: Technical, temporary,	33
Transportation	Public transport, Traffic jam, Accidents, Crashes, Long delay, Services, Hazardous, Roads, Cancellation.	Verb: see, take, Noun: car, crash, plane, train, Adjective: fast, dangerous,	258
Health	Flu, Fever, Virus, Disease, Illness.	Verb: spread, circulate, Noun: influenza, rate, season, Adjective: medical, serious,	45
Crime	Shooting, Theft, Damage, Kidnapping, Homicide, Murder, Manslaughter, Drugs, Threat, Fight, Money laundering, Sexual assault, Illegal, Fraud, Alcohol, Corruption, Internet Crimes.	Verb: witness, report, arrest, Noun: victim, blood, abuse, Adjective: vulnerable, brutal,	341
Terrorism	Terrorist Activities, Explosion, Explosives, Weapons, Hostage, Armed robbery, Bomb, Attacks, Violence, Stabbing, Suicide, Hacking.	Verb: release, support, Noun: email, Syria, knife, Adjective: suspicious, explosive,	230
Politics	Riots, Protests, Political insults, Celebrities, Occasions, News.	Verb: organise, group, Noun: chaos, looting, arson, Adjective: corrupt, violent,	256
Others	Religious, Financial, Social incidents, Death, Rumour.	Verb: spread, die, claim, confirm Noun: truth, correction, rumour, Adjective: false, incorrect,	129

region and we used the middle east as our location, collecting tweets from users who chose one of the middle east countries as their location. Nearly 425,000 unique hashtags appear in the 40 million tweet corpus from roughly 18,000,000 distinct user accounts.

England Riots 2011 Our second dataset consists of 1.6 million tweets and was generated during the 2011 riots in England, which began as an isolated incident in Tottenham on August 6 but quickly spread across London and to other cities in England and gave rise to levels of looting, destruction of property, and violence not seen in England for more than 30 years [MPS 2012]. This event was selected because of a publicly available record of intelligence and incidents reported during this period that provides us with a gold standard evaluation dataset. Data were purchased from Twitter reseller Gnip from August 6 to August 12, 2011, using the following quer: #londonriots OR #tottenham OR #enfield OR #birminghamriots OR #UKRiots OR #Croydon OR #hackney OR #tottenhamriots OR #tottenhamshooting OR #Londonriots OR #riotcleanup OR #rioting OR #manchesterriots OR #liverpoolriots OR #bullring OR #enfieldriots OR #croydonriots OR #Londonsburning OR #prayforlondon. We usually select the most popular hashtags that attract the users' attention. This is reflected as peaks in the use of these hashtags of tweeting rates. In the process of selecting these hashtags, the system only considers sudden increases with respect to the recent tweeting activity using these hashtags.

Table III. F-measure for Different Classification Algorithms

	Naive Bayes	Support Vector Machines	Logistic Regression
Accuracy	86.13	83.93	80.13
Precision	83.64	80.84	78.91
Recall	87.95	86.54	82.90
F-measure	85.43	83.86	80.22

6.1.2. Evaluation Matrix. We used standard classification metrics; precision, recall, and F-measure to measure the effectiveness of our framework. We have also implemented two well-known information retrieval metrics, namely, *Precision@K* and *NDCG* [Croft et al. 2009], to evaluate the overall performance of the event detection task. *Precision@K* reports the fraction of correctly identified events out of the top-k selected clusters, averaged over all hours, whereas the NDCG metric ranks the top events relative to their ideal ranking as well as NDCG supports graded judgments and rewards relevant documents in the top ranked list.

6.2. Framework Evaluation

6.2.1. Classification. The aim of the first experiment is to elect the best classifier from leading machine-learning algorithms for the purpose of identifying event and non-event tweets. The classification algorithms used in the experiment were Naive Bayes [Lewis 1998], a statistical classifier based on the Bayes' theorem; Logistic Regression [Friedman et al. 1998], a generalized linear model to apply regression to categorical variables; and SVMs [Joachims 1998], which aims at maximizing (maximum margin) the minimum distance between two classes of data using a hyperplane that separates them.

Using a systematic sample extracted from the Middle East dataset, three annotators manually labelled 5,000 tweets into two classes, "Event" and "Non-Event," to create a training dataset. Event instances outnumber the non-event ones as the training set consisted of 1,900 Non-Event tweets and 3,100 event-related tweets. Agreement between our three annotators, measured using Cohen's kappa coefficient, was substantial (kappa = 0.807). A tenfold cross validation was used to train and test the classifiers. Table III shows a comparison of classifiers with unigram presence, which indicates that the Naive Bayes classifier produces the best results.

6.2.2. Online Clustering. Following the classification output, we employed three more human annotators to manually label 1,600 clusters, randomly selected from the top-20 fastest-growing clusters according to hourly message volume at the end of each hour in October (800 clusters) and November (800 clusters) 2015. October data were used for training and refining the clustering algorithm, and November was used to test and evaluate the clustering output. The agreement between annotators was calculated using Cohen's kappa (kappa = 0.782), which indicates an acceptable level of agreement. For testing, we only used the clusters on which all annotators agreed (602).

Recall that our clustering algorithm presented in Section 4.4 relies on the Threshold τ. To tune the clustering threshold τ for a specific dataset, we run the clustering algorithm on a subset of labelled training data. We evaluate the algorithm's performance on the training data using a range of thresholds and identify the threshold setting that yields the highest-quality solution according to a given clustering quality metric (here we implement the f-measure). Threshold values for the online clustering algorithm were varied from 0.10 to 0.90 at graded increments of 0.05% with a total of 17 tests in order to find the best cutoff of $\tau = 0.45$ (63 character difference). Figure 2 illustrates the F-measure scores for different thresholds where the best performing threshold $\tau = 0.45$

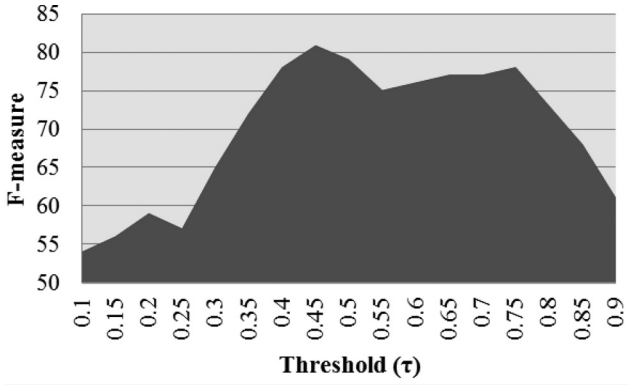

Fig. 2. F-measure of the online clustering algorithm over different thresholds.

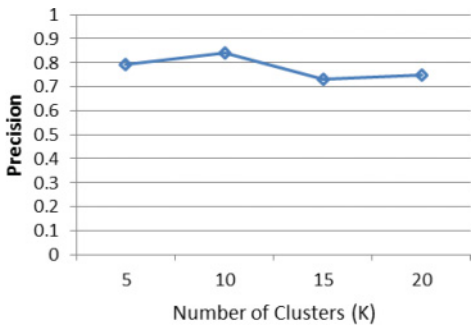

Fig. 3. *Precision@K* of our classification-clustering framework.

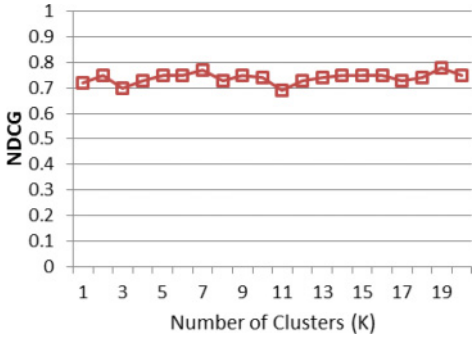

Fig. 4. *NDCG* at *k* of our classification-clustering framework.

seems to be reasonable because it allows some similarity between posts but does not allow them to be nearly identical.

As shown in Figures 3 and 4, our proposed framework is effective and performs well both in the *NDCG* and *Precision@K* evaluation measures. In fact, our framework discovers many real-world events such as the refugee crisis and its implications, disasters and terrorist attacks (e.g., Paris attacks, Beirut bombings, etc.), the war against Islamic State in Iraq and Syria (ISIS), as well as many other events and stories. Some large events identified by our system are shown in Table IV.

6.3. Feature Selection

To enhance the event detection system, we used feature selection to refine our model, focussing now on "disruptive events," which were selected from the human annotated clusters. We hypothesized that not all features are expected to lead to better system performance or contribute equally to improved machine classification and/or clustering accuracy.

For the clustering evaluation process, one of the annotators' tasks was to label or categorize the cluster based on the topic of the cluster as politics, finance, sport, entertainment, technology, culture, disruptive event, and other-event. The other-event category represents all other events that are not related to the above categories. Then we run a one-vs-all strategy using a Naive Bayes classifier where the disruptive event

Table IV. Examples of Automatically Detected Global Events by Our Framework in November 2015

Date	Event	Event Keywords
12 Nov 2015	• Beirut bombings • refugee crisis	• Beirut, bombings, two, twin blasts, Suicide, explosives, 43 dead, ISIS, 239 injured, attacks, Syria, suspects, Hezbollah, #BeirutBombings • refugees, migrants, Europe, asylum, Germany, Syria, Afghan, Iraq, Greece, Macedonia, fence, police, Border, #migrants, #MigrantCrisis, #refugees
13 Nov 2015	• Shooting in Mount Hebron, West Bank • Attacks in Baghdad • Paris attacks	• Shooting, gunman, shot, car, family, 2 killed, 2 injured, teen, unidentified, run • Terrorist, attacks, Baghdad, suicide, bombing, 19 killed, 33 wounded, ISIS, #Baghdad, #Iraq • Shooting, restaurant, bar, explosions, gunshots, Bataclan theatre, hostages, Stade de France, 130 killed, more 200 injured, #ParisAttacks, #prayforparis, #notafraid, #porteouverte

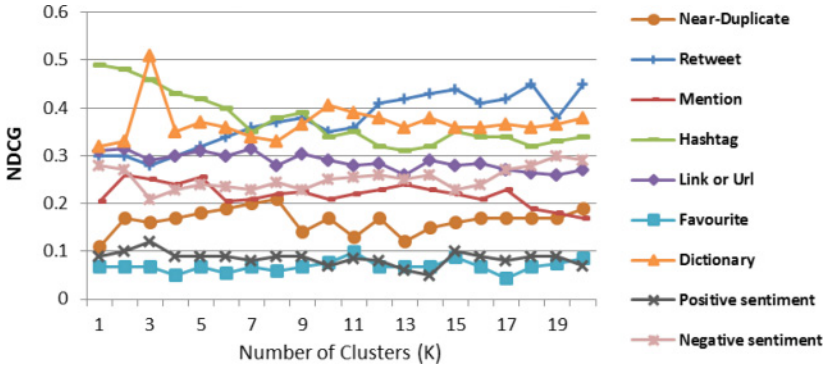

Fig. 5. Performance of various proposed features.

class is fitted against all the other classes in order to identify the performance of the disruptive event features.

We investigate the discriminative power of the textual features, allowing us to remove the least-discriminative features to reduce the computational workload required to compute the results. The results are shown in Figure 5, which illustrates the NDCG scores for each feature.

The near-duplicate measure, the favourite ratio and the positive sentiment ratio are the least-discriminative features, which suggest that they appear in all different types of posts, not only in disruptive events. The dictionary-based model, the retweet ratio, and Hashtag ratio are the most discriminative, suggesting that references to present time and references to descriptive terms (e.g., live, breaking etc.) are good discriminators. The retweet ratio suggests that other users pick up on event commentaries and propagate them further through the network. Linking content features such as Hashtags and URLs are also very predictive of events, suggesting that tweets reporting events provide evidence or further information (via Uniform Resource Locator (URL)) or are bound to an event and made more discoverable via a self-defined topic discriminator in form of a Hashtag.

Another important observation in Figure 5 is that the negative sentiment model outperforms the positive sentiment model in NDCG scores. Hence, negative sentiment posts have high adoption rate regarding reporting disruptive events. This is due to the fact that reporting disruptive events usually involves negative terms and sentiment. Another possible reason is that posts with negative sentiment are more likely to be retweeted, as shown in Hecht et al. [2011], Ma et al. [2013], and Thelwall et al. 2011].

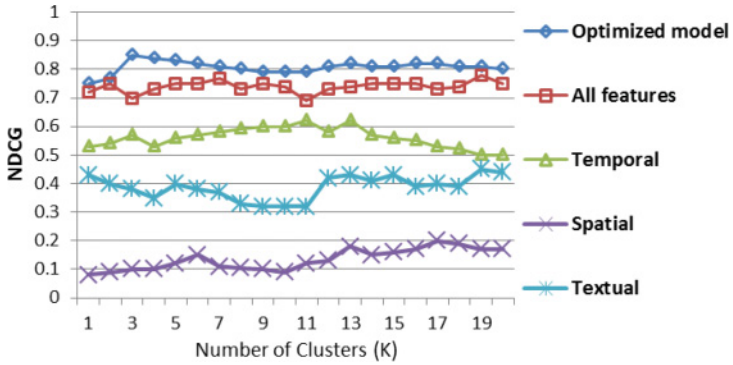

Fig. 6. Comparison of different models for the event identification task according to the NDCG scores.

Table V. Number of Small-Scale Real-World Events

Incident Type	Number of Real-world events
Car accident	6,248
Fire Incident	3,907
Shooting	143
Stabbing	77
Protest	109

Figure 6 compares the performance of various models: First, we use individual feature models: temporal, spatial, and textual (the textual model uses all the features from Figure 5). For the next model, we use a combination of all features. Finally, we use the temporal feature, the spatial feature, and only the most effective textual features from Figure 5 (above 0.25 in NDCG evaluation measure) to build our optimized model.

The temporal feature model substantially outperforms spatial and textual models, obtaining a performance score of about 13.2% over textual features and about 38.7% on average compared with spatial features. Hence, the temporal feature is the most effective in detecting events. Using the textual feature model, we are still able to obtain a reasonable performance of, on average, 40% content about an event and provide situational awareness information about that event. However, it is emphatically not the case when using the spatial feature in isolation, thus leading to the conclusion that spatial features are weak indicators to be implement on their own.

A combination of all three features results in the best performance, because it gives the best of all three set of features with a much better performance, but further investigation by removing unnecessary textual features (such as near-duplicate measure, favourite ratio, mention ratio and the positive sentiment ratio) yields the best model performance (average 0.802 NDCG score). The number of small-scale events identified by our system differentiated by incident type is shown in Table V.

6.4. Case Study: Reading the Riots

In order to further validate our approach, we evaluated it against other leading approaches using the 2011 riots dataset. We used the model produced using the training set, which was evaluated in Section 6. We do not train specifically on the riots data; thus we are testing the generality of our model for a real-world example. Our evaluation is based on high-quality ground-truth data from public Metropolitan Police Service (MPS) reports. On August 4, Mark Duggan was shot in Tottenham by police officers. On the evening of August 6, following a peaceful protest march to a Tottenham police station, organised by the victim's friends and family, the first outbreaks of public

Table VI. Comparison of Approaches for Disruptive Event Detection

	Numberof real-world events identified				
Incident Type	Police Intelligence	Ours	Becker et al.	Spatial LDA	Zubiaga et al.
Car Accident	—	285	108	74	92
Fire Incident	311	214	121	186	127
Shooting	4	3	1	3	0
Stabbing	5	4	0	3	1
Protest	187	143	106	163	32

disorder occurred. Then they quickly spread across London and to other cities in England, and the levels of crimes and offences increased dramatically, including looting, violence, burglary, arson, and other disorder-related offences, which makes this case study and our collected dataset ideal for large-scale event detection (riot) and smaller disruptive event detection—from small-scale looting incidents of local shops to one of the largest arsons in Europe [MPS 2012]. In terms of social media, the MPS is clear that its capability for using social media networks as engagement was in its infancy at that time [MPS 2012].

We compare the output of our framework with similar existing methods, namely Spatial LDA [Pan and Mitra 2011], unsupervised [Becker et al. 2011a], and [Zubiaga et al. 2012] methods. Spatial LDA [Pan and Mitra 2011] combines an LDA model [Blei et al. 2003] with temporal segmentation and spatial clustering. Becker et al. [2011a] use an unsupervised clustering technique to group topically similar tweets together and computed features (temporal, social, topical, and Twitter-specific) that can be used to train a classifier to distinguish between event and non-event clusters. Zubiaga et al. [2012] explores the real-time summarization of scheduled events using a two-step system: (i) sub-event detection and (ii) tweet selection. The first step is based on peaks detection (reflected as peaks in the histogram of tweeting rates) with an enhancement of two ideas: the sudden increase in the tweeting rate and the outlier detection. The tweet selection step selects a representative tweet after ranking all tweets that were sent during the sub-event. They use the Kullback-Leibler divergence weighting scheme for the tweet ranking.

All three methods have successfully been applied to event detection, and thus we aim to outperform these using our proposed temporal TF-IDF and online clustering algorithms. Table VI presents the performance of the comparative experiments in terms of the number of real-world events (as reported to MPS) detected, system Precision, system Recall, and the F-measure. Precision is defined as the fraction of the retrieved documents that are relevant. Recall is defined as the fraction of the relevant documents retrieved to the total number of relevant documents should have been returned, and the F-measure is defined as a harmonized mean of precision and recall [Weng and Lee 2011; Schulz et al. 2015; Pan and Mitra 2011; Becker et al. 2011a].

	Ours			Becker et al.			Spatial LDA			Zubiaga et al.		
Incident Type	P	R	F	P	R	F	P	R	F	P	R	F
Fire Incident	74.64%	68.81%	71.61%	39.77%	38.91%	39.34%	60.09%	59.81%	59.95%	42.26%	40.84%	41.54%
Shooting	57.41%	75.00%	65.04%	30.22%	25.00%	27.36%	52.75%	75.00%	61.94%	8.43%	0	0
Stabbing	63.64%	80.00%	70.89%	3.55%	0	0	45.18%	60.00%	51.55%	18.29%	20%	19.07%
Protest	77.82%	76.47%	77.14%	53.85%	56.69%	55.23%	38.78%	33.67%	36.04%	32.67%	17.11%	22.46%

According to Table VI, our proposed methodology is effective and outperforms other approaches. This is the case even though the topics in the *Riots 2011* dataset are about disruptive events described by a diverse vocabulary and often comprising relatively few posts per incident. The MPS did not include car accidents and vehicle damages related to the riots, and, hence, we could not compute the recall measure. However, the

number of events detected indicate that our framework is able to detect 4 times more real-world incidents compared to Spatial LDA and at least twice as good as Becker et al.

We offer the following explanation as to how the systems we tested could be impaired: First, not all events reported in the MPS report using traditional intelligence are reported in social media and vice versa. Second, the Twitter API only allows 1% of the total number of tweets for researchers, which means that we fail to report the 99% of online conversations. Conversely, 1% is in fact a huge corpus of tweets per day for sampling and researching purposes but, however, not enough to cover all disruptive events reported. The presence of rumours and false information during the 2011 England riots and generally during emergencies and disasters is another issue and effects the reported results negatively. The detection of rumours in social media is beyond the scope of this article and is reserved for future work. By studying the lifecycle of several rumours as well as by investigating the propagation, we may be able to effectively identify social media rumours.

In addition, classification after clustering has a crucial impact on performance in terms of quality, especially with moving time windows. Many of the small clusters are filtered out, since they do not exceed the predefined thresholds and are considered non-relevant events (noise). This eliminates many of them together with noise, which confuses the scoring and ranking of event detection. This explains why the Becker et al. approach performance is smaller than ours. The results in Table VI also show that the Spatial LDA approach outperforms the Becker et al. system only in larger events, such as fire incidents. However, it fails to achieve such results in other cases due to the fact that tweets are short, and a collection of tweets per hour may contain many more topics in multiple small-scale cases such as car accidents or small group protests. The approache in Zubiaga et al. [2012] and similar systems like those in Shen et al. [2013] and Yajuan et al. [2012] are limited to scheduled events such as soccer games; they also require the starting time in order for the system to start looking for new sub-events. This explains why the performance of the Zubiaga et al. approach is worse than the results reported in this article.

Visualisations are arguably well suited to displaying real-time disruptive events sensed from social streams. We visualize the real-time output from our system alongside the post-event visualisation provided by MPS in their public report [MPS 2012] in Figure 7. For space limitation, we only present results of the Enfield borough, although the MPS report [MPS 2012] presents the results for three case studies (Enfield, Croydon, and Wandsworth). As can be seen from Figure 7, most of the disruptive events, including looting, arson, violence, and so on, have been successfully identified and monitored in real time and in some cases our system provides information ahead of traditional intelligence. Furthermore, Table VII present the time difference between the disruptive incidents being identified by our framework and the corresponding police information. The columns show the time of the events being discovered by the summarization of our system, the time of the intelligence that was reported by officials, and how much Twitter leads police intelligence. Entries marked in bold occur first.

From Table VII, we observe that Twitter information extracted by our system leads police sources most of the time, and police sources lead only twice and by 10min in both cases. The delay can result from the time for posting a tweet by a user, the time to index the post in Twitter servers, and the time to make queries by our system. In fact, our system detected all of the disruptive events that were reported by officials far faster than them, on average 23 mins. The task of identifying accurate intelligence during the disorder is much more valuable if it is received in real time to enable decision makers to move ahead of such events. These results support the hypothesis that information extracted from social media can be used effectively as a valuable additional source of

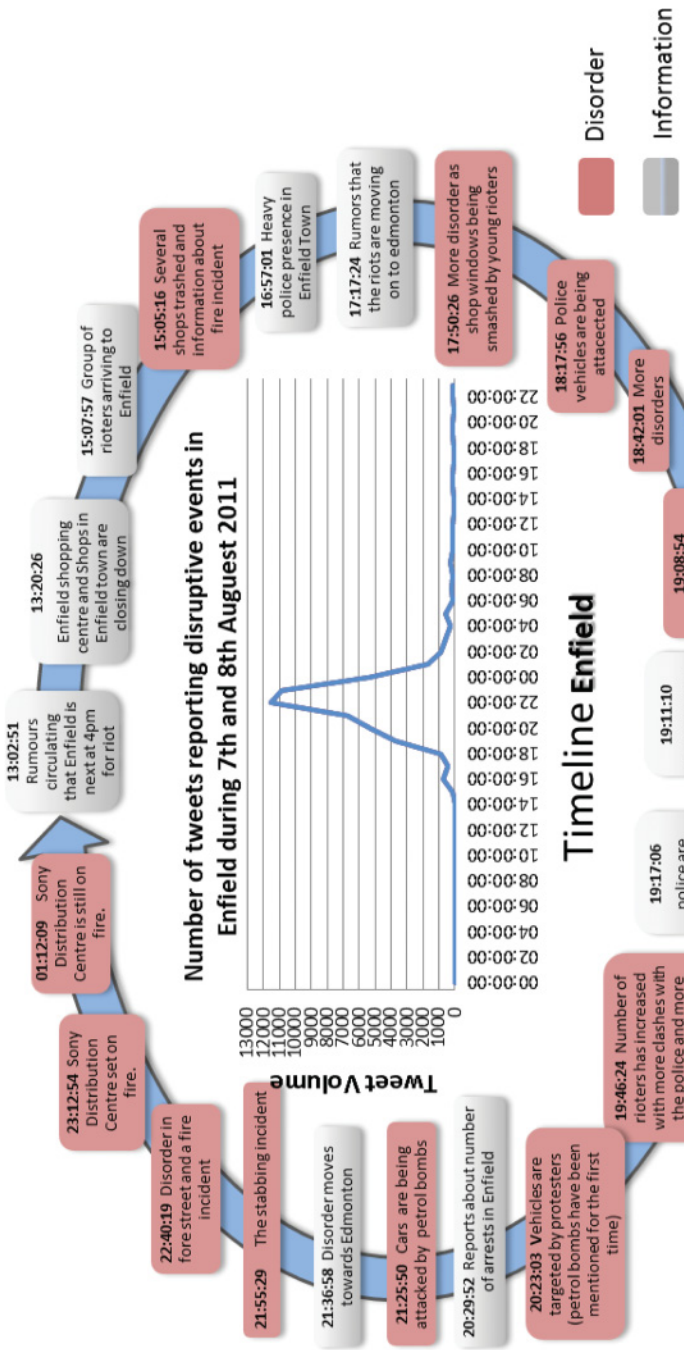

Fig. 7. Comparison of disruptive events obtained by our framework (top) and MPS (bottom) for Enfield borough.

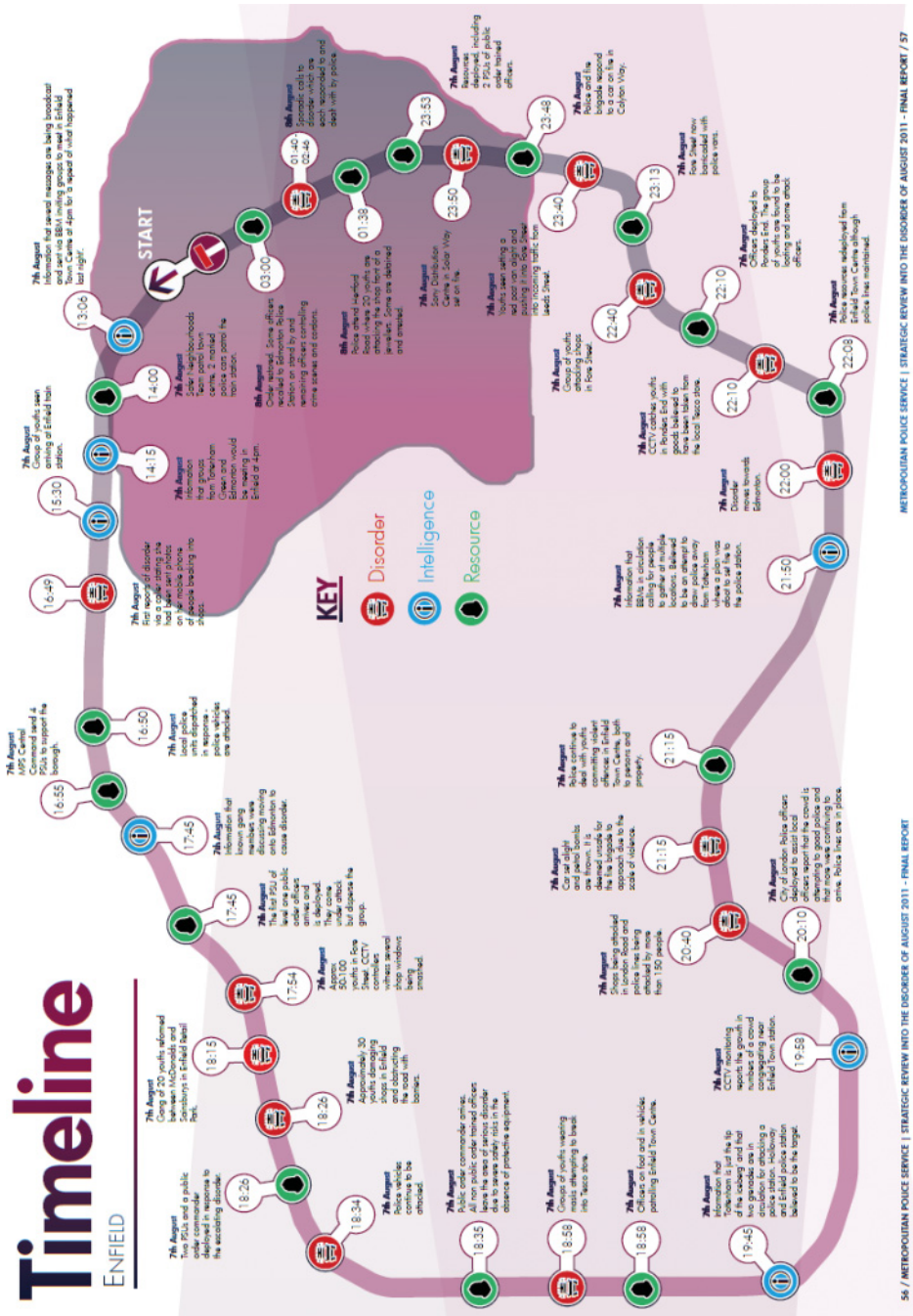

Fig. 7. Continued

Table VII. Disruptive Event Exploration Using Police Intelligence and Summarization by Our System for Enfield Borough on August 7, 2011 (+ When Twitter Leads)

Police Intelligence	Summarization by our system	Time/Police	Time/Our system	Lead
Information that several messages are being broadcast to meet in Enfield Town Centre at 4pm for a repeat of what happened last night.	Rumours circulating Enfield is TONIGHT. #Tottenham #Riots	13:06	**13:02**	+0:04
Information that groups from Tottenham Green and Edmonton would be meeting in Enfield at 4pm.	#rumour has it #enfield riot k.o's at 4!	14:15	**13:19**	+1:04
Group of youths seen arriving at Enfield train station.	The rioters are now in Enfield and Edmonton. #londonriots	15:30	**15:07**	+0:23
First reports of disorder via a caller stating she had been sent photos on her mobile phone of people breaking into shops.	Ok its officially kicking off in #Enfield Town, one fire and hmv has been smashed in, people coming from all over london to #loot.	16:49	**15:37**	+1:12
Information that known gang members were discussing moving onto Edmonton to cause disorder.	not feeling the rumors that the rioters are looking to move to edmonton and #enfield town. DON'T YOU PEOPLE THINK YOU'VE DONE ENOUGH!!!!	17:45	**17:17**	+0:28
The first PSU of level one public order officers arrives and is deployed. They come under attack.	ok so 9 police vans just drove past my house! ok make that 10! #enfield	17:45	**17:39**	+0:06
Approximately 30 youths damaging shops in Enfield and obstructing the road with barriers.	RT Police car wrecked in Enfield - most rioters looked under 16, lots of young girls throwing concrete slabs through shop windows. #enfield	18:26	**17:50**	+0:36
Police vehicles continue to be attacked.	police car trashed RT @XXXXX: BREAKING: This just happened at #EnfieldTown; Police outnumbered once again; http://yfrog.com/kf4rlauj	18:34	**18:17**	+0:17
Groups of youths wearing masks attempting to break,into Tesco store.	Police horse vans in #enfield tesco car park http://yfrog.com/h7eyhirj	**18:58**	19:08	−0:10
CCTV monitoring reports the growth in numbers of a,crowd congregating near Enfield Town station.	#Enfield Police attacking riotmob with batons and,dogs in the town. Over 230+ riot mobs in #Enfield town	19:58	**19:49**	+0:09
Car set alight and petrol bombs are thrown. It is deemed unsafe for the fire brigade,to approach due to the scale of violence.	*ALERT* Protestors are throwing petrol bombs on passing cars on the A10 from #Tottenham to #Enfield. Avoid the road.	**21:15**	21:25	−0:10
Not reported in the official report as it might not be relevant to the Riots	Teenager stabbed outside #Edmonton WorkingMen's Conservative Club. Medics on scene. #Enfield		21:58	
Disorder moves towards Edmonton.	I hear edmonton is next #enfield	22:00	**21:46**	+0:14
CCTV catches youths in Ponders End with goods believed to have been taken from the local Tesco store.	#Enfield disturbances now spreading to Ponders End #PondersEnd	22:10	**21:36**	+0:34
Group of youths attacking shops in Fore Street.	Carphone warehouse getting smashed up in #edmonton, ridiculous!!	22:40	**21:54**	+0:46
Youths seen setting a red post van alight and,pushing it into Fore Street into incoming traffic from Leeds Street.	Car near fore street about to explode, about 50 man standing off with police. #Edmonton	23:40	**23:12**	+0:28
Sony Distribution Centre in Solar Way set on fire.	40 firefighters at a fire in a warehouse on Solar Way in Enfield. #LondonRiots #Enfield	23:50	**23:19**	+0:31

intelligence as well as to bridge that gap between the use of "big data" and modern policing in order to maintain situational awareness and enhance public safety and decision making.

7. CONCLUSION

In this article, we have presented an integrated framework for detecting real-world events, both large and small, using the Internet-enabled social networking site Twitter. Event detection was performed in several stages: data collection, preprocessing, classification, clustering, and summarization. We have also presented several experiments on various features and show how they can be implemented to discriminatively distinguish between events, particularly disruptive events. The results indicate that it is not adequate to consider temporal, spatial, or content-based aspects in isolation. Rather, a combination of features covering all these aspects leads to a robust system that encourages the best event detection results. Extensive experiments were conducted to evaluate the effectiveness of the proposed framework using large real-world datasets. Our experiments suggest that our framework yields better performance than many leading approaches in real-time event detection, and using a real-world ground truth published by the MPS after the 2011 riots in England, we showed our system to detect events far quicker than they were reported to MPS. These promising results do not necessarily enable us to "predict a riot" but can provide actionable insights before they were received during the events.

There are many directions for future work. One of the main directions is to improve the location detection and disambiguation process for small-scale events. Another direction is to consider more features in the context of event discovery such as social network features (community influence detection), visual features (images and video), and semantic features. We intend to further evaluate the summarization output to not only map onto real events but to provide qualitatively useful output for decision making. Finally, the detection of rumors in the social media, the analysis of the distinctive characteristics of rumors, and the way in which they propagate in the microblogging communities will be addressed in the future. Spammer detection in various online social networking platforms is another interesting task that is reserved for future work.

REFERENCES

Fabian Abel, Claudia Hauf, Geert Houben, Richard Stronkman, and Ke Tao. 2012. Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web (WWW'14 Companion)*. ACM, 305–308.

Manoj K. Agarwal, Krithi Ramamritham, and Manish Bhide. 2012. Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. *Proc. VLDB Endow.* 5, 10 (June 2012), 980–991. DOI: http://dx.doi.org/10.14778/2336664.2336671

Nasser Alsaedi and Pete Burnap. 2015. Arabic event detection in social media. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'15)*. 384–401.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2014. A combined classification-clustering framework for identifying disruptive events. In *Proceedings of the 6th ASE International Conference on Social Computing (SocialCom'14)*.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2015. Identifying disruptive events from social media to enhance situational awareness. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15)*.

Hila Becker, Mor Naaman, and Luis Gravano. 2011a. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

Hila Becker, Mor Naaman, and Luis Gravano. 2011b. Selecting quality twitter content for events. In *Proceedings of the 5th International Conference on Weblogs and Social Media*.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (HLT'11)*. Association for Computational Linguistics, Stroudsburg, PA, 389–398.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Mach. Learn. Res.* 3 (March 2003), 993–1022.

Alexander Boettcher and Dongman Lee. 2012. EventRadar: A real-time local event detection scheme using twitter stream. In *Proceedings of the 2012 IEEE International Conference on Green Computing and Communications (GreenCom)*. 358–367. DOI:http://dx.doi.org/10.1109/GreenCom.2012.59

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* 30, 1–7 (1998), 107–117. DOI:http://dx.doi.org/10.1016/S0169-7552(98)00110-X

Pete Burnap, Amir Javed, Omer Rana, and Malik Shahzad Awan. 2015. Real-time classification of malicious URLs on twitter using machine activity data. In *Proceedings of the 2015 ACM International Conference on Advances in Social Networks Analysis and Mining (SNAM'15)*. ACM, New York, NY.

Pete Burnap, Matthew Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Soc. Netw. Anal. Min.* 4 (2014), 206.

Soudip Roy Chowdhury, Muhammad Imran, Muhammad Rizwan Asghar, Sihem Amer-Yahia, and Carlos Castillo. 2013. Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM'10)*.

Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic summarization of events from social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM 2013)*.

Mário Cordeiro. 2012. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Doctoral Symposium on Informatics Engineering, DSIE*.

Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice* (1st ed.). Addison-Wesley.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004: Short Papers (HLT-NAACL-Short'04)*. Association for Computational Linguistics, Stroudsburg, PA, 149–152.

Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. 2015. Multiscale event detection in social media. *Data Min. Knowl. Discov.* 29, 5 (2015), 1374–1405. DOI:http://dx.doi.org/10.1007/s10618-015-0421-2

Gunes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 22, 1 (2004), 457–479.

Atefeh Farzindar and Khreich Wael. 2015. A survey of techniques for event detection in twitter. *Comput. Intell.* 31, 1 (Feb. 2015), 132–164. DOI:http://dx.doi.org/10.1111/coin.12017

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 1998. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28 (1998), 2000.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 237–246. DOI:http://dx.doi.org/10.1145/1978942.1978976

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.* 47, 4, Article 67 (June 2015), 38 pages. DOI:http://dx.doi.org/10.1145/2771588

Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)*. ACM, 159–162. DOI:http://dx.doi.org/10.1145/2567948.2577034

Akshaya Iyengar, Tim Finin, and Anupam Joshi. 2011. Content-based prediction of temporal boundaries for events in twitter. In *Proceedings of the 3rd IEEE International Conference on Social Computing*. 186–191.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML'98)*. Springer-Verlag, London, UK, 137–142.

George Karypis, Rajat Aggarwal, Vipin Kumar, and Shashi Shekhar. 1997. Multilevel hypergraph partitioning: Application in VLSI domain. In *Proceedings of the 34th Annual Design Automation Conference (DAC'97)*. ACM, New York, NY, 526–529. DOI:http://dx.doi.org/10.1145/266021.266273

David D. Lewis. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning (ECML'98)*. Springer-Verlag, London, UK, 4–15.

Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. 2012. TEDAS: A twitter-based event detection and analysis system. In *ICDE*. IEEE Computer Society, 1273–1276.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, New York, NY, 131–140. DOI: http://dx.doi.org/10.1145/1526709.1526728

Zongyang Ma, Aixin Sun, and Gao Cong. 2013. On predicting the popularity of newly emerging hashtags in twitter. *J. Assoc. Inf. Sci. Technol.* 64, 7 (2013), 1399–1410.

Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 227–236. DOI: http://dx.doi.org/10.1145/1978942.1978975

Michael Mathioudakis and Nick Koudas. 2010. TwitterMonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD'10)*. ACM, New York, NY, 1155–1158. DOI: http://dx.doi.org/10.1145/1807167.1807306

Donald Metzler, Congxing Cai, and Eduard Hovy. 2012. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'12)*. Association for Computational Linguistics, Stroudsburg, PA, 646–655.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP'04)*. Association for Computational Linguistics, 404–411.

Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. 2002. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 3 (March 2002), 301–312. DOI: http://dx.doi.org/10.1109/34.990133

United kingdom Metropolitan Police Service MPS. 2012. 4 Days in August: Strategic Review into the Disorder of August 2011 - final report. Retrieved January 1, 2016 from http://www.met.police.uk/foi/pdfs/priorities_and_how_we_are_doing/corpo rate/4_days_in_august.pdf.

Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI'12)*. ACM, New York, NY, 189–198. DOI: http://dx.doi.org/10.1145/2166966.2166999

Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. 1998. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of the Advances in Digital Libraries Conference (ADL'98)*. IEEE Computer Society, Washington, DC, 12–.

Andrei Olariu. 2014. Efficient online summarization of microblogging streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. 236–240.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*.

Saša Petrović Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimptonand. 2013. Can twitter replace newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*.

Chi-Chun Pan and Prasenjit Mitra. 2011. Event detection with spatial latent dirichlet allocation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*. ACM, New York, NY, 349–358. DOI: http://dx.doi.org/10.1145/1998076.1998141

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*. Association for Computational Linguistics, Stroudsburg, PA, 181–189.

Swit Phuvipadawat and Tsuyoshi Murata. 2010. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. 120–123.

M. F. Porter. 1997. An algorithm for suffix stripping. In *Readings in Information Retrieval*, Karen Sparck Jones and Peter Willett (Eds.). Morgan Kaufmann, San Francisco, CA, 313–316.

Dragomir R. Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multidocument summarization using MEAD. *First Document Understanding Conference* (2001).

Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, and Ali R. Hurson. 2006. TF-ICF: A new term weighting scheme for clustering dynamic data streams. In *Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06)*. IEEE Computer Society, Washington, DC, 258–263. DOI:http://dx.doi.org/10.1109/ICMLA.2006.50

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24, 5 (Aug. 1988), 513–523. DOI:http://dx.doi.org/10.1016/0306-4573(88)90021-0

Hassan Sayyadi and Louiqa Raschid. 2013. A graph analytical approach for topic detection. *ACM Trans. Internet Technol.* 13, 2, Article 4 (Dec. 2013), 23 pages. DOI:http://dx.doi.org/10.1145/2542214.2542215

Emmanouil Schinas, Georgios Petkos, Symeon Papadopoulos, and Y. Kompatsiaris. 2012. CERTH @ mediaeval 2012 social event detection task. In *Proceedings of the MediaEval 2012 Workshop*. 6–7.

Axel Schulz, Benedikt Schmidt, and Thorsten Strufe. 2015. Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT'15)*. ACM, New York, NY, 3–12. DOI:http://dx.doi.org/10.1145/2700171.2791038

David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2010. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?, Horizon, in CSCW 2010 (2010).

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*. Association for Computational Linguistics, Stroudsburg, PA, 685–688.

Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. 2013. A participant-based approach for event summarization using twitter streams. In *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. 1152–1162.

Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*. ACM, New York, NY, 7–16. DOI:http://dx.doi.org/10.1145/2145204.2145212

Nicholas A. Thapen, Donal Stephen Simmie, and Chris Hankin. 2015. The early bird catches the term: Combining twitter and news data for event detection and situational awareness. *CoRR* abs/1504.02335 (2015).

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *J. Am. Soc. Inf. Sci. Technol.* 62, 2 (Feb. 2011), 406–418. DOI:http://dx.doi.org/10.1002/asi.21462

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.* 43, 6 (Nov. 2007), 1606–1618. DOI:http://dx.doi.org/10.1016/j.ipm.2007.01.023

Konstantinos N. Vavliakis, Andreas L. Symeonidis, and Pericles A. Mitkas. 2013. Event identification in web social media through named entity recognition and topic modeling. *Data Knowl. Eng.* 88 (2013), 1–24.

Sarah Vieweg, Carlos Castillo, and Muhammad Imran. 2014. Integrating social media communications into the rapid assessment of sudden onset disasters. In *Proceedings of the 6th International Conference on Social Informatics*. 444–461.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 1079–1088. DOI:http://dx.doi.org/10.1145/1753326.1753486

Maximilian Walther and Michael Kaisser. 2013. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR'13)*. Springer-Verlag, Berlin, 356–367. DOI:http://dx.doi.org/10.1007/978-3-642-36973-5_30

Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. 2011. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. ACM, New York, NY, 2541–2544. DOI:http://dx.doi.org/10.1145/2063576.2064014

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

Matthew Williams and Pete Burnap. 2015. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *Br. J. Criminol.* (2015), 1–28.

Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. 2013. A preliminary study of tweet summarization using information extraction. In *Proceedings of the Conference of the Association of Computational Linguistics and Workshop on Language in Social Media (LASM'13)*. 20–29.

Duan Yajuan, Chen Zhumin, Wei Furu, Zhou Ming, and Heung Y. Shum. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*. 763–780.

Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. 2012. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, 370–378. DOI:http://dx.doi.org/10.1145/2339530.2339591

Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark A. Cameron, Bella Robinson, and Robert Power. 2015. Using social media to enhance emergency situation awareness: Extended abstract. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI*. 4234–4239.

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*. ACM, New York, NY, 319–320. DOI:http://dx.doi.org/10.1145/2309996.2310053

Received March 2016; revised July 2016; accepted September 2016