Geoinformatica (2016) 20:765–795
DOI 10.1007/s10707-016-0263-0

Automatic targeted-domain spatiotemporal event detection in twitter

Ting Hua[1] · Feng Chen[2] · Liang Zhao[1] · Chang-Tien Lu[1] · Naren Ramakrishnan[1]

Received: 9 November 2015 / Revised: 21 June 2016 /
Accepted: 28 June 2016 / Published online: 8 August 2016

Abstract Twitter has become an important data source for detecting events, especially tracking detailed information for events of a specific domain. Previous studies on targeted-domain Twitter information extraction have used supervised learning techniques to identify domain-related tweets, however, the need for extensive manual labeling makes these supervised systems extremely expensive to build and maintain. What's more, most of these existing work fail to consider spatiotemporal factors, which are essential attributes of target-domain events. In this paper, we propose a semi-supervised method for Automatical Targeted-domain Spatiotemporal Event Detection (ATSED) in Twitter. Given a targeted domain, ATSED first learns tweet labels from historical data, and then detects on-going events from real-time Twitter data streams. Specifically, an efficient label generation algorithm is proposed to automatically recognize tweet labels from domain-related news articles, a customized classifier is created for Twitter data analysis by utilizing tweets' distinguishing features, and a novel multinomial spatial-scan model is provided to identify geographical locations for detected events. Experiments on 305 million tweets demonstrated the effectiveness of this new approach.

Keywords Social media · Data mining · Spatiotemporal

✉ Ting Hua
tingh88@vt.edu

Liang Zhao
zhaoliangvaio@gmail.com

Naren Ramakrishnan
naren@cs.vt.edu

[1] Department of Computer Science, Virginia Tech, Falls Church, VA, USA

[2] Department of Computer Science, University at Albany-SUNY, Albany, NY, USA

1 Introduction

Online social microblogs such as Twitter have become a major medium for information sharing. The rich up-to-date sensing data in Twitter allows important events to be discovered and tracked prior to their inclusion in standard news bulletins. When a social event occurs, traditional media usually take hours or even days to report the related news, while the corresponding information may begin to spread immediately after the occurrence in social media like Twitter [29, 32]. For example, Fig. 1 depicts the number of tweets and news reports related to a spatiotemporal event (a protest held by local residents) that happened around 12 noon on January 12th, 2013 in Mexico. Number of event-related tweets immediately increased after the event began (12 noon), while the first news report was published at 2 pm, 2 hours later than the tweet burst.

Although detecting events from formal texts has been extensively studied [3, 5], analyzing messages from Twitter requires more sophisticated techniques. First, newswire texts are relatively long and well written, while Twitter messages are short and written in a much more informal style. It is therefore unrealistic to simply apply traditional news-text based event detection methods on Twitter data. What's more, events mentioned in news documents have already been identified as being of general importance, but in the case of Twitter data, nearly half of all tweets are actually non-event related babbles discussing the minutia of daily life.

In previous studies of Twitter event detection, most researchers have adopted *general-domain event detection* approaches to extract popular open-domain events, without imposing specific constraints on event type. These methods generally utilize unsupervised learning techniques, such as clustering [13, 31], topic modeling [33], and burst detection [11], all of which can catch breaking news yet will not normally identify relatively small-scale spatiotemporal events. However, different users may demand different information from Twitter. For instance, companies need feedback about their products from customers, governments seek data related to social events (such as crime [14], civil unrest, and disease outbreaks [2, 28]), and scientists are interested in collecting tweets about natural disasters [26] or climate changes . We call these demands related to tracking information in a specific domain *targeted-domain event detection*. Existing *targeted-domain event detection* methods have applied supervised learning techniques (e.g., SVM) to differentiate event-related tweets from non-event relevant contexts [14, 26]. However, these methods suffer from the following shortcomings: **1) Highly relying on manually-labelled data.** To build a training dataset for supervised learning, these technologies require extensive human input to label tweet data correctly, and to maintain good system performance, these label datasets must be updated regularly. Each day, more than 200 million active Twitter users publish over 400

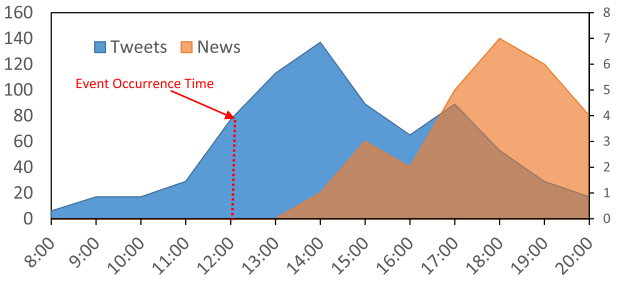

Fig. 1 Number of tweets and news reports related to a protest event occurring at around 12 pm on January, 12, 2013 in Mexico

Date	Location	Top 5 Event Phrases
1/10/2012	Mexico/Mexico City	peasant, camp, Tabasco, protest, 900
1/10/2012	Brazil/São José	GM, protest, layoff, Jaime, close
1/12/2012	Mexico/Mexico City	dogs, protest, #yosoycan26,march,Zocalo
1/13/2012	Mexico/Hermosillo	Padrés, tax, protest, Sonora, congress

Fig. 2 ATSED output example

million tweets.[1] This huge volume of data makes periodical updates extremely expensive and even unrealistic. **2) Inability to utilize Twitter's distinct features.** Classifiers designed by existing methods usually treat Twitter data as a set of plain textual documents, without any consideration of Twitter network properties such as "mentions", "hashtags", and "replies". In Twitter data, "hashtags" can be used to denote tweets about the same topics, one user can "mention" another user, a tweet can be "replied" by another tweet. **3) Restricted ability to estimate event location.** Existing methods usually predict event location through single location terms that either involve user locations [14] or GPS tags [26], discarding all other types of geopolitical terms. Instead, our proposed multinomial spatial scan considers all possible Twitter location terms, including registered locations in user profiles, GPS information, and geo-tags mentioned in the tweet content.

In this paper, we propose a semi-supervised approach for detecting spatiotemporal events from Twitter, named Automatical Targeted-domain Spatiotemporal Event Detection (ATSED). Figure 2 is an illustrative example of our model output. Given historical news reports related to a specific domain, such as "civil unrest", ATSED can yield a set of real-time "civil unrest" events detected from Twitter, which are consisted of key information such as location, date, and brief description. First, utilizing the knowledge learned from news reports, ATSED can automatically generate labels from historical Twitter data. These Twitter labels are then served as training data for a classifier specially designed for Twitter data analysis. Next, the trained classifier can be applied to real-time Twitter data steams to identify event related tweets. Finally, event locations are extracted from event-related tweets through a novel multinomial spatial-scan method. In summary, this article makes the following contributions:

- **Methodology for automatic label generation.** Labels are generated from historical tweets, which are first ranked by various similarities to news documents, and then separated into positive and negative examples through an EM inferring algorithm. This method eliminates the need of using manually selected label data, and therefore reduces the cost associated with human input.
- **Customized text classifier for Twitter data.** To better analyze Twitter data, we utilize distinct Twitter features, such as hashtags, mentions, and replies to cluster tweets before classification. This attempt enables classification based on tweet groups rather than single tweets, which therefore greatly improves classification accuracy.
- **Multinomial spatial-scan location estimation.** We extend spatial scan statistics with multinomial distribution by combining factors from various location items (e.g., user-profile locations or geo-tags). This approach makes maximum usage of all Twitter geographical information.

[1] https://blog.twitter.com/2013/celebrating-twitter7.

– **Extensive experimental evaluation and performance analysis.** Our method was extensively evaluated on a real world dataset containing 305 million tweets. Compared to existing state-of-the-art methods, our method clearly demonstrated its effectiveness.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 provides a general overview of the proposed ATSED system and formally defines the problem. Section 4 describes the detailed algorithms utilized for label generation. Section 5 presents the proposed event detection methods. The performance analysis is discussed in Section 6, and the paper is concluded in Section 7.

2 Related work

This section reviews research directions related to our work. The first branch consists of detection methods that have been widely used in tracking events from news stream. Recently, event detection on social media streams becomes a hot research topic. Existing event detection algorithms can be broadly classified into two categories: general domain and targeted domain approaches. Besides, in aspect of automatical label generation, our approach is related to distant supervision and transfer learning.

2.1 Event detection in newswire documents

Much research has focused on detecting events from formal texts, such as news articles, blogs, and emails. Some of these approaches group documents into events based on their semantic similarity. Brants et al. [3] built an event detection system based on incremental TF-IDF model, identifying events by caculating the Hellinger distance between new texts and previous documents. Kumaran et al. [10] took a different approach, detecting new events by extending cosine similarity and the vector space model to include story categorization and the use of named entities. Other researchers have sought to first identify event-related features and then cluster feature bursts into events. For example, Fung et al. [5] proposed a way to identify events that consist of a set of bursty features appearing simultaneously. Their model treats bursty features as a time-series of probability, and then groups strongly interrelated bursty features into bursty events. Bursty features are first evaluated by their distributions, and strongly interrelated bursty features are then grouped to create bursty events.

While news event detection methods work well for formally written news articles, they are incapable of detecting events from social media data like tweets. Tweets are very short and often written informally with abbreviations and mistakes. More sophisticated technologies that can handle noisy Twitter data are therefore desired.

2.2 General-domain event detection in twitter

In order to detect emerging general events in Twitter steams, *general-domain event detection* usually applies unsupervised learning techniques, such as topic modeling, burst detection, and clustering. Topic modelling is a particularly popular solution, since event detection in Twitter data is similar to the problem of topic detection in formal texts. For example, Yin et al. [33] developed topic modeling techniques to detect geographic topic clusters in local

regions. Cataldi et al. [4] proposed the use of a graph for topic detection, using emerging terms in tweets posted by authoritative users. Ritter et al. [25] prefer to focus on extracting events from noisy Twitter data and then generating event categories based on latent variable models. Another alternative is to detect events through spatial-temporal word bursts. Lappas et al. [11] examined ways to discover terms that burst in geographical neighborhoods within a certain time period, taking into account content, structural, and temporal signals. Clustering technologies have also been utilized for event detection. For example, a recent study applied wavelet analysis for noise filtering in Twitter and identified word groups with high correlations as indicators of an event [31]. Petrovic et al. [20] detected breaking news from Twitter data by building a nearest-neighbor tweet network and summarizing connected tweets into events.

Our goal differs from that of *general-domain event detection* as we are seeking to detect events in a particular domain, such as earthquakes, disease outbreaks, social unrest, or crimes. From the perspective of detection, our work is most closely related to *targeted-domain event detection* and its approaches.

2.3 Targeted-domain event detection in twitter

Supervised learning methods are commonly used in *targeted-domain event detection*. Typically, a classifier is trained via manually labeled data to identify tweets in the targeted domains and then clustering techniques are applied to analyze the events' locations. Sakaki et al. [26] first trained a SVM classifier to recognize tweets about "earthquake", and then built a Kalman filtering model to detect the geographic regions of these events. Similarly, Li et al. [14] focused on crime event detection from Twitter, by training a classifier with crime domain keywords and Twitter-specific features (e.g., hashtags). Popescu et al. [22] utilized targeted named entities and a decision-tree strategy to decide whether corresponding snapshots do indeed represent an event. Becker et al. [1] began by clustering similar tweets, and then applied a manually trained classifier to identify different events, based on features such as hashtags and retweets. Zhang et al. [34] utilized labeled documents from a source domain to help build latent semantic space for short texts in the target domain. Unlike the method presented here, their methods all require extensive labeled data in the source domain.

Due to their supervised nature, existing methods aimed at detecting targeted events usually require expensive human effort to create suitable labeled data. Our previous work proposed a method that capable of identifying trustworthy tweets [35]. In this work, we attempt to build an appropriate label dataset automatically, and utilize these automatically generated data for detecting spatiotemporal events.

In summary, traditional event detection methods are suitable for news documents, but works poorly in noisy Twitter data. Most of previous work on social media detection are general-domain approaches. General-domain event detection methods are to identify breaking news, which are most popular events during a certain period of time, regardless the specific event type. There exist only few detecting methods that are able to recognize events of targeted event types, which are most closely related to our proposed ATSED. But unlike ATSED, none of these targeted-domain detection systems are capable of automatically detecting social media events without pre-given human labeled data. And few of these previous work focused on spatiotemporal event detection and made poor utilization of tweets' location information.

2.4 Distant supervision and transfer learning

Transfer learning techniques usually first extract the knowledge from the source domain and then utilize the knowledge for tasks in the targeted domain [19]. There exist some approaches adopted transfer learning technologies for Twitter text mining. Jin et al. [7] developed a variation of LDA to jointly learn topics from both short and long texts. The knowledge shared by the two datasets is controlled by different settings of Dirichlet priors. Zhang et al. [34] first learned a latent semantic space from source dataset, and then mapped the target dataset to the space for the further mining tasks. Phan et al. [21] enriched Twitter with hidden topics learned from external data source such as Wikipedia and MEDLINE. This model is designed to find long texts related to given short texts, oppositely, our work aims to extract short tweet labels from given long articles.

Distant supervision methods heuristically label corpus using supervision from known knowledge base [27]. Mintz et al. [16] use existing relations in external knowledge base as training data. For each entities pair, they collected all the sentences mentioning them in text, and use their relation type in knowledge base as label. Based on these generated labels, they trained a classifier to learn relations. Purver et al. [23] used some heuristical intuition (emotional marker) to generate noisy labels first, and then examined the classifiers trained by these pseudo labels.

The intuition behind distant supervision, transfer learning, and our proposed method is that: some hidden patterns and relationships are shared by source and target datasets, and the learned knowledge from the source is likely to appear in the target data in some way. Our method can be view as distant supervision as we generated pseudo labels with heuristical rules first, and demonstrated our good performance despite the imperfect labels. Most distant supervision methods are proposed to the relationship between entities or words [16, 24], under the supervision of large knowledge base. But our goal here is to study the relationship between events and words, and the supervisor is external document dataset (similar to transfer learning to some extent).

3 Framework and problem formulation

This section first introduces the framework of ATSED, then formally describes some key concepts used in this paper, and finally define the tasks of this paper based on these concepts.

3.1 Framework

Our framework consists of two main components: *label generation* and *spatiotemporal event detection*. The input data sources contain: historical Twitter data, historical news articles, and real time Twitter streams. Historical Twitter data and news articles are used by *label generation* component to produce pseudo labels. *Spatiotemporal event detection* module then trains classifier through these labels and detects events from real time Twitter data.

In the *label generation* component, tweet labels are generated utilizing historical news articles. Based on the labels generated from the historical data by the *label generation* module, the *spatiotemporal event detection* module can now move on to identify on-going events related to the targeted interest from real-time Twitter streams.

The *label generation* module can produce both positive and negative tweet examples with knowledge learned from given news report documents. First, the submodule *feature*

extraction detects domain-feature *domain words* and event-feature *event words* from news reports. Next, the *domain words* and *event words* are utilized as queries to search Twitter data. Then, a *relevancy ranking* method is proposed to evaluate tweets' relevancy to the given event, based on the spatial, temporal, and textual similarities between tweets and event-related news documents. Tweets with high relevancy scores are considered as candidates for *positive examples*, while tweets with low scores are potential *negative examples*. Finally, an expectation maximization (EM) *label refinement* algorithm is provided to further separate the positive and negative examples.

The *Twitter classifier* submodule combines clustering and classification. Tweets in the real-time Twitter data stream are first clustered into mini-tweet-groups, utilizing tweets' social ties such as hashtags, mentions, and replies. Next, clustered tweet groups are input into the trained classifier (using historical labels from the *label generation* module), which identifies the positive and negative classes for tweets. In the *location estimation* submodule, an extended spatial scan approach is harnessed to cluster tweets in the positive class into different spatiotemporal events. As a result, each event detected by ATSED is represented by location, timestamp, and event-related tweets.

3.2 Problem formulation

Corresponding to the framework introduced above, targeted-domain Twitter event detection can be formally defined in terms of two tasks, *label generation* and *spatiotemporal event detection*, beginning with a few key concepts as follows.

First, different from trivial daily life events, events mentioned in our paper are something "significant". These events should be discussed in public media and associated with some news articles, since they are significant.

Definition 1 (**Spatiotemporal event**) An spatiotemporal event $x = (l, t)$ is a significant real-world incident that happened at location l and time t. Domain $\mathbf{X}_p$ is defined as a set of events falling into the same domain p, such as music, sports, civil unrest, etc.

Definition 2 (**Article**) The article set of targeted domain p is designated $\mathbf{A}_p$, while the set of open-domain articles (containing various topics) is designated $\mathbf{A}$. An article $a_x \in \mathbf{A}_p$ denotes a news report document about event x. Notice that one event may be associated with multiple news reports, so we merge these documents into one article.

Suppose we are interested in detecting events in the targeted-domain "civil unrest". For example, the event "dog protest" happened on January, 12, 2013 in Mexico.[2] A segment of the event-related news article is as follows (the original Spanish text has been translated into English using Google Translate):

> *Accompanied by a dozen of dogs, about 150 people of the movement YoSoyCan26 marched around the Zocalo of Mexico City, and insisted 57 dogs that were captured as the homicides in Cerro de la Estrella be freed.*

Besides news articles, when an event occurs, there could also exist some tweets that relevant to the given event. Among these event-related tweets, some are truly relevant to the given event. For example, tweet "With protests in the Zocalo, # YoSoyCan26 requires

[2] http://www.milenio.com/cdb/doc/noticias2011/fcd1c695e4a21d7edcae432c9f931ecd?$quicktabs_1 = 2$.

Iztapalapa dogs to be free."[3] is a positive tweet to event "dog protest". In contrast, *negative examples* are tweets that share some features with positive ones yet are in fact irrelevant to the given event. For example, the tweet "I do not understand social networks. Fuss over a dog, I have not seen it to help people in the street."[4] has some positive features (e.g., "dog" and "street"), but fails to provide any information related to the given protest event.

Definition 3 (**Tweet**) A tweet $y = (d, l, t)$ contains textual document d, location l and time-stamp t. Twitter data stream $\mathbf{Y}$ is therefore defined as a set of tweets.

Definition 4 (**Positive tweet**) A tweet $y^{(x)} = (d, l, t)$ containing textual document d, location l and time-stamp t is a positive tweet to event x, if it is truly related to event x.

Definition 5 (**Negative tweet**) A tweet $\bar{y}^{(x)} = (d, l, t)$ contains textual document d, location l and time-stamp t.

With concepts of "event", "article", and "tweets", we can further define the concept "label" used in this paper, which consists of event, event news article, and event tweets.

Definition 6 (**Label**) A label z is defined as $(x, \mathbf{Y}^{(x)}, \bar{\mathbf{Y}}^{(x)})$, where x is an event, $\mathbf{Y}^{(x)}$ is the set of tweets related to event x, and $\bar{\mathbf{Y}}^{(x)}$ are irrelevant tweets. The label set $\mathbf{Z}_p = \{(x, \mathbf{Y}^{(x)}, \bar{\mathbf{Y}}^{(x)}) | x \in \mathbf{X}_p\}$ for target domain p consists of labels generated from events X_p in domain p.

Given a list of historical events and corresponding newswire documents, the task of *label generation* is to determine the set of tweets related to each event.

Task 1 (**Label generation**) *Given an event set X_p and a news article set $\mathbf{A}_p$, where each event $x_i \in \mathbf{X}_p$ has a corresponding news article $a_{x_i} \in \mathbf{A}_p$, the goal of label generation is to find label set $\mathbf{Z}_p = \{(x, \mathbf{Y}^{(x)}, \bar{\mathbf{Y}}^{(x)}) | x \in \mathbf{X}_p\}$, from historical tweets $\mathbf{Y}$.*

Note that, both the tweets and news articles used in the *label generation* module consist of historical data. In contrast, *spatiotemporal event detection* discovers newly emerging events in the targeted domain, therefore Twitter data used in *spatiotemporal event detection* consist of real-time data streams.

Task 2 (**Spatiotemporal event detection**) *Given a label set $\mathbf{Z}_p$ (product of Task 1) and real-time Twitter stream $\mathbf{Y}'$, the event detection algorithm aims to identify an on-going event set $\mathbf{X}'_p$ for targeted domain p from Twitter data stream $\mathbf{Y}'$. Each spatiotemporal event $x' \in \mathbf{X}'_p$ consists of location l', time t', and event-related tweets $\mathbf{I}_p^{(x')}$.*

4 Automatic label generation

In this section, we discuss Automatic Label Generation (ALG) algorithm in detail. First, ALG extracts feature terms from news reports, then ranks tweets based on their similarities

[3] https://twitter.com/BicitlanRadio/status/290232591246823425.

[4] https://twitter.com/revistaeneo/status/290185989815676930.

to the news reports, and finally splits the tweet set into positive and negative examples through an EM based refinement algorithm.

4.1 Feature extraction

The goal of *feature extraction* is to obtain features that can identify a specific event in the targeted domain. Although tweets and news articles are quite different in writing style, they are likely to share some semantic features when describing the same event, which are referred to as *domain words* and *event words* in this paper. *Domain words* are those most representative words for events occurring in a certain domain. For example, the words "protest" and "march" may be *domain words* for "civil unrest" events. *Event words* are words that can distinguish a particular event from other events in the same domain. In the above mentioned news article ("dog protest" event), the words "YoSoyCan26" and "Zocalo" are *event words* which are highly relevant to the specific event. To identify "*domain words*" and "*event words*", we define *domain weight* and *event weight* as follows.

Definition 7 (**Domain weight**) Domain weight $C(w_i, p)$ quantifies the ability of word w_i in representing targeted domain p. Given targeted-domain news article set $\mathbf{A}_p = \cup_{i=1}^{n} a_{x_i}$ and an open-domain document set $\mathbf{A}$, $C(w_i, p)$ is computed as the product of two parts, namely the normalized term frequency $f(w_i, \mathbf{A}_p)$ of word w_i in open-domain set $\mathbf{A}_p$, and the inverse document frequency of w_i in targeted-domain set $\mathbf{A}$:

$$C(w_i, p) = \frac{f(w_i, \mathbf{A}_p)}{\max\{f(w, \mathbf{A}_p) : w \in \mathbf{A}_p\}} \times \lg\left(\frac{|\mathbf{A}|}{|\{a \in \mathbf{A} : w_i \in a\}| + 1}\right). \tag{1}$$

Definition 8 (**Event weight**) Event weight $E(w_i, x)$ quantifies the ability of word w_i in distinguishing event x from other events in the same domain. It is computed as the product of two parts, the term frequency of word w_i in event article a_x, and the inverse document frequency of w_i in document set $\mathbf{A}_p$:

$$E(w_i, x) = \frac{f(w_i, a_x)}{\max\{f(w, a_x) : w \in a_x\}} \times \lg\left(\frac{|\mathbf{A}_p|}{|\{a \in \mathbf{A}_p : w_i \in a\}| + 1}\right). \tag{2}$$

At the beginning, we compute domain weight and event weight for all words in $\mathbf{A}_p$. Namely, both *domain words* set and *event words* set are equal to set $\mathbf{A}_p$. MAD algorithm [30] is adopted to decide thresholds that can remove trivial words. After applying the hard threshold filtering, only words with values (domain weight or event weight) bigger than the thresholds are kept in the corresponding set. Taking "domain words" for example, domain weight threshold η_c can be calculated as follows.

$$\delta_c = median(|f(w, \mathbf{A}_p) : \forall w \in \mathbf{A}_p|), \tag{3}$$

$$\eta_c = \delta_c + \alpha_c \times median(|f(w, \mathbf{A}_p) - \delta_c, \forall w \in \mathbf{A}_p\ |). \tag{4}$$

As shown in Eq. 4, parameter α_c determines the value of threshold η_c. When α_c is set to be too small (e.g., 0.1), trivial words such as "yesterday", "adult", and "down" are selected as domain words. Oppositely, a large value of α_c will remove important words. As suggested by Leys et al. [12], value of α_c can be set as $1/Q(0.75)$, where $Q(0.75)$ is the 0.75 quantile of the distribution. Therefore, we set α_c to be 3.97 ($\eta_c = 0.087$), which returns a medium-size domain word set that contains 52 words. Similarly, threshold δ_e computed by the MAD algorithm is to remove trivial words from the *event words* set.

The *domain words* and *event words* that have been extracted from news reports can now be used as queries to search Twitter data. Only tweets containing at least one *domain word* or one *event word* are retrieved and sent to the next module, *relevancy ranking*.

4.2 Relevancy ranking

The *relevancy ranking* module evaluates the relevancy between tweets and events. To compute this, "total" relevancy is factorized as a product of three similarity subfactors: textual, spatial, and temporal similarity.

4.2.1 Textual similarity

As shown in Eq. 5, the textual similarity $\phi_{x,y}$ between event x and tweet y is defined as the product of tweet words' *domain weight* sum and *event weight* sum:

$$\phi_{x,y} = \sum_{w_i \in (d_y \cap \mathbf{W_C^{(P)}})} C(w_i, p) \times \sum_{w_i \in (d_y \cap \mathbf{W_E^{(x)}})} E(w_i, x), \tag{5}$$

where d_y is the context of tweet y. Only words in the *domain word* set $\mathbf{W_C^{(P)}}$ are considered when calculating the *domain weight* sum, and only words in the *event word* set $\mathbf{W_E^{(x)}}$ of event x are considered when computing the *event weight* sum. The rationale behind the formula is as follows.

- Sum of domain/event word weights. A tweet is more likely to be event-related, if it contains more *domain words* and *event words*. To accumulate effects of individual words, both the first and the second term in Eq. 5 take the form of word weight sum.
- product of weight sums. Only tweets containing both *domain words* and *event words* in a sufficient way are qualified to be event-related. A tweet with many *domain words* but few *event words* may discuss other events in the same domain. While a tweet with many *event words* (e.g., event location name) but few *domain words* may relate to events in other domains (e.g., something that also happened in the same location). To balance the effects of *domain words* and *event words*, Eq. 5 multiply domain weight sum with event weight sum.

4.2.2 Spatial similarity

The spatial similarity between event x and tweet y is decided by two factors: 1) the distance between tweet location l_y and event occurrence location l_x, and 2) the spatial influence scope of tweet y. The first factor is to relate event and tweet in the same location. An event and a tweet are more likely to be relevant if they are close in distance. The second factor further enhances the event-relevancy for tweets with high textual-similarity scores. Intuitively, within the same distance to event occurrence location, tweets of higher textual-similarity scores are more likely to be event-related. Therefore, a tweet y's spatial influence for event x is modeled as a Gaussian distribution $\varphi_{x,y} = \mathrm{N}(l_y, \sum_{x,y})$, centered at tweet y's location l_y, with influence scope $\sum_{x,y} = \begin{pmatrix} \phi_{x,y} & 0 \\ 0 & \phi_{x,y} \end{pmatrix}$, where $\phi_{x,y}$ is the textual similarity defined in Eq. 5.

4.2.3 Temporal similarity

After the initial burst of tweets upon the occurrence of a particular event, the number of event-related tweets usually decreases as a Poisson process [26]. In other words, the possibility of tweet y being related to event x decreases as time goes by, which indicates the likelihood that an individual tweet related to the event also decreases following a Poisson process. Therefore, temporal similarity between tweet y and event x can be described as an exponential distribution:

$$\rho_{x,y} = \lambda e^{-\lambda |t_x - t_y|}, \tag{6}$$

where t_x is the occurrence time of event x and t_y is the publishing time of tweet y.

By integrating the textual, spatial, and temporal similarities, the event-tweet relevancy $\psi_{x,y}$ is ranked by the following function:

$$\psi_{x,y} = \phi_{x,y} \cdot \varphi_{x,y} \cdot \rho_{x,y}. \tag{7}$$

For a tweet y, we choose event x^* that maximizes event-tweet relevancy $\psi_{x,y}$ as its most correlated event:

$$x^* = g(y) = \arg\max_{x \in \mathbf{X}_p} \psi_{x,y}. \tag{8}$$

Correspondingly, for each event x, its related tweet set $\tilde{\mathbf{Y}}^{(x)}$ is identified through an inverse process, that each element tweet $y^{(x)}$ in set $\tilde{\mathbf{Y}}^{(x)}$ satisfies $g(y^{(x)}) = x$.

4.2.4 Label refinement

The initial event-tweet pairs obtained using the procedure outlined above contain a great deal of noisy data. Although top ranked tweets are indeed highly related to the corresponding events (*positive examples*), many of the low ranked tweets are in fact irrelevant (*negative examples*). However, it is difficult to set a uniform threshold suitable for all events to separate the positive and negative tweets. One alternative is to cluster tweets based on their similarities, by assuming that positive examples are more similar to each other than negative ones. Suppose we have a set of *positive tweets* and a set of *negative examples*. Based on these existing label sets, the labels of other tweets can be inferred based on their similarities. However, the assumed positive and negative sets of existing labels are actually unknown. This turns out to be an inference dependent problem: the inference of a single tweet's label depends on the existing positive and negative sets, while constructing positive and negative sets depends on the assignment of each tweet. Therefore, an EM-based inference algorithm is developed and applied here to solve the "inference dependency" problem.

For an event-tweets pair $(x, \tilde{\mathbf{Y}}^{(x)})$, each tweet $y_j^{(x)} \in \tilde{\mathbf{Y}}^{(x)}$ is represented by a n-dimensional feature vector $v_j^{(x)}$, where n is the total number of words in the event related tweets set $\tilde{\mathbf{Y}}^{(x)}$. An element $v_{jw}^{(x)} \in v_j^{(x)}$ is set to be h, if word $w \in \tilde{\mathbf{Y}}^{(x)}$ appears h times in tweet $y_j^{(x)}$.

The tweets' relevancy distribution is modeled as Q-Gaussian mixtures, in which the qth Gaussian is denoted as $\mathbf{G}_q = N(\mu_q, \Sigma_q)$ with mixing coefficient θ_q. The goal is to maximize the likelihood function:

$$p(\tilde{\mathbf{Y}}^{(x)}) = \prod_{j=1}^{n} p\left(v_j^{(x)}\right) = \prod_{j=1}^{n} \sum_{q=1}^{Q} \theta_q \cdot \mathrm{N}\left(v_j^{(x)} | \mu_q, \Sigma_q\right). \tag{9}$$

E-step In the E-step, given the estimates of parameters μ and Σ, the probability of $v_j^{(x)}$ belonging to Gaussian $\mathbf{G_q}$ is calculated as follows:

$$p\left(\mathbf{G}_q|v_j^{(x)}\right) = \frac{\theta_q \mathrm{N}\left(v_j^{(x)}|\mu_q, \Sigma_q\right)}{\sum_{m=1}^{Q} \theta_m \cdot \mathrm{N}\left(v_j^{(x)}|\mu_m, \Sigma_m\right)}. \tag{10}$$

M-step In the M-step, by taking partial derivatives of Eq. 9, estimations of parameter are renewed as follows:

$$\mu_q^* = \frac{\sum_{j=1}^{n} p\left(\mathbf{G}_q|v_j^{(x)}\right) v_j^{(x)}}{\sum_{j=1}^{n} p\left(\mathbf{G}_q|v_j^{(x)}\right)}, \tag{11}$$

$$\sum_q^* = \frac{\sum_{j=1}^{n} p\left(\mathbf{G}_q|v_j^{(x)}\right) v_j^{(x)} \left(v_j^{(x)} - \mu_q^*\right)\left(v_j^{(x)} - \mu_q^*\right)^T}{\sum_{j=1}^{n} p\left(\mathbf{G}_q|v_j^{(x)}\right)}, \tag{12}$$

$$\theta_q^* = \frac{\sum_{j=1}^{n} p\left(\mathbf{G}_q|v_j^{(x)}\right)}{n}. \tag{13}$$

First, tweets set $\tilde{\mathbf{Y}}^{(x)}$ is split into Q parts with a descending order based on the initial Gaussian mixtures. Then E-step and M-step are conducted iteratively. When convergence is achieved, the Gaussian group with the maximum value of relevancy score is selected as the positive examples set $\mathbf{Y}^{(x)}$, while tweets in other Gaussians are treated as negative examples $\bar{\mathbf{Y}}^{(x)}$. This accomplishes **Task 1** *label generation*, as for each event $x_i \in \mathbf{X}_p$, label $z_i = (x_i, \mathbf{Y}^{(x_i)}, \bar{\mathbf{Y}}^{(x_i)})$ can be generated from the historical Twitter stream $\mathbf{Y}$ through the above process.

5 Spatiotemporal event detection

In this section, we discuss the detection of newly emerging spatiotemporal events from real-time Twitter data streams. A tweet classifier is first trained by using historical event-tweets labels generated according to the previous section. Then, tweets of positive class (output of the classifier) are grouped into geo-clusters (events) by applying a multinomial spatial scan method.

5.1 Tweet classifier

Different from traditional text classifier, our proposed tweet classifier consists of two parts: social-ties clustering and mini-tweet-group classification. We first clustered tweets into mini-groups based on their social-ties, and then conduct SVM-based classification to these tweet-mini-groups.

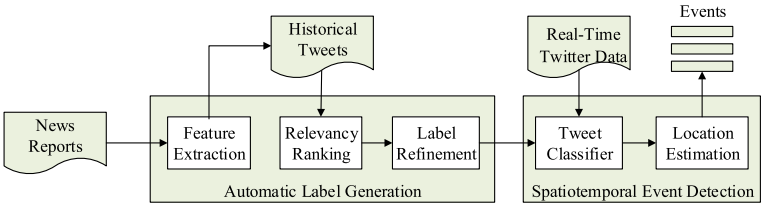

Fig. 3 ATSED system architecture

5.1.1 Social ties clustering

This clustering process is applied to both event-tweets labels (training data) and the incoming Twitter data stream (testing data). The basic idea here is that tweets sharing common social ties (e.g., mentions, replies, and hashtags) are more likely to be about the same topic. To cluster tweets through social ties, a tweet-tie heterogeneous graph is built and then split into small subgraphs by applying graph partition.

As shown in Fig. 3, tweets are connected by social ties to create a tweet-tie heterogeneous graph $\Lambda = (\mathbf{Y}, \mathbf{E})$. $\mathbf{Y}$ is the tweet set, denoted as small nodes in Fig. 4, and E is the edge set, where each edge e_{ij} is the number of shared social-ties between tweet y_i and y_j.

Our goal is to partition the entire graph Λ into a set of subgraphs $\mathbf{P}$ such that connections are strong within one subgraph yet are weak across different subgraphs. The modularity of such partitioning is defined as [15]:

$$\mathbf{M} = \frac{1}{2\sum_i k_i} \sum_{i,j} \left(e_{ij} - \frac{k_i k_j}{\sum_i k_i} \right) p_i p_j, \tag{14}$$

where k_i is the degree of tweet (node) y_i and p_i is the index of the subgraph. To partition graph Λ is equivalent to maximize M. In fact, $\mathbf{M}$ can be rewritten in the form of a modularity matrix $\mathbf{B}$'s eigenvalue β_i and the corresponding eigenvector u_i:

$$B_{ij} = e_{ij} - \frac{k_i k_j}{\sum_i k_i}, \tag{15}$$

$$M = \sum_i \left(u_i^T \mathbf{P} \right)^2 \beta_i. \tag{16}$$

Therefore, maximizing $\mathbf{M}$ is approximated by calculating $\mathbf{B}$'s largest eigenvalue β_1 and the corresponding eigenvector u_1. In this way, graph Λ is split into two subgraphs based on the signs of the elements in the first eigenvector u_1. This process is repeated until $\mathbf{M}$ can

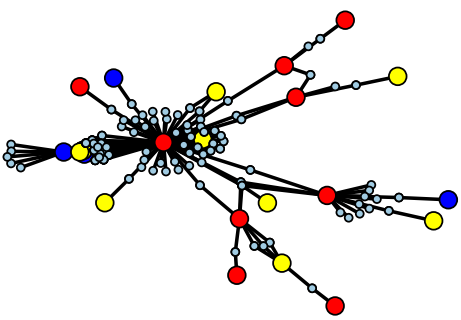

Fig. 4 Example of tweet-tie heterogeneous graph. *Big* nodes represent social-ties: *red* nodes are hashtags, *blue* nodes are mentions, and *yellow* nodes are retweets. *Small* nodes denote tweets

no longer be increased by further divisions. Each resulting subgraph corresponds to a tweet subset $\mathbf{Y}_j$ of the original tweet set $\mathbf{Y}$ ($\mathbf{Y} = \bigcup_j \mathbf{Y}_j$), which is referred as a mini-tweet-group.

5.1.2 Mini-tweet-group classification

Given event-tweets labels that are in a specific domain p, $\mathbf{Z}_p = \{(x, \mathbf{Y}^{(x)}, \bar{\mathbf{Y}}^{(x)}) | x \in \mathbf{X}_p\}$, a classifier is trained based on a support vector machine. An essential step in this training is feature selection. First, rare words that appear less frequently than some threshold value ζ (calculated from historical statistics) are filtered out, unless they are hashtags, mentions or links. Second, common words such as "love" and "people" should also be removed from the feature set. Although these words are frequently mentioned in positive tweets, they are also likely to be more frequent in total Twitter space.

$$\tau_w = \frac{n_w}{n} / \frac{N_w}{N}. \tag{17}$$

In Eq. 17, n_w and n denotes the appearance times of word w and the total number of words in the positive tweets set, respectively, while N_w and N represent the occurrences of word w and count of all words in the entire tweet space. τ_w of trivial words such as "love" and "people" are bigger than one. Thus, considering both frequency threshold and feature score, feature set $\mathbf{W}_F$ can be denoted as:

$$\mathbf{W}_F = \{w | \forall w \in \cup_i \mathbf{Y}^{(x_i)}, \tau_w(t) > 1, n_w > \zeta\}. \tag{18}$$

The feature vector π_j of mini-tweet-group $\mathbf{Y}_j$ is a $|\mathbf{W}_F|$-dimensional vector, and each element π_{jk} in π_j is defined as:

$$\pi_{jk} = \begin{cases} 1, & if\ w_k \in \mathbf{Y}_j,\ w_k \in \mathbf{W}_F, \\ 0, & if\ w_k \notin \mathbf{Y}_j,\ w_k \in \mathbf{W}_F. \end{cases} \tag{19}$$

In the training process, social ties clustering is first applied to historical labels $\mathbf{Z}_p$. For each event x_i, clustering on positive set $\mathbf{Y}^{(x_i)}$ and negative set $\bar{\mathbf{Y}}^{(x_i)}$ is conducted separately. Then, if a mini-tweet-group $\mathbf{Y}_j$ is in positive example set $\mathbf{Y}^{(x_i)}$, its classifier indicator s_j is set to be 1; if $\mathbf{Y}_j$ is within negative examples set $\bar{\mathbf{Y}}^{(x_i)}$, then $s_j = 0$. Our goal for training is to minimize the objective in Eq. 20 to obtain optimal values for weight ω [8], where $C > 0$ is a penalty parameter:

$$\min_{\omega} \frac{1}{2}\omega^T\omega + C \sum_j (\max(0, 1 - s_j \omega^T \pi_j)). \tag{20}$$

For the next step in the testing process, the trained classifier is applied to classify mini-tweet-groups from the real-time Twitter data stream $\mathbf{Y}'$. Specifically, a mini-tweet-group $\mathbf{Y}'_j$ is predicted to be positive if $\tilde{\omega}^T \pi'_j > 0$, and negative otherwise, where $\tilde{\omega}$ is the optimal solution of Eq. 20. Finally, all tweets in the positive class are merged into a domain-related tweet set, denoted by $\mathbf{I}_p$.

5.2 Event location estimation

From the above sections, tweets in the targeted domain (the event-related tweet set $\mathbf{I}_p$) may contain discussions about several different events. The next step is to apply location estimation technologies on targeted-domain tweets to distinguish different events happening during the same time period.

One tweet may contain multiple location indicators, such as the geo-tags generated by GPS, location mentions in the content, and user pre-given locations in the profile. To make the best use of all the location information, a multinomial spatial-scan method is proposed to detect significant spatial clusters, treating each tweet's location as a multinomial variable. Suppose there are K cities in one country, then location $\tilde{\beta}_y$ of tweet y can be represented by a K-dimensional vector $(\beta_1, ..., \beta_K)$, where $\sum_{k=1}^{K} \beta_k = 1$, and $\beta_k \geq 0$. Element β_{yk} in vector $\tilde{\beta}_y$ denotes the probability that tweet y is related to city k. For each tweet y, a location weight vector $\tilde{\beta}_y$ can be computed through the following process.

1. Extract the initial geo-location vector $\tilde{\mathbf{h}}_y$. Each element h_{yi} in $\tilde{\mathbf{h}}_y$ is a longitude-latitude (coordinates) pair (u_{yi}, v_{yi}) converted from the geo-terms contained in the original tweet, such as profile locations, geo-tags, and location mentions. The length of vector $\tilde{\mathbf{h}}_y$ is decided by the number of geo-terms in the original tweet y.
2. Construct the city-level location vector $\mathbf{G}_y$. For a city k, its spatial scope is represented as pair $(\mathbf{U}_k, \mathbf{V}_k)$, where $\mathbf{U}_k$ is a longitude region (u_{k1}, u_{k2}) and $\mathbf{V}_k$ is a latitude region (v_{k1}, v_{k2}). City k covers a geo-term $h_{yi} = (u_{yi}, v_{yi})$ in $\tilde{\mathbf{h}}_y$, if $u_{yi} \in \mathbf{U}_k$ and $v_{yi} \in \mathbf{V}_k$. The value of g_{yk} is therefore decided by the number of geo-terms city k covers.
3. Calculate the location weight vector $\tilde{\beta}_y$. Given the city-level location vector $\mathbf{G}_y$, element $\beta_{yk} \in \tilde{\beta}_y$ is then calculated as $\beta_{yk} = g_{yk} / \sum_{k=1}^{K} g_{yk}$.

Given a real-time Twitter stream $\mathbf{Y}$ and event-related tweets set $\mathbf{I}_p$, we can now aggregate the count of event-related tweets at the city-level and apply a fast subset scan [18] to identify a set $\Omega = \{L_1, ..., L_H\}$, that contains H candidate city clusters with Kulldorff's statistics [9]:

$$K_r = (C_{\mathbf{A}} - C_{\mathbf{R}}) \lg \left(\frac{C_{\mathbf{A}} - C_{\mathbf{R}}}{B_{\mathbf{A}} - B_{\mathbf{R}}} \right) + C_{\mathbf{R}} \lg \left(\frac{C_{\mathbf{R}}}{B_{\mathbf{R}}} \right) - C_{\mathbf{A}} \lg \left(\frac{C_{\mathbf{A}}}{B_{\mathbf{A}}} \right). \tag{21}$$

In Eq. 21, $C_{\mathbf{A}}$ and $B_{\mathbf{A}}$ refer to the total count and base in the country, respectively, where set $\mathbf{A}$ contains all cities in the country. $C_{\mathbf{A}}$ is computed via the event-related tweets set $\mathbf{I}_p$ such that $C_{\mathbf{A}} = \sum_{m} \sum_{k \in \mathbf{A}} \beta_k$, where k is a city in country $\mathbf{A}$ and m is the number of tweets in set $\mathbf{I}_p$. Correspondingly, the country-level base $B_{\mathbf{A}}$ is calculated through Twitter stream $\mathbf{Y}$ that $B_{\mathbf{A}} = \sum_{n} \sum_{k \in \mathbf{A}} \beta_k$, where k is a city in country $\mathbf{A}$ and n is the number of tweets in set $\mathbf{Y}$. Similarly, $C_{\mathbf{R}}$ and $B_{\mathbf{R}}$ refer to the count and base in the spatial region $\mathbf{R}$, which is a set of neighboring cities. $C_{\mathbf{R}}$ is then calculated using the targeted-domain tweet set $\mathbf{I}_p$ such that $C_{\mathbf{R}} = \sum_{m} \sum_{k \in \mathbf{R}} \beta_k$, and $B_{\mathbf{R}}$ is calculated using the original tweets set $\mathbf{Y}$ with $B_{\mathbf{R}} = \sum_{n} \sum_{k \in \mathbf{R}} \beta_k$. To reduce the computational cost, we only consider regions with a count $C_{\mathbf{R}}$ greater than a specified minimum count C_{min} and a base $B_{\mathbf{R}}$ larger than a specified minimum base number B_{min}.

The above process yields the candidate city cluster set Ω. Randomization testing is then conducted on Ω to obtain the significant cluster subset $\Omega' = \{L'_1, ..., L'_h\}$ of Ω $(h \leq H)$. Empirically, parameter H is usually set to be greater than the maximum number of potential clusters that may exist, and the insignificant clusters are filtered out later by randomization testing. Only those clusters with empirical p-values smaller than a given threshold P_v (e.g., 0.05) are retained in the result subset Ω'.

Finally, each element $L'_i \in \Omega'$ is converted into an event x'_i, which is the eventual output of the *event detection* module. Specifically, location cluster L'_i can be represented as a

location-tweets pair $(\mathbf{R}'_i, \mathbf{I}^{(i)})$, where $\mathbf{R}'_i$ is a set of neighboring cities and $\mathbf{I}^{(i)}$ is the corresponding tweet set. Finally, as the solution to **Task 2**, the earliest timestamp of tweet in $\mathbf{I}^{(i)}$ is used as the event date $t_{x'_i}$, the center coordinates of $\mathbf{R}'_i$ is extracted as event location $l_{x'_i}$, and tweet set $\mathbf{I}^{(i)}$ is treated as event-related tweet set $\mathbf{I}^{(x'_i)}$.

6 Results

In this section, we first introduce the datasets used for evaluation, and then compare ATSED with five existing algorithms. Next, the effectiveness of each component in ATSED is validated. Finally, two case studies from ATSED output are discussed. All experiments were performed on a computer with one 3.20 GHz Intel Xeon CPU and 18.0 GB RAM.

6.1 Datasets and evaluation metrics

Two datasets are used in our experiments, one is Twitter dataset and the other is GSR dataset. Both of them consist of data from July 2012 to May 2013 of 10 countries in Latin America. These datasets were separated into two parts: 1) Data from July 2012 to December 2012 were utilized as the label generation data source for ATSED and as the training set for the supervised comparison methods, and 2) Data for January 2013 to May 2013 were used as the testing set for validating all the methods.

The Twitter dataset was collected through Twitter API.[5] Tweets' contexts were stemmed and stop-words were removed. Location terms were extracted from the original Twitter data, including GPS geo-tags, location mentions, and user profile locations. Twitter locations used in *label generation* module are inferred location, with the priority as: location mentions > GPS geo-tags > user profile locations. While *spatiotemporal event detection* module can use all these location information to estimate the location of detected events. In total, 305 million tweets were collected.

Detection results were validated against a labeled events set named "Gold Standard Report" (GSR).[6] Each GSR event consists of date, location, and corresponding news reports. A real world event was selected as a GSR event if it was reported by local news outlets or by influential international media. Table 1 lists the detailed information about events of each country.

Results of all the methods were validated through GSR events. A detected event is regarded as "matching" a GSR event, if it satisfied following two conditions: 1) the event time detected is the same as that recorded in GSR; and 2) the event location detected is within the same city as that recorded in GSR.

Generally, two types of metrics are used in our evaluation: relevance and timeliness metrics. Specifically, relevance metrics include precision, recall, and F-score: "precision" quantifies the fraction of detected events that are matches to GSR events, "recall" quantifies the percentage of GSR events that are correctly detected, "F-score" represents the harmonic mean of precision and recall. Timeliness metric "lead time" measures the delays between event time reported by Twitter event detection methods and the earliest publish date of news media. A positive value of "lead time" means detected event comes earlier than news, while

[5] https://dev.twitter.com/rest/public.

[6] http://www.mitre.org/.

Table 1 Distribution of events in 10 Latin countries

Country	News source[7]	#Training events	#Testing events
Argentina	Clarín; La Nación; Infobae	365	318
Brazil	O Globo; O Estado de São Paulo; Jornal do Brasil	451	361
Chile	La Tercera; Las Últimas Notícias; El Mercurio	252	229
Colombia	El Espectador; El Tiempo; El Colombiano	298	213
Ecuador	El Universo; El Comercio; Hoy	275	123
El Salvador	El Diáro de Hoy; La Prensa Gráfica; El Mundo	180	127
Mexico	La Jornada; Reforma; Milenio	1217	811
Paraguay	ABC Color; Ultima Hora; La Nacíon	563	387
Uruguay	El País; El Observador	124	104
Venezuela	El Universal; El Nacional; Ultimas Notícias	678	557

"News source" shows the news agencies utilized as sources for the GSR dataset

a negative values denotes this event is first reported by news media rather than Twitter streams.

6.2 Methods for comparison

We compared ATSED with 5 popular event detection methods, including two supervised algorithms, Earthquake Detection [26] and TEDAS [14], and three unsupervised methods, Topic Modeling [33], Graph Partition [31], and Spatial Temporal Burst [11]. Detailed experimental settings for these methods were as follows.

- **Earthquake Detection** [26]: This work designed a SVM classifier to distinguish earthquake-related tweets for event detection. Three features are mentioned in the paper for classification training: statistical, keyword, and word context. All three features were test in our evaluation, and keyword feature was chosen for its best performance (measured in F-value).
- **TEDAS** [14]: TEDAS is another supervised event detection system based on SVM. There are two pairs of tunable parameters (α, β) and (α', β') in this paper, which are priors to punish words with low frequencies. The well recommended settings $\beta = \beta' = 10$ provided by the authors were followed in our experiments. Due to the low percentage of civil unrest content, α and α' were assigned with a small value 0.1 to capture the sparse data.
- **Topic Modeling** [33]: The implementation code applied here was provided by the authors. Hashtags were treated as tags and tweet geotags were deemed to be the corresponding geographic regions.
- **Graph Partition** [31]: The authors employed MAD algorithm [12] to deal with the skewness of the signal strength distribution. In our experiment, various settings for the MAD threshold (1, 5, 10, 20, 30, 40) were evaluated and a value of 20 is chosen as it produced the best performance.

[7] In addition to domestic Top 3 news outlets, the following global news outlets are also included: The New York Times; The Guardian, The Wall Street Journal, The Washington Post, The International Herald Tribune, The Times of London, Infolatam.

– **Spatiotemporal Burst** [11]: The implementation code was provided by the authors.[8] For our experiment, *domain words* were used as the input queries for the spatiotemporal search engine. The tunable temporal window size was set to 6 as recommended in the original work. We also evaluated other values, including 12 and 24, but observed similar results.

We created a manual label set, which was used as training data for the two supervised comparison methods (Earthquake and TEDAS). Tweets that were definitely related to "civil unrest" were picked up as positive, for example "With protests in the Zocalo, # YoSoyCan26 requires Iztapalapa dogs to be free", and those containing some keywords but that were definitely irrelevant to "civil unrest" were deemed to be negative, such as "Measures should be taken to protest trees against winter damage". To strengthen the quality of the training data set, each tweet was assigned to three different annotators. In total, 11,533 tweets were collected for the training, of which about 46 % were "civil unrest related" (positive examples), and 54 % were non-related (negative examples).

All the comparison methods and baselines returned the event-related tweet content, time, and location. However, in addition to the targeted "civil unrest" events, Topic Modeling and Graph Partition, also returned events that were actually on other topics. In order to ensure a fair comparison, a SVM classifier trained by the manual label set was adopted to identify "civil unrest" events from the general event set.

6.3 Parameter settings

This section gives the settings of all the parameters used in ATSED system.

– Domain weight threshold η_c. In the *feature extraction* module, threshold α_c in Eq. 4 defines the score boundary η_c between important domain words and trivial ones. As suggested by Leys et al. [12], value of α_c can be set as $1/Q(0.75)$, where $Q(0.75)$ is the 0.75 quantile of the distribution. To maintain a balance between word importance and quantity, α_c was set to be 3.97 ($\eta_c = 0.087$), which returned a medium-size domain word set with 52 words. Event weight threshold δ_e can be set in a similar way.
– Temporal coefficient λ. As introduced in the *relevancy ranking* module, Possion parameter λ has a significant impact on temporal similarity. Figure 5 illustrates the fitting process of parameter λ. X-axis denotes the temporal distance of tweet and event, where "0" means tweet publish date and event occurrence date are on the same day. Y-axis shows daily event-related tweets number, normalized by their amount sum. To estimate value of λ, 500 events were sampled and fitted to an exponential distribution. On average, $\lambda = 0.48$ with $R^2 = 0.81$ was chosen as default setting in our experiment.
– Gaussian mixture coefficient Q. As illustrated in Fig. 6, there is a trade-off between average relativity score and positive set size. The left-y-axis denotes the proportion of positive tweets. The right-y-axis is the average event-tweet relativity score of positive tweets. A larger value of parameter Q will produce a smaller positive set, with tweets of higher relativity score. Oppositely, a smaller value of Q will involve more tweets into the positive set, in the cost of relativity score decrease. To balance quantity and quality of positive tweets, we set Q to be 4 in this paper, which is the closet value to the intersection point of these two curves.

[8] http://www.cs.ucr.edu/tlappas/scripts/STBurst.rar.

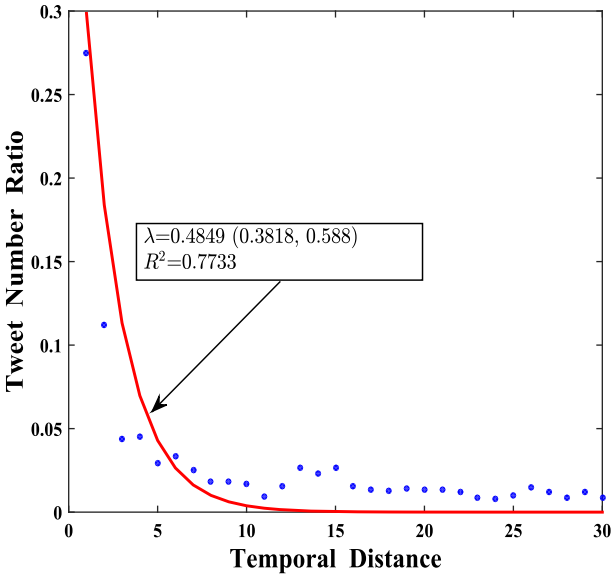

Fig. 5 Temporal decay pattern of event-tweets. *Blue* nodes are actual values, while *red line* denotes the fitted model

– Word frequency threshold ζ. Similar to domain weight threshold η_c, MAD method [12] is used to calculate the value of ζ. Following the principle suggested in [12], ζ is set to be 93 to filter non-trivial words.
– Parameters in location estimation module. There are three tunable parameters that may affect the final performance of location estimation: minimal count C_{min},

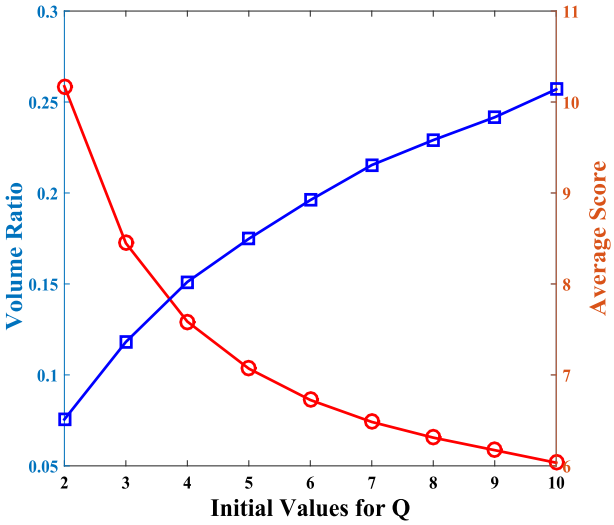

Fig. 6 Trade-off between average relativity score and positive set size for parameter Q

minimal base B_{min}, and p-value P_v. No obvious differences are observed when we change p-value P_v from 0.01 to 0.1 or change minimal base B_{min} from 10 to 50. The key parameter affecting final performance is minimal count C_{min}, which will be discussed in section "Evaluation of the Extended Spatial Scan".

6.4 Performance analysis

In this part, we first compared the overall performance of ATSED with 5 previous methods, and then separately evaluated the effectiveness of each component in ATSED.

6.4.1 Overall relevance evaluation

ATSED was compared with five existing methods, and results are listed in Table 2. ATSED achieved the best overall performance that it obtains the highest F-score in 7 out of the 10 countries. TEDAS was the second best method, achieving the highest F-score in the remaining 3 countries. The performance of the supervised learning method Earthquake was comparable with that of ATSED in precision, but failed to match ATSED in recall and F-score. Spatial Temporal Burst performed relatively well in large countries such as Brazil, but poorly in small countries like Uruguay. Graph Partition and Topic Modeling yielded the worst overall results, which suggests that, even with SVM-based content filtering, unsupervised methods designed for detecting general topics are still insufficient for detecting events in targeted domains.

In general, the two supervised methods (Earthquake and TEDAS) and ATSED performed generally better than any of the unsupervised methods (Graph Partition, Topic Modeling, and Spatial Temporal Burst). To achieve further analysis, Fig. 5 compares the temporal performance for the two supervised methods (Earthquake, TEDAS) and ATSED. Three observations were made based on the data reported in Table 2 and Fig. 7.

1. Overall performance. Both Table 2 (spatial comparison) and Fig. 7 (temporal comparison) indicate that, ATSED, a semi-supervised approach was able to achieve comparable precision to that of the supervised systems using manual labels, and outperformed them with much better recall and F-score.
2. Spatial Performance. ATSED performed stably in different countries, while Earthquake and TEDAS clearly functioned unstably across different countries. Although TEDAS worked better than ATSED in small countries such as Paraguay and Uruguay, it falled short in large countries like Mexico and Venezuela, which generate more than 32 % of the total Twitter data in Latin America.
3. Temporal Performance. ATSED also yielded a stable temporal performance, while Earhquake and TEDAS fluctuated over different time periods. For the February data, Earthquake and TEDAS both suffered sharp decreases in recall and F-score, but ATSED maintained good performance in all three metrics.

In summary, ATSED outperformed all of the other methods in both effectiveness and robustness, clearly demonstrating its ability to yield better results and work more stably across various countries and time periods. Several reasons may account for ATSED's excellent performance. First, our use of automatically generated labels may contribute to the superior overall performance as they enable ATSED to generate a large amount of high-quality labels for countries with different languages, while it is hard to collect sufficient labels with equivalent diversity manually. Second, far beyond the traditional text-based

Table 2 Spatial performance comparison among Twitter event detection methods (Precision, Recall, F-score)

Dataset	ATSED	Graph partition	Earthquake	Topic modeling	TEDAS	ST burst
Brazil	0.48, 0.85, **0.61**	0.55, 0.34, 0.42	0.65, 0.19, 0.30	0.46, 0.09, 0.15	0.39, 0.20, 0.27	0.80, 0.45, 0.58
Colombia	0.80, 0.92, **0.86**	0.68, 0.29, 0.41	0.55, 0.49, 0.52	0.26, 0.39, 0.31	0.66, 0.41, 0.50	0.87, 0.48, 0.62
Uruguay	0.53, 0.34, 0.41	0.28, 0.23, 0.25	0.86, 0.11, 0.20	0.22, 0.06, 0.09	0.88, 0.56, **0.68**	0.11, 0.06, 0.08
El Salvador	0.64, 0.62, **0.63**	0.35, 0.07, 0.1	0.32, 0.06, 0.10	0.40, 0.05, 0.09	0.71, 0.36, 0.48	0.30, 0.12, 0.17
Mexico	0.69, 0.86, **0.77**	0.72, 0.23, 0.35	0.51, 0.19, 0.28	0.34, 0.08, 0.12	0.56, 0.20, 0.29	0.76, 0.43, 0.55
Chile	0.64, 0.77, **0.70**	0.83, 0.39, 0.53	0.46, 0.19, 0.27	0.42, 0.48, 0.45	0.96, 0.36, 0.53	0.67, 0.69, 0.68
Paraguay	0.50, 0.85, 0.63	0.76, 0.19, 0.30	0.40, 0.10, 0.16	0.86, 0.07, 0.13	0.88, 0.67, **0.76**	0.34, 0.12, 0.18
Argentina	0.57, 0.78, **0.66**	0.88, 0.14, 0.24	0.63, 0.57, 0.60	0.38, 0.42, 0.40	0.51, 0.64, 0.57	0.63, 0.73, 0.67
Venezuela	0.87, 0.86, **0.87**	0.46, 0.21, 0.29	0.87, 0.22, 0.35	0.47, 0.37, 0.41	0.79, 0.28, 0.42	0.82, 0.33, 0.47
Ecuador	0.74, 0.38, 0.50	0.30, 0.22, 0.25	0.78, 0.60, 0.68	0.67, 0.04, 0.08	0.55, 0.92, **0.69**	0.29, 0.26, 0.27

Numbers in bold show the best F-score values in corresponding countries

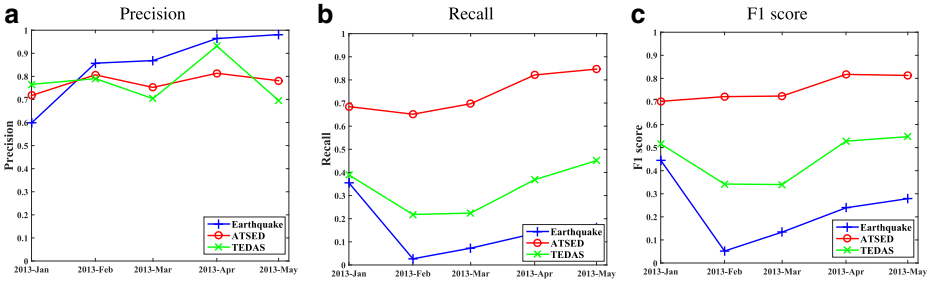

Fig. 7 Temporal performance comparison of ATSED, Earthquake, and TEDAS

classifier, the classifier incorporated in ATSED can have a beneficial effect on the final results as it takes into account the social ties among tweets. Utilizing an extended spatial scan can also enhance ATSED's output by improving quality of the location data. In the following sections, we will further evaluate the effect of each component in ATSED separately.

6.4.2 Timeliness evaluation

To further evaluate how soon the newly emerging events can be detected, Fig. 8 shows the comparison of timeliness metric "lead time" among the three best performers: Earthquake, TEDAS, and our proposed ATSED. In general, our ATSED achieves the best "lead time" of 2.42 days, TEDAS is the second best with 2.34 days ahead of news reports, and Earthquake performs worst with overall "lead time" of 2.04 days.

1. Twitter comes earlier than news. "Civil unrest" events generally appear first in Twitter that even the worst performer Earthquake can detect events 2.04 days prior to the news report. This is because 75 % of "civil unrest" events are planned in advance [17], and social media such as Twitter plays a key role in organizing protests, especially in the

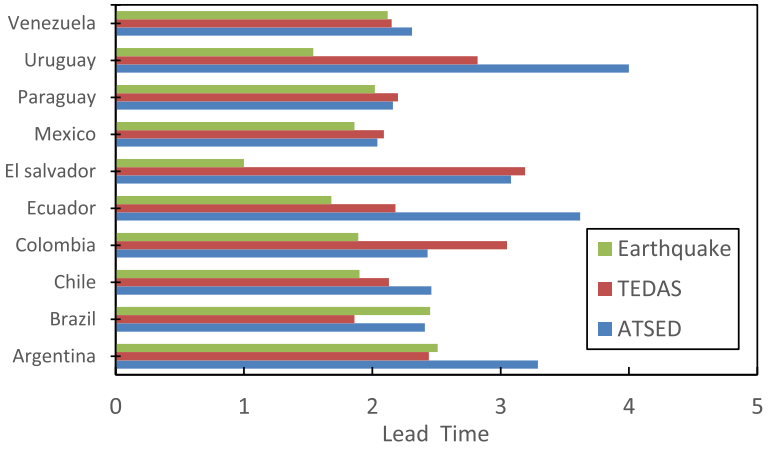

Fig. 8 Lead time comparison of ATSED, earthquake and TEDAS

Table 3 Sample tweets for the baseline method and ATSED

Tweets by baseline	1. Northern Ireland live another **march** day: **Demonstrators protest** since December by a decree ... http://t.co/O2K9hMIq
	2. #EnImágenes Students **protest** in several states against the judgment of the Supreme Court http://t.co/clj5XraS
	3. RT @FilosofiaTipica: People change. Love hurts. Friends **leave**. Things sometimes go wrong. But remember that life goes on.
Positive tweets by ATSED	1. With **protests** in the **Zocalo**, **#YoSoyCan26** requires **government** to free **dogs** of **Iztapalapa**. http://t.co/XPsQ90po#AMLO
	2. **#YoSoyCan26 march** in solidarity with Socket for victims' families in **Cerro de la Estrella** and **demand** liberty for **dogs**.
	3. RT @politicosmex: To people of **Mexico**, **dogs** are murderers is incredulous : **Government** of the capital is asked to clarify the truth ...http://t.co/m5UbmJXT
Negative tweets by ATSED	1. According to reports from the **authorities** in **Iztapalapa**, six **people** are murdered in three offices...http://t.co/qvhCsEhl
	2. Your charger does not work anymore? You have a broken **dog**? **Bring** it to us in Tampico Altamira tree http://t.co/fSOj8U2C
	3. RT @ CristhianH23: **Dogs** killers, untouchable aliens, fair **elections**, less unemployment, peaceful **marches**, united **people** #Mé

Domain words are denoted by bold style and *event words* are marked with underlining. The tweets, originally in Spanish, have been translated into English using Google Translate

early stages.[9] Detecting events from Twitter can provide "beforehand information" for civil unrests, while traditional news media only produce "morning-after" reports.

2. Organized protests come earlier than spontaneous protest. Note that ATSED can obtain better "lead time" in countries such as Uruguay and Argentina than Brazil. We studied Brazilian protests and found that they were more spontaneous compared to other countries: for instance, the initial protests were triggered by bus fare and soon developed into protests against government, most of which were not organized.

6.4.3 Evaluation of label generation

The effectiveness of the automatic label generation (ALG) component was demonstrated through the high quality of the tweet labels. The above mentioned "dog protest" in Mexico was taken as the case study here, as it was a small scale protest that would normally be hard to identify. The top 3 ranked example labels generated by ATSED are listed in Table 3. For comparison, the top ranked tweets retrieved by the keywords matching method [6] are also listed in the table, using words most relevant to "civil unrest", such as "protest" and "march".

From the results shown in Table 3, tweets obtained through the keyword matching baseline method contain following noises.

1. Tweets irrelevant to the given targeted domain. Some tweets were completely unrelated to the topic "civil unrest". Consider Tweet #3 for example. Its original Spanish text was :"La gente cambia. El amor duele. Los Amigos se marchan. Las cosas aveces van

[9] https://goo.gl/8wfhkN.

mal. Pero recuerda que la vida sigue". Although this did contain one civil unrest keyword "marchan" (which becomes "march" after stemming), this tweet was in fact about people's feelings, rather than "civil unrest" events.
2. Tweets irrelevant to the specific event. Within those tweets that were indeed related to "civil unrest", most reflect influential protests that occurred in countries outside Mexico. For example, Tweet #1 was actually about a protest in Northern Ireland, and Tweet #2 mentioned a protest that happened in Venezuela. Small events such as the "dog protests" were submerged in these "big events".

In contrast, the positive tweets retrieved by ATSED were highly related to the "dog protest" event. These tweets can be summarized into two types.

1. Tweets referred to the protest itself. For example, tweets #1 and #2 contained highly ranked "civil unrest" domain words, such as "protesta" (protest) and "marcha" (march), as well as important event words, for example, "perros" (dogs) and "Iztapalapa" (location name).
2. Tweets related to events that triggered the protest. The reason for the protest was not mentioned in the news report, but can be revealed according to Tweet #3: citizens were protesting to gain the freedom of innocent dogs that had been captured by government officials as suspects in the killing of 4 people. Besides the event words, these tweets also contained middle-ranked domain words such as "Gobierno" (government) and "México", which were weak indications for "civil unrest" when appearing alone, but became stronger when they co-occurred in the same tweet.

In addition, as shown in Table 3, ATSED also provided negative examples, which can be generally divided into three types as follows.

1. Low textual score tweets. For example, domain words (authorities and people) and event words (Iztapalapa) contained in Tweet #1 are low weight words and result in a poor textual score.
2. Low spatial score tweets. For instance, Tweet #2 had a relatively high textual score, as it contained the strong event word "perro" and the domain word "Tráelo". However, its spatial score was low because the location it provided was the city of "Tampico", which was about 500 kilometers away from the event location (Mexico City).
3. Low temporal score tweets. Tweet #3 had a strong textual score as it contained both "dogs" and "marches", but a weak temporal score as it was published on Jan 19, one week after the event date.

"Precision@K" is used to quantitatively evaluate the quality of generated labels. It is calculated as the ratio of tweets that truly relevant to the targeted domain "civil unrests" among those top K ranked.

$$Precision@K = \mathbf{D}_T \bigcap \mathbf{D}_{topK} / K \tag{22}$$

where $\mathbf{D}_T$ are the ground truth of positive labels from our manual label set mentioned in Section 6.3, $\mathbf{D}_{topK}$ are top K tweets ranked by methods. Specifically, we selected a mixture label set consisted of 1,000 positive tweets and 5,000 negative tweets, and ranked these tweets through random selection, keyword matching, and our proposed ATSED. The results list in Table 4 show that labels generated through ATSED outperform other methods in almost all stages. ATSED beats other methods because it can assign weights for words with knowledge learned from news. The outputs of keyword matching method are acceptable when K is small (e.g., $K = 50$), however, its performance drops quickly as K increases and tends to the output of random selection in the final stages.

Table 4 Labels quality evaluation through "Precision@K"

	P@50	P@100	P@150	P@200	P@250	P@300	P@350
Random selecting	0.18	0.19	0.16	0.17	0.17	0.15	0.15
Keyword matching	0.63	0.46	0.32	0.25	0.22	0.19	0.18
ATSED	0.84	0.79	0.77	0.74	0.73	0.74	0.71

6.4.4 Evaluation of the tweet classifier

ATSED's tweet classifier was compared with that of two supervised methods, Earthquake and TEDAS. To ensure a fair comparison among the tweet classifier components, labels generated by ATSED were used as training data for both Earthquake and TEDAS. Given the same training dataset, any differences among the three methods will depend mainly on the design of the Twitter text classifier. Table 5 compares the performance achieved by each of the three methods. The data in the table reveal that:

1. Using labels generated by ATSED improved detection performance for both Earthquake and TEDAS. Comparing Tables 2 and 5 reveals that these two methods exhibited obvious increases in recall and F-scores in most countries, accompanied by slight decreases in precision. With respect to the F-score, Earthquake performed better than before in 8 countries, and TEDAS achieved gains in 6 countries. Compared to human analysts, ATSED inevitably produced some noisy labels, which may have been responsible for the small reduction in precision. However, ATSED can easily generate a large amount of relevant labels, which boost both recall and F-score. Creating a manual label set of equivalent size would be extremely expensive and time-consuming.
2. When using the same training data, ATSED outperformed both Earthquake and TEDAS in all ten countries. This observation strongly indicates the effectiveness of our proposed Twitter classifier. Without considering tweets' distinct features (e.g., hashtags, mentions), Earthquake still turned in the worst performance. While both TEDAS and ATSED took into account Twitter terms as additional features for the SVM classifier,

Table 5 Performance comparison for Twitter text classifiers (Precision, Recall, F-score)

Dataset	ATSED	Earthquake	TEDAS
Brazil	0.48, 0.85, **0.61**	0.39, 0.28, 0.32↑	0.70, 0.53, 0.60↑
Colombia	0.80, 0.92, **0.86**	0.29, 0.41, 0.34	0.72, 0.51, 0.60↑
Uruguay	0.53, 0.34, **0.41**	0.52, 0.25, 0.38↑	0.27, 0.44, 0.33
El Salvador	0.64, 0.62, **0.63**	0.45, 0.09, 0.16↑	0.52, 0.58, 0.55↑
Mexico	0.69, 0.86, **0.77**	0.62, 0.36, 0.46↑	0.77, 0.55, 0.64↑
Chile	0.64, 0.77, **0.70**	0.69, 0.71, 0.70↑	0.71, 0.50, 0.59↑
Paraguay	0.50, 0.85, **0.63**	0.46, 0.39, 0.42↑	0.49, 0.79, 0.60
Argentina	0.57, 0.78, **0.66**	0.58, 0.66, 0.62↑	0.42, 0.74, 0.53
Venezuela	0.87, 0.86, **0.87**	0.51, 0.42, 0.46↑	0.80, 0.45, 0.58↑
Ecuador	0.74, 0.38, **0.50**	0.16, 0.44, 0.23	0.16, 0.52, 0.25

Upward arrows denote performance improvements over the original results shown in Table 2. Numbers in bold show the best F-score values for each country

ATSED clustered tweets based on social ties first, which increased the efficiency and precision of the ensuring classification.

6.4.5 Evaluation of the extended spatial scan

The new multinomial spatial scan model was also compared with the original spatial scan method [18]. Three tunable parameters are shared by the two methods: cut off threshold p-value P_v, minimal count number C_{min}, and minimal base number B_{min}. No significant difference was observed between the two models, when adjusting either p-value P_v or the minimal base number B_{min}. To obtain the best performance, we set $P_v = 0.05$ and $B_{min} = 20$ for both the methods. However, ATSED was sensitive to the parameter minimal count number C_{min}. Figure 9 plots the precision and recall of the two methods, when the minimal count number C_{min} is changed from 2 to 10. We can therefore make the following observations.

1. In both these methods, recall decreased with increasing C_{min}. For all C_{min} values, our proposed multinomial spatial scan always achieved better recall than the original model. As C_{min} increased from 2 to 10, little difference was observed in the term of distance between the two recall curves.
2. Increasing C_{min} led to an increase in precision. At the start point ($C_{min} = 2$), the original model had a better precision score. However, our proposed multinomial model obtained a much greater increase rate than the original spatial scan. Therefore, as C_{min} increased, the advantage of the original model narrowed and finally disappeared.
3. After C_{min} reached 6, both the methods became stable and no more changes were observed. In the stable state, with precision close to 1, our multinomial spatial scan model still maintained good recall above 0.5, while original spatial scan only achieved 0.4.

In general, the multinomial spatial-scan contributed better recall with little lost in precision. For the metric of recall, our extended spatial scan consistently provided a clear advantage over the original spatial scan. As for precision, the multinomial spatial scan yielded a comparable precision to that of the original spatial scan when $C_{min} < 6$, and achieved the same precision when $C_{min} \geq 6$.

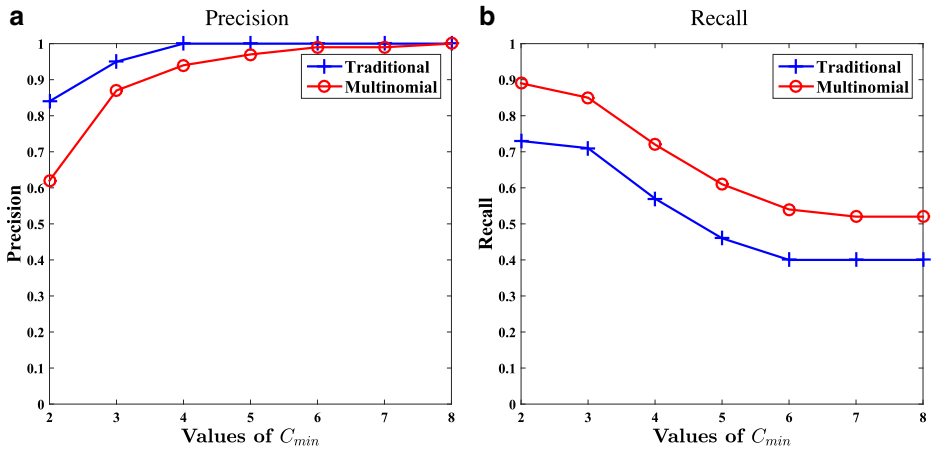

Fig. 9 Comparison of multinomial and original spatial scan performance

6.5 Case study

Several interesting partterns were observed in the ATSED output results. Figure 10 describes three events detected by ATSED on Jan, 20th, 2013 in Mexico. In the figure, each detected event is represented by a location point (red circle), a summary word cloud, and a corresponding ground-truth GSR description. Although these events happened simultaneously in the same country, ATSED successfully distinguished all three and captured their different social focus. As shown in the word cloud, the "Hermosillo" protesters were demanding a reduction in their "vehicle tax", while the event in "Mexico City" was mainly about a "parking" issue, and teachers in "Oaxaca" were marching to protest against "education reform". The cases in Fig. 10 reveals that ATSED can identify spatial events at the city-level, while most previous Twitter event detection technologies can only detect events at the country-level.

Figure 11 plots the trends of 3 popular hashtags found in the detected tweets from the ATSED output. All 3 hashtags were related to "teacher" protests: "#SNTE" was the hottest topic among the "civil unrest" tweets at the beginning of March, but "#CETEG" and "#CNTE" became more popular from April onwards. These data patterns were caused by several interesting facts. The head of the National Union of Education Workers (SNTE) was arrested for corruption on Feb, 28th. The scandal stimulated protests against "SNTE" in the following month, and resulted in the popularity of "#SNTE" in March. As "SNTE"

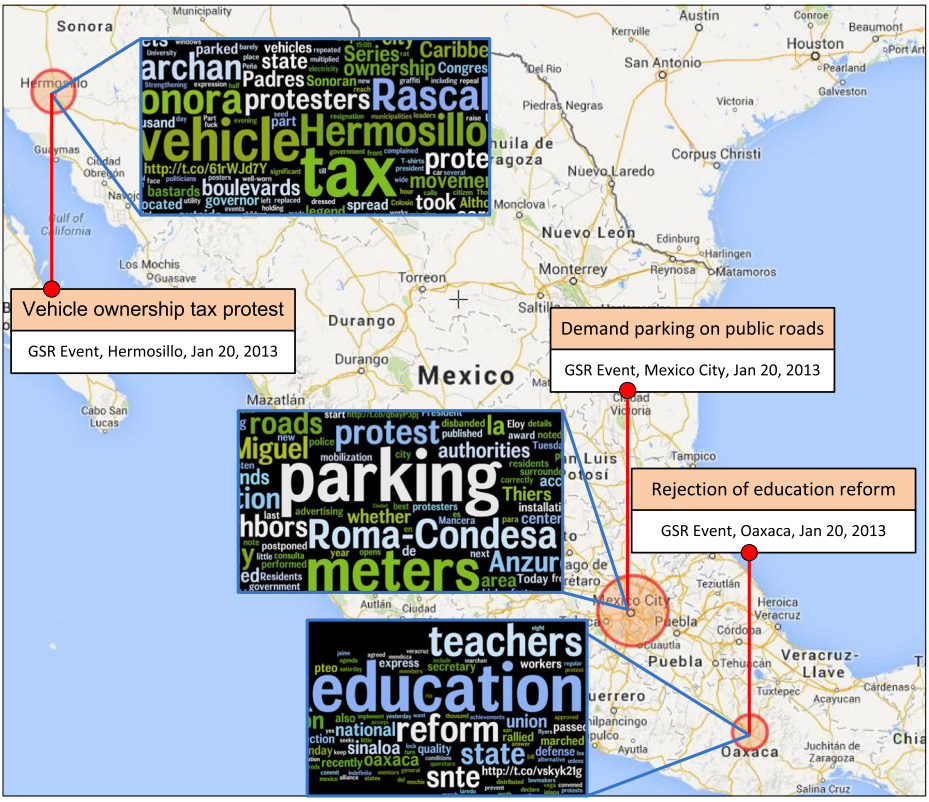

Fig. 10 Case study on spatial factors of ATSED event detection results

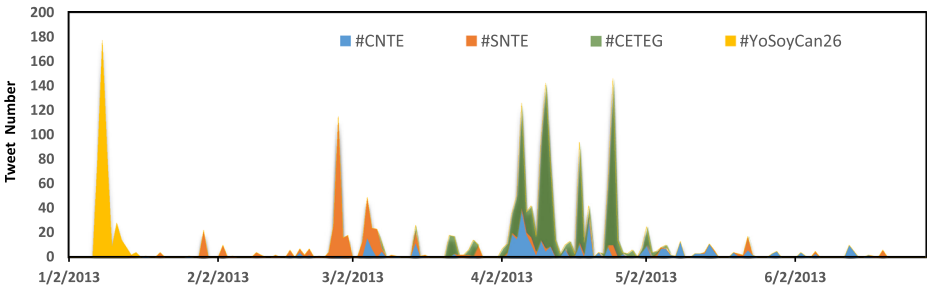

Fig. 11 Case study on temporal factors of ATSED event detection results

suffered from the negative impact of the corruption event, other teacher organizations emerged rapidly and after April, "#SNTE" almost disappeared from the tweet data, being replaced by two new teacher unions, the Guerrero State Coordinator of Education Workers (CETEG) and the National Education Workers Coordinator (CNTE). It requires extensive human efforts to manually relabel training data to keep up with events on the ground in traditional methods, but ATSED is capable of updating its training dataset periodically. Trends in these three hashtags demonstrated ATSED's ability to capture the dynamics of Twitter data.

7 Conclusion

This paper provided a model named ATSED to detect spatiotemporal events of targeted domains from Twitter streams. Beyond the civil unrest events studied in this paper, ATSED can also handle spatiotemporal events of other targeted domains (e.g., sports, politics, environment). Previous Twitter event detection methods usually require manually labeled data for training, instead, ATSED can generate high-quality label data automatically. Based on these labels, a SVM-based classifier that utilizing Twitter social-ties is trained and applied to real-time Twitter streams to recognize event-related tweets. To enhance the estimation accuracy of event locations, all terms of Twitter location information are considered in multinomial spatial scan component of ATSED. The experimental results have shown that ATSED effectively improved detection performance, compared to existing Twitter event detection approaches. And further evaluation demonstrated that each part of ATSED contributes probably to the integral performance.

Acknowledgments Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

1. Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on twitter. In: Proceedings of the 5th international AAAI conference on weblogs and social media. AAAI, pp 438–441

2. Bhattacharya I (2013) Google trends for formulating GIS mapping of disease outbreaks in India. Int J Geoinform 9. Springer
3. Brants T, Chen F, Farahat A (2003) A system for new event detection. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. ACM, pp 330–337
4. Cataldi M, Di Caro L, Schifanella C (2010) Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the 10th international workshop on multimedia data mining. ACM, pp 1–10
5. Fung GPC, Yu JX, Yu PS, Lu H (2005) Parameter free bursty events detection in text streams. In: Proceedings of the 31st international conference on very large data bases. VLDB Endowment, pp 181–192
6. Hu M, Liu S, Wei F, Wu Y, Stasko J, Ma KL (2012) Breaking news on twitter. In: Proceedings of the 21st SIGCHI conference on human factors in computing systems. ACM, pp 2751–2754
7. Jin O, Liu NN, Zhao K, Yu Y, Yang Q (2011) Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on information and knowledge management. ACM, pp 775–784
8. Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European conference on machine learning. Springer, pp 137–142
9. Kulldorff M (1999) Spatial scan statistics: models, calculations, and applications. In: Scan statistics and applications. Springer, pp 303–322
10. Kumaran G, Allan J (2004) Text classification and named entities for new event detection. In: Proceedings of the 27th annual ACM SIGIR conference on research and development in information retrieval. ACM, pp 297–304
11. Lappas T, Vieira MR, Gunopulos D, Tsotras VJ (2012) On the spatiotemporal burstiness of terms. In: Proceedings of the VLDB endowment, vol 5. VLDB Endowment, pp 836–847
12. Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49:764–766. Elsevier
13. Li C, Sun A, Datta A (2012) Twevent: segment-based event detection from tweets. In: Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 155–164
14. Li R, Lei KH, Khadiwala R, Chang KCC (2012) Tedas: a twitter-based event detection and analysis system. In: Proceedings of the 28th international conference on data engineering. IEEE, pp 1273–1276
15. Mark N (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103:8577–8582. National Acad Sciences
16. Min B, Grishman R, Wan L, Wang C, Gondek D (2013) Distant supervision for relation extraction with an incomplete knowledge base. In: HLT-NAACL. ACL, pp 777–782
17. Muthiah S, Huang B, Arredondo J, Mares D, Getoor L, Katz G, Ramakrishnan N (2015) Planned protest modeling in news and social media. In: Proceedings of the 29th AAAI conference on artificial intelligence. AAAI, pp 3920–3927
18. Neill DB (2012) Fast subset scan for spatial pattern detection. J R Stat Soc Ser B (Stat Methodol) 74:337–360. Wiley Online Library
19. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359. IEEE
20. Petrović S., Osborne M, Lavrenko V (2010) Streaming first story detection with application to twitter. In: Proceedings of the 2010 annual conference of the North American chapter of the association for computational linguistics. ACL, pp 181–189
21. Phan XH, Nguyen LM, Horiguchi S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on world wide web. ACM, pp 91–100
22. Popescu AM, Pennacchiotti M, Paranjpe D (2011) Extracting events and event descriptions from twitter. In: Proceedings of the 20th international conference companion on world wide web. ACM, pp 105–106
23. Purver M, Battersby S (2012) Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics. ACL, pp 482–491
24. Ritter A, Clark S, Etzioni O et al (2011) Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. ACL, pp 1524–1534
25. Ritter A, Mausam, Etzioni O, Clark S (2012) Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1104–1112
26. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web. ACM, pp 851–860

27. Settles B (2010) Active learning literature survey, vol 52. University of Wisconsin, Madison, p 11
28. Signorini A, Segre AM, Polgreen PM (2011) The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PloS One 6:e19467. Public Library of Science
29. Tufekci Z, Wilson C (2012) Social media and the decision to participate in political protest: observations from Tahrir Square. J Commun 62:363–379. Wiley Online Library
30. Walker HM (1931) Studies in the history of the statistical method. The Williams and Wilkins Company, pp 24–25
31. Weng J, Lee BS (2011) Event detection in twitter. In: Proceedings of the 5th international AAAI conference on weblogs and social media. AAAI, pp 401–408
32. Wilson C, Dunn A (2011) Digital media in the Egyptian revolution: descriptive analysis from the Tahrir data sets. Int J Commun 5:1248–1272. USC Annenberg Press
33. Yin Z, Cao L, Han J, Zhai C, Huang T (2011) Geographical topic discovery and comparison. In: Proceedings of the 20th international conference on World wide web. ACM, pp 247–256
34. Zhang D, Liu Y, Lawrence RD, Chenthamarakshan V (2011) Transfer latent semantic learning: microblog mining with less supervision. In: Proceedings of the 25th AAAI conference on artificial intelligence. AAAI, pp 561–566
35. Zhao L, Hua T, Lu CT, Chen R (2015) A topic-focused trust model for Twitter. In: Journal of Computer Communications, vol 76. Springer, pp 1–11

Ting Hua is a doctoral candidate in computer science and a research assistant in the Spatial Data Management Lab at Virginia Tech. Contact her at tingh88@vt.edu.

Feng Chen is an assistant professor of computer science at SUNY of Albany.

Liang Zhao is an assistant professor at George Mason University. Contact him at zhaoliangvaio@gmail.com.

Chang-Tien Lu received his MS degree in computer science from the Georgia Institute of Technology in 1996 and PhD degree in computer science from the University of Minnesota in 2001. He is a professor in the Department of Computer Science, Virginia Polytechnic Institute and State University. He served as Program Co-Chair of the 18th IEEE International Conference on Tools with Artificial Intelligence in 2006, and General Co-Chair of the 20th IEEE International Conference on Tools with Artificial Intelligence in 2008 and the 17th ACM International Conference on Advances in Geographic Information Systems in 2009. He also served as Secretary (2008-2011) and then Vice Chair (2011-2014) of the ACM Special Interest Group on Spatial Information (ACM SIGSPATIAL). His research interests include spatial databases, data mining, geographic information systems, and intelligent transportation systems. He is an ACM Distinguished Scientist (2015).

Naren Ramakrishnan, Discovery Analytics column editor, is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of the university Discovery Analytics Center. Contact him at naren@cs.vt.edu.