# Local Event Discovery from Tweets Metadata

Mohammed Hasanuzzaman
ADAPT Centre, School of Computing, DCU
Dublin 9, Ireland

Andy Way
ADAPT Centre, School of Computing, DCU
Dublin 9, Ireland

## ABSTRACT

We present a two-step strategy that addresses fundamental deficiencies in social media-based event detection and achieves effective local event by taking advantage of geo-located data from Twitter. While previous work has mainly relied on an analysis of tweet text to identify local events, we show how to reliably detect events using meta-data analysis of geo-tagged tweets. The first step of the method identifies several spatio-temporal clusters within the dataset across both space and time using metadata to form potential candidate events. In the second step, it ranks all the candidates by the amount of hashtag/entity inequality. We used crowdsourcing to evaluate the proposed approach on a data set that contains millions of geo-tagged tweets. The results show that our framework performs reasonably well in terms of precision and discovers local events faster.

## CCS CONCEPTS

• **Information systems** → **Clustering**; **Social networks**;

**ACM Reference Format:**
Mohammed Hasanuzzaman and Andy Way. 2017. Local Event Discovery from Tweets Metadata. In *K-CAP 2017: K-CAP 2017: Knowledge Capture Conference, December 4–6, 2017, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3148011.3154477

## 1 INTRODUCTION

A local event is an activity that attracts a fair amount of attention of people within a local area for a specific duration of time. Early detection of local events (*e.g.*, protests, accidents and disasters, crimes, sports and games) can be of utmost interest and importance to business and administrative decision makers as it can buy extra precious time for them to make informed decisions. Despite its practical importance, the task is particularly challenging due to several reasons: (a) *Integrating diverse types of data* from geo-tagged tweets such as location, time, and text (b) *Overwhelming noise.* Almost 40% of the tweets are pointless babbles[1]. As truly interesting local events buried in massive irrelevant tweets, it is nontrivial to accurately identify them.

In this paper we put forward a simple and effective two-step strategy, STED, based on the intuition that a local event usually leads to a considerable amount of tweets in the locality, especially around the occurring place (*e.g.,* many audiences of a concert may post tweets on the spot to express their feelings at a given point in time). As such, tweets that are spatio-temporally and topically close form a spatio-temporal cluster, which can be considered a potential local event. Motivated by the above, the first step of our methodology finds all spatio-temporal clusters in a query window[2] as candidate events. Specifically, it looks for the increase in tweets regardless of their content within a space-time context and evaluates whether this increase is due to random variation.

It is anticipated that not every detected spatio-temporal cluster represents a local event but rather contains garbage data. The second step of our methodology ranks all candidate clusters based on the concentration of *hashtags/entities* inside it. In general, most of the existing trend sensing studies use bursty keywords or hashtag peaks to identify trendy topics on Twitter. However, only looking at these peaks across the network might lead to spurious results. Therefore, we define the *Twitter Gini* index that captures hashtag and entity distributions within each cluster to sense the level of public attention for a particular topic.

Finally, we perform our experiments on a large corpus of geo-tagged tweets and evaluate the results using a crowdsourcing platform. Experimental results demonstrate that our methodology performs reasonably well in terms of precision and detects local events much faster.

## 2 RELATED WORK

We provide an overview of the few important event detection approaches mainly at the global and local level.

Global event detection targets to identify events that are bursty and unusual in the entire tweet stream. To detect global events in Twitter, clustering of the twitter stream is applied in [2, 14]. Along other line, a number of term-pivot approaches have been proposed [9, 17]. There has also been extensive work [11, 13] on the detection of specific types of events. The above methods are all designed for detecting events that are bursty in the entire stream. As mentioned before, a local event is usually bursty in a small geographical region instead of the entire stream. Therefore, directly applying these methods to the geo-tagged tweet stream would miss many local events.

A few studies have investigated local event detection [1, 16]. However, the most relevant work in line of our research is by Cheng *et al.* [4], where they adopted spatial scan statistic proposed by Kulldorff [6] to detect local events from geo-tagged tweets. Their approach suffers from several drawbacks. First, it considers a circular scanning window with variable size to define the potential cluster (geographical span of candidate events) area. However, most geographical areas are non-circular. Therefore, it's difficult to detect

[1] http://www.cnbc.com/id/32446935. Last Access: 02/05/2016.

[2] A query window represents a specific geographical area within a given period.


*K-CAP 2017, December 4–6, 2017, Austin, TX, USA*

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3154477

non-circular clusters, as those along the river [15]. In addition, even if the null hypothesis is rejected, the circular spatial scan statistic tends to detect a larger cluster than the true cluster by absorbing surrounding regions. Second, no method to filter out irrelevant spatio-temporal cluster. As mentioned before, it is expected that not all detected spatio-temporal cluster represents a local event.

## 3 APPROACH

Our local event detection methodology has two main steps: candidate event generation and candidate ranking. In this section, we will introduce these two steps in details.

### 3.1 Problem Description

Let $C = (c_1, c_2, c_3, ..., c_n)$ be a continuous stream of geo-tagged tweets and each tweet $c$ is a three element tuple $\langle d_c, z_c, H_c \rangle$, where $d_c$ is the timestamp, $z_c$ is the geo-location, and $H_c$ is the bag of keywords.

For each tweet, we extract hashtag(s), if there is any, and entities using a freely available tool namely twitter-nlp[3] as keywords. Consider a query time window $Q$=$[d_s, d_e]$, where $d_s$ and $d_e$ are the start and end timestamps satisfying $d_{c_1} \leq d_s < d_e \leq d_{c_n}$. Local event detection in $Q$ consists of two sub-tasks: first, generates from $C$ all the candidate events that occur during $Q$; second, ranks all the candidates according to its spatio-temporal burstiness.

### 3.2 Candidate Event Generation

The first step of our methodology is based on Scan Statistic originally designed for epidemic cluster detection [6].

The algorithm looks through a corpus of geotagged tweets, systematically scanning over localized regions of time and space for unusual spikes in the volume of tweets. For a given region on the ground at a given time, the algorithm calculates how many geotagged tweets to expect in normal, default situations. If the actual volume of tweets in a time span exceeds the calculated number by a significant amount, then a candidate event form. Overall event candidates generation process is defined as follows:

It can be defined by a cylindrical window with a circular (or elliptic) geographic base and with height corresponding to time. The base is defined exactly as for the purely Spatial Scan Statistic, while the height reflects the time period of potential clusters. The cylindrical window is then moved in space and time, so that for each possible geographical location and size, it also visits each possible time period. Overall event candidates generation process is defined as follows:

Suppose we have hourly/minutely (or any other unit of time) number of tweets generated within postcode areas[4] for the query window, $c_{zd}$ is the observed number of tweets in postcode area $z$ during the time $d$. The total number of tweets generated within the study region ($C$) is

$$C = \sum_{z} \sum_{d} c_{zd} \tag{1}$$

For each postcode area and time, we calculate the expected number of tweets $\mu_{zd}$ conditioning on the observed tweets:

$$\mu_{zd} = \frac{1}{C}(\sum_{z} c_{zd})(\sum_{d} c_{zd}) \tag{2}$$

This is the proportion of all tweets that occurred in postcode area $z$ times the total number of tweets during the time $d$. The expected number of tweets $\mu_A$ in a particular scan window shape $A$ is the summation of these expectations over all the postcode-hours/minute within that shape:

$$\mu_A = \sum_{(z,d) \in A} \mu_{zd} \tag{3}$$

Assume $c_A$ be the observed number of tweets in the scan window shape $A$. Conditioned on the marginals, when there is no space-time interaction, $c_A$ is distributed according to the hyper-geometric distribution with mean $\mu_A$ and probability function.

$$P(C_A) = \frac{\binom{\sum_{z \in A} c_{zd}}{c_A} \binom{C - \sum_{z \in A} c_{zd}}{\sum_{d \in A} c_{zd} - c_A}}{\binom{C}{\sum_{d \in A} c_{zd}}} \tag{4}$$

When both $\sum_{z \in A} c_{zd}$ and $\sum_{d \in A} c_{zd}$ are small compared to C, $c_A$ is approximately Poisson distributed with mean $\mu_A$ [5]. Based on this, Poisson Generalized Likelihood Ratio (GLR) is used to measure the abnormal level of the number of tweets in the scan window shape $A$.

Most widely used scan window shape is the circle, as it is the most compact shape that can be obtained. However, with the circular scanning window of variable size to define the potential cluster area, it is difficult to correctly detect some non-circular clusters. Therefore, we have used non-circular window shape (ellipse), which one would often expect to be the case for this study.

$$GLR = \left(\frac{c_A}{\mu_A}\right)^{c_A} \left(\frac{C - c_A}{C - \mu_A}\right)^{C - c_A} \tag{5}$$

Finally, statistical significance is evaluated using Monte Carlo hypothesis testing [3].

### 3.3 Candidate Ranking

We have measured the concentration of keywords (hashtags and entities) inside a given candidate, and consider the value as candidate's final ranking score. We know that users quickly converge on a few hashtags to use for an event [8]. Similarly, entities inside tweet text can also be very informative. Therefore, the high concentration of a unique hashtag/entity in a given candidate event is likely to represent a local event. To measure this phenomenon, we defined *Twitter Gini* index. It is defined in the Equation 6.

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1) x_i^{'}}{n^2 \mu} \tag{6}$$

where $n$ is the number of distinct hashtags and entities in the candidate event, $x_i^{'}$ is the count of a given hashtag/entity $i$ and $\mu$ is the average number of tweets a hashtag/entity can appear in within the candidate event. Note that hastags/entities are in ascending order by their counts.

[3] https://github.com/aritter/twitter-nlp

[4] The granularity of locations can vary from building, streets, regions to the entire city or state

| Top K | P@1 | P@2 | P@3 | P@4 | P@5 |
|---|---|---|---|---|---|
| Precision | 0.35 | 0.33 | 0.30 | 0.29 | 0.26 |
| Query Interval (Hour) | 2 | 3 | 4 | 5 | 6 |
| Precision | 0.26 | 0.28 | 0.31 | 0.35 | 0.37 |

**Table 1: Local event detection precision by STED.**

## 4 EXPERIMENTS

**Data Set.** As the main source of information, we used geo-tagged tweets that are collected using the Twitter streaming API[5]. The project only uses English tweets that originate from geo-enabled users within a bounding box set of coordinates containing the Greater London (UK) during the period 7th January to 18th January 2013.

### 4.1 Results

A total of 334 spatio-temporal clusters are detected by the candidate event generation step considering hourly temporal aggregation. However, the dominant portion of detected clusters is insignificant or most likely to have occurred by chance with p-value more than 0.05. Only 158 number of significant spatio-temporal clusters are produced by the first step of our methodology.

It is important to note that these 158 clusters are significant only within the context of scan statistics. It is expected that the majority of these clusters may contain noise within the purview of our study, and are not attributable to local events. However, this cannot be concluded unless these clusters are processed further. In order to further analyze these 158 significant spatio-temporal clusters, *Twitter Gini index* is calculated for each. It is observed that most of these spatio-temporal clusters belong to very low Gini index under 0.4. Gini index of 0 have also been encountered quite a few times. As the Gini score is the important feature to discover relevant local events, clusters with low scores were dropped. Finally, 88 out of 158 clusters with Gini index of more than 0.4 were picked up as local events. It is important to note that the Gini coefficient of 0.4 is considered as threshold for this study. It is chosen arbitrarily, but can be optimized. The highest observed *Hashtag Gini* index is 0.83333. Table 2 represents a sample of randomly selected detected clusters.

### 4.2 Evaluation

Evaluating event detection from microblogs has traditionally been challenging due to a lack of ground truth. One of the most influential approaches in previous studies for such evaluation is based on human judgment [7, 12]. Even then, however, it is difficult to estimate a recall rate, since there is no assurance that the ground truth contains every relevant event.

**Manual evaluation:** We randomly generate twenty five (25) non-overlapping query windows of five 2-hour window, five 3-hour window, five 4-hour window, five 5-hour window and five 6-hour window. For each query, we run the our method and upload the top 5 retrieved local events in the crowdsourcing service of CrowdFlower platform[6] for evaluation. In particular, we represent each event with top-5 most representative tweets and top-5 keywords to the annotators.

For each event, annotators were asked the question *"Do three or more of these tweets look like related to the same local event as each other?"*. The available answers were *"Yes", "No",* and *"Doubtful"*. Each unit was annotated by at least 10 annotators. Each annotator required to be an English speaker in the UK, as the UK originated English language tweets we used in our experimental corpus. Apart from it, each annotators were presented with detailed annotation instructions.[7] Events which received a majority of "Yes" vote (atleast 5 or more) are considered as gold-standard [10]. The annotation phase resulted in a gold standard event dataset for our method. Finallt, we measure the precision of our method after gathering judgments and presented in Table 1.

The precision of STED declines with k. This phenomena demonstrates the usefulness of our ranking module. One of the simplest methods to evaluate ranking modules is determining the precision of the first k top-ranked results for some fixed k. First half shows that our ranking module can effectively identify local events and put them in the top positions of the result list.

**Average lag time:** We searched though different sources such as web, local newspapers, event diaries and collected the actual occurrence date and time for a subset of gold-standard events. We calculate the time gap between these event's occurrence time and discovery time by our proposed method. The average lag time for our method is 3.25 hours.

### 4.3 Case Study

We randomly selected two detected local events and presented in Table 3. In particular, each event is represented with its corresponding date of occurrence, hashtag/entity inside the cluster, top three tweets selected to describe the event, and event cluster ranking score. The second row in Table 3 represents a protest march to save Lewisham Hospital from closure and to protect the NHS. The third row is about a seminar to share knowledge and experience and also to explore the potential of social media in local government with fellow councilors at Facebook HQ in the London, UK. Examining the results of STED, one can see the generated spatio-temporal clusters are of high quality: the tweets in each cluster are topically close. Interestingly, STED can group the tweets that discuss about the same topic using different keywords (e.g., "demos" and "march").

### 4.4 Local Event Types

A closer look at the events detected by our system showed they fell into a relatively small number of well-defined categories. We hired a professional linguist to categorize events that were voted to be an actual local event by at least 5 or more CrowdFlower annotators. Top 5 event categories with their share of local events are shown in Table 4. Not surprisingly, miscellaneous event (misc. event) which consists of local protests/demonstrations, snowfall, fireworks, or local festivals, is the most dominant event category in the detected local events. We also detected many events (convention category) which only attract the attention from a certain group of people. For

[5]https://dev.twitter.com/streaming/overview. Last accessed on 02-06-2015

[6]http://www.crowdflower.com/

[7]Detail annotation guideline is out of the scope of this paper

| Sl. No. | Coordinates | Time Frame | p-value | Gini Score |
|---|---|---|---|---|
| 01 | 51.495955 N, 0.130729 W | 2013/1/(08-08) | $1.0 \times 10^{-16}$ | 0.7142 |
| 02 | 51.522980 N, 0.138890 W | 2013/1/(10-10) | $4.5 \times 10^{-6}$ | 0.6470 |
| 03 | 51.4167 N, 0.0333 W | 2013/1/(11-11) | $1.5 \times 10^{-5}$ | 0.6944 |
| 04 | 51.521164 N, 0.071525 W | 2013/1/(17-17) | $8.3 \times 10^{-4}$ | 0.7692 |
| 05 | 51.523003 N, 0.127974 W | 2013/1/(12-12) | $2.3 \times 10^{-3}$ | **0.8333** |

**Table 2: Details of detected clusters attributable to local events (randomly selected sample).**

| Sl. No. | Date | Hashtag/Entity | Example Tweet | Gini Score |
|---|---|---|---|---|
| 01 | 2013-1-19 | #SaveLewishamAE #SaveLewishamHospital Lewisham | 1. Massive demos outside to save Lewisham 2. @****-why I'm marching #SaveLewishamAE 3.PSA. please save hospital's emergency room https://**** they saved my son's life once | 0.6944 |
| 02 | 2013-1-12 | #cllrcamp Facebook | 1. LET'S DO THIS #cllrcamp @ Facebook UK 2. @*** kicking of lightening talks #cllrcamp 3. @** #cllrcamp goodluck to all those taking part. It should be a great day. | 0.4743 |

**Table 3: Detected local events by STED (randomly selected sample)**

these events, only the users who attend or directly related to the event will discuss it. For example, the third row in Table 3 represents a social media analytics seminar at Facebook HQ, London. These events may have only 10-15 of related tweets and are very hard to detect by the tweet spike detection based methods.

| Local Event Types | Fractions in Gold standard |
|---|---|
| misc. event | 33.2% |
| show | 23.7 % |
| sports | 11.3 % |
| traffic | 7.4 % |
| convention | 10.2 % |

**Table 4: Local event types detected by STED.**

## 5 CONCLUSION

We proposed to use Twitter as a social sensor that can identify local events of different categories faster. The use of our approach is not limited to Twitter. Rather, any geo-tagged social media stream (*e.g.*, Facebook, Instagram photo tag) can use our approach to discover interesting local events as well. We plan to follow up this work by integrating stream-centric processing frameworks like Apache Storm or Apache Spark Streaming to support real-time event discovery.

## ACKNOWLEDGMENTS

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

## REFERENCES

[1] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. 2013. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment* 6, 12 (2013), 1326–1329.

[2] Charu C Aggarwal and Karthik Subbian. 2012. Event Detection in Social Streams.. In *SDM*, Vol. 12. SIAM, 624–635.

[3] James W Buehler, Ruth L Berkelman, David M Hartley, and Clarence J Peters. 2003. Syndromic surveillance and bioterrorism-related epidemics. *Emerging infectious diseases* 9, 10 (2003), 1197.

[4] Tao Cheng and Thomas Wicks. 2014. Event detection using Twitter: a spatio-temporal approach. *PloS one* 9, 6 (2014), e97807.

[5] Peacock B Evans M, Hastings N. 2000. *Statistical distributions, 3rd Ed.* New York: Wiley. 221 p.

[6] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.

[7] Florian Kunneman and Antal van den Bosch. 2014. Event detection in Twitter: A machine-learning approach based on term pivoting. In *Grootjen, F.; Otworowska, M.; Kwisthout, J.(ed.), Proceedings of the 26th Benelux Conference on Artificial Intelligence*. Nijmegen, the Netherlands: Radboud University, 65–72.

[8] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 251–260.

[9] Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 155–164.

[10] Gözde Özbal, Carlo Strapparava, Rada Mihalcea, and Daniele Pighin. 2011. A comparison of unsupervised methods to associate colors with words. In *Affective Computing and Intelligent Interaction*. Springer, 42–51.

[11] Ana-Maria Popescu and Marco Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1873–1876.

[12] Yanxia Qin, Yue Zhang, Min Zhang, and Dequan Zheng. 2013. Feature-Rich Segment-Based News Event Detection on Twitter.. In *IJCNLP*. 302–310.

[13] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[14] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*. ACM, 42–51.

[15] Toshiro Tango and Kunihiko Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics* 4, 1 (2005), 11.

[16] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. 2011. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2541–2544.

[17] Jianshu Weng and Bu-Sung Lee. 2011. Event Detection in Twitter. *ICWSM* 11 (2011), 401–408.