Leveraging the Power of Informative Users for Local Event Detection

Hengtong Zhang, Fenglong Ma, Yaliang Li, Chao Zhang, Tianqi Wang, Yaqing Wang, Jing Gao, Lu Su
Department of Computer Science and Engineering, SUNY Buffalo, Buffalo, NY USA
Tencent Medical AI Lab, Palo Alto, CA USA
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL USA
{hengtong, fenglong, twang47, yaqingwa, jing, lusu}@buffalo.edu, yaliangli@tencent.com, czhang82@illinois.edu

***Abstract*—Detecting local events (e.g., protests, accidents) in real-time is an important task needed by a wide spectrum of real-world applications. In recent years, with the proliferation of social media platforms, we can access massive geo-tagged social messages, which can serve as a precious resource for timely local event detection. However, existing local event detection methods either suffer from unsatisfactory performances or need intensive annotations. These limitations make existing methods impractical for large-scale applications. Through the analysis of real-world datasets, we found that the informativeness level of social media users, which is neglected by existing work, plays a highly critical role in distilling event-related information from noisy social media contexts. Motivated by this finding, we propose an unsupervised framework, named LEDetect, to estimate the informativeness level of social media users and leverage the power of highly informative users for local event detection. Experiments on a large-scale real-world dataset show that the proposed LEDetect model can improve the performance of event detection compared with the state-of-the-art unsupervised approach. Also, we use case studies to show that the events discovered by the proposed model are of high quality and the extracted highly informative users are reasonable.**

I. Introduction

Local events are unusual activities (e.g., protests, incidents, disasters, sports games, accidents) occurred in a local area within a specific time period [1], which have adverse social and economic impacts on people's everyday life. Thus detecting these local events in real time is of great importance and need for many real-world applications. For example, the timely alarm of emergent disasters or accidents can not only help city residents to plan their trip routes and avoid possible dangers, but also assist local goverments in better allocating rescue resources and law enforcement. Decades ago, however, this task is nearly impossible because the real-time messages from witnesses and reliable sources were not available. While in recent years, with the proliferation of social media platforms like Twitter, Instgram and Facebook, we are able to get access to a considerable amount of social media messages with explicit geo-locations. When a local event occurs, a large number of geo-tagged messages will be posted by the participants as well as witnesses, discussing the situations and their experiences. These messages can provide us with a comprehensive view of the local events and serve as precious first-hand resources for the task of local event detection.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

Fig. 1: Motivating Example. Blue circles denote the messages related to flight delay, while the yellow circles are check-ins and irrelevant uninformative messages.

Despite the importance of real-time detection of local events from social media streams, there are very few efforts [1]–[3] put to solve this problem. The state-of-the-art unsupervised method GeoBurst [1] proposes to cluster social media messages based on keywords and/or geo-distances to identify event candidates (i.e., geo-topic clusters), and then uses heuristic ranking functions to select relative meaningful events based on some hand-crafted features. However, this method achieves unsatisfactory performance, which is far from enough for real-world applications.

The major limitation of the existing work is that they treat all the social media users equally, and fail to take into consideration their abilities of providing informative messages. On social media, some users, though being active in sharing with others, very few of their posts are informative. Some other users, who are silent most of the time, can always provide useful information whenever they talk. In many cases, event-related messages from highly informative users are buried by massive irrelevant ones, and cannot be distilled by existing methods.

To clearly show the limitation of existing work and motivate the insight of this work, we conduct a data analysis on the geo-tagged tweets collected from LaGuardia Airport in the

city of New York. The event is a flight delay incident on June 3 2017. In Figure 1, each circle represents the location of a geo-tagged tweet. Blue circles denote the messages relevant to the flight delay incident, while the yellow circles are check-ins and irrelevant uninformative messages like "I am at LGA" or "I am at Terminal C at LGA". When we employ the aforementioned GeoBurst algorithm [1] on this dataset, the geo-tagged tweets in the airport hall are clusetered as an event candidate (enclosed by a green dashed circle in Figure 1). As can be seen, the few meaningful event-related tweets are buried by a large collection of noisy and irrevelant messages that share a large ratio of common keywords with them. As a result, GeoBurst would likely return uninformative messages instead of the ones that are indeed describing this important local event. Therefore, if we can automatically identify and rely more on these informative social media users in different local districts, we will be able to distill unpopular but meaningful events from massive tweet data. Inspired by this case study, the goal of this paper is to *estimate the informativeness level of social media users from their posts and leverage the power of highly informative users to improve the performance of local event detection*.

Towards this end, we propose LEDetect (**L**ocal **E**vent **Detect**), an user-informativeness-aware framework that enables highly comprehensive local event detection from geo-tagged social media stream. To the best of our knowledge, this is the first work that: (1) recognizes the importance of user informativeness in the task of local event detection; (2) proposes a fully unsupervised model to estimate the informativeness level of social media users and leverage the estimated informativeness scores for local event detection. The basic idea of the proposed LEDetect framework can be summarized as follows. First, it learns low-dimensional embeddings that map all the geo-regions, time, and message keywords into the same space. These embeddings will serve as the knowledge to capture the innate correlations among these heterogeneous entities (i.e. geo-regions, time and keywords). Built upon these multimodal embeddings, we propose an uncertainty-aware optimization model to estimate the informativeness level of users by evaluating their abilities of providing informative and trustworthy information related to local events. At the same time, we propose a self-adaptive density-based clustering algorithm that jointly considers the location and semantic correlations among social media messages to cluster individual messages into meaningful geo-topic event candidates. Finally, we utilize the inferred user informativeness scores and a set of features to distill and summarize meaningful local events from the event candidates. Compared with existing unsupervised local event detection approaches, our proposed method can achieve a much better performance.

Our main contributions are summarized as follows:

- We recognize the importance of user informative level in the task of online local event detection through data analysis on a real-world dataset.
- We propose an unsupervised model, which utilize social media messages to estimate the informativeness level of social media users. This enables the framework to utilize the power of highly informative social media users for more accurate and comprehensive local event detection.
- We conduct extensive experiments on a large-scale geo-tagged tweet dataset. The results shows that the proposed LEDetect framework significantly improves the detection performance, compared with the state-of-the-art unsupervised event detection method.

II. Related Work

Global Event Detection. The goal of global event detection is to extract events that are emerging and significant in the entire social network stream. Based on the methodology, existing work can be classified into two categories. The first line of approaches is document-based approaches [4]–[6]. In this line of work, social media messages are treated as documents and the approaches group documents with similar semantics to event clusters. Then the confident events are extracted based on the semantic coherence within each cluster. Another line of of approaches are feature-based methods [7], [7]–[12]. In this line of work, features like keyword bursty scores are proposed and used to discover events. The above methods are all designed for detecting global events that are bursty in the entire stream. Directly applying these methods to the geo-tagged tweet stream would miss many local events.

Local Event Detection. Compared with global event detection, there are few approaches proposed [1]–[3], [13], [14] for the task of local event detection. Watanabe et al. [13] identify the tweets discussing the same topic within a short time window at a specific place to discover potential events. The authors also propose an automatic tagging method to exploit the information within the tweets without geo-tags. Feng et al. [14] propose a hierarchically clustering spatio-temporal hashtags over the Twitter stream to get potential local events. Abdelhaq et al. [2] first discover bursty words in social media messages and select localized words based on spatial entropy. After that, the approach clusters localized words and ranks the clusters based on hand-crafted features. Zhang et al. [1] propose the state-of-the-art unsupervised local event detection framework GeoBurst. GeoBurst first detects geo-topic clusters by jointly considering the location and semantic correlations among social media messages. After that GeoBurst selects spatio-temporal bursty clusters as local events. The authors of [3] propose a supervised framework TrioVecEvent, which uses multimodal embedding method to represent the keywords, geo-region and time within each social media message. Based on these representations, social media messages are clustered into geo-topic groups. Then TrioVecEvent trains a classifier to find the real local events using human annotation and a set of bursty and concentration features. Compared with this paper, no existing work considers the informativeness of social media users in terms of providing relevant information of local events.

Other Applications. Our work is also related to some other applications. For instance, spatio-temporal topic modeling and storyline generation [15]–[19] employ probabilistic models to

explain the relation among location, time and topics. These methods can also get a distribution of the emerging topics over time. Nevertheless, they can not clearly detect, locate and summarize each event.

III. Methodologies

We start by defining problem of the local event detection.

Definition 1 (Geo-Tagged Social Media Message): Geo-tagged social media messages are generated by users on social media platforms with geo-location tags. Each message d is represented as a tuple $(t_d, l_d, \mathbf{x}_d, u)$, where t_d is its post time, l_d stands for its geo-coordinates, $\mathbf{x}_d$ is a bag of message keywords, and u denotes the provider of d.

Definition 2 (User Informative Score): User informative scores are denoted as $R = \{r_1, r_2, \cdots, r_U\}$, in which r_u is the score of the u-th source. In this paper, r_u characterizes user u's ability of providing trustworthy information related to meaningful local events.

Definition 3 (Local Event Detection): Let $D = (d_1, ..., d_n)$ be a continuous stream of geo-tagged social media messages. Consider a time window $Q = [t_s, t_e]$ where t_s and t_e are the start and end timestamps. The goal of this paper is to:

- Extract all the local events that occur during Q;
- Estimate the informativeness scores of each user u.

A. Framework Overview

The proposed framework is shown in Figure 2. We first learn low-dimension embeddings for geo-locations, time, and keywords using massive geo-tagged social media data. After that, we develop an efficient user informativeness evaluation method to identify highly informative users by analyzing their messages. Specifically, we divide the geo-tagged social media stream into spatio-temporal units. The proposed model takes the embeddings of all the social messages in a single unit as inputs, and uses an optimization-based model to estimate user informativeness in terms of the ability of providing trustworthy information for local event detection.

At the same time, using the learned multimodal embeddings, we develop a self-adaptive density-based clustering model. The model jointly considers geographical locations and semantic embeddings to extract coherent geo-topic clusters (i.e., event candidates) within each query window. In addition, the clustering method is able to adjust the clustering criteria based on the sample density, which is suitable for constantly changing social network contexts.

Finally, we discover local events from event candidates according to:

- social media messages from highly informative users within each candidate clusters;
- a set of features (e.g. semantic/spatio concentration and bustiness).

B. Multimodal Embedding

The multimodal embedding module aims to project all the spatial, temporal, and textual units within the social media

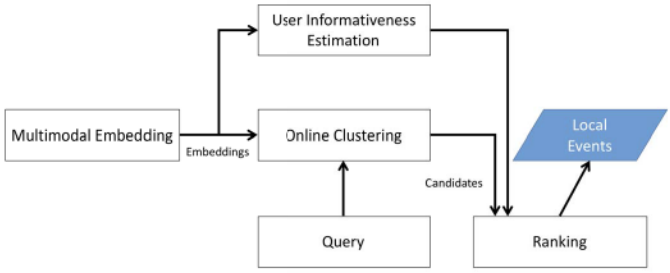


Fig. 2: The framework of LEDetect.

streams into the same low-dimensional space. While the keywords are natural textual units for embedding, the space and time are continuous and there are no natural embedding units. To address this issue, we break the geographical space into equal-size regions ($300m \times 300m$) and consider each region as a spatial unit. Similarly, we break one day into 24 hours and consider every hour as a basic temporal unit. For each time-slot, we keep newly arrived messages in a cache C and periodically update the embeddings. The existing embeddings are used as initializations to make the training process more efficient. The embedding algorithm is the same as [3], [20], so we abbreviate the details here. The semantic embedding (i.e., $\mathbf{v}^{(txt)}$ of each social message by averaging the embeddings of the keywords within the message. For the simplicity of computation, we normalize all the message, geo-region and time embeddings (i.e. $\mathbf{v}^{(txt)}$, $\mathbf{v}^{(loc)}$ and $\mathbf{v}^{(time)}$, respectively).

C. User Informativeness Modeling

Based on the learned multimodal embeddings, we can estimate the informativeness scores for online users by analyzing their social messages. For a specific spatio-temporal unit, there may exist a collection of local events. For each local event, there are a set of related social messages around the location where the event occurs. Mathematically, let t and s denote the current time window and geo-region, respectively, the overall semantics in the spatio-temporal unit can be viewed as a single density function f_{st} over the embeddings of all the social messages. Intuitively, there can be more than one peaks in f_{st}, where each peak stands for a potential event. With the embeddings of all the social messages, the most straightforward way of estimating f_{st} is:

$$f_{ts} = \sum_{x \in \mathcal{X}_{tsu}} \Phi(\mathbf{v}_{tsx}),$$

where $\mathbf{v}_{tsx}$ is a embedding vector, $\Phi(\mathbf{v}_{tsu})$ is a component function that maps embeddings of social media messages from real-valued vectors to functional space and $\mathcal{X}_{tsu}$ is the set of messages from user u at region s in time window t. Specifically, we define this mapping as $\Phi : \mathbf{v} \in \mathbb{R}^d \rightarrow K(\cdot, \mathbf{v})$, where K is a positive symmetric function. In this paper, $K(\mathbf{x}, \mathbf{v})$ is defined as: $K(\mathbf{x}, \mathbf{v}) = \cos(\mathbf{x}, \mathbf{v})$, if $\mathbf{x}^T\mathbf{v} \geq 0$, and 0 otherwise.

However, the above f_{ts} does not consider the importance or quality of users. Intuitively, users have different level of

abilities of providing trustworthy and informative information for local events. Motivated by [21], for a specific user u, we denote its informativeness as r_u. In order to incorporate informativeness, we extend the above f_{ts} to a weighted aggregation of all the semantic functional components. Specifically, in the t-th time window at region s, we can obtain:

$$f_{ts} = \sum_{x \in \mathcal{X}_{ts}} w_{tsx} \Phi(\mathbf{v}_{tsx}),$$

where $\mathbf{v}_{tsx}$ is a embedding vector, $\mathcal{X}_{ts}$ is the set of messages in the t-th time window at region s and $\{w_{tsx}\}$ is a set of normalized user informativeness weights with $\sum_{x \in \mathcal{X}} w_{tsx} = 1$. This weighted aggregation of the social message embeddings from different users can be viewed as a summarization of overall embeddings in the spatio-temporal unit.

Moreover, the highly-informative users should provide meaningful information related to local events, instead of routine and trivial messages. Hence, we design a term which favors the messages that are relatively rare in current spatio-temporal context, while penalizes the messages that are commonly seen. Specifically, the term is defined as:

$$M - [(\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(time)} + \alpha \cdot (\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(loc)}],$$

As one can see, the term is inversely proportional to the similarity between the keywords in the massage and its context (i.e. time and geo-regions). α is a tradeoff coefficient that controls the ratio of spatial unusualness and temporal unusualness. M is a constant that adjusts the value range of the whole term. By applying this term to all the messages, users that provide unusual but trustworthiness messages are identified. Finally, the constraint reflects the distribution of users' informativeness scores.

Based on all these designs, we define our optimization objective function. Particularly, we need to find a set of functions f_{ts} and a set of numbers $r_j \in \mathbb{R}^+, j = 1, \cdots, |\mathcal{U}|$,which can minimize the total loss function:

$$J_w = \sum_{t=1}^{|\mathcal{T}|} \sum_{s=1}^{|\mathcal{S}|} \frac{1}{m_{ts}} \sum_{u=1}^{|\mathcal{U}|} r_u \sum_{x \in \mathcal{X}_{tsu}} \Big(M - [(\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(time)} + \alpha \cdot (\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(loc)}] \Big) \cdot ||\Phi(\mathbf{v}_x^{(txt)}) - f_{ts}||^2,$$

$$s.t. \sum_{u=1}^{|\mathcal{U}|} n_u \exp(-r_u) = 1,$$

where $\mathcal{T}$ stands for the set of timestamps, $\mathcal{S}$ stands for the set of geo-regions in the local area, $|\mathcal{U}|$ denotes the number of users. m_{ts} is the number of users who post social messages in the t-th time slot at geo-region s. n_u is the number of social messages provided by the u-th user. $\mathcal{X}_{tsu}$ is the set of social messages from user u in time window t at region s. $\mathbf{v}_j^{(txt)}$, $\mathbf{v}_j^{(loc)}$, and $\mathbf{v}_j^{(time)}$ are the keywords, geo-region and time representation of the j-th message, respectively. M is a shifting constant that is set to 2 in this paper.

In J_w, $||\Phi(\mathbf{v}_x^{(txt)}) - f_{ts}||^2$ measures the semantic distance between f_{ts} and the functional component transformed from the embedding of x-th message (i.e. $\Phi(\mathbf{v}_x^{(txt)})$). The loss term evaluates the semantic deviation of current message compared with the weight-aggregated peak that x is describing. To minimize the overall loss, the informativeness of user u tends to be larger if the messages from u are close to the weight-aggregated function f_{ts}, and visa versa. In a nutshell the embedding representation of a social message from high-informative user should near the semantic center of the corresponding weight-aggregate peak.

To minimize the loss function with constraint, we use Lagrangian multiplier to convert the problem into the following new loss function. Specifically, the update rule for r_u and f_{ts} are as follows.

Update of r: Suppose $f_{tx}, (t, s) \in (\mathcal{T}, \mathcal{S})$ are fixed, r_u is updated based on the following equation:

$$r_u = -\log \left(\frac{\frac{1}{n_u} \sum_{t=1}^{|\mathcal{T}|} \sum_{s=1}^{|\mathcal{S}|} H(t, s, u)}{\sum_{u'=1}^{U} \sum_{t=1}^{|\mathcal{T}|} \sum_{s=1}^{|\mathcal{S}|} H(t, s, u')} \right) \tag{1}$$

where:

$$H(t, s, u) = \frac{1}{m_{ts}} \sum_{x \in \mathcal{X}_{tsu}} \Big(M - [(\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(time)} + \alpha \cdot (\mathbf{v}_x^{(txt)})^T \mathbf{v}_x^{(loc)}] \Big) \cdot ||\Phi(\mathbf{v}_x^{(txt)}) - f_{ts}||^2. \tag{2}$$

Here, $||\Phi(\mathbf{v}_x^{(txt)}) - f_{ts}||^2$ is defined as:

$$\begin{aligned} &||\Phi(\mathbf{v}_x^{(txt)}) - f_{ts}||^2 \\ &= K(\mathbf{v}_x^{(txt)}, \mathbf{v}_x^{(txt)}) - 2 \sum_{j \in U_{ts}} w_{tsj} \sum_{l \in \mathcal{X}_{tsj}} K(\mathbf{v}_x^{(txt)}, \mathbf{v}_l^{(txt)}) \\ &\quad + \sum_{j, j' \in U_{ts}} w_{tsj} w_{tsj'} \sum_{l \in \mathcal{X}_{tsj}} \sum_{l' \in \mathcal{X}_{tsj}} K(\mathbf{v}_l^{(txt)}, \mathbf{v}_{l'}^{(txt)}) \end{aligned} \tag{3}$$

where U_{ts} denotes the users that provide messages in region s at time slot t. $\mathcal{X}_{tsj}$ stands for the set of messages from user j in region s at time slot t.

Update of f: In addition, when all the r_u are fixed, we have the update rule for f:

$$f_{ts} = \sum_{x \in X_{tsu}} w_{tsx} \Phi_i(\mathbf{v}_x), \tag{4}$$

where $w_{tsx} = \frac{r_{U(x)}}{\sum_{j'=U(x), x \in X_{ts}} r_{j'}}$ and $U(x)$ is a function that maps the index of message x to its provider, $\mathcal{X}_{ts}$ stands for all the messages in time window t at region s.

In this paper, the user informativeness is evaluated in batches, we select the users with top 1% informativeness scores as trustworthy users.

D. Event Candidate Generation

Intuitively, when a local event occurs at a location l, there is a collection of related social media messages around l. Hence, we can witness the phenomenon that the density of messages around the center of an event candidate is relatively high, surrounded by low-density messages that are less relevant.

Based on this intuition, we derive Algorithm 1, which is a variant of [22] to find the event candidates.

Batch Mode Clustering: Algorithm 1 details the process of clustering social media messages in time window t into self-coherent geo-topic message clusters, which denote event candidates. We define the distance function between two messages i and j as:

$$d(i,j) = cos(\mathbf{v}_i^{(txt)}, \mathbf{v}_j^{(txt)}) + \beta \cdot dist((x_j, y_j), (x_i, y_i)) \quad (5)$$

where $dist((x_1, y_1), (x_2, y_2))$ is the euclidean distance of i and j's coordinates. This indicates that we consider both geographical and semantical relevance, which helps discern semantically different events happening at the same location as well as semantically similar events happening at different locations.

As shown in Algorithm 1, we first set the 1st percentile of the pairwise distances (i.e., the value that is larger than 1% of the distance values) as a distance threshold dc_t, and then denote the number of neighbor messages of x with distances smaller than dc_t as $\rho(x)$ (i.e., local message density for x). After that, for each message x, we find its nearest neighbor with higher density and record their distance as $\delta(x)$. $\rho(x)$ and $\delta(x)$ are the references that we used to find the centers for geo-topic clusters.

Intuitively, the social media message x describing a local event that occurs right at the geo-location embedded in x should have a higher local density than its surrounding messages. We regard x as the geo-topic representative message for its corresponding event. On the other hand, the representative message of other events should have a relative large distance from x. Based on such intuition, we propose to use score $\eta(x) = \rho(x) \times \delta(x)$ to evaluate the chance for message to be a geo-topic cluster center (line 13-18 in Algorithm 1). Here, we regard the social message x with $\eta(x)$ greater than two times of the standard deviation for all the possible η scores as cluster center. Finally, if a message x is not a cluster center, it is assigned to the cluster whose center is nearest to x. Thus we manage to cluster social messages into geo-topic clusters.

Online Update: Here we present the way to update the clusters with time window shifting continuously. Assume that the newly arrived social media messages are in the time span Δt, the goal here is to find the local events in Δt. The online update algorithm updates the clustering assignment with as little cost as possible. Specifically, the algorithm first calculates the local density score and finds the neighbors with higher density for each social message in time window Δt. After that, we pick out the social media messages in time window Δt, whose $\eta'(x) = \rho'(x) \times \delta'(x)$ is greater than two times of the standard deviation of all the possible η' in time window δ. Finally, each social message in Δt is assigned to the nearest event candidate center.

Algorithm 1 Social Media Message Clustering

Input: Local geo-tagged social media message at time window t
Output: Message clusters which stand for event candidates
1: **for** each time window $t \in \mathcal{T}$ **do**
2: Calculate pairwise distances among messages.
3: Find the message distance boundary dc_t, which is the 1st percentile of pairwise distances.
4: *// Calculate local density for each message.*
5: **for** each message x in time window t **do**
6: Calculate density score $\rho(x)$ using dc_t.
7: **end for**
8: *// Get nearest neighbor with higher density for each message.*
9: **for** each message x in time window t **do**
10: Find x's nearest neighbor $dn(x)$ whose density score is higher than x.
11: Save x's distance to $dn(x)$ as $\delta(x)$.
12: **end for**
13: *// Cluster Assignment*
14: **for** each message x in time window t **do**
15: **if** $\eta(x) = \rho(x) \times \delta(x)$ is two times greater than all the possible η score **then**
16: Mark x as a candidate event center.
17: **end if**
18: **end for**
19: **for** each message x in time window t **do**
20: **if** x is not a candidate event center **then**
21: Assign x to the cluster with center $dn(x)$
22: **end if**
23: **end for**
24: **end for**

E. Event Filtering

Up to now, we have obtained a set of local event candidates as well as user informativeness scores. As discussed before, not all the candidates are relevant to a meaningful local event. Thus, we introduce the following approach to determine the confident local events. First, we define a set of features to capture the characteristics of each event candidate:

1) **Spatial concentrations**: This feature quantifies how concentrated a event candidate C is over the spatial dimension. The value of feature $score_{sp}(C)$ is computed as the sum of: 1) the standard deviation of the longitudes; and 2) the standard deviation of the latitudes.
2) **Semantic concentration**: This feature evaluates the degree of semantic coherence of C:

$$score_{se}(C) = \frac{\sum_{x \in C} \cos(\mathbf{v}_x^{(txt)}, \bar{\mathbf{v}}^{(txt)})}{|C|},$$

where $\bar{\mathbf{v}}^{(txt)}$ is the mean value of semantic embedding in C.
3) **Volume**: We define it as the number of messages in a time window divided by the time span of the time window.

Then we propose two schemes to filter out real local events based on the features and user informativeness:

A. Feature-based filtering: In this scheme, we compute the ranking score of candidate C via the following ranking function:

$$score(C) = \mathbb{I}(|C| > \Omega_{volume}) \cdot (\lambda_1 score_{sp}(C) + \lambda_2 score_{se}(C))$$

where $\mathbb{I}(x)$ equals 1 if expression x is true and 0 otherwise. Ω_{volume} is a predefined threshold of the number of messages in a time window. After the score of each candidate C is

computed, we select the top-K candidates as real local events. In this paper, K is set to 5.

B. User-based filtering: In this scheme, we select the candidate C that is supported by one or more trustworthy users as real local events.

Finally, from these clusters representing real local events, the representative tweets are selected as follows. For each cluster C, we keep:

1) the top 10 messages with semantic embedding $\mathbf{v}_x^{(txt)}$ nearest to $\bar{\mathbf{v}}^{(txt)}$;
2) the messages from trustworthy users.

IV. Experiment

A. Experiment Settings

Baseline method For comparison, we implement the state-of-the-art *unsupervised* local event detection method *GeoBurst* [1] as the comparison baseline method. The approach utilizes keyword co-occurrence graph to calculate word similarities and further detects local event candidates. After that, event candidates are ranked via spacial and temporal burstiness. Here, we do not include *TrioVecEvent* [3] as a comparison method because that work proposes a *supervised* model, which is different from the setting of this paper.

Dataset We take our experiments on a real-world dataset, which is collected via Twitter Streaming API during 2017.5.1 – 2017.7.1 in New York. The dataset consists of 13.5 million geo-tagged tweets. After removing the tweets without entities or noun phrases, we obtain 1.8 million tweets. We remove the keywords that appear less than 50 times in the corpus.

To evaluate the effectiveness of the proposed method, we generate 50 three-hour non-overlapping query time windows. For every query, we run the baseline and the proposed model to discover local events. We manually label *whether each event is indeed a local event or not*. The verification procedure is done according to the *reports of local news media* [1], and *open social events records released by New York City government* [2][3].

Metrics To quantify the performance, we adopt the following metrics:

- **Precision** The detection precision is: $P = \frac{N_{true}}{N_{report}}$, where N_{true} is the number of true local events and N_{report} is the total number of reported events.
- **Pseudo Recall** As the total number of events in each query window is difficult to determine, we aggregate the true positives of different methods. Let N_{total} be the total number of local events detected by all the methods. We compute the pseudo recall of each method as: $R = \frac{N_{true}}{N_{total}}$.
- **Pseudo F1-Score** We also report the pseudo F1 score of each method, which is computed as: $F1 = \frac{2PR}{P+R}$.

Parameters and Environment In the proposed model, α is set to 0.10, while β is set to 3. λ_1 and λ_2 are set to 1 and 2, respectively. The parameters of the baseline method is set as the original paper recommend. The experiments are carried out on a computer with Intel Core i7 2.7GHz and 16 GB memory. The proposed method is implemented in Python.

[1] http://nydailynews.com/
[2] https://data.ny.gov/
[3] https://data.cityofnewyork.us/

B. Quantitative Results

The performance of the proposed LEDetect as well as the baseline method, are shown in Table I. As one can see, LEDetect significantly outperforms the baseline method. The reason of such an improvement is two-fold: (1) user quality is extremely important in finding trustworthy events and filtering out noisy information; (2) the proposed framework successfully identifies informative users that can provide informative messages for local event detection. With the help of these highly informative users, the proposed method can return a much larger amount of highly trustworthy event-related social media messages than existing methods, which describe a larger collection of local events. A more detailed study on the cases where the proposed model works much better than GeoBurst is shown in the next section.

TABLE I: Performance of different methods.

Method	Precision	Recall	F1
GeoBurst	0.1034	0.1858	0.1329
LEDetect	0.1943	0.8584	0.3168

C. Case Study

Case Study on Event Detection We perform a case study to compare the local events detected by the proposed method and the state-of-the-art method GeoBurst. We choose the query window to be 7 – 10 PM on June 1st 2017 and list the local events detected by both methods in Figure 3. For each event, we show the representative tweets, and offer a visualization of the geo-locations[4] in tweets.

From the visualization, we can see that the proposed LEDetect method can discover both bursty events (i.e., the baseball game at the Citi Field between team Mets and Brewers in New York city) and less popular but meaningful events (i.e., the Construction at Sprain Brook State Parkway). We can also see that the local events discovered by LEDetect is very geo-coherent. On the contrary, GeoBurst discovers the baseball game in Citi Field, but fails to detect the parkway construction. The reason why the proposed method manages to discover the less popular local event is that we utilize the trustworthy information from highly-informative users like @511NYnj. Even with few messages, we are confident that the candidate indeed describes a real and meaningful local event. Nevertheless, as the existing method does not take the power of highly-informative users into consideration, it cannot accomplish a comprehensive result.

Moreover, we found that both methods detect the event of President Trump withdrawing from Paris climate summit. This event is a huge event that draws global attention. Hence, we can find 'dense' discussions on this topic all over the places. Although this event is not local, it is meaningful and

[4] Note: we down sample the tweets for better visualization.

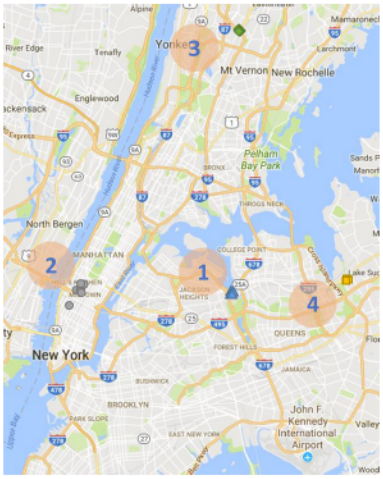
	Results of GeoBurst	Local Event	Visualization
#1	▪ Checking Out The B Squad Of The NYC w mkandrach Citi Field ▪ great day for some baseball Happy June Citi Field ▪ My best friend in the entire world thankful humbled Citi Field ▪ Vacation continues Today Im at Citi Field to watch the Mats take on the Brewers baseball	Yes	
#2	▪ Im at Mad Sq Eats madsqparknyc in New York NY ▪ We start real work with new partner in New York just the best the real fashion is not ▪ lil trip down memory lane VaynerMedia in New York NY ▪ Sundaes and Cones in New York NY	No	
#3	▪ Dedication Vibes Freedom Creation susiemcreative New York ▪ tbt to the crew during tatompa s bachelor weekend in nyc New York ▪ days short days and long days TheLoveBomb podcast niconiconico atwillradio New York ▪ Yall gon show the same love New York	No	
#4	▪ Trump proving to the world once again hes full of covfefe F** him ▪ Hey Trump prove to me you actually listen we have more solar works than coal we mo ▪ Trump look like a terror talking about terror ▪ sad on so many levels whatabouthenextgeneration Trump set to withdraw from Paris accord	No	
	Results of LEDetect	**Local Event**	**Visualization**
#1	▪ Today Im at Citi Field to watch the Mats take on the Brewers baseball ▪ Gorgeous day for a ballgame Citi Field ▪ Wow even Mr Met cant behave these days Citi Field ▪ Checked in Citi Field for the brewers vs mets game w	Yes	
#2	▪ Im at Wellington Hotel wellingtonnyc in New York NY ▪ Im at Worldwide Plaza in New York NY ▪ First stop as always Times Square New York City ▪ Just posted a video Times Square New York City	No	
#3	▪ SprainBrook open FINALLY. ▪ Cleared Construction on SprainBrookStateParkway NB from Exit Bronx River Parkway to Exit NY 100 (From trustful user @511nyNJ) ▪ Updated Construction on SprainBrookStateParkway NB from Exit Bronx River Parkway to Exit NY 100 (From trustful user @511nyNJ)	Yes	
#4	▪ I really dont care about the climate but withdrawing from the Paris Agreement is just fucking stupid ▪ Just waiting for realDonaldTrump to promise no homework and soda machines in the cafeteria Paris ▪ If this was the case the Paris Agreement wouldnt be necessary ParisAgreement ▪ Here is someone who speaks coherently gives the true argument for Paris while throwing intellectual shade	No	

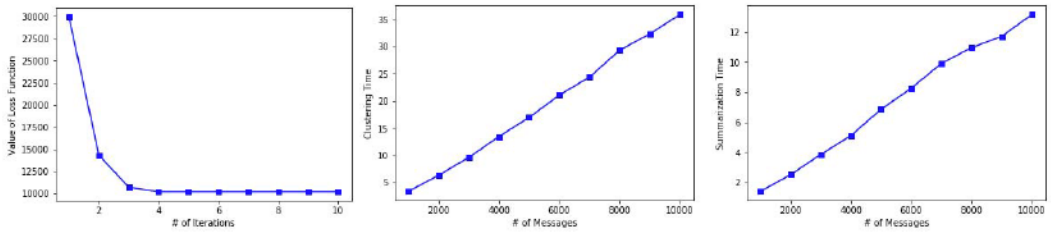
Fig. 3: Local Events Discovered by GeoBurst and LEDetect.

significant. Finally, the only false positive local event reported is #2. However, we can see that these messages are generated automatically by mobile apps, which can be removed using straightforward anti-spammer rules.

Case Study on User Informativeness Analysis In addition, we perform a case study to demonstrate that the estimated user informativeness is interpretable. Table II lists the examples of users with high informative scores (Top-20). From Table II, we can see that these users are mostly public service accounts and celebrities, which are likely to provide trustworthy information. Also, the two ordinary users are all highly-active users that retweet extremely high volume of local news events. These phenomena show that the proposed method successfully captures the characters of users who actively provide informative messages on local events.

TABLE II: Examples of Users with High Informative Scores

Users	Descriptions
@fleurdeliselle	Highly-active Twitter user in Brooklyn
@Rontu	Highly-active users that retweets news events.
@511nyNJ	Traffic & transit updates for the New Jersey area provided by New York State 511.
@mattstopera	A senior editor at BuzzFeed.
@511NY	Traffic & transit updates for all of New York State.



(a) Convergence of User informativeness Analysis Model.

(b) Running time of Batch Mode Clustering.

(c) Running time of Local Event Summarization.

Fig. 4: Visualization of running time under different settings.

D. Convergence and Efficiency Analysis

First, we study the convergence of the proposed user informativeness analysis model. We randomly select 800 query windows, and apply the optimization algorithm. Figure 4(a) shows the value of the loss function as the number of iterations increases. We observe that the loss function quickly converges after less than 5 iterations. Then, we illustrate the efficiency of the proposed method in terms of generating event candidates in Figure 4(b). As one can see, the increase rate of batch mode clustering is linear. We also illustrate the time of summarizing local events from the event candidates in Figure 4(c). The time usage of summarizing a corpus with 10000 social messages is less only than 15 sec. These results show that the proposed model is capable for large-scale applications.

V. Conclusions

In this work, we propose an unsupervised framework LEDetect to achieve accurate and comprehensive online local event detection from social media streams. Building upon the multimodal embeddings of the geo-location, time and text, LEDetect identifies highly-informative users via social message analysis and leverages their power to ensure a comprehensive and accurate detection of the underlying location events. The extensive experiments have demonstrated that the proposed method can achieve improved performance over the state-of-the-art method. Efficient studies and case illustrations demonstrate that the proposed framework can generate high quality local summarization with high efficiency.

Acknowledgment

This work was sponsored in part by US National Science Foundation under grants IIS-1553411, CNS-1742845, CNS-1652503 and CNS-1737590. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

[1] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han, "Geoburst: Real-time local event detection in geo-tagged tweet streams," in *Proc. of SIGIR*, 2016, pp. 513–522.

[2] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *Proc. of VLDB*, vol. 6, no. 12, pp. 1326–1329, 2013.

[3] C. Zhang, L. Liu, D. Lei, Q. Yuan, H. Zhuang, T. Hanratty, and J. Han, "Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams," in *Proc. of KDD*, 2017, pp. 595–604.

[4] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in *Proc. of SDM*, 2012.

[5] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proc. of SIGIR*, 1998.

[6] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proc. of GIS*, 2009, pp. 42–51.

[7] Q. He, K. Chang, and E.-P. Lim, "Analyzing feature trajectories for event detection," in *Proc. of SIGIR*, 2007, pp. 207–214.

[8] C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets," in *Proc. of CIKM*, 2012, pp. 155–164.

[9] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in *Proc. of VLDB*. VLDB Endowment, 2005, pp. 181–192.

[10] J. Weng and B.-S. Lee, "Event detection in twitter." *Proc. of ICWSM*, vol. 11, pp. 401–408, 2011.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proc. of WWW*, 2010, pp. 851–860.

[12] J. Krumm and E. Horvitz, "Eyewitness: Identifying local events via space-time signals in twitter feeds," in *Proc. of SIGSPATIAL*, 2015, p. 20.

[13] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in *Proc. of CIKM*, 2011, pp. 2541–2544.

[14] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang, "Streamcube: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream," in *In Proc. ICDE*, 2015, pp. 1561–1572.

[15] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proc. of WWW*, 2012.

[16] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang, "Topicsketch: Real-time bursty topic detection from twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2216–2229, 2016.

[17] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in *Proc. of KDD*, 2013.

[18] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proc. of KDD*, 2006.

[19] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec, "Information cartography: creating zoomable, large-scale maps of information," in *Proc. of KDD*, 2013.

[20] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han, "Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning," in *Proc. of WWW*, 2017.

[21] M. Wan, X. Chen, L. M. Kaplan, J. Han, J. Gao, and B. Zhao, "From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach," in *Proc. of KDD*, 2016, pp. 1885–1894.

[22] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.