Computational Intelligence

AN INTERNATIONAL JOURNAL

Computational Intelligence, Volume 31, Number 1, 2015

A SURVEY OF TECHNIQUES FOR EVENT DETECTION IN TWITTER

FARZINDAR ATEFEH AND WAEL KHREICH

NLP Technologies Inc., Montreal, QC, Canada

Twitter is among the fastest-growing microblogging and online social networking services. Messages posted on Twitter (tweets) have been reporting everything from daily life stories to the latest local and global news and events. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. This article provides a survey of techniques for event detection from Twitter streams. These techniques aim at finding real-world occurrences that unfold over space and time. In contrast to conventional media, event detection from Twitter streams poses new challenges. Twitter streams contain large amounts of meaningless messages and polluted content, which negatively affect the detection performance. In addition, traditional text mining techniques are not suitable, because of the short length of tweets, the large number of spelling and grammatical errors, and the frequent use of informal and mixed language. Event detection techniques presented in literature address these issues by adapting techniques from various fields to the uniqueness of Twitter. This article classifies these techniques according to the event type, detection task, and detection method and discusses commonly used features. Finally, it highlights the need for public benchmarks to evaluate the performance of different detection approaches and various features.

Received 29 November 2012; Revised 17 April 2013; Accepted 11 July 2013

Key words: event detection, event identification, microblogs, monitoring social media, Twitter data stream.

1. INTRODUCTION

Microblogging is a broadcast medium that allows users to exchange small digital content such as short texts, links, images, or videos. Although it is a relatively new communication medium compared with traditional media, microblogging has gained increased attention among users, organizations, and research scholars in different disciplines. The popularity of microblogging stems from its distinctive communication services such as portability, immediacy, and ease of use, which allow users to instantly respond and spread information with limited or no restrictions on content. Virtually any person witnessing or involved in any event is nowadays able to disseminate real-time information, which can reach the other side of the world as the event unfolds. For instance, during recent social upheavals and crises, millions of people on the ground turned to Twitter to report and follow significant events.

Twitter is currently the most popular and fastest-growing microblogging service, with more than 140 million users producing over 400 million tweets per day—mostly mobile—as of June 2012.[1] Twitter enables users to post status updates, or *tweets*, no longer than 140 characters to a network of *followers* using various communication services (e.g., cell phones, e-mails, Web interfaces, or other third-party applications). While some users consider the 140-character constraint as a severe limitation, many argue that it is the feature that sets Twitter apart—short information is easier to consume and faster to spread. Even though tweets are limited in size, Twitter is updated hundreds of millions of times a day by people all over the world, and its content varies tremendously based on user interests and behaviors

Address correspondence to Wael Khreich, NLP Technologies Inc., 52 Le Royer, Montreal, QC, Canada; e-mail: wael@nlptechnologies.ca

[1] http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/ (accessed on November 8, 2012).

(Java et al. 2007; Krishnamurthy et al. 2008; Zhao and Rosson 2009). Therefore, Twitter streams contain large and diverse amount of information ranging from daily-life stories to the latest local and worldwide news and events (Hurlock and Wilson 2011; Kwak et al. 2010).

Online social media sites (Facebook, Twitter, Youtube, etc.) have revolutionized the way we communicate with individuals, groups, and communities and altered everyday practices (Boyd and Ellison 2007). Several recent workshops, such as Semantic Analysis in Social Media (Farzindar and Inkpen 2012), are increasingly focusing on the impact of social media on our daily lives. For instance, Twitter has changed the way people and businesses perform, seek advice, and create "ambient awareness" (a sort of virtual omnipresence) and reinforced the weak and strong tie of friendship (Geser 2011; McFedries 2007; Thompson 2008). Unlike other media sources, Twitter messages provide timely and fine-grained information about any kind of event, reflecting, for instance, personal perspectives, social information, conversational aspects, emotional reactions, and controversial opinions.

Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets. Tweets can be seen as a dynamic source of information enabling individuals, corporations, and government organizations to stay informed of "what is happening now." For instance, people would be interested in getting advice, opinions, facts, or updates on news or events (Java et al. 2007; Krishnamurthy et al. 2008; Zhao and Rosson 2009). Companies are increasingly using Twitter to advertise and recommend products, brands, and services; to build and maintain reputations; to analyze users' sentiment regarding their products (or those of their competitors); to respond to customers' complaints; and to improve decision making and business intelligence (Farzindar 2012; Jansen et al. 2009; Jiang et al. 2011; Liu et al. 2012; Pak and Paroubek 2010). Twitter has also emerged as a fast communication channel for gathering and spreading breaking news (Amer-Yahia et al. 2012; Phuvipadawat and Murata 2010; Sankaranarayanan et al. 2009), for predicting election results, and for sharing political events and conversations (Small 2011; Tumasjan et al. 2010). It has also become an important analytical tool for crime prediction (Wang et al. 2012) and monitoring terrorist activities.[2]

In general, events can be defined as real-world occurrences that unfold over space and time (Allan et al. 1998; Troncy et al. 2010; Xie et al. 2008; Yang et al. 1998). Event detection from conventional media sources has been long addressed in the Topic Detection and Tracking (TDT) research program, which mainly aims at finding and following events in a stream of broadcast news stories (Allan et al. 1998; Yang et al. 1998). However, event detection from Twitter streams pose new challenges that are different from those faced by the event detection tasks in traditional media. In contrast with the well-written, structured, and edited news releases, Twitter messages are restricted in length and written by anyone. Therefore, tweets include large amounts of informal, irregular, and abbreviated words, large number of spelling and grammatical errors, and improper sentence structures and mixed languages. In addition, Twitter streams contain large amounts of meaningless messages (Hurlock and Wilson 2011), polluted content (Lee et al. 2011), and rumors (Castillo et al. 2011), which negatively affect the performance of the detection algorithms.

This article provides a survey of techniques found in the literature for event detection from Twitter streams. These techniques are classified according to the event type (specified

[2] http://www.nextgov.com/big-data/2012/07/pentagon-seeks-predict-terrorism-monitoring-facebook-twitter/57085/ (accessed on November 8, 2012).

or unspecified), detection task (retrospective or new event detection), and detection method (supervised or unsupervised) as well as to the target application (Tables 1 and 2). Commonly used feature representations are also presented and discussed (Table 2). Event detection from Twitter streams is a vibrant research area that draws on techniques from various fields such as machine learning, natural language processing, data mining, information extraction and retrieval, and text mining. This article does not provide an exhaustive review of existing approaches but rather selects representative techniques to give the reader a perspective on the main research directions.

Section 2 provides background information on Twitter and presents the motivation and challenges associated with event detection from Twitter streams. Section 3 briefly describes relevant techniques for event detection from traditional media outlets. In Section 4, the techniques for detecting real-world events from Twitter data streams are described and categorized according to the type of event, detection task, and detection methods with the commonly used feature representations. Finally, Section 5 provides a general discussion, followed by the conclusion.

2. TWITTER

2.1. Service Overview

As described in Section 1, Twitter is currently the most popular fastest-growing microblogging service. It is the eighth most popular site in the world according to the 3-month Alexa traffic rankings.[3] This popularity is also reflected in the increasingly large number of research papers published about Twitter in various fields. Twitter's core function allows users to post and read short messages. Similar to blogging sites, Twitter allows sharing any kind of information, thoughts, opinions, and ideas to keep people updated or informed of happenings. However, Twitter encourages concise description of ideas via tweets—short text messages that are no longer than 140 characters—which allow for effective and timely communication of information. These tweets are automatically posted as a streams (and publicly accessible) on the user's profile on Twitter and instantly sent to the user's network of followers. Messages can be posted on Twitter through various communication services such as cell phones, emails, Web interfaces, or other third-party applications. In particular, tweets can be easily posted via mobile devices, such as short message service messages, providing an efficient medium for instant information dissemination and consumption. The portability, immediacy, and ease of use are among the main reasons behind Twitter's popularity.

Twitter is also a unique online social networking service that allows people to create profiles, communicate, and connect with other people on the service. The social relationship on Twitter is asymmetric and can be conceptualized as a directed social network or *follower* network (Brzozowski and Romero 2011). A user can follow any other user without requiring an approval or a reciprocal connection from the followed users. Twitter does not impose any limits on the number of followers to a user account; however, one user account can typically follow up to 2000 users. There is, however, a user-dependent limit to follow additional users based on the ratio of followers to be followed.[4] By default, posted messages are available to anyone. Although users can modify their privacy settings to only update their followers and to decide who can follow them, these are not commonly used. Users can consume tweets

[3] http://www.alexa.com/siteinfo/twitter.com (accessed on November 8, 2012).

[4] https://support.twitter.com/entries/68916 (accessed on November 8, 2012).

TABLE 1. Taxonomy of Event Detection Techniques in Twitter.

References	Type of event		Detection method		Detection task		Application
	Specified	Unspecified	Supervised	Unsupervised	NED	RED	
Sankaranarayanan et al. (2009)		x	x	x	x		Breaking-news detection
Phuvipadawat and Murata (2010)		x		x	x		Breaking-news detection
Petrović et al. (2010)		x		x	x		General (unknown) event detection
Becker et al. (2011a)		x	x	x	x		General (unknown) event detection
Long et al. (2011)		x		x	x		General (unknown) event detection
Weng and Lee (2011)		x		x	x		General (unknown) event detection
Cordeiro (2012)		x		x	x		General (unknown) event detection
Popescu and Pennacchiotti (2010)	x		x		x		Controversial news events about celebrities
Popescu et al. (2011)	x		x		x		Controversial news events about celebrities
Benson et al. (2011)	x		x			x	Musical event detection
Lee and Sumiya (2010)	x			x	x		Geosocial event monitoring
Sakaki et al. (2010)	x		x		x		Natural disaster events monitoring
Becker et al. (2011)	x		x			x	Query-based event retrieval
Massoudi et al. (2011)	x			x		x	Query-based event retrieval
Metzler et al. (2012)	x			x		x	Query-based structured event retrieval
Gu et al. (2011)	x			x		x	Query-based structured event retrieval

TABLE 2. Summary of Detection Techniques and Feature Representations.

References	Detection techniques	General features	Twitter-specific features
Sankaranarayanan et al. (2009)	Naive Bayes classifier and online clustering	Term vector	Hashtags and timestamps
Phuvipadawat and Murata (2010)	Online clustering	Term vector, proper nouns (conventional NER)	Hashtags, #followers, #retweets and timestamps
Petrović et al. (2010)	Online clustering (based on locality sensitive hashing)	#tweets, #users and entropy of messages	–
Becker et al. (2011a)	Online clustering and support vector machine classifier	Term vector	Hashtags, multi-word hashtags with special capitalization, retweets, replies and mentions.
Long et al. (2011)	Hierarchical divisive clustering	Word frequency and entropy	Probability of word occurring in hashtags
Weng and Lee (2011)	Discrete wavelet analysis and graph partitioning	Individual words	–
Cordeiro (2012)	Continuous wavelet analysis and latent Dirichlet allocation	–	Hashtag occurrences
Popescu and Pennacchiotti (2010)	Gradient boosted decision trees	Correlation of target events (or entities) with the Web and traditional news media	Proportion of nouns, verbs, questions, bad words, etc.; #tweet, #retweets, #replies, #tweets per user, hashtags; proportion of tweets and hashtags involving buzziness, sentiment, controversy

TABLE 2. *Continued.*

References	Detection techniques	General features	Twitter-specific features
Popescu et al. (2011)	Gradient boosted decision trees	Part-of-Speech tagging and regular expressions (in addition to the features used by Popescu et al. (2011))	Relative positional information, length of snapshot, category, language (in addition to the features used by Popescu et al. (2011))
Benson et al. (2011)	Factor graph model and conditional random fields	Term vectors for artist names (extracted from Wikipedia) and for city venue names.	Word shape, patterns for emoticons, time references, venue types
Lee and Sumiya (2010)	Statistical modeling of normal crowd behavior	–	#Tweet, #Crowd, #MovingCrowd based on geotags
Sakaki et al. (2010)	Support vector machine classifier	–	#Words, #keywords and the words surrounding users query
Becker et al. (2011)	Recursive query construction	Term frequency and co-location	Hashtags and URL
Massoudi et al. (2011)	Generative language modeling		Emoticons, post length, shouting, hyperlinks, capitalization, recency, #reposts and #followers
Metzler et al. (2012)	Temporal query expansion technique		Burstiness score based on the frequency of query term occurrence
Gu et al. (2011)	Event modeling (ETree)	Term vector and n-gram models	Replies to tweets

by reading their *timeline*—a chronological streams of incoming tweets from everyone they follow.

Twitter has evolved over time and adopted suggestions originally proposed by users to make the platform more flexible. It currently provides different ways for users to converse and interact by referencing each other in posted messages in a well-defined markup vocabulary. Placing the "@" symbol before a user name (also called a handle, "@username") creates a *mention* or a *reply* link to the referenced user's account. A mention is used anywhere in the message for signaling that the mentioned user is also registered on Twitter. A reply is a special mention from one user in response to another user's message starting with the replied-to @username (Honeycutt and Herring 2009). Mentions are displayed in the referenced user accounts to keep track of messages mentioning their names. Twitter also allows users to forward or *retweet* someone else's tweet to their followers. It is commonly carried out by using the RT prefix before the user name that originated the message, "RT @username." Retweeting is a common practice on Twitter to share useful or interesting information while giving credit to the original user (Boyd et al. 2010). Topics on Twitter can be categorized by a *hashtag*, which is any keyword preceded by a hash sign "#" (e.g., #nlptechnologies). Hashtags were developed as a means to create groupings on Twitter. Twitter users can use hashtag to indicate the subject of their messages, to collate tweets from different users on a shared subject, and to regularly track specific events in real time.

Twitter provides an application programming interface (API),[5] which allows developers to programmatically access the public data streams as well as many features of the service. For instance, Twitter streaming API provides filtering by location, keywords, author, and others. The availability of Twitter data has motivated significant research work in various disciplines and led to numerous applications and tools.

2.2. Twitter as a Source of Information

Twitter is becoming the microphone of the masses, which altered news production and consumption (Murthy 2011). Many real-world examples have shown the effectiveness and the timely information reported by Twitter during disasters and social movements. Representative examples include the bomb blasts in Mumbai in November 2008 (Oh et al. 2011), the flooding of the Red River Valley in the United States and Canada in March and April 2009 (Starbird et al. 2010), the U.S. Airways plane crash on the Hudson river in January 2009, the devastating earthquake in Haiti in 2010, the demonstrations following the Iranian Presidential elections in 2009, and the "Arab Spring" in the Middle East and North Africa region (Khondker 2011; Khan 2012).

Several studies have analyzed Twitter's user intentions (Java et al. 2007; Krishnamurthy et al. 2008; Zhao and Rosson 2009; Kwak et al. 2010; Kaplan and Haenlein 2011). For instance, Java et al. (2007) categorized user intentions on Twitter into daily chatter, conversations, sharing information, and reporting news. They also identified Twitter users as information sources, friends, and information seekers. Krishnamurthy et al. (2008) presented similar classification of user intentions and also included evangelists and spammers that are looking to follow anyone. According to Kaplan and Haenlein (2011), people are motivated by the concept of ambient awareness—being updated about even the most trivial matters in other peoples' lives and by the platform for virtual exhibitionism and voyeurism provided for both active contributors and passive observers. Many research efforts have also focused on Twitter user motivations in specific environments such as at work (Zhao and

[5] https://dev.twitter.com/docs/streaming-apis.

Rosson 2009; Lovejoy and Saxton 2012), during conferences (Reinhardt et al. 2009), and in politics (Tumasjan et al. 2010; Small 2011).

2.3. Challenges

Tweets have reported everything from daily life stories to latest local and worldwide events. Twitter content reflects real-time events in our life and contains rich social information and temporal attributes. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information. However, Twitter streams contain large amounts of meaningless messages (*pointless babbles*) (Hurlock and Wilson 2011) and rumors (Castillo et al. 2011). These are important to build users' social networks (Kaplan and Haenlein 2011) and may help to understand people's reactions to events, however they negatively affect event detection performance. Nevertheless, Twitter characteristics and popularity are particularly alluring for spammers and other content polluters (Lee et al. 2011), to spread advertisements, pornography, viruses, and phishing, or just to compromise system reputation (Benevenuto et al. 2010).

A major challenge facing event detection from Twitter streams is therefore to separate the mundane and polluted information from interesting real-world events. In practice, highly scalable and efficient approaches are required for handling and processing the increasingly large amount of Twitter data (especially for real-time event detection). Other challenges are inherent to Twitter's design and usage. These are mainly due to the short length of tweet messages, the frequent use of (dynamically evolving) informal, irregular, and abbreviated words, the large number of spelling and grammatical errors, and the use of improper sentence structure and mixed languages. Such data sparseness, lack of context, and diversity of vocabulary make the traditional text analysis techniques less suitable for tweets (Metzler et al. 2007). In addition, different events may enjoy different popularity among users and can differ significantly in content, number of messages and participants, periods, inherent structure, and causal relationships (Nallapati et al. 2004).

3. EVENT DETECTION IN TRADITIONAL MEDIA

This section provides a brief overview of event detection techniques applied to traditional media outlets. Some of these techniques have been adapted to event detection in Twitter. They are generally classified into document-pivot and feature-pivot techniques depending on whether they rely on document or temporal features.

3.1. Document-Pivot Techniques

Event detection has long been addressed in the TDT program (Allan 2002), an initiative sponsored by the Defense Advanced Research Projects Agency, concerned with event-based organization of textual news document streams. The motivation for the TDT research initiative was to provide core technology for news monitoring tools from multiple sources of traditional media (including newswire and broadcast news) to keep users updated about news and developments. Originally, the TDT consisted of three main tasks: segmentation, detection, and tracking. These tasks attempt to segment news text into cohesive stories, detect new (unforeseen) events, and track the development of a previously reported event.

According to TDT, the objective of event detection is to discover new or previously unidentified events, where each event refers to a specific thing that happens at a specific time and place (Allan et al. 1998). Event detection consists of three major phases: data preprocessing, data representation, and data organization or clustering (Yang et al. 1998;

Yang et al. 2002). Data preprocessing involves filtering out stopwords (on, of, are, etc.) and applying words stemming and tokenization techniques.

Traditional data representation for event detection involves the *term vectors* or *bag of words* whose entries are nonzero if the corresponding terms appear in the document. Each term in the vector is typically weighted using the classical term frequency–inverse document frequency (*tf-idf*) approach (Salton 1988), which evaluates how important a word is to a document in a corpus. However, the term vector model is subject to the curse of dimensionality when the text is long. More importantly, the temporal order of words and the semantic and syntactic features of the text (e.g., named entity and grammar) are discarded by the term vectors; thus, the model may not capture the similarity (or dissimilarity) among related (or unrelated) events. For instance, it would be difficult for the term vector to distinguish between two different airplane crashes. In fact, Allan, Lavrenko, Marlin, and Swan (2000) presented an upper bound for full-text similarity, which leads to exploring other data representation techniques such as semantical and contextual features (Allan, Lavrenko, and Jin 2000).

The *named entity vector* is an alternative data representation (Kumaran and Allan 2004), which attempts to extract information answering the 4Ws questions: who, what, when, and where (Mohd 2007). Both term and named entity vectors have been integrated into a *mixed vector* (Yang et al. 1999; Kumaran and Allan 2004). Probabilistic representations have also been applied including language models (Leek et al. 2002) and more advanced probabilistic frameworks that incorporate both content and time information (Li et al. 2005). Similarity between events is measured using traditional metrics such as the Euclidean distance, Pearson's correlation coefficient, and cosine similarity. More recently, other similarity measures have been proposed such as the Hellinger distance (Brants et al. 2003) and the clustering index (Jo and Lee 2007).

In TDT, event detection has been broadly divided into two categories: retrospective event detection (RED) and new event detection (NED)[6] (Yang et al. 1998; Allan et al. 1998). RED focuses on discovering previously unidentified events from accumulated historical collections (Yang et al. 1998), while NED involves the discovery of new events from live streams in (near) real time (Allan et al. 1998). Clustering-based algorithms (Berkhin 2002; Aggarwal and Zhai 2012) have been mainly employed for both RED and NED tasks.

Retrospective event detection involves iterative clustering algorithms that require the entire document collection, to organize the documents into topic clusters. Hierarchical clustering approaches, such as the bottom-up hierarchical agglomerative clustering (HAC) (Jain and Dubes 1988), have been widely employed for this task. At first, each single data point is represented by a cluster, and then closest clusters are merged on the basis of the similarity measure until all data points become a single cluster or some termination criteria are satisfied. Several variations of the HAC algorithm have been employed in TDT tasks. For instance, Yang et al. (1998) employed the group average clustering to detect news events from accumulated news stories. A two-layer HAC approach based on affinity propagation has also been proposed to reduce the false positives; however, at the expense of increased complexity (Dai et al. 2010). The traditional k-means algorithm and its variants, such as k-median and k-means++, have also been applied (Bouras and Tsogkas 2010).

New event detection has been characterized as a *query-free* retrieval tasks because the event information is not known a priori and hence cannot be expressed as an explicit query (Allan et al. 1998). In contrast to RED, NED must provide decisions (new or old events) as documents arrive. Therefore, the employed clustering approaches are typically based on

[6] Also known as first-story detection or novelty detection.

incremental (greedy) algorithms that process the input streams sequentially and merge an event with the most similar one or create a new cluster if the similarity measure exceeds a predefined threshold (Allan et al. 1998). In practice, such approach could be time and resource intensive and may be unfeasible without employing specialized techniques for improved system efficiency (Luo et al. 2007). For instance, using a sliding time window over old stories and comparing the new story with the most recent number of stories would alleviate the resource requirements (Yang et al. 1998; Papka 1999; Luo et al. 2007). The underlying assumption is that the occurrence of stories related to the same event should be close in time. Other techniques for improved system efficiency include limiting the number of terms per document, limiting the number of total terms kept, and employing parallel processing (Luo et al. 2007).

These event detection techniques are typically called *document-pivot* techniques, because they detect events by clustering documents on the basis of their textual similarity. However, the TDT line of research assumes that all documents are relevant and contain some old or new events of interest (Allan, Lavrenko, and Jin 2000). This assumption is clearly violated in Twitter data streams, where relevant events are buried in large amounts of noisy data (Becker et al. 2011b; Castillo et al. 2011; Hurlock and Wilson 2011; Lee et al. 2011). In addition, these techniques are not designed to handle the speed and scale requirements of social media.

3.2. Feature-Pivot Techniques

Trend detection tasks over textual data collection generally aim to identify topic areas that were previously unseen or rapidly growing in importance within the corpus (Kontostathis et al. 2004). Recently, there has been a significant interest in bursty event detection techniques in traditional media (Kleinberg 2002; Fung et al.2005; He, Chang, and Lim 2007; He, Chang, Lim, and Zhang 2007; Wang et al. 2007; Goorha and Ungar 2010). These *feature-pivot* techniques model an event in text streams as a bursty activity, with certain features rising sharply in frequency as the event emerges. An event is therefore conventionally represented by a number of keywords showing burst in appearance counts (Kleinberg 2002). The underlying assumption is that some related words would show an increased usage as an event occurs. Different from traditional RED and NED approaches, these techniques analyze feature distributions and discover events by grouping bursty features with identical trends.

In his seminal work, Kleinberg (2002) proposed in infinite-state automaton to model the arrival times of documents in a streams to identify bursts that have high intensity over limited durations of time. The states of the probabilistic automaton correspond to the frequencies of individual words, while the state transitions capture the burst, which correspond to a significant change in word frequency. Fung et al. (2005) modeled word appearance as binomial distribution, identified the bursty words according to a heuristic-based threshold, and grouped bursty features to find bursty events. He, Chang, Lim, and Zhang (2007) applied spectral analysis using discrete Fourier transformation (DFT) to categorize features for different event characteristics (e.g., important or not, and periodic or aperiodic events). DFT converts the signals from the time domain into the frequency domain, such that a burst in the time domain corresponds to a spike in the frequency domain. However, DFT cannot identify the period of a bursty event. Therefore, He, Chang, and Lim (2007) employed Gaussian mixture models to identify feature bursts and their associated periods. Snowsill et al. (2010) presented an online approach for detecting events in news streams based on statistical significant tests of n-gram word frequency within a time frame. An incremental suffix tree data structure was applied to reduce the time and space constraints required for online detection.

Direct application of feature-pivot techniques to Twitter streams may not be suitable, because the temporal distributions of features are very noisy and not all bursts are relevant events of interest.

4. EVENT DETECTION IN TWITTER

This section describes various techniques proposed for event detection from Twitter streams. Table 1 presents a taxonomy of these techniques according to the type of event, detection task, and detection methods. Depending on the type of events, these techniques are classified into unspecified and specified event detection. According to the detection task and target application, they are classified into RED and NED techniques. Furthermore, depending on the detection method, the presented techniques are also categorized into supervised and unsupervised (or a combination of both) techniques (Table 1). In addition, the main event detection methods are further illustrated in Table 2 along with the feature representations, which are divided into general and Twitter-specific features.

4.1. Unspecified versus Specified Event

Depending on the available information on the event of interest, event detection can be classified into specified and unspecified techniques, as shown in columns 2 and 3 of Table 1. Because no prior information is available about the event, the unspecified event detection techniques rely on the temporal signal of Twitter streams to detect the occurrence of a real-world event. These techniques typically require monitoring for bursts or trends in Twitter streams, grouping the features with identical trend into events, and ultimately classifying the events into different categories. On the other hand, the specified event detection relies on specific information and features that are known about the event, such as a venue, time, type, and description, which are provided by the user or from the event context. These features can be exploited by adapting traditional information retrieval and extraction techniques (such as filtering, query generation and expansion, clustering, and information aggregation) to the unique characteristics of tweets.

The following subsections describe the techniques for unspecified and specified event detection. These techniques are then further classified according to the detection task (Section 4.2) and detection methods (Section 4.3)

4.1.1. Unspecified Event Detection. The nature of Twitter posts reflect events as they unfold; hence, these tweets are particularly useful for unknown event detection. Unknown events of interest are typically driven by emerging events, breaking news, and general topics that attract the attention of a large number of Twitter users. Because no event information is available, unknown events are typically detected by exploiting the temporal patterns or signal of Twitter streams. New events of general interest exhibit a burst of features in Twitter streams yielding, for instance, a sudden increased use of specific keywords. Bursty features that occur frequently together in tweets can then be grouped into trends (Mathioudakis and Koudas 2010). In addition to trending events, endogenous or nonevent trends are also abundant on Twitter (Naaman et al. 2011). Techniques for unspecified event detection in Twitter must therefore discriminate trending events of general interest from the trivial or nonevent trends (exhibiting similar temporal pattern) using scalable and efficient algorithms. The techniques described in the following text attempted to address these challenges.

Sankaranarayanan et al. (2009) proposed a news processing system based on Twitter, called TwitterStand, to capture tweets that correspond to late breaking news. They employ a naive Bayes classifier to separate news from irrelevant information and an online clustering

algorithm based on weighted term vector according to tf-idf and cosine similarity to form clusters of news. In addition, hashtags are used to reduce clustering errors. Clusters are also associated with time information for management and for determining the clusters of interest. Other issues addressed include removing the noise and determining the relevant locations associated with the tweets.

Phuvipadawat and Murata (2010) presented a method to collect, group, rank, and track breaking news from Twitter. They first sample tweets (through Twitter streaming API) using predefined search queries, for example, "#breakingnews" and "#breaking news" keyword, and index their content with Apache Lucene.[7] Messages that are similar to each other are then grouped together to form a news story. Similarity between messages are based on tf-idf with an increased weight for proper noun terms, hashtags, and usernames. Proper nouns are identified using the Stanford Named Entity Recognizer (NER) trained on conventional news corpora. They use a weighted combination of number of followers (reliability) and the number of retweeted messages (popularity) with a time adjustment for the freshness of the message to rank each cluster. New messages are included in a cluster if they are similar to the first message and to the top-k terms in that cluster. The authors stress the importance of proper nouns identification to enhance the similarity comparison between tweets and hence improve the overall system accuracy. An application based on the proposed method called Hot-streams has been developed.

Petrović et al. (2010) adapted the online NED approach proposed for news media (Allan, Lavrenko, and Jin 2000), which is based on cosine similarity between documents to detect new events that have never appeared in previous tweets. They focused on improving the efficiency of online NED algorithm and proposed a constant time and space approach based on an adapted variant of the locality sensitive hashing methods (Gionis et al. 1999), which limits the search to a small number of documents. However, they did not consider replies, retweets, and hashtags in their experiments or the significance of newly detected events (e.g., trivial or not). Results have shown that ranking according to the number of users is better than ranking according to the number of tweets and considering entropy of the message reduces the amount of spam messages in output.

Becker et al. (2011a) focused on online identification of real-world event content and its associated Twitter messages using an online clustering technique, which continuously clusters similar tweets and then classifies the clusters content into real-world events or nonevents. These nonevents involve Twitter-centric topics, which are trending activities in Twitter that do not reflect any real-world occurrences (Naaman et al. 2011). Twitter-centric activities are difficult to detect, because they often share similar temporal distribution characteristics with real-world events. Their clustering approach is based on a classical (threshold-based) incremental clustering algorithm that has been proposed for NED in news documents (Allan et al. 1998). Each message is represented as a tf-idf weight vector of its textual content, and cosine similarity is used to compute the distance from a message to cluster centroids. In addition to traditional preprocessing steps such as stop-word elimination and stemming, the weight of hashtag terms are doubled because they are considered a strong indication of the message content. The authors combined temporal, social, topical, and Twitter-centric features. The temporal features rely on term frequency that appear in the set of messages associated with a cluster over time. The social features include the percentage of messages containing users interaction (i.e., retweets, replies, and mentions) out of all messages in a cluster. The topical features are based on the hypothesis that event clusters tend to revolve around a central topic, whereas nonevent clusters often center around various

[7] http://lucene.apache.org.

common terms (e.g., "sleep" or "work") that do not reflect a single theme. The Twitter-centric features are based on the frequency of multiword hashtags with special capitalization (e.g., #BadWrestlingNames). Because the clusters constantly evolve over time, the features are periodically updated for old clusters and computed for newly formed ones. Finally, a support vector machine (SVM) classifier is trained on a labeled set of cluster features and used to decide whether the cluster (and its associated messages) contains real-world event information.

Long et al. (2011) adapted a traditional clustering approach by integrating some specific features to the characteristics of microblog data.[8] These features are based on "topical words," which are more popular than others with respect to an event. Topical words are extracted from daily messages on the basis of word frequency, word occurrence in hashtag, and word entropy. A (top-down) hierarchical divisive clustering approach is employed on a co-occurrence graph (connecting messages in which topical words co-occur) to divide topical words into event clusters. To track changes among events at different time, a maximum-weighted bipartite graph matching is employed to create event chains, with a variation of Jaccard coefficient as similarity measures between clusters. Finally, cosine similarity augmented with a time interval between messages is used to find the top-k most relevant posts that summarize an event. These event summaries are then linked to event chain clusters and plotted on the time line. For event detection, the authors found that top-down divisive clustering outperforms both k-means and traditional hierarchical clustering algorithms.

Weng and Lee (2011) proposed an event detection based on clustering of *discrete* wavelet signals built from individual words generated by Twitter. In contrast with Fourier transforms, which have been proposed for event detection from traditional media (Section 3.2), wavelet transformations are localized in both time and frequency domain and hence able to identify the time and the duration of a bursty event within the signal. Wavelets convert the signals from the time domain to time-scale domain, where the scale can be considered as the inverse of frequency. Signal construction is based on time-dependent variant of document frequency–inverse document frequency (DF-IDF), where DF counts the number of tweets (document) containing a specific word, while IDF accommodates word frequency up to the current time step. A sliding window is then applied to capture the change over time using the H-measure (normalized wavelet entropy). Trivial words are filtered out on the basis of (a threshold set on) signals cross-correlation, which measure similarity between two signals as function of a time lag. The remaining words are then clustered to form events with a modularity-based graph partitioning technique, which splits the graph into subgraphs each corresponding to an event. Finally, significant events are detected on the basis of the number of words and the cross-correlation among the words related to an event.

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). Instead of individual words, hashtags are used for building wavelet signals. An abrupt increase in the number of a given hashtag is considered a good indicator of an event that is happening at a given time. Therefore, all hashtags were retrieved from tweets and then grouped in intervals of 5 minutes. Hashtag signals are constructed over time by counting the hashtag mentions in each interval, grouping them into separated time series (one for each hashtag), and concatenating all tweets that mention the hashtag during each time series. Adaptive filters are then used to remove noisy hashtag signals, before applying the continuous wavelet transformation and getting a time-frequency representation of the

[8] The authors applied their approach to Sina (http://t.sina.com.cn), a popular microblog in China.

signal. Next, wavelet peak and local maxima detection techniques are used to detect peaks and changes in the hashtag signal. Finally, when an event is detected within a given time interval, LDA is applied to all tweets related to the hashtag in each corresponding time series to extract a set of latent topics, which provide an improved summary of event description.

4.1.2. Specified Event Detection. Specified event detection includes known or planned social events. These events could be partially or fully specified with the related content or metadata information such as location, time, venue, and performers. The techniques described later attempt to exploit Twitter textual content or metadata information or both, using a wide range of machine learning, data mining, and text analysis techniques.

Popescu and Pennacchiotti (2010) focused on identifying controversial events that provoke public discussions with opposing opinions in Twitter, such as controversies involving celebrities. Their detection framework is based on the notion of a Twitter snapshot, a triplet consisting of a target entity (e.g., Barack Obama), a given period (e.g., 1 day), and a set of tweets about the entity from the target period. Given a set of Twitter snapshots, an event detection module first distinguishes between event and nonevent snapshots using a supervised gradient boosted decision trees (Friedman 2001), trained on manually labeled data set. To rank these event snapshots, a controversy model assigns higher scores to controversial-event snapshots, on the basis of a regression algorithm applied to a large number of features. The employed features are based on Twitter-specific characteristics including linguistic, structural, buzziness,[9] sentiment, and controversy features, and on external features such as news buzz and Web-news controversy. These external features require time alignment of entities in news media and Twitter sources, to capture entities that are trending in both sources because they are more likely to refer to real-world events. The authors have also proposed to merge the two stages (detection and scoring) into a single-stage system by including the event detection score as an additional feature into the controversy model, which yielded an improved performance. Feature analysis of the single-stage system revealed that the event score is the most relevant feature because it discriminates event from nonevent snapshots. Hashtags are found to be important semantic features for tweets, because they help identify the topic of a tweet and estimate the topical cohesiveness of a set of tweets. Nevertheless, external features based on news and the Web are also found useful; hence, correlation with traditional media helps validate and explain social media reactions. In addition, the linguistic, structural, and sentiment features also provide considerable effects. The authors concluded that a rich, varied set of features is crucial for controversy detection.

In a successive work, Popescu et al. (2011) employed the same framework described earlier, but with additional features to extract events and their descriptions from Twitter. The key idea is based on the *importance* and the *number* of the entities to capture commonsense intuitions about event and nonevent snapshots. As observed by the authors: "Most event snapshots have a small set of important entities and additional minor entities while nonevent snapshots may have a larger set of equally unimportant entities." These new features are inspired from the document aboutness system (Paranjpe 2009) and aim at ranking the entities in a snapshot with respect to their relative importance to the snapshot. This includes relative positional information (e.g., offset of term in snapshot), term-level information (term frequency, Twitter corpus IDF), and snapshot-level information (length of snapshot, category, language). Opinion extraction tools such as an off-the-shelf part-of-speech (POS) tagger and regular expressions have also been applied for improved event and main entity

[9] Approximated by the number of tweets in a snapshot referring to an entity over the average number of tweets in the N previous snapshots referring to the same entity.

extraction. The number of snapshots containing action verbs, the buzziness of an entity in the news on a given day, and the number of reply tweets are among the most useful new features found by the authors.

Benson et al. (2011) present a novel approach to identify Twitter messages for concert events using a factor graph model, which simultaneously analyzes individual messages, clusters them according to event type, and induces a canonical value for each event property. The motivation is to infer a comprehensive list of musical events from Twitter (based on artist–venue pairs) to complete an existing list (e.g., city event calender table) by discovering new musical events mentioned by Twitter users that are difficult to find in other media sources. At the message level, this approach relies on a conditional random field (CRF) to extract the artist name and location of the event. The input features to CRF model include word shape; a set of regular expressions for common emoticons, time references, and venue types; a bag of words for artist names extracted from external source (e.g., Wikipedia); and a bag of words for city venue names. Clustering is guided by term popularity, which is an alignment score among the message term labels (artist, venue, none) and some candidate value (e.g., specific artist or venue name). To capture the large text variation in Twitter messages, this score is based on a weighted combination of term similarity measures, including complete string matching, and adjacency and equality indicators scaled by the inverse document frequency. In addition, a uniqueness factor (favoring single messages) is employed during clustering to uncover rare event messages that are dominated by the popular ones and to discourage various messages from the same events to cluster into multiple events. On the other hand, a consistent indicator is employed to discourage messages from multiple events to form a single cluster. The factor graph model is then employed to capture the interaction between all components and provide the final decision. The output of the model consists of a musical event-based clustering of messages, where each cluster is represented by an artist–venue pairs.

Lee and Sumiya (2010) present a geosocial local event detection system based on modeling and monitoring crowd behaviors via Twitter, to identify local festivals. They rely on geographical regularities deduced from the usual behavior patterns of crowds using geotags. First, Twitter geotagged data are collected and preprocessed over a long period for a specific region (Fujisaka et al. 2010). The region is then divided into several regions of interest (ROI) using the k-means algorithm, applied to the geographical coordinates (longitudes/latitudes) of the collected data. Geographical regularities of crowd within each ROI are then estimated from historical data based on three main features: the number of tweets, users, and moving users within an ROI. Statistics for these features are then accumulated over historical data using 6-hour time interval to form the estimated behavior of crowd within each ROI. Finally, unusual events in the monitored geographical area can be detected by comparing statistics from new tweets with those of the estimated behavior. The authors found that an increased user activity (moving inside or coming to an ROI) combined with an increased number of tweets provides strong indicator of local festivals.

Sakaki et al. (2010) exploited tweets to detect specific types of events such as earthquakes and typhoons. They formulated event detection as a classification problem and trained an SVM on a manually labeled Twitter data set comprising positive events (earthquakes and typhoons) and negative events (other events or nonevents). Three types of features have been employed: the number of words (statistical), the keywords in a tweet message, and the words surrounding users queries (contextual). Analysis of the number of tweets over time for earthquakes and typhoons data revealed an exponential distribution of events. Parameters of the exponential distribution are estimated from historical data and then used for computation of a reliable wait time (during which more information is being gathered from related tweets) before raising an alarm. Experiments have shown that the

statistical features provided the best results, while a small improvement in performance has been achieved by the combination of the three features. The authors have also applied Kalman filtering and particle filtering (Fox et al. 2003) for estimation of earthquake center and typhoon trajectory from Twitter temporal and spatial information. They found that particle filters outperformed Kalman filters in both cases, because of the inappropriate Gaussian assumption of the latter for this type of problems.

Becker et al. (2011) presented a system for augmenting information about planned events with Twitter messages, using a combination of simple rules and query building strategies. To identify Twitter messages for an event, they begin with simple and precise query strategies derived from the event description and its associated aspects (e.g., combining time and venue). An annotator is then asked to label the results returned by each strategy for over 50 events that provide high-precision tweets. To improve recall, they employ term-frequency analysis and co-location techniques on the resulting high-precision tweets to identify descriptive event terms and phrases, which are then used recursively to define new queries. In addition, they build queries using URL and hashtag statistics from the high-precision tweets for an event. Finally, they build a rule-based classifier to select among this new set of queries and then use the selected queries to retrieve additional event messages. In a related work, Becker et al. (2011b) proposed centrality-based approaches to extract high-quality, relevant, and useful Twitter messages related to an event. These approaches are based on the observation that the most topically central messages in a cluster are more likely to reflect key aspects of the event than other, less central cluster messages. The techniques from both works have recently been extended and incorporated into a more general approach that aims at identifying social media contents for known events across different social media sites (Becker et al. 2012).

Massoudi et al. (2011) employed a generative language modeling approach based on query expansion and microblog "quality indicators" to retrieve individual microblog messages. However, the authors only considered the existence of a query term within a specific post and discarded its local frequency. The quality indicators include part of the blog "credibility indicators" proposed by Weerkamp and de Rijke (2008) such as emoticons, post length, shouting, capitalization, and the existence of hyperlinks, extended with specific microblog characteristics such as a recency factor, and the number of reposts and followers. The recency factor is based on difference between the query time and the post time. The values provided with these microblog-specific indicators are averaged into a single value and are weight combined with the credibility indicators to compute the overall prior probability for a microblog post. The query expansion technique selects top-k terms that occur in a user-specified number of posts close to the query date. The final query is therefore a weighted mixture of the original and expanded query. The combination of the quality indicator terms and the microblog characteristics has been shown to outperform each method alone. In addition, tokens with numeric or nonalphabetic characters have turned out beneficial for query expansion.

Rather than retrieving individual microblog messages in response to an event query, Metzler et al. (2012) proposed retrieving a ranked list (or timeline) of historical event summaries. The search task involves temporal query expansion, timespan retrieval, and summarization. In response to a user query, this approach retrieves a ranked set of timespans[10] on the basis of the occurrence of the query keywords. A burstiness score is then computed for all terms that occur in messages posted during each of the retrieved timespans. This score is based on the frequency of term occurrence within the retrieved timespan

[10] The authors suggest dividing the microblog streams into hourly based timespans.

to that within the entire microblog archive. The idea is to capture terms that are heavily discussed and trending during a retrieved timespan, because they are more likely to be related to the query. The scores for each term are aggregated (using geometric mean) over all retrieved timespans, and the k highest weighted terms are considered for query expansion. The expanded query is now used to identify the 1000 highest scoring timespans, with respect to the term expansion weight and to the cosine similarity between the burstiness of the query terms and the burstiness of the timespan terms. Adjacent timespans (contiguous in time) are then merged into longer time interval to form the final ranked list. To produce a short summary for each retrieved time interval, a small set of query-relevant messages posted during the timespan are then selected. These relevant messages are retrieved as top ranked message according to a weighted variant of the query likelihood scoring function, which is based on the burstiness score for expansion terms and a Dirichlet smoothed language modeling estimate for each term in the message. The authors showed that their approach is more robust and effective than the traditional relevance-based language models (Lavrenko and Croft 2001) applied to the collected Twitter corpus and to English Gigaword corpus.

Gu et al. (2011) proposed an event modeling approach called ETree for event modeling from Twitter streams. ETree employs n-gram-based content analysis techniques to group a large number of event-related messages into semantically coherent information blocks, an incremental modeling process to construct hierarchical theme structures, and a life cycle-based temporal analysis technique to identify potential causal relationships between information blocks. The n-gram model is used to detect frequent key phrases among a large number of event-related messages, where each phrase represents an initial information block. Semantically coherent messages are merged into the corresponding information block. The weighted cosine similarity is computed between each of the remaining messages (that does not include any key phrase) and each information block, and the messages with high similarities are merged into the corresponding information block. In addition, replies to tweets are also merged into the corresponding information block. An incremental (top-down) hierarchical algorithm based on weighted cosine similarity is proposed to construct and update the theme structures, where each theme is considered as a tree structure with information blocks as leaf nodes and subtopics as internal nodes. For instance, when a new tweet becomes available, it may be assigned to an existing theme or node or may become a new theme (in this case, the hierarchy must be reconstructed). Finally, casual relationships between information blocks are computed on the basis of content (weighted cosine) similarity and temporal relevance. Temporal information is based on the time boundaries of each information block as well as on the temporal distribution reflecting the number of messages posted within each period. The authors show that the n-gram-based block identification generates coherent information blocks with high coverage. An event is considered coherent if more than half of its information blocks are relevant, while the coverage of an event is defined as the percentage of messages that are captured into one of the identified information blocks. In addition, ETree is shown more efficient compared with its nonincremental version and to TSCAN—a widely used algorithm that derives major themes of events from the eigenvectors of a temporal block association matrix (Chen and Chang Chen 2008).

4.2. New versus Retrospective Event

Similar to event detection from conventional media, described in Section 3, event detection in Twitter can also be classified into RED and NED depending on the task and application requirements as well as on the type of event.

Because NED techniques involve continuous monitoring of Twitter signals for discovering new events in near real time, they are naturally suited for detecting unknown real-world

events or breaking news, as shown in Table 1, column 6. In general, trending events on Twitter could be aligned with real-world breaking news. However, sometimes a comment, person, or photo related to real-world breaking news may become more trending on Twitter than the original event. One such example is Bobak Ferdowsi's hairstyle, which became viral on social media while NASA's Curiosity rover was landing on Mars[11] Although the NED approaches do not impose any assumption on the event, they are not restricted to unspecified event detection. When the monitoring task involves specific events (natural disasters, celebrities, etc.) or a specific information about the event description (e.g., geographical location), these information could be integrated into the NED system, for instance, by using filtering techniques (Sakaki et al. 2010) or exploiting additional features such as the controversy (Popescu and Pennacchiotti 2010) or the geotagged information (Lee and Sumiya 2010), to better focus on the event of interest. Most NED approaches could also be applied to historical data to detect and analyze past events.

While most research focused on NED to exploit the timely information provided by Twitter streams, recent studies have shown an interest in RED from Twitter's historical data (Table 1, column 7). Existing microblog search services, such as those offered by Twitter and Google, only provide limited search capabilities that allow to retrieve individual microblog posts in response to a query (Metzler et al. 2012). The challenges in finding Twitter messages relevant to a given user query are mainly due to the sparseness of the tweets and the large number of vocabulary mismatch (which is dynamically evolving). For example, relevant messages may not contain any query term, or new abbreviation terms or hashtags may emerge with the event. Traditional query expansion techniques rely on terms that co-occur with query terms in relevant documents. In contrast, event retrieval from Twitter data focused on temporal and dynamic query expansion techniques. Recent research efforts have started to focus on providing more structured and comprehensive summaries of Twitter events.

4.3. Detection Methods and Features

Event detection from Twitter streams draws on techniques from different fields, which are extensively covered in the literature, including machine learning and data mining (Murphy 2012; Hastie et al. 2009), natural language processing (Manning and Schütze 1999; Jurafsky and Martin 2009), information extraction (Hogenboom et al. 2011), text mining (Hogenboom et al. 2011; Aggarwal 2011), and information retrieval (Baeza-Yates and Ribeiro-Neto 2011). In this section, the major directions of the survey approaches for event detection from Twitter are discussed.

In general, machine learning tasks involve learning a mapping function: $f(X) \rightarrow Y$, from an input space X to an output space Y. These tasks are typically divided into supervised and unsupervised learning approaches. In supervised learning, a labeled set of N input–output pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is provided at training time to learn $f(X)$, while unsupervised learning relies solely on the input data $\{(x_1, \ldots, y_n)\}$ to discover interesting patterns (i.e., estimate both $f(X)$ and Y).

The remainder of this section categorizes the event detection techniques into supervised and unsupervised learning or a combination of both approaches (as shown in Table 1) and discusses the feature representation for each approach, which are classified into general and Twitter-specific features as illustrated in Table 2.

[11] http://www.examiner.com/article/photos-mohawk-guy-bobak-ferdowsi-s-hair-goes-viral-as-curiosity-lands-on-mars.

4.3.1. Unsupervised Detection Approaches. Similar to event detection from conventional media (Section 3), most techniques for unspecified event detection from Twitter streams rely on clustering approaches, as described in Section 4.1.1 and also shown in Tables 1 and 2. Clustering approaches are naturally suitable for unspecified NED from Twitter, because they are unsupervised in that they require no labeled data for training. However, given the increasingly large amounts of data and the real-time nature of Twitter streams, clustering algorithms for NED must be efficient and highly scalable. In addition, they should not require any prior knowledge (e.g., the number of clusters), because Twitter data are dynamically evolving and new events arise over time. Ideally, these clustering algorithms must process and analyze new tweets as they become available using a single or few passes within a limited amount of time and memory and be able to provide decisions at any given time. Therefore, partitioning clustering techniques such as K-means, K-median, and K-medoid or other approaches based on the expectation–maximization algorithm (Aggarwal and Zhai 2012; Berkhin 2002) are also not suitable because they require a prior knowledge of the number of clusters (K).

Several threshold-based online (or incremental) approaches for clustering Twitter streams have been mainly adopted from the NED in the TDT (Becker et al. 2011a; Petrović et al. 2010; Phuvipadawat and Murata 2010; Sankaranarayanan et al. 2009). As described in Section 3, incremental clustering approaches are appropriate for grouping of continuously generated text, by setting a maximum similarity between new tweets and any of the existing clusters. When the similarity is greater than a preset threshold, the new tweet is considered similar to and merged with the closest cluster; otherwise, it is considered as a new event and a new cluster is formed. Although some work has focused on improving the efficiency of the original online clustering algorithm (Petrović et al. 2010), little research efforts have been devoted to threshold settings and fragmentation issues. Thresholds are typically set empirically using the training (or validation) data and assumed to generalize to unseen messages. Fragmentation occurs when tweets that talk about the same event are grouped in different clusters. Fragmentation is inherent to incremental clustering and depends on threshold settings. To alleviate this issue, Sankaranarayanan et al. (2009) suggested a periodic second pass to merge similar or duplicate clusters, while Petrović et al. (2010) proposed comparing new tweets with a fixed number of most recent clusters.

Graph-based clustering algorithms have also been proposed for the NED task. A hierarchical divisive clustering approach is used on a co-occurrence graph (connecting messages according to word co-occurrence) to divide topical words into event clusters (Long et al. 2011). A modularity-based graph partitioning technique is used to form events by splitting the graph into subgraphs each corresponding to an event (Weng and Lee 2011). The power iteration method employed in the PageRank algorithm (Ipsen and Wills 2006) is proposed to alleviate the computational burden associated with finding the largest eigenvalue of the modularity matrix (Weng and Lee 2011). In general, hierarchical clustering algorithms do not scale to the large size of the data because they require the full similarity matrix, which contains the pairwise similarity between groups (Becker et al. 2010; Cordeiro 2012). In addition, scalable graph partitioning algorithms may not capture the highly skewed event distribution of Twitter data as they are biased toward balanced partitioning (Becker et al. 2010).

New event detection approaches that are based on unsupervised clustering typically employ shallow feature representations, including the weighted term vector according to tf-idf computed over some period (Petrović et al. 2010; Weng and Lee 2011) and augmented with other features such as hashtag occurrence (Becker et al. 2011a; Sankaranarayanan et al. 2009). In addition to word and hashtag frequency, some authors have included other features such as word entropy (Long et al. 2011) and proper names

(Phuvipadawat and Murata 2010). Because of their discriminating power, and for improved efficiency, the hashtag features have been recently used without any additional features (Cordeiro 2012). Cosine similarity is most commonly used within these online clustering algorithms to compute the distance between the (augmented) term vectors and the centers of clusters. The approaches based on spectral analysis and weighted graphs (Cordeiro 2012; Long et al. 2011; Weng and Lee 2011) inherently detect the bursty events associated with the previously considered features over time. On the other hand, NED approaches that are only based on clustering typically exploit Twitter social network information, which reflect the bursty events, to rank the resulting event clusters and their associated messages before presenting the most relevant real-world events. This includes ranking event clusters according to the number of tweets, retweets, users, and followers (Petrović et al. 2010; Phuvipadawat and Murata 2010).

New event detection of specific type of events could be addressed similar to anomaly detection techniques (Khreich et al. 2009), which rely on modeling normal user behavior and detecting any deviation from this baseline profile. This alternate unsupervised learning approach has been shown effective in detecting local festival events, by learning the normal behavior of users in a given location over some period, and in detecting deviations as possible events (Lee and Sumiya 2010). Boxplot statistics such as the median, quartiles, and minimum and maximum values have been used to learn the geographical regularities of a crowd within a region of interest using features such as the number of tweets and the number of users, and user-movement based on geotag information (Lee and Sumiya 2010). From this perspective, online and incremental learning of sequential models such as hidden Markov models (Khreich et al. 2012b) would provide a more robust solution for modeling the temporal behavior of the data.

In contrast to traditional query expansion techniques, which rely on query-term co-occurrence, the unsupervised approaches for RED focus on temporal and dynamic query expansion techniques. Because of the temporal and social aspects in Twitter streams, people searching Twitter show more interest in timely information (e.g., related to news or events) and social information related, for instance, to other users or popular trends (Teevan et al. 2011). The proposed temporal query expansion techniques are based on the number of posts close to the query date to retrieve individual messages (Massoudi et al. 2011) or on past (possibly distant) timespans to extract various messages that are related to a specific event (Metzler et al. 2012). Queries are extended on the basis of features such as the relative term frequency, hashtags, existence of hyperlinks and emoticons, number of replies and followers, post length, abbreviation, and capitalization that co-trend with the query terms in relevant timespans. This allows queries to adapt to new terms, abbreviations, or hashtags that may emerge during specific events.

Providing a structured and comprehensive summary of Twitter events, which goes beyond simple message retrieval, is currently an active area of research. As a postprocessing step to event detection, it would provide a more comprehensive view of the content than a list of tweets. Effective summarization and structuring of tweets is also important for several microblog-related applications ranging from trend detection to microblog retrieval and sentiment analysis (Efron 2011; Kim et al. 2011; Sharifi et al. 2010). A "phrase reinforcement" algorithm is proposed to summarize multiple tweets related to the same event by finding the most commonly used phrase that encompasses the topic (Sharifi et al. 2010). Long et al. (2011) return the k most relevant and diverse posts to capture the event context based on cosine similarity between posts within a given time interval, while Cordeiro (2012) returns the set of hashtags related to the events based on an LDA topic model. A more structured view of events is provided by the ETree framework (Gu et al. 2011). ETree identifies the major aspects of the event, the key message clusters, and their hierarchical structure

and causal relationships by using the term vector, number of replies, and employing n-gram analysis, incremental modeling, and a life cycle–based temporal analysis.

4.3.2. Supervised Detection Approaches. While most NED approaches for unspecified events involve unsupervised clustering of new tweets (as shown in Table 1), the NED techniques that focus on detecting a specified type of event mainly rely on supervised learning approaches. Although manually labeling a large number of twitter messages is a labor-intensive and time-consuming task, it is more feasible for specified events than for unspecified events (as illustrated in Table 3). When some event descriptions are known, filtering techniques could be used to reduce the amounts of irrelevant messages and make it easier for a human expert to annotate a data set of "reasonable" size. Furthermore, filtering according to specific event descriptions, such as keywords, location, or time, would also reduce the amount of Twitter messages that must be processed during system operation and allow the detection algorithm to focus on a restricted set of tweets.

Several supervised classification algorithms have been proposed for NED of specified events, including naive Bayes (Becker et al. 2011a; Sankaranarayanan et al. 2009), SVM (Becker et al. 2011a; Sakaki et al. 2010), and gradient boosted decision trees (Popescu and Pennacchiotti 2010; Popescu et al. 2011). These classifiers are typically trained on a small set of Twitter messages collected over a few weeks or months and then filtered and labeled according to the target event as, for instance, an event or nonevent (Becker et al. 2011a; Sankaranarayanan et al. 2009), an earthquake or non-earthquake event (Sakaki et al. 2010), and a controversial or noncontroversial event (Popescu and Pennacchiotti 2010; Popescu et al. 2011). The labeling procedure usually involves two human annotators with specific domain knowledge. An agreement measure, such as Cohen's Kappa measure (Carletta 1996), is then used to evaluate the level of interannotator agreement. Ambiguous events with a high level of disagreement are discarded.

In addition to filtering out part of the irrelevant messages, when the detection task involves specified events, additional features (other than word or hashtag frequency) could be included in the detection algorithm for improved system accuracy. These features may vary widely depending on the target event and its description. For instance, in addition to word frequency, Sakaki et al. (2010) considered special keywords mentioning an "earthquake", its variant, or related words (e.g., "shaking") as well as the contextual information surrounding these keywords. For detecting controversial events about celebrities, Popescu and Pennacchiotti (2010) employed a large set of linguistic, structural, burst, sentiment, and controversy features from Twitter and external features such as news buzz and Web-news controversy. For improved event detection accuracy and event description quality, Popescu et al. (2011) augmented this set with more sophisticated features based on information retrieval and natural language processing techniques, such as relative positional information, POS tagging, and main entity extraction.

4.3.3. Hybrid Detection Approaches. While supervised classification and unsupervised clustering approaches have been applied separately for NED, a combination of both approaches has also been proposed (Table 1). Some approaches employ classification or detection techniques to identify relevant or important tweets before clustering (Sankaranarayanan et al. 2009). A trained classifier used to discriminate between events and nonevents would help reduce the amount of (noisy) data provided for clustering and hence improve system efficiency. However, this approach is sensitive to the classification accuracy and threshold settings; relevant real-world events could be discarded before reaching the

TABLE 3. Summary of Data Sets and Evaluation Metrics.

	Collection	Corpus size	Temporal scope	Evaluation
Sankaranarayanan et al. (2009)	GardenHose BirdDog	–	–	Qualitative
Phuvipadawat and Murata (2010)	Streaming API (predefined search queries targeting breaking news)	10 tweets	–	Qualitative
Petrović et al. (2010)	Streaming API	163.5 M tweets	6 months	Average precision over 1000 threads from system output manually labeled as relevant, neutral, or spam
Becker et al. (2011a)	Streaming API (tweets from NYC-based users)	2.6 M tweets	1 month	Macro-averaged F_1 over 300 tweet clusters manually labeled as real-world event, Twitter-centric activity, other nonevent, or ambiguous
Long et al. (2011)	Sina API	22 M posts	2.5 months	Precision
Weng and Lee (2011)	REST API	4.3 M tweets	1 month	Precision on 21 system-detected events (tweets from 1k most followed Singapore-based users + their followers)
Cordeiro (2012)	Spritzer	13.6 M tweets	8 days	Visual illustration
Popescu and Pennacchiotti (2010)	Firehose	740 K snapshots	7 months	Average precision, area under ROC curve over 10-fold cross-validation

TABLE 3. *Continued.*

	Collection	Corpus size	Temporal scope	Evaluation
Popescu et al. (2011)	Firehose	5040 snapshots	–	Average precision, mean reciprocal rank over
Benson et al. (2011)	Tweets from NYC-based users	4.7 M tweets	3 weekends	Precision/recall over records scraped from NYC.com music event guide
Lee and Sumiya (2010)	Search API (tweets geotagged as from Japan)	21.6 M tweets	1.5 months	Precision/recall over system results manually judged against 15 prespecified festivals in a 3-day period
Sakaki et al. (2010)	Search API (queries: {typhoon} and {earthquake + shaking})	597 tweets	–	Precision/recall
Becker et al. (2011)	Twitter API	–	–	Qualitative
Massoudi et al. (2011)	Queries based on trending topics	110 M tweets	6 months	Standard IR metrics over manually assessed pool of top 20 results of each query
Metzler et al. (2012)	Streaming API	46.6 M tweets	6 months	Precision@10 for 50 different event type queries
Gu et al. (2011)	Search API (queries for 20 manually selected events across seven categories)	3.5 M tweets	5 months	Coverage/coherence

clustering stage. Other approaches first proceed with a clustering and then attempt to classify whether a cluster contains relevant information about real-world events. Because the clusters constantly evolve over time, the features are periodically updated for old clusters and computed for newly formed ones.

Another hybrid approach proposed to identify Twitter messages corresponding to concert events uses a factor graph model, which simultaneously extracts the artist name and location of the event using a supervised CRF classifier and then clusters them according to event type and induces a canonical value for each event property (Benson et al. 2011). This novel approach could be seen as a specific named entity recognizer. In this line of research, Ritter et al. (2011) developed more general NLP tools for Twitter text, including a POS tagger, a shallow parser, and a named entity recognizer based on supervised CRF models.

5. DISCUSSION

As discussed in the previous section and presented in Tables 1 and 2, event detection techniques from Twitter mainly rely on unsupervised and supervised detection approaches. While the benefit of unsupervised clustering approaches is that they do not require labeled data, several options are still available for potential optimization. For instance, setting the thresholds of incremental clustering algorithms should be based on more advanced and possibly adaptive techniques rather than simply relying on static values computed from small data sets during system design. Alternative techniques for improved clustering efficiency and scalability may be considered, such as blocking or canopy techniques (Bilenko et al. 2006; McCallum et al. 2000; Reuter and Cimiano 2012). Candidate retrieval or blocking methods alleviate the scalability issue by selecting a subset of object pairs with large similarity values (e.g., between messages or between messages and clusters), leaving out the remaining pairs as dissimilar, and hence reducing the number of messages that are considered as potential events (Bilenko et al. 2006; Reuter and Cimiano 2012). Performing clustering in two stages is another alternative. First, an approximate distance measure is used to efficiently (and roughly) divide the data into overlapping subsets or "canopies." Then, a more rigorous clustering stage using expensive distance measurements is applied to the messages that occur in a common canopy (McCallum et al. 2000). Cluster fragmentation is also another issue that deserves more focus. In addition to the temporal aspect of the messages, other features such as the location proximity (using geotags) could be used as another indicator that messages are related to the same event.

On the other hand, the proposed supervised event detection approaches generally assume a static environment. A single classifier is typically trained off-line on a relatively small batch of Twitter data labeled manually. The classifier is then deployed for detecting events directly or combined with a clustering approach. These techniques are therefore restricted in scope, because limited labeled data are available for training and Twitter is a continuously evolving environment. For instance, users may leave or join the service, active users may become inactive, new terms, abbreviations, and hashtags may emerge. A static classifier is prone to both false positive and negative errors when a concept drift occurs in the data streams. Techniques such as incremental learning (Joshi and Kulkarni 2012) and ensemble methods (Khreich et al. 2012a; Kuncheva 2004; Polikar 2006) may be employed to account for unseen events and adapt to changes that may occur over time.

Other approaches that have proved useful when dealing with sparse labeled data include semi-supervised learning (Chapelle et al. 2006; Zhu Updated on July 19, 2008) and transfer learning (Pan and Yang 2010; Pan et al. 2012). Semi-supervised learning exploits a small amount of labeled data together with the large amount of unlabeled data to build

classifiers (Chapelle et al. 2006; Zhu Updated on July 19, 2008). Transfer learning methods are designed to extract useful knowledge from different but related domains (Chapelle et al. 2006; Zhu Updated on July 19, 2008). For instance, transferring existing knowledge in Wikipedia documents to help classify Twitter messages. Following this line of research, Meij et al. (2012) proposed a method for automatically mapping tweets to Wikipedia articles to facilitate semantic mining. This approach is based on a combination of high-recall concept ranking and high-precision machine learning including random forests or gradient boosted regression trees.

People are increasingly searching the Web not only to find documents but also to make decisions. Because queries are inherently ambiguous and difficult to interpret in isolation, recent advances in Web search technologies are relying on the notion of relevance to reduce ambiguity. Therefore, several learning to rank models have been proposed for information retrieval tasks to capture document relevance by combining various global, context-specific, and user-specific features (Li et al. 2008). In addition to traditional information retrieval methods, the previously described techniques for semi-supervised and transfer learning may be used for feature generation, ranking model selection, and labeled data collection for model training and evaluation. The challenge with RED resides in designing features that capture the importance of an event with respect to specific keywords, while taking into account the temporal, geospacial, and social network specific to the user.

Performance evaluation of different approaches and features is a major issue facing event detection in Twitter. In typical information retrieval tasks, precision, recall, and F-measures are common performance metrics. The precision is the number of relevant events detected over the total number of events detected, while the recall is the number of relevant events detected over the total number of relevant events that exist in the data streams. F-measures are weighted harmonic means of precision and recall. Recall is generally difficult to compute for large and noisy data sets, because manual enumeration of all relevant events that exist in a given Twitter streams is time consuming for small sets and infeasible for larger ones. Therefore, as illustrated in Table 3, some of the work surveyed in this article only focused on precision measures such as average precision or precision@K (which capture the fraction of correctly detected events out of the top-K detected ones), while others only presented few examples of the detected events. Representative data sets and common testbed are highly required for evaluation of different detection techniques and feature representations that are proposed for event detection from Twitter. Therefore, sharing (and if possible merging) the labeled data sets online (such as those presented in Table 3), as well as using crowd sourcing services such as Amazon's Mechanical Turk for larger scale labeling, would provide more representative data for evaluation.

Most techniques focused on English language. In addition to stop-words and trivial nonalphanumeric words, all non-English words are filtered out. However, depending on the event location, Twitter messages could be written in mixed languages or in completely non-English languages. Accurate translation of all non-English tweets would increase the computational load the amounts of data to process by downstream applications. There is a growing body of mining translingual knowledge from textual data found on the Web for statistical machine translation (Nie et al. 2012). In general, techniques for both RED and NED do not always require high-quality text translations. Parallel or even comparable corpora can be directly used to train models or learn term similarity measures for query translation.

This survey focused on event detection from a single source of social media information. More robust solutions would be provided by integrating and combining event information from multiple social sources, such as Facebook, Flicker, and Youtube (Chen and Roy 2009; Kennedy et al. 2007; Kinsella et al. 2011; Mirkovic et al. 2011; Rattenbury et al. 2007; Stefanidis et al. 2013; Tang et al. 2012). Missing information from one source

may be available in the other. Clustering and classification approaches could be applied to various social media sources simultaneously using their common feature representations included in the metadata (title, user, time, tags, location, etc.). As an alternative, these detection approaches could be applied to each social media source to further exploit site-specific features and then be merged using learned similarity measures (Becker et al. 2010). Furthermore, multiple text streams could be indexed and aligned by the same set of time points called coordinated text streams (Wang et al. 2007), to determine the interesting events and associations between different streams. These approaches and additional features can help detect new events from complementary social media sites.

Although the research community is making progress toward specific event detection subtasks, simultaneously monitoring and analyzing the events and activities from different social media services remain a challenge. A considerable effort is still required to achieve efficient and reliable event detection systems, such as designing better feature extraction and query generation techniques, more accurate filtering and detection algorithms, improved techniques to combine and analyze information from multiple sources (social and traditional media) and multiple languages, and enhanced summarization and visualization approaches. Many organizations and research scholars are actively developing new systems and algorithms to overcome these challenges to exploit this rich and continuous flow of user-generated content.

6. CONCLUSION

Event detection aims at finding real-world occurrences that unfold over space and time. As a fast-growing microblogging and online social networking service, Twitter provides unprecedentedly valuable user-generated content that can be transformed into actionable and situational knowledge. More importantly, messages posted on Twitter—currently exceeding 400 million tweets per day—could reveal information about real-world events as they unfold. However, event detection from Twitter data must efficiently and accurately uncover relevant information about events of general or specific interest, which is buried within a large amount of mundane information (e.g., meaningless, polluted, and rumor messages). This article provides a survey of techniques proposed for event detection from Twitter data. These techniques are classified according to the type of target event into specified or unspecified event detection. Depending on the detection task and target application, these techniques are also classified into RED or NED. Nevertheless, they are also categorized according to the detection methods that involve supervised, unsupervised, and hybrid approaches. General and Twitter-specific feature representations corresponding to each category are also presented and discussed. Finally, this article highlights major issues and open research challenges, in particular, the need for publicly available testbeds for comprehensive evaluation of performance and objective comparison of different detection approaches.

REFERENCES

AGGARWAL, C. C. 2011. An introduction to social network data analytics. *In* Social Network Data Analytics. *Edited by* C. C. AGGARWAL. Springer: New York, pp. 1–15.

AGGARWAL, C. C., and C. ZHAI. 2012. A survey of text clustering algorithms. *In* Mining Text Data. *Edited by* C. C. AGGARWAL, and C. ZHAI. Springer: New York, pp. 77–128.

ALLAN, J. 2002. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers: Norwell, MA.

ALLAN, J., J. CARBONELL, G. DODDINGTON, J. YAMRON, and Y. YANG. 1998. Topic detection and tracking pilot study final report. *In* Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, pp. 194–218.

ALLAN, J., V. LAVRENKO, and H. JIN. 2000. First story detection in TDT is hard. *In* Proceedings of the Ninth International Conference on Information and Knowledge Management, CIKM '00, ACM, New York, NY, pp. 374–381.

ALLAN, J., V. LAVRENKO, D. MARLIN, and R. SWAN. 2000. Detections, bounds, and timelines: UMass and TDT–3. *In* Proceedings of Topic Detection and Tracking (TDT–3), Vienna, VA, pp. 167–174.

ALLAN, J., R. PARKA, and LAVRENKO V. 1998. On-line new event detection and tracking. *In* Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, ACM, New York, NY, pp. 37–45.

AMER-YAHIA, S., S. ANJUM, A. GHENAI, A. SIDDIQUE, S. ABBAR, S. MADDEN, A. MARCUS, and M. EL-HADDAD. 2012. MAQSA: A system for social analytics on news. *In* Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12, ACM, New York, NY, pp. 653–656.

BAEZA-YATES, R. A., and B. RIBEIRO-NETO. 2011. Modern Information Retrieval the Concepts and Technology Behind Search (2nd ed.). Pearson Education Ltd.: Harlow, England.

BECKER, H., F. CHEN, D. ITER, M. NAAMAN, and L. GRAVANO. 2011. Automatic identification and presentation of Twitter content for planned events. *In* International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

BECKER, H., D. ITER, M. NAAMAN, and L. GRAVANO. 2012. Identifying content for planned events across social media sites. *In* Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, ACM, New York, NY, pp. 533–542.

BECKER, H., M. NAAMAN, and L. GRAVANO. 2010. Learning similarity metrics for event identification in social media. *In* WSDM'10, ACM, New York, pp. 291–300.

BECKER, H., M. NAAMAN, and L. GRAVANO. 2011a. Beyond trending topics: Real-world event identification on Twitter. *In* ICWSM, Barcelona, Spain.

BECKER, H., M. NAAMAN, and L. GRAVANO. 2011b. Selecting quality Twitter content for events. *In* International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

BENEVENUTO, F., G. MAGNO, T. RODRIGUES, and V. ALMEIDA. 2010. Detecting spammers on Twitter. *In* Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), Redmond, WA.

BENSON, E., A. HAGHIGHI, and R. 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 of *HLT '11*, Association for Computational Linguistics, Stroudsburg, PA, pp. 389–398.

BERKHIN, P. 2002. A survey of clustering data mining techniques. Technical report, Yahoo!, Inc.

BILENKO, M., B. KAMATH, and R. J. MOONEY. 2006. Adaptive blocking: learning to scale up record linkage. *In* Proceedings of the Sixth International Conference on Data Mining, ICDM '06, IEEE Computer Society, Washington, DC, pp. 87–96.

BLEI, D. M., A. Y. NG, and M. I. JORDAN. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, **3**: 993–1022.

BOURAS, C., and V. TSOGKAS. 2010. Assigning Web news to clusters. *In* Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services, ICIW '10, IEEE Computer Society, Washington, DC, pp. 1–6.

BOYD, D., S. GOLDER, and G. LOTAN. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. *In* 2010 43rd Hawaii International Conference on System Sciences (HICSS), pp. 1–10.

BOYD, D. M., and N. B. ELLISON. 2007. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, **13**(1): 210–230.

BRANTS, T., F. CHEN, and A. FARAHAT. 2003. A system for new event detection. *In* Research and Development in Information Retrieval, New York, NY, pp. 330–337.

BRZOZOWSKI, M. J., and D. M. ROMERO. 2011. Who should I follow? Recommending people in directed social networks. *In* Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, The AAAI Press.

CARLETTA, J. 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, **22**(2): 249–254.

CASTILLO, C., M. MENDOZA, and B. POBLETE. 2011. Information credibility on Twitter. *In* Proceedings of the 20th International Conference on World Wide Web, WWW '11, ACM, New York, NY, pp. 675–684.

CHAPELLE, O., B. SCHÃŰLKOPF, and B. ZIEN. 2006. Semi-Supervised Learning. Adaptive Computation and Machine Learning series. MIT Press: Cambridge, MA.

CHEN, C. C., and M. CHANG CHEN. 2008. TSCAN: A novel method for topic summarization and content anatomy. *In* Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, ACM, New York, NY, pp. 579–586.

CHEN, L., and A. ROY. 2009. Event detection from Flickr data through wavelet-based spatial analysis. *In* Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, NY, pp. 523–532.

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

DAI, X. Y., Q. C. CHEN, X. L. WANG, and J. XU. 2010. Online topic detection and tracking of financial news based on hierarchical clustering. *In* 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, pp. 3341–3346.

EFRON, M. 2011. Information search and retrieval in microblogs. Journal of the American Society of Information Science and Technology, **62**(6): 996–1008.

FARZINDAR, A. 2012. Industrial perspectives on social networks. *In* EACL 2012 - Workshop on Semantic Analysis in Social Media.

FARZINDAR, A., and D. INKPEN. 2012. Proceedings of the Workshop on Semantic Analysis in Social Media. Association for Computational Linguistics, Avignon, France.

FOX, D., J. HIGHTOWER, L. LIAO, D. SCHULZ, and G. BORRIELLO. 2003. Bayesian filtering for location estimation. IEEE Pervasive Computing, **2**: 24–33.

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, **29**: 1189–1232.

FUJISAKA, T., R. LEE, and K. SUMIYA. 2010. Discovery of user behavior patterns from geo-tagged micro-blogs. *In* Proceedings of the 4th International Conference on Uniquitous Information Management and Communication, ICUIMC '10, ACM, New York, NY, pp. 36:1–36:10.

FUNG, G. P. C., J. XU YU, P. S. YU, and H. LU. 2005. Parameter free bursty events detection in text streams. *In* Proceedings of the 31st International Conference on Very Large Data Bases, VLDB '05, VLDB Endowment, pp. 181–192.

GESER, H. 2011. Has tweeting become inevitable? Twitter's strategic role in the world of digital communication. Online Publications.

GIONIS, A., P. INDYK, and R. MOTWANI. 1999. Similarity search in high dimensions via hashing. *In* Proceedings of the 25th International Conference on Very Large Data Bases, VLDB '99. Morgan Kaufmann Publishers Inc.: San Francisco, CA, pp. 518–529.

GOORHA, S., and L. UNGAR. 2010. Discovery of significant emerging trends. *In* Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, NY, pp. 57–64.

GU, H., X. XIE, Q. LV, Y. RUAN, and L. SHANG. 2011. ETree: Effective and efficient event modeling for real-time online social media networks. *In* Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference, Vol. 1, pp. 300–307.

HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer: Stanford, CA.

HE, Q., K. CHANG, and E.-P. LIM. 2007. Analyzing feature trajectories for event detection. *In* Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, ACM, New York, NY, pp. 207–214.

HE, Q., K. CHANG, E.-P. LIM, and J. ZHANG. 2007. Bursty feature representation for clustering text streams. *In* SIAM International Conference on Data Mining.

HOGENBOOM, F., F. FRASINCAR, U. KAYMAK, and F. DE JONG. 2011. An overview of event extraction from text. *In* Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011), Vol. 779 of CEUR Workshop Proceedings. *Edited by* M. VAN ERP, W. R. VAN HAGE, L. HOLLINK, A. JAMESON, and R. TRONCY. CEUR-WS.org: Koblenz, Germany; pp. 48–57.

HONEYCUTT, C., and S. C. HERRING. 2009. Beyond microblogging: conversation and collaboration via Twitter. *In* 42nd Hawaii International Conference on System Sciences, HICSS '09, pp. 1–10.

HURLOCK, J., and M. WILSON. 2011. Searching Twitter: separating the tweet from the chaff. *In* International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

IPSEN, I. C. F., and R. S. WILLS. 2006. Mathematical properties and analysis of Google's pagerank. Boletín de la Sociedad Española de Matemática Aplicada, **34**: 191–196.

JAIN, A. K., and R. C. DUBES. 1988. Algorithms for Clustering Data. Prentice-Hall: Upper Saddle River, NJ.

JANSEN, B. J., M. ZHANG, K. SOBEL, and A. CHOWDURY. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, **60**(11): 2169–2188.

JAVA, A., X. SONG, T. FININ, and B. TSENG. 2007. Why we Twitter: Understanding microblogging usage and communities. *In* Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD '07, ACM, New York, NY, pp. 56–65.

JIANG, L., M. YU, M. ZHOU, X. LIU, and T. ZHAO. 2011. Target-dependent Twitter sentiment classification. *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, pp. 151–160.

JO, T., and M. LEE. 2007. The evaluation measure of text clustering for the variable number of clusters. *In* Proceedings of the 4th International Symposium on Neural Networks: Part II–Advances in Neural Networks, ISNN '07. Springer-Verlag: Berlin, Heidelberg, pp. 871–879.

JOSHI, P., and P. KULKARNI. 2012. Incremental learning: areas and methods – a survey. International Journal of Data Mining & Knowledge Management Process, **2**(5): 43–51.

JURAFSKY, D., and J. H. MARTIN. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). Prentice Hall: Upper Saddle River, NJ.

KAPLAN, A. M., and M. HAENLEIN. 2011. The early bird catches the news: Nine things you should know about micro-blogging. Business Horizons, **54**(2): 105–113.

KENNEDY, L., M. NAAMAN, S. AHERN, R. NAIR, and T. RATTENBURY. 2007. How Flickr helps us make sense of the world: Context and content in community-contributed media collections. *In* Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, ACM, New York, NY, pp. 631–640.

KHAN, A. A. 2012. The role social of media and modern technology in arabs spring. Far East Journal of Psychology and Business, **7**(1): 56–63.

KHONDKER, H. H. 2011. Role of new media in the Arab Spring. Globalizations, **8**(5): 575–679.

KHREICH, W., E. GRANGER, A. MIRI, and R. SABOURIN. 2012a. Adaptive ROC-based ensembles of HMMs applied to anomaly detection. Pattern Recognition, **45**(1): 208–230.

KHREICH, W., E. GRANGER, A. MIRI, and R. SABOURIN. 2012b. A survey of techniques for incremental learning of HMM parameters. Information Sciences, **197**: 105–130.

KHREICH, W., E. GRANGER, R. SABOURIN, and A. MIRI. 2009. Combining hidden Markov models for anomaly detection. *In* International Conference on Communications (ICC), Dresden, Germany, pp. 1–6.

KIM, H. D., K. GANESAN, P. SONDHI, and C. ZHAI. 2011. Comprehensive review of opinion summarization. Technical report, University of Illinois at Urbana-Champaign.

KINSELLA, S., A. PASSANT, and BRESLIN. J. 2011. Topic classification in social media using metadata from hyperlinked objects. *In* Advances in Information Retrieval, Vol. 6611 of *Lecture Notes in Computer Science*. *Edited by* P. CLOUGH, C. FOLEY, C. GURRIN, G. JONES, W. KRAAIJ, H. LEE, and V. MUDOCH. Springer Berlin: Heidelberg, pp. 201–206.

KLEINBERG, J. 2002. Bursty and hierarchical structure in streams. *In* Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, ACM, New York, NY, pp. 91–101.

KONTOSTATHIS, A., L. GALITSKY, W. M. POTTENGER, S. ROY, and D. J. PHELPS. 2004. A survey of emerging trend detection in textual data mining. *In* A Comprehensive Survey of Text Mining: Clustering, Classification and Retrieval. *Edited by* M. W. BERRY. Springer: New York.

KRISHNAMURTHY, B., P. GILL, and M. ARLITT. 2008. A few chirps about Twitter. *In* Proceedings of the First Workshop on Online Social Networks, WOSN '08, ACM, New York, NY, pp. 19–24.

KUMARAN, G., and J. ALLAN. 2004. Text classification and named entities for new event detection. *In* Research and Development in Information Retrieval, Sheffield, UK, pp. 297–304.

KUNCHEVA, L. I. 2004. Classifier ensembles for changing environments. *In* Multiple Classifier Systems, Vol. 3077, Cagliari, Italy, pp. 1–15.

KWAK, H., C. LEE, H. PARK, and S. MOON. 2010. What is Twitter, a social network or a news media? *In* Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, pp. 591–600.

LAVRENKO, V., and W. B. CROFT. 2001. Relevance based language models. *In* Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, ACM, New York, NY, pp. 120–127.

LEE, K., B. EOFF, and J. CAVERLEE. 2011. Seven months with the devils: A long-term study of content polluters on Twitter. *In* International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

LEE, R., and K. SUMIYA. 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. *In* Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10, ACM, New York, NY, pp. 1–10.

LEEK, T., R. SCHWARTZ, and S. SISTA. 2002. Probabilistic approaches to topic detection and tracking. *In* Topic Detection and Tracking. *Edited by* J. ALLAN. Kluwer Academic Publishers: Norwell, MA, pp. 67–83.

LI, P., C. BURGES, and Q. WU. 2008. McRank: Learning to rank using multiple classification and gradient boosting. *In* Advances in Neural Information Processing Systems 20. *Edited by* J. PLATT, D. KOLLER, Y. SINGER, and S. ROWEIS. MIT Press: Cambridge, MA, pp. 897–904.

LI, Z., B. WANG, M. LI, and W. MA. 2005. A probabilistic model for retrospective news event detection. *In* Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, ACM, New York, NY, pp. 106–113.

LIU, K. L., W. LI, and M. GUO. 2012. Emoticon smoothed language models for Twitter sentiment analysis. *In* Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON.

LONG, R., H. WANG, Y. CHEN, O. JIN, and Y. YU. 2011. Towards effective event detection, tracking and summarization on microblog data. *In* Web-Age Information Management, Vol. 6897 of *Lecture Notes in Computer Science*. *Edited by* H. WANG, S. LI, S. OYAMA, X. HU, and T. QIAN. Springer: Berlin/Heidelberg, pp. 652–663.

LOVEJOY, K., and G. D. SAXTON. 2012. Information, community, and action: How nonprofit organizations use social media. Journal of Computer-Mediated Communication, **17**(3): 337–353.

LUO, G., C. TANG, and P. S. YU. 2007. Resource-adaptive real-time new event detection. *In* Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07, ACM, New York, NY, pp. 497–508.

MANNING, C. D., and H. SCHÜTZE. 1999. Foundations of Statistical Natural Language Processing. MIT Press: Cambridge, MA.

MASSOUDI, K., M. TSAGKIAS, M. DE RIJKE, and W. WEERKAMP. 2011. Incorporating query expansion and quality indicators in searching microblog posts. *In* Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11. Springer-Verlag: Berlin, Heidelberg, pp. 362–367.

MATHIOUDAKIS, M., and N. KOUDAS. 2010. TwitterMonitor: Trend detection over the Twitter stream. *In* SIGMOD Conference, Indianapolis, IN, pp. 1155–1158.

MCCALLUM, A., K. NIGAM, and L. H. UNGAR. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. *In* Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, ACM, New York, NY, pp. 169–178.

MCFEDRIES, P. 2007. Technically speaking: All A-Twitter. IEEE Spectrum, **44**(10): 84.

MEIJ, E., W. WEERKAMP, and M. DE RIJKE. 2012. Adding semantics to microblog posts. *In* Proceedings of the fifth ACM International Conference on Web Search and Data Mining, WSDM '12, ACM, New York, NY, pp. 563–572.

METZLER, D., C. CAI, and E. H. HOVY. 2012. Structured event retrieval over microblog archives. *In* HLT-NAACL, pp. 646–655.

METZLER, D., S. DUMAIS, and C. MEEK. 2007. Similarity measures for short segments of text. *In* Proceedings of the 29th European Conference on IR Research, ECIR'07. Springer-Verlag: Berlin, Heidelberg, pp. 16–27.

MIRKOVIC, M., D. CULIBRK, S. PAPADOPOULOS, C. ZIGKOLIS, Y. KOMPATSIARIS, G. MCARDLE, and V. CRNOJEVIC. 2011. A comparative study of spatial, temporal and content-based patterns emerging in Youtube and Flickr. *In* Computational Aspects of Social Networks (CASoN), 2011 International Conference, pp. 189–194.

MOHD, M. 2007. Named entity patterns across news domains. *In* Proceedings of the 1st BCS IRSG conference on Future Directions in Information Access, FDIA'07, British Computer Society, Swinton, UK, pp. 5–5.

MURPHY, K. P. 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning Series. Mit Press: Cambridge, MA.

MURTHY, D. 2011. Twitter: microphone for the masses? Media, Culture & Society, **33**(5): 779–789.

NAAMAN, M., H. BECKER, and GRAVANO. L. 2011. Hip and trendy: characterizing emerging trends on Twitter. Journal of the American Society of Information Science and Technology, **62**(5): 902–918.

NALLAPATI, R., A. FENG, F. PENG, and J. ALLAN. 2004. Event threading within news topics. *In* Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04, ACM, New York, NY, pp. 446–453.

NIE, J. Y., J. GAO, and G. CAO. 2012. Translingual mining from text data. *In* Mining Text Data. *Edited by* C. C. AGGARWAL, and C. ZHAI. Springer: New York, pp. 323–359.

OH, O., M. AGRAWAL, and H. RAO. 2011. Information control and terrorism: Tracking the Mumbai terrorist attack through Twitter. Information Systems Frontiers, **13**: 33–43.

PAK, A., and P. PAROUBEK. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *In* Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta.

PAN, S. J., and Q. YANG. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, **22**: 1345–1359.

PAN, W., E. ZHONG, and Q. YANG. 2012. Transfer learning for text mining. *In* Mining Text Data. *Edited by* C. C. AGGARWAL, and C. ZHAI. Springer: New York, pp. 223–257.

PAPKA, R. 1999. On-line new event detection, clustering and tracking. Ph.D. thesis, Department of Computer Science, University of Massachusetts, Amherst.

PARANJPE, D. 2009. Learning document aboutness from implicit user feedback and document structure. *In* Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, NY, pp. 365–374.

PETROVIĆ, S., M. OSBORNE, and V. LAVRENKO. 2010. Streaming first story detection with application to Twitter. *In* Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pp. 181–189.

PHUVIPADAWAT, S., and T. MURATA. 2010. Breaking news detection and tracking in Twitter. *In* IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 3, Toronto, ON, pp. 120–123.

POLIKAR, R. 2006. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, **6**(3): 21–45.

POPESCU, A. M., and M. PENNACCHIOTTI. 2010. Detecting controversial events from Twitter. *In* Proceedings of the 19th ACM international Conference on Information and Knowledge Management, CIKM '10, ACM, New York, NY, pp. 1873–1876.

POPESCU, A. M., M. PENNACCHIOTTI, and D. PARANJPE. 2011. Extracting events and event descriptions from Twitter. *In* Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pp. 105–106.

RATTENBURY, T., N. GOOD, and M. NAAMAN. 2007. Towards automatic extraction of event and place semantics from Flickr tags. *In* Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, ACM, New York, NY, pp. 103–110.

REINHARDT, W., B. GUNTER, E. MARTIN, and C. COSTA. 2009. How people are using Twitter during conferences. *In* Proceedings of 5th EduMedia Conference, Salzburg, Vienna, pp. 145–156.

REUTER, T., and P. CIMIANO. 2012. A systematic investigation of blocking strategies for real-time classification of social media content into events. *In* International AAAI Conference on Weblogs and Social Media, Dublin, Ireland.

RITTER, A., M. SAM CLARK, and O. ETZIONI. 2011. Named entity recognition in tweets: An experimental study. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, pp. 1524–1534.

SAKAKI, T., M. OKAZAKI, and Y. MATSUO. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. *In* Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, pp. 851–860.

SALTON, G. 1988. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley: Boston.

SANKARANARAYANAN, J., H. SAMET, B. E. TEITLER, M. D. LIEBERMAN, and J. SPERLING. 2009. TwitterStand: News in tweets. *In* Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09, ACM, New York, NY, pp. 42–51.

SHARIFI, B., M. A. HUTTON, and J. 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, pp. 685–688.

SMALL, T. A. 2011. What the hashtag? Information, Communication & Society, **14**(6): 872–895.

SNOWSILL, T., F. NICART, M. STEFANI, T. DE BIE, and N. CRISTIANINI. 2010. Finding surprising patterns in textual data streams. *In* Cognitive Information Processing (CIP), 2010 2nd International Workshop, Tuscany, Italy, pp. 405–410.

STARBIRD, K., P. PALEN, A. L. HUGHES, and S. VIEWEG. 2010. Chatter on the red: What hazards threat reveals about the social life of microblogged information. *In* Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work, CSCW '10, ACM, New York, NY, pp. 241–250.

STEFANIDIS, A, A. CROOKS, and J. RADZIKOWSKI. 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal, **78**(2): 319–338.

TANG, J., X. WANG, H. GAO, X. HU, and H. LIU. 2012. Enriching short text representation in microblog for clustering. Frontiers of Computer Science in China, **6**(1): 88–101.

TEEVAN, J., D. RAMAGE, and M. R. MORRIS. 2011. #TwitterSearch: A comparison of microblog search and web search. *In* Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, ACM, New York, NY, pp. 35–44.

THOMPSON, C. 2008. Brave New World of Digital Intimacy. New York Times. Online. Accessed on November 1, 2012.

TRONCY, R., B. MALOCHA, and A. T. S. FIALHO. 2010. Linking events with media. *In* Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, ACM, New York, NY, pp. 42:1–42:4.

TUMASJAN, A., T. O. SPRENGER, P. G. SANDNER, and I. M. WELPE. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *In* Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM. The AAAI Press: Washington, DC.

WANG, X., M. S. GERBER, and D. E. BROWN. 2012. Automatic crime prediction using events extracted from Twitter posts. *In* Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, SBP'12. Springer-Verlag: Berlin, Heidelberg, pp. 231–238.

WANG, X., C. ZHAI, X. HU, and R. SPROAT. 2007. Mining correlated bursty topic patterns from coordinated text streams. *In* Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, ACM, New York, NY, pp. 784–793.

WEERKAMP, W., and M. DE RIJKE. 2008. Credibility improves topical blog post retrieval. *In* ACL, Columbus, OH, pp. 923–931.

WENG, J., and B.-S. LEE. 2011. Event detection in Twitter. *In* ICWSM, Barcelona, Spain.

XIE, L., H. SUNDARAM, and M. CAMPBELL. 2008. Event mining in multimedia stream. Proceedings of the IEEE, **96**(4): 623–647.

YANG, Y., J. G. CARBONELL, R. D. BROWN, T. PIERCE, B. T. ARCHIBALD, and X. LIU. 1999. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, **14**(4): 32–43.

YANG, Y., T. PIERCE, and J. CARBONELL. 1998. A study of retrospective and on-line event detection. *In* Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, ACM, New York, NY, pp. 28–36.

YANG, Y., J. ZHANG, J. CARBONELL, and C. JIN. 2002. Topic-conditioned novelty detection. *In* Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, ACM, New York, NY, pp. 688–693.

ZHAO, D., and M. B. ROSSON. 2009. How and why people Twitter: The role that micro-blogging plays in informal communication at work. *In* Proceedings of the ACM 2009 International Conference on Supporting Group Work, GROUP '09, ACM, New York, NY, pp. 243–252.

ZHU, X. Updated on July 19, 2008. Semi-supervised learning literature survey, University of Wisconsin Madison Technical Report TR 1530.