Mining Twitter Features for Event Summarization and Rating

Deepa Mallela
Dept. of Computer Science
Boise State University
Boise, Idaho, USA
deepamallela@u.boisestate.edu

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

Maria Soledad Pera
Dept. of Computer Science
Boise State University
Boise, Idaho, USA
solepera@boisestate.edu

ABSTRACT

We present CEST, a generic method for detection and rich summarization of events occurring in a city. CEST exploits Twitter metadata, does not need prior information on events, and is event category and structure agnostic. We developed CEST to process unstructured documents and take advantage of shorthand notations, hashtags, keywords, geographical and temporal data, as well as sentiment within tweets to both detect and summarize arbitrary events without prior knowledge. We also introduce a novel strategy that analyzes sentiment and tweeting behavior over time to create a qualitative score that captures events' overall appeal to attendees.

CCS CONCEPTS

• **Information systems → Web and social media search**; *Social networking sites*; *Spatial-temporal systems*; *Web mining*;

KEYWORDS

Twitter, event, summarization, spatio-temporal analysis

ACM Reference format:
Deepa Mallela, Dirk Ahlers, and Maria Soledad Pera. 2017. Mining Twitter Features for Event Summarization and Rating. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
https://doi.org/10.1145/3106426.3106487

1 INTRODUCTION

An *event* can refer to any activity occurring within a time interval that receives attention from people. While information about upcoming events is often well-structured and readily available, identifying past or current events can be more challenging. Having said that, an overview of past or current events can be of use, specially if such an overview could capture not only structured information, but also emergent features such as discussions around the events and other reactions from people at the event. Looking for a source of such rich social data, Twitter, which is one of the largest growing Social Networking Sites (SNS) or microblogs, seems the natural choice. The immediacy of its data, the metadata associated with every tweet and user, the richness of expression, being publicly available, and the fact that it acts as a real-time sensor to actions around the globe [9, 25], have attracted the attention of researchers for event detection and mining purposes [2, 19, 25, 27]. Moreover, Twitter has a strong mobile and here-and-now aspect, adding to its immediacy and ability to gather live data from users, making it a good source for timely, fresh, and localized event information without imposing any conscious burden on users.

An event category (or type) represents the motive behind the event occurrence, e.g., political versus sports events. Different types of events are associated with different activities, e.g., a debate can occur within a political campaign, whereas a goal happens in a soccer game. To date, most Twitter event detection techniques [6, 12, 27, 34] have focused on identifying a given event category, and thus are not applicable to detect all types of events. Furthermore, most of these techniques rely on prior knowledge, e.g., specific locations where events can take place or vocabulary that describes specific events, or they use bursty topics without considering the spatial aspect in deep detail, so they cannot detect generic events. In addition to event detection, there is an interest to generate summaries from tweets that convey important aspects of each event throughout its lifetime [28]. However, most techniques that explore Twitter event summarization either depend on prior information about events, are only applicable to structure-rich events, or identify representative posts without considering polarity [4, 8, 28].

To address the limitations of existing event detection and summarization techniques, we developed CEST (City Event Summarization using Twitter), a tool that detects and summarizes events that occurred within a city based on Twitter data. In building CEST, we introduce novel approaches for identifying different categories of events that are discussed on Twitter without prior information about them and create qualitative descriptions based on text and varying users' sentiments along lifetimes of events. To the best of our knowledge, this is the first work that focuses on summarizing and rating events based on tweets posted across a city without manual intervention. Our main contributions include:
(1) *Studying Twitter usage around Points-of-Interest (POIs)*, by observing the spatio-temporal distribution of tweets within a city. We do so to validate our choice of data source and to understand Twitter usage patterns around potential event locations.
(2) *Detecting events based on spatio-temporal characteristics of tweets*, by applying a simple, yet effective, event detection technique that identifies arbitrary events occurring in an area or city we monitor, irrespective of predefined categories or types.
(3) *Creating a concise profile of an event by aggregating tweets that are related to the event*, by developing a summarization technique that generates overviews that capture details of arbitrary events in a city without prior knowledge about type or structure of events.
(4) *Inferring an overall rating for each event*, by analyzing the variation of users' sentiments expressed across the timeline of an event.

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3106487

The novelty of CEST lies in the study of Twitter use around POIs, the summarization technique highlighting noteworthy facts about an event, including the generation of a numerical score to quantitatively capture the overall appeal of a detected 'synthetic' event, and then combining it with a generic event detection approach into a functioning integrated system. CEST shares traits with some of its counterparts. For example, CEST performs content analysis to discover possible events, examines burstiness in tweeting activity to identify key "moments" of an event, and considers the dynamic nature of information. However, CEST leverages these traits in a single tool; identifies facts to be included in the corresponding summary, as opposed to selecting representative tweets; and includes a rating score generated as a result of sentiment change over time. Furthermore, CEST considers content, time progression, and geographical locations in order to isolate tweets that discuss events.

2 TWITTER USAGE AROUND POIS

We study Twitter usage around POIs to understand its characteristics, such as density and spatial distribution of tweets at a location. This initial study then informs the development of CEST.

The benefits of exploiting Twitter data to discover information about events has been well-documented [2, 29]. Stilo et al. [29] recently reported that event detection is typically treated as a topic modeling problem, which is a constraint, given that "not all topics are events". As an alternative, they propose a time-based clustering strategy for event detection. From our requirements' view, their strategy is yet to consider the valuable metadata of the geo-location of tweets. We do so by associating geo-location metadata with possible POIs and treating POIs as locations that people find appealing and where they go and send messages from, such as a tourist attraction or a theater. We assume that POIs can correspond to events by being the venues where events take place and thus we define events as real-life spatio-temporal phenomena that people are messaging about, which puts constraints in terms of size and duration and currently only considers single-location events.

Given that events tend to be short-lived and rarely re-occur [20], there is not known correlation between SNS usage at events and POIs. Few research works focus on SNS usage around POIs. Some researchers verified that users tend to post pictures around POIs [17], while others examined users' interests towards POIs based on their SNS activity in the vicinity [36] or explored SNS data to analyze travel patterns around local tourist destinations [37]. Existing techniques that analyze SNS usage around POIs [17, 30] consider geotagged data generated on Flickr or FourSquare, but no work has studied Twitter usage around POIs.

2.1 DGP Clustering Analysis

To examine SNS usage around POIs, we developed DGP (see Algorithm 1), an algorithm that considers two data sources: a geotagged dataset collected from a given SNS, e.g., geotagged tweets, and a set of POIs, e.g., set of specific venues or locations.

For analysis purposes, we built two small datasets: *Cities Tweets* and *Cities POIs*, described in Table 1. *Cities Tweets* includes a set of tweets collected on September 26, 2014 through Twitter's limited (1% of overall) streaming API by extracting the tweets that occur within the geographical coordinates of three sample cities. We require the explicit tweet geocoordinates, as the estimation of the Twitter bounding box may give us containment within the cities, but not the exact coordinates we need for the analysis. We selected these cities based on their range in population size. *Cities POIs* includes POIs (determined using tourist and city development websites) of the sample cities. As a POI usually corresponds to an area as opposed to a single coordinate [3], we draw a default bounding box around a POI with a radius of 0.35km.

Algorithm 1 DGP – Distribution of Geotagged data around POIs

```
input: Geotagged Dataset - GD, POI Dataset - PD
clusters = DBScan(GD)
clusterboundingboxes = null
for cluster in clusters do
   clusterboundingbox = Draw Bounding Box with
      extreme points in cluster
   add clusterboundingbox to clusterboundingboxes
end for
for POI in PD do
   for clusterboundingbox in clusterboundingboxes do
      if POIcoordinates is within clusterboundingbox then
         POI is detected
      end if
   end for
end for
```

City	New York	Chicago	Seattle
Population size	8.4 mil	2.7 mil	652,000
Metropolitan pop.	20.2 mil	9.52 mil	3.61 mil
# of POIs	365	67	62
# of Tweets	42,142	10,925	6,589
# of Flickr entries	39,811	20,801	19,379

Table 1: *Cities POIs*, *Cities Tweets*, and *Cities Flickr* datasets

For analysis of geotagged tweets, we used DBSCAN for spatial clustering and tuned the parameters based on an empirical study conducted across tweet samples of different cities. We selected parameters that yielded suitable clusters with a maximum of relevant points (*Eps*=0.002 and *MinPts*=5). Sample results are shown in Figure 1. Each resulting cluster is a collection of geographical points and more than one cluster may be created around a POI. To simplify the clusters and their spread of points, we generate bounding boxes for them. Then the overlap of cluster bounding boxes with each POI bounding box is calculated to measure SNS use for each POI.

To quantify city-wide SNS-at-POI usage, we compute *POI coverage* as the proportion of POIs detected in *Cities Tweets* $|GD|$ with respect to the number of known POIs (*Cities POIs*) in the city $|PD|$: $\textit{POI coverage} = |GD|/|PD|$. As shown in Figure 2, there is more than 73% coverage of POIs in New York, i.e., around three-quarters of POIs have nearby Twitter usage within a day. Coverage is around 45% and 68% for Chicago and Seattle. However, examining *Cities POIs*, we noticed that POIs in New York were centralized within few locations, which increases the likelihood of finding more tweets in that area. This leads DGP to find more than one POI in a single cluster. Given that there are fewer tweets available from Chicago and Seattle, fewer clusters around POIs in these cities were created.

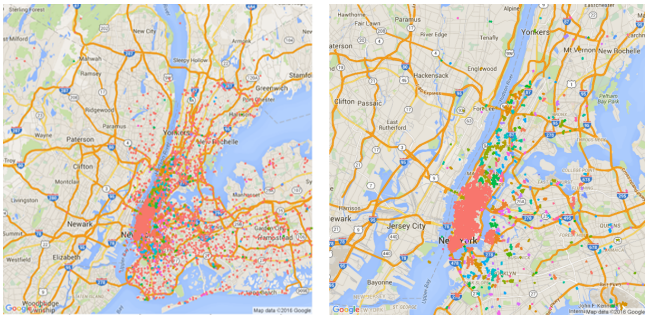

Figure 1: Tweets posted around New York City on September 26, 2014 (*left*) along with the corresponding clusters generated using Spatial Clustering (*right*)

2.2 Dynamic SNS Comparison

To put the Twitter analysis in context, we replicated the study using Flickr data, since its usage around POIs has already been demonstrated [17, 37]. We created *Cities Flickr* by collecting Flickr data from New York, Chicago, and Seattle also on September 26, 2014. As illustrated in Figure 2, Flickr was used around at least 70% of the POIs across all three cities. Even though New York has a higher population than Chicago, Flickr usage around POIs of Chicago is higher. We noticed from *Cities POIs* that Chicago has a higher percentage of scenic locations than New York, which may be a factor that can influence the results. Overall, usage around POIs is higher for Flickr than Twitter. However, the manner in which Twitter and Flickr samples were collected, including the volume of data points collected due to API rate limits and percentage of geotagging across different SNS (i.e., 2% for Twitter versus 4% for Flickr), can have an effect on the computed results[1].

We observed that not all tweets from a city correspond to POIs. To understand SNS usage across the area of a POI and confirm the validity of the POI coverage, we examined the relevant average area covered by SNS around a POI across different samples. We use a precision metric that computes the overlap of *POI bounding box* in *Cities POIs* with that of the *cluster bounding boxes* where a POI was detected in the samples. The precision computed across the clusters of samples of the three different cities is shown in Table 2. The results indicate that while SNS usage does not occur across the whole area around a POI, people are active within the region. We are aware that Twitter usage is independent of POIs. However, this preliminary study was designed to understand the relation of POIs and tweets as well as the general distribution of tweets. The study also allowed us to conclude that Twitter is sufficiently used around places of activities to allow event detection.

3 CEST: DETECT & SUMMARIZE EVENTS

We aim to develop a strategy that can detect any arbitrary event from Twitter without any prior knowledge. For an event to be detectable, it is necessary to identify a sufficient number of tweets and then verify the event. This leads to the challenge of determining which tweets are actually related and pertain to an event. In our approach, we use tweets with geolocation information to initially

[1]Due to the dynamic nature of SNS data, we conducted further experiments using a random sampling bootstrap technique and verified the validity of our results.

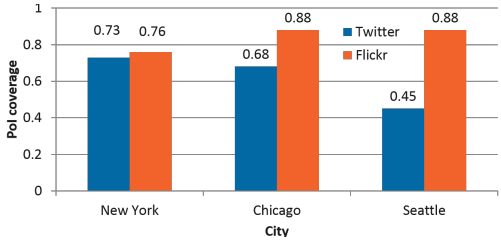

Figure 2: City POI coverage based on Twitter and Flickr

City	New York	Chicago	Seattle
Twitter	0.326	0.188	0.477
Flickr	0.843	0.689	0.786

Table 2: Precision of area coverage of POIs

detect spatio-temporal events, but include a spreading mechanism that identifies additional non-geotagged tweets related to those events for additional information. We are thus able to deal with events with both low- and high-density of tweets.

3.1 Twitter Data Collection

To validate and test CEST, we created a 9-month Twitter data collection, using Twitter's Streaming API[2], which allows tweet collection with and without filter parameters. Without parameter values, it collects a 1% random sample of all posted tweets. With parameters, it only draws from tweets that meet the specified criteria, not changing the sample size, but allowing the retrieval of more focused data. We created two datasets: *Public Tweets*, with tweets collected without filter parameter values; and *NY Tweets*, with tweets collected by setting a 'location bounding box' filter parameter to the coordinates of New York City. *Public Tweets* contains 4.2TB from Jan 2015 – May 2015 and *NY Tweets* contains 54GB from Aug 2015 – Nov 2015. For scalability, we stored these large datasets on Hadoop and used distributed processing. We also use some smaller datasets for aspects of our research, which we present later.

3.2 Event Detection Approach

Not every event occurs at a predefined location or POI. Recognizing events within a region over a period of time is therefore done through a specific pipeline. Previous work detected events by analyzing patterns of Twitter data generation [19, 25, 27]. However, these strategies often focused on specific categories of events, e.g., disaster-related events. We aim instead to develop a technique for detecting arbitrary events with no prior knowledge about what the event may be, which is needed to capture varied real-life events across a city. Our event detection technique, shown in Figure 3, depends upon spatial, temporal, and textual features.

We collected tweets (posted from the region specified by a bounding box) by 24-hour intervals, each denoted TD, which we partition into two subsets: TG as the set of tweets with geolocation, and TN as the remaining tweets in TD that can deliver additional descriptive details. To detect events of a city, we first determine regions with observed Twitter activity using TG. For each region, we extract

[2]https://dev.twitter.com/streaming/overview

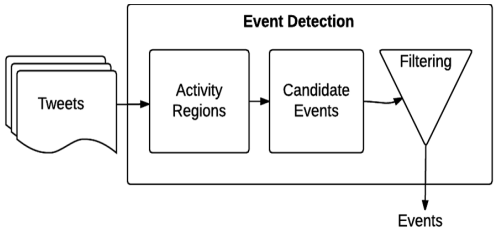

Figure 3: Overview of CEST's event detection approach

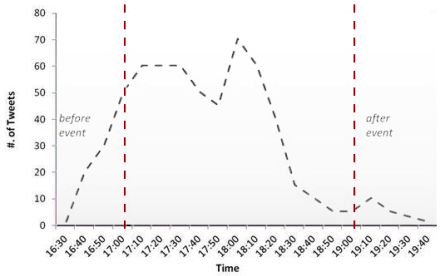

Figure 4: Frequency of tweets at a sample event: *Toronto Blue Jays versus Boston RedSox baseball game; June 2015*

the set of tweets that potentially correspond to an event, which we treat as a *candidate event*. These are filtered to identify those actually corresponding to real life events. Discussions corresponding to these events are incorporated as a result of further examining TN.

Determining Activity Regions. A base assumption is that places where multiple users congregate often represent an event [27], and that the spatial metadata of tweets reflects the physical locations of users [9]. We identify regions of activity where events can potentially occur by grouping tweets in TG based on the physical proximity of their coordinates using DBSCAN, as used previously in 2. Clustered tweets in TG are used for further processing.

Extracting Candidate Events. Not all clusters represent events and not all tweets in a cluster are related, i.e., multiple events can occur at a single location at different times. Since events in our definition last for a finite period of time and users tend to tweet at a higher rate within the event time bounds, (as exemplified in Figure 4), we use time series analysis on tweet activity in the clusters. Tweet frequency variations over time allow us to isolate subsets of tweets that may correspond to distinct events. Each cluster is monitored at regular intervals to determine if users are active in that region and show a minimum of activity. We overlay regular time slots over a cluster and use empirically determined values for *time chunk* as the minimum duration of an event and *MinPts* as the minimum activity within a time chunk. Tweet clusters that fulfill these limits are taken as *candidate event*. Contiguous candidate time slots of a cluster are merged to a single *candidate event*.

Filtering Candidate Events. We further analyze the frequency and user variation of tweets over time in candidate events, since depending on usage, there are certain scenarios that may lead our approach to mistakenly treat a cluster as a candidate event. For example, users and automatic systems may generate continuous streams of tweets that can lead our time-based strategy to mistakenly treat this tweet cluster as a candidate event. To further filter out candidates that do not represent an event, we define a set of criteria with empirically determined thresholds. A *candidate event* is no longer treated as so if: (i) 30% of associated tweets are generated from a single exact geocoordinate, as these are likely automated, (ii) 50% of tweets for the cluster are posted by only one user, to account for single users posting from coordinates in close proximity, such as their home or work, as well as a very limited number of users posting at high frequency, (iii) less than 5 distinct users tweeting about the event, since events should be shared by multiple users and show variety, or (iv) content among tweets is too dissimilar or exactly matching (based on a simple word-based Jaccard similarity). This latter criteria ensures that tweets are varied yet related, in terms of their content, and avoids treating as a candidate event a cluster created due to retweets, e.g., people who retweet commercial information or promotions.

Event Extension and Propagation. Since only 2% of tweets are geotagged [27], we are missing a majority of tweets that may discuss events and offer insight into the nature of the event, which is important for our summarization approach. For this reason, we examine non-geotagged tweets in TN and assign them to the corresponding candidate event, based on their timestamp, i.e., they have to be within the time bounds of the event, and textual content, i.e., they are sufficiently similarity to the candidate event tweets– based an adapted Jaccard distance on the frequency distribution of the terms per tweet, weighted towards the most representative ones.

3.3 Summarization Approach

A concise summary of an event detected from tweets offers users a quick overview of the most important event features and can be a better alternative than reading through streams of tweets. Significant research has been done on summarizing structured documents, such as blogs, news articles or longer documents [13, 16]. However, tweets are short and unstructured, so existing summarization strategies that apply to Twitter data usually work on tweets pertaining to specific types of events [8, 10, 31]. These techniques observe bursts within an event timeline to extract important moments of an event based on topic and keyword variations, or detect and summarize events by domain-specific keywords or hashtags. Alternatively, other strategies simply focus on identifying representative posts which are treated as summaries [4].

We rather aim at a general method that can work with events varying in their structure and associated actions. For example, a touch-down in a football game versus a keynote at a conference. Our generic event summarization strategy is applicable to create a brief overview of any type of event, even when no prior knowledge about unfolding events exists. To do that, we analyze temporal, spatial and user behavior corresponding to that event as well as text content and further metadata of tweets within a collection that discusses an event, denoted ET, i.e., a tweet cluster as identified in 3.2 which represents a candidate event. We pay special attention to the subset of tweets denoted EGT, which contains tweets with geolocation information. EGT is specifically analyzed to determine the location and extent of an event; the full set ET is used to extract the event's facts and details through a mix of methods. CEST's overall summary generation process is shown in Figure 5. We structure our distinctive summary constituents as detailed below.

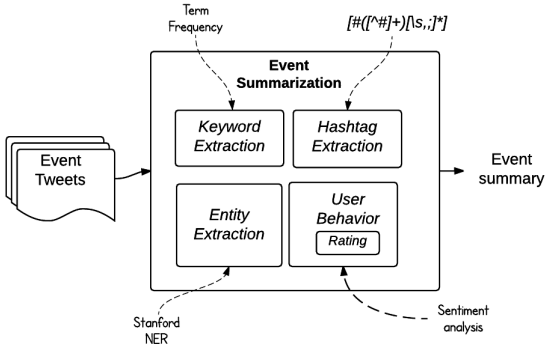

Figure 5: Overview of CEST's summarization strategy

3.3.1 Content Analysis.

Named-entity recognition (NER) is a subtask of information extraction that seeks to identify and classify elements in texts into pre-defined categories, such as "location" and "person". Events consist of a sequence of moments, each of which may refer to actors, e.g., a player in a sport event, or objects, e.g., a venue for the event. For entity recognition, we use the well-known Stanford NER[3]. As a large number of entities related to an event can be discussed in ET, we include in the event summary of only the most representative[4] based on their frequency distribution, to capture the entities that gained the most attention from people at that event.

Title, hashtags, and keywords provide a quick overview about the main purpose of an event. Tweets do not explicitly provide a title for the event they discuss. However, we can infer that information to a large degree. Given that Twitter users tend to use hashtags to refer to issues or topics pertaining to an event, we rely on the most frequently-used hashtags in ET to capture the title of the event. Hashtags are easily detected by their prefix of the '#' symbol. Based on hashtags' frequency distribution, the top-2 are used as the summary's event title. To showcase other prominently-discussed topics, we also include in the summary the top-k representative, i.e., most frequent, hashtags. Keywords in tweets may also provide descriptive opinions about different aspects of an event. We tokenize tweets in ET and analyze their frequency distribution, excluding hashtags and stopwords. The top-k most highly-mentioned keywords are deemed representative and added to the summary.

Location and time. To estimate the duration of an event, we use the timestamps of the first and last tweet in ET. Furthermore, we determine the event location based on the minimum bounding box estimated using the geographical coordinates in EGT. As mentioned in 2.1, an event location may not be represented by a single point. This estimation includes all tweets from users tweeting about the event within set thresholds (cf. 3.2) and will change in size depending on the type of event.

3.3.2 Behavior Analysis.

Participants sharing and expressing their opinions about an event are the source of all information we can collect. The number of distinct users (in ET) can reflect the popularity of an event [5] and is included in the summary.

[3]http://stanfordnlp.github.io/CoreNLP/

[4]As summaries are meant to (i) reduce the work load for the interpreter, (ii) maintain coherence and coverage, and (iii) include important aspects of the story [11], we include top-k representative terms, hashtags, and keywords.

Sentiment	# of Dictionary Words	# of Emoticons
Positive	2,252	199
Negative	5,082	228

Table 3: Dictionary created for analyzing users' opinions

User Behavior and people's interactions with Twitter and their opinions towards an event can dynamically change over the duration of the event. Their behavior can convey how they reacted to different moments of an event [21]. To capture these reactions, we analyze the non-stagnant behavior of users across the timeline of the event. Tweets in ET are partitioned into 5 minute time-chunks (each denoted TC_i). These time chunks form the basis for the analysis of tweets' burstiness and sentiment change over time.

Burstiness of tweets over an event's duration indicate important moments where higher participation of users is observed [8]. To illustrate the frequency distribution of tweets, we visualize the number of tweets in each TC_i as a graph in the event summary.

The *polarity* of user opinions indicates how people reacted to different moments at the event. To capture the varying polarity of users' opinions we rely on sentiment analysis. There are two well-known directions for measuring the sentiment: lexical approaches and supervised machine learning. The former rely on the creation of dictionaries comprised of words tagged with respect to their polarity, whereas the latter learn the patterns provided in a training dataset. Similar to the strategy detailed in [7], we favor a lexical approach, as it does not depend upon the existence of sufficient training data, which is a constraint given the short length and vague vocabulary commonly used on Twitter. However, our technique considers polarity not as a static value, but as a trait that can change over the lifetime of an event.

Tweets often include shorthand notations (e.g., gn for good night), negations (e.g., not good) and emoticons (e.g., :)), which play an important role in determining user sentiment. To handle these special terms, we created a dictionary that includes sentiment words previously compiled [15] and that has been used widely for sentiment analysis [24]. We expanded this list with shorthand notations and emoticons, which we manually compiled from Twitter data samples. Details on the sentiment dictionary are shown in Table 3. We believe the large number of emoticons and shorthand notations can be of special use and interest to the research community.[5] To showcase the sentiment development, we count the number of positive and negative terms in tweets that correspond to TC_i. The distribution of positive and negative sentiment across an event timeline is illustrated as a graph in the event summary.

3.3.3 Rating Generation & Analysis.

Especially for longer-duration events, a time-series overview of features of the event can be a relevant way for users to better understand it. A content-based summary that includes the development of the discussion in the tweets throughout the event's lifetime can be very informative, but there is another aggregation step that can be taken: estimating a rating that would capture attendee preference based on their tweets. We argue for a measure of the appeal of an event to its attendees, which should be a single qualitative appeal score. Numeric ratings as an assessment are common for objects such as a book or a movie, in terms of quality or quantity or

[5]The full list will be made available on GitHub.

both of that object. In fact, ratings are used by many applications, such as ranking factors for search and recommendation [26], and can also provide an easy overview for users that are dealing with a large number of events. Unfortunately, people do not use numeric ratings to qualify events on Twitter, which is why there is a need for inferring ratings by proxy through the sentiment expressed towards an issue or a moment of an event in individual tweets. We could then combine ratings inferred from tweets throughout an event into a single quantitative measure.

There are a number of strategies available for inferring rating scores [18, 23, 24, 33]. However, they are based on probabilistic models or machine learning, and depend upon analyzing sentiment expressed in single reviews, which tend to be well-formed (devoid of shorthand notations or emoticons) and are longer than the average tweet. Amazon reviews on average are about 582 characters[6], with a lower mean, compared to a maximum of 140 on Twitter. Alternatively, other work examines comments, which can be similar to tweets in length, but only generate binary rating inferences (i.e., "like" or "dislike"). None of the existing strategies consider time series analysis for inference, which is key to reveal overall trends of behavior [32] and thus the appeal of an event.

While our case is also an instance of the rating-inference problem [23], in contrast to previous literature, we propose to use tweets instead of reviews as the basis of the inference, and additionally, we explicitly consider the temporal aspect. It has been shown before that Twitter is a suitable corpus for sentiment analysis [22], but no work has looked at estimating ratings from it. Similarly, while some work has explored the temporal aspect for summarization [14], it has not yet been used to infer any form of rating.

We propose a novel strategy for rating inference that examines opinions conveyed on event-related tweets and quantifies the varying sentiments expressed by users at an event both by their temporal distribution and a single rating. Opinions are not static across an event. Hence, estimating a rating using static tweets but disregarding their time would not be effective. We have observed that the type and direction of opinion changes matter. For example, if users' opinions are positive at the start of an event but turn to be negative towards the end, it indicates that the event does not have a positive impact on people. On the other hand, when people have not expressed positive opinions in the beginning but they become extremely active and positive towards the end, it indicates that people enjoyed that event. With that in mind, we designed our rating strategy so that it explicitly considers *sentiment variation* across the timeline of an event. In order to capture the degree to which sentiment expressed during TC_i influences the overall rating of the event, we associated each TC_i with a monotonically rising weight that reflects the order in which TC_i occurs in the timeline of the event. For example, time-chunks TC_1 and TC_{10} are associated with weights 1 and 10, respectively, which indicates that sentiment expressed on tweets posted during TC_{10} influence the overall rating of the event more than the sentiment expressed on tweets posted during TC_1. This leads to the following calculation:

$$event\_rating = 5 * (\sum_{i=1}^{n} W_{TC_i} * S_{TC_i}) / \sum_{i=1}^{n} W_{TC_i} \quad (1)$$

where n is the number of time-chunks in an event, 5 is a normalization factor to bound ratings to a 0–5 scale, W_{TC_i} denotes the weight capturing the importance of TC_i in determining the rating of the event, and S_{TC_i} is a score that reflects the polarity of the sentiment expressed in tweets in TC_i, which is computed as follows:

$$S_{TC_i} = \begin{cases} 1, & \text{if } \sum_{\forall t \in TC_i} W_{p,t} - \sum_{\forall t \in TC_i} W_{n,t} \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where t is a tweet in TC_i and $W_{p,t}$ and $W_{n,t}$ denote the number of positive and negative sentiment terms in t. The number of possible sentiment terms for different time-chunks is not known in advance. Thus, we normalize S_{TC_i} by setting it to be 1 if the number of positive terms outnumbers the negative ones and 0 otherwise. $W_{p,t}$ and $W_{n,t}$ are graphed over time for the polarity variation visualization. These visualizations are shown on an example in 4.3 and Figure 6.

4 EVALUATION

Given that a number of strategies considered in our methodology are well-established, we do not evaluate each component. Instead, we focus our evaluations on CEST's event detection and rating-inference strategies. We also discuss CEST in practice in Section 4.3.

4.1 Event Detection Evaluation

We verify the accuracy of our event detection method by comparing the event extracted by our method to an external event dataset. We compiled a comprehensive list of 27 event sites for our target cities for broad coverage. Out of all events from these sites, we identified 65 events from the sports category that took place on different cities in one day. We then applied our event detection method based on cluster generation and time analysis, as described in 3.2, to Twitter stream data for that day and had it extract events. We compared the resulting events and their features with the preselected 65 events. Since our approach does not necessarily lead to the same titles of events as dedicated event pages, especially on noisy Twitter data, we could not do automatic matching. We therefore manually assessed the accuracy of the approach. Our event detection methodology identifies 56 out of 65 events, corresponding to a 0.86 recall[7].

The non-identified events are due to a couple of scenarios in which our filtering criteria were not met. For example, less than 5 users tweeted at events at a higher rate to promote the events, which caused the content similarity criteria to not be satisfied. In other events, users were not sufficiently active, i.e., tweets were posted at a lower rate than 5 tweets for 15 minutes, which led some criteria unfulfilled and thus the system discarded candidate events. It is an option for future work to improve low-frequency event detection and make thresholds more dynamic for such events.

4.2 Rating Evaluation via Product Reviews

Techniques used in creating event summaries, such as sentiment analysis and entity detection, are well-established [24]. However, our novel rating generation strategy based on event-related tweets needs to be validated to support its use in CEST. There is no previous work on estimating ratings for events from Twitter data, and

[6] http://minimaxir.com/2014/06/reviewing-reviews/

[7] We repeated this experiment over other sampled days and events with similar results.

Product domain	# of Reviews	# of Products
Baby	2,033,757	6,962
Watches	751,916	10,318
Beauty	2,772,616	29,004

Table 4: Amazon reviews dataset

	Absolute Error	
Product domain	All reviews	Only sentiment reviews
Baby	0.939	0.886
Watches	0.897	0.792
Beauty	0.810	0.717

Table 5: Absolute error for different product domains

therefore no baseline algorithm or benchmark dataset exist that can be used to evaluate our rating-generation mechanism.

We argue that an "indirect" analysis of the rating-generation strategy is possible by restating the problem as finding a dataset with similar characteristics: containing short texts with user opinions discussing an item along with a corresponding known rating, which can be used as a gold standard. An interesting source that fits these requirements is product reviews, which give us access to gold standard "ratings" for products together with time-stamped textual data with changing sentiment. While missing the explicit event character, reviews change over time and have a temporal distribution, which can roughly emulate an event. This evaluates mainly the aggregate rating, not individual sentiment. They have textual content plus a star rating that we can use to verify our own rating approach on the text. In addition, we can emulate the temporal nature of tweets by using the temporal distribution of reviews for a product, while simulating the product as a long-lived event. In short, we treat products as events with corresponding reactions for the evaluation purposes. This is an admittedly unorthodox, but viable and interesting way to test our ranking strategy on time-ordered sets of tweets in the absence of a ground truth.

We use different product domains from the Amazon Review Dataset [1] due to their variety and size (Table 4). Each dataset contains a list of products and corresponding attributes: *'review'*, *'userId'*, *'time'*, and *'rating'* (on a 1–5 scale). To emulate the temporal nature of tweets, we consider each review as belonging to a different time-chunk. To evaluate our strategy, we use *Absolute Error* to compute the difference between actual and predicted ratings.

As shown in Table 5, the error computed for products using reviews demonstrates that the rating predicted using our strategy differs by less than 1 star units with respect to average users' rating (i.e., rating provided by users to the products in the dataset). We have observed that the computed Absolute Errors are consistent across different domains, which further conveys that our strategy can predict a rating close to users' provided ratings. Our rating generation uses the polarity of a tweet or review content to infer rating. However, we identified that some of the reviews for products do not include words that express sentiment. We have also noticed that reviews are longer than tweets, which causes the ratio of non-sentiment words in reviews to be higher compared to tweets. To further evaluate the correctness of our approach, we applied our rating strategy only on reviews that include sentiment words. In doing so, we saw a consistent decrease in Absolute Error, which further demonstrates that our rating strategy performs better when applied on reviews with sentiment. Based on the results reported in this section, we verified the viability of our rating generation strategy. The findings ensure that this strategy can evaluate the appeal of users towards an event using Twitter data.

2015 U.S. Open Tennis women's quarter final
Labor Day Parade
Construction on #ELine BOTHDIR from Forest Hills-71st Avenue Station to Queens
Salman @ powerHouse book launch # Manhattan Bridge
West Indian American Day Carnival
Yankee Stadium baseball
Free Beer Tasting #actors #beer #laborday

Table 6: Top real-time events detected in NYC 8.Sep.2015

4.3 System Demonstration on New York Data

To show the validity of the approach and the successful integration of all described modules, we apply CEST to a set of tweets from New York City (NYC), sampled during September 8th, 2015. It consists of 123,112 tweets; 22,197 with geolocation and 100,915 without. CEST detected 83 candidate events, a number that is reduced to 74, after the filtering stage. Each set of tweets was manually examined to confirm it refer to a real-life event. As shown in 6, CEST finds events from a wide range of domains and categories.

Out of the 74 detected events, we demonstrate the results of applying CEST to the example of the *2015 U.S. Open Tennis women's quarter final*. The corresponding summary in Figure 6 is based on the information extracted from the 778 detected tweets in the NYC tweet dataset that discuss the US Open event. (The 502 tweets that have a geolocation associated with them were used by CEST to infer the location bounding box coordinates where the event took place.) Among the 331 hashtags used in tweet discussions about this event, the two most frequently-used ones are *#usopen* and *#usopen2015*. Both date and duration are correctly recognized. More than 300 unique users in the New York area contributed tweets. These users employ different hashtags, with *#tennis*, *#nyc*, and *#venus* as the most frequently-used. The most frequently discussed entities during the game are *Serena* and *Arthur Ashe*; respectively denoting a tournament player and the tournament venue. Similarly, *tennis*, *usopen* and *championships* are identified as recurrent keywords.

Besides the directly extracted features, CEST analyzes users' behavior along the lifetime of the event. We observe from the graphs and corresponding tweets that people are excited about Venus and Serena Williams playing together. We also notice that the positive sentiment within the game increases with the frequency of tweets, which illustrates that users are enjoying the game. In fact, at the time of the peak in both graphs, the hashtag *#Serena* was the most discussed and individuals' sentiments are extremely positive compared to other moments, which reflects her victory in the match. Finally, CEST computes a single qualitative rating for the event, which conveys the overall attitude of event attendees towards the game. Both players were toughly competing in the game with many positive tweets, but also certain moments with negative comments. The rating is calculated as 3.36, based on the dynamically changing sentiments of users across the lifetime of the whole event.

Figure 6: CEST's summary for the 2015 US Open finals

5 CONCLUSION & FUTURE WORK

We proposed DGP (Distribution of Geotagged data around POI) to understand and identify Twitter usage around different POIs in a city. We then presented CEST, a city event summarization tool that analyzes tweets posted from a given city. It automatically identifies events from Twitter streams and provides rich summaries of activities happening within a city, including sentiment variations and event rankings. CEST was built by combining multiple techniques for event detection and summarization with rich features over the lifetime of an event, including aggregated textual and spatio-temporal features, dynamic sentiments, and ratings.

We introduced a strategy that captures the overall appeal of an event based on analysis of sentiment and user behavior across an event's lifetime. Furthermore, we proposed a novel evaluation method for ratings of sets of tweets by exploring the cross-domain use of product reviews as a stand-in for tweets for an evaluation with no available ground truth. We also showed that CEST's strategies are generic and agnostic to categories, allowing it to capture emerging events from the Twitter stream without prior knowledge.

There are some limitations to address in future work. Given that one of the main contributions is our rating-inference strategy, we would like to conduct more detailed evaluation and refinement. This would require we turn to human judges to work on a testset of pre-defined events. We would also like to make the developed techniques more accessible by feeding it into other applications. For example, by combination with interactive exploration and map views, to give a full city overview of local events [35] or include it as part of recommendation approaches [20]. Lastly, we are interested in linking CEST to external knowledge sources, such as event calendars, not only to acquire more structured event information but also to examine dynamic aspects of event discussions and provide feedback on how events were perceived.

REFERENCES

[1] *Web data: Amazon reviews, http://snap.stanford.edu/data/web-Amazon-links.html.*
[2] Charu C Aggarwal and Karthik Subbian. 2012. Event Detection in Social Streams.. In *SDM12*. SIAM.
[3] Dirk Ahlers. 2015. Granularity as a Qualitative Concept for GIR. In *GIR '15*.
[4] Nasser Alsaedi, Peter Burnap, and Omer Farooq Rana. 2016. Automatic summarization of real world events using Twitter. (2016).
[5] Sebastien Ardon, Amitabha Bagchi, Anirban Mahanti, Amit Ruhela, Aaditeshwar Seth, Rudra Mohan Tripathy, and Sipat Triukose. 2013. Spatio-Temporal and Events Based Analysis of Topic Popularity in Twitter. In *CIKM'13*. ACM.
[6] Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31, 1 (2015), 132–164.
[7] Seyed-Ali Bahrainian and Andreas Dengel. 2013. Sentiment analysis and summarization of twitter data. In *CSE 2013*. IEEE.
[8] Deepayan Chakrabarti and Kunal Punera. 2011. Event Summarization Using Tweets. *ICWSM* 11, 66–73.
[9] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *CIKM*.
[10] David Corney, Carlos Martin, and Ayse Göker. 2014. Spot the ball: Detecting sports events on Twitter. In *ECIR'14*.
[11] Murthy Devarakonda, Dongyang Zhang, Ching-Huei Tsou, and Mihaela Bornea. 2014. Problem-Oriented Patient Record Summary: An Early Report on a Watson Application. In *Healthcom'14*. IEEE.
[12] Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding Bursty Topics from Microblogs *(ACL '12)*. 536–544.
[13] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *SIGIR*.
[14] Ruifang He, Yang Liu, Guangchuan Yu, Jiliang Tang, Qinghua Hu, and Jianwu Dang. 2016. Twitter summarization with social-temporal context. *WWW*, 1–24.
[15] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *SIGKDD*. ACM, 168–177.
[16] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *AAAI-CAAW-06*.
[17] Ickjai Lee, Guochen Cai, and Kyungmi Lee. 2013. Mining Points-of-Interest Association Rules from Geo-tagged Photos. In *HICSS*. 1580–1588.
[18] Cane Wing-Ki Leung, Stephen Chi-Fai Chan, Fu-Lai Chung, and Grace Ngai. 2011. A probabilistic rating inference framework for mining user preferences from reviews. *WWW* 14, 2 (2011), 187–215.
[19] Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *CIKM'12*. ACM.
[20] Sean MacLachlan, Stacey Donohue, Nevena Dragovic, and Maria Pera. 2016. "One Size Doesn't Fit All": Helping Users Find Events from Multiple Perspectives. *RecSys Workshop on Recommenders in Tourism (RecTour)*.
[21] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing Sporting Events Using Twitter. In *IUI'12*. ACM.
[22] Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.. In *LREc*.
[23] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.
[24] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2, 1-2 (2008), 1–135.
[25] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW'10*. ACM.
[26] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW'01*. ACM.
[27] Axel Schulz, Benedikt Schmidt, and Thorsten Strufe. 2015. Small-scale incident detection based on microposts. In *HT'15*. ACM.
[28] Beaux Sharifi, Mark-Anthony Hutton, and Jugal K Kalita. 2010. Experiments in microblog summarization. In *SocialCom*. IEEE, 49–56.
[29] Giovanni Stilo and Paola Velardi. 2016. Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery* 30, 2 (2016), 372–402.
[30] Panagiotis Symeonidis, Alexis Papadimitriou, Yannis Manolopoulos, Pinar Senkul, and Ismail Toroslu. 2011. Geo-social recommendations based on incremental tensor reduction and local path traversal. In *SIGSPATIAL*. ACM, 89–96.
[31] Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ujwal Gadiraju. 2013. Groundhog day: near-duplicate detection on twitter. In *WWW'13*.
[32] Mike Thelwall. 2014. Sentiment analysis and time series with Twitter. *Twitter and society* (2014), 83–95.
[33] Ivan Titov and Ryan McDonald. 2008. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *ACL'08*, Vol. 8. 308–316.
[34] Jianshu Weng and Bu-Sung Lee. 2011. Event detection in Twitter. *ICWSM*.
[35] Chaolun Xia, Raz Schwartz, Ke Xie, Adam Krebs, Andrew Langdon, Jeremy Ting, and Mor Naaman. 2014. CityBeat: Real-time Social Media Visualization of Hyper-local City Data. In *WWW '14 Companion*.
[36] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of- interest recommendation. In *SIGIR*.
[37] Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Mining travel patterns from geotagged photos. *TIST* 3, 3 (2012), 56.