7-2016

Can Instagram posts help characterize urban micro-events?

Kasthuri JAYARAJAH
Singapore Management University, kasthurij.2014@phdis.smu.edu.sg

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Citation

JAYARAJAH, Kasthuri and MISRA, Archan. Can Instagram posts help characterize urban micro-events?. (2016). *2016 19th International Conference on Information Fusion (FUSION): Heidelberg, Germany, July 5-8*. 130-137. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3639

Can Instagram Posts Help Characterize Urban Micro-Events?

Kasthuri Jayarajah, Archan Misra
School of Information Systems, Singapore Management University
kasthurij.2014@phdis.smu.edu.sg, archanm@smu.edu.sg

***Abstract*—Social media content, from platforms such as Twitter and Foursquare, has enabled an exciting new field of "social sensing", where participatory content generated by users has been used to identify unexpected emerging or trending events. In contrast to such text-based channels, we focus on image-sharing social applications (specifically Instagram), and investigate how such urban social sensing can leverage upon the additional multi-modal, multimedia content. Given the significantly higher fraction of geotagged content on Instagram, we aim to use such channels to go beyond identification of long-lived events (e.g., a marathon) to achieve finer-grained characterization of multiple micro-events (e.g., a person winning the marathon) that occur over the lifetime of the macro-event. Via empirical analysis from a corpus of Instagram data from 3 international marathons, we establish the need for novel data pre-processing as: (a) semantic annotation of image content indeed provides additional features distinct from text captions, and (b) an appreciable fraction of the posted images do not pertain to the event under consideration. We propose a framework, called EiM, that combines such preprocessing with clustering-based event detection. We show that our initial prototype of EiM shows promising results: it is able to identify many micro-events in the three marathons, with spatial and temporal resolution that is less than 1% and 10%, respectively, of the corresponding ranges for the macro-event.**

I. Introduction

Social media channels, such as Twitter and Instagram, provide a powerful crowd-sourced, *participatory* sensing channel for urban event detection and understanding. Most existing research has focused on the automated *identification* of such transient or unexpected events (e.g., Twitcident), typically using either statistical clustering [1] or generative models [2] on textual content from Twitter. In this paper, we focus on the opportunities and challenges of using multi-modal content from Instagram (a social image-sharing application) to *characterize* such urban events.

Instagram represents a rapidly growing and globally distributed image & video-centric social media channel (close to 500 million users). Instagram offers an interesting new modality for urban social sensing because of two factors: *(a)* its dominant content type is image-based (as opposed to the dominance of text in Twitter), thus offering new possibilities for applying image processing for content understanding, and *(b)* a significantly higher percentage of the Instagram posts are geo-tagged (with the coordinates of the user at the time of posting the image), providing easier and more reliable indicators of an event's spatial distribution. Moreover, content posted on Instagram is inherently *multi-modal* with posts containing short textual "captions" (a combination of regular words and hashtags) accompanying the images.

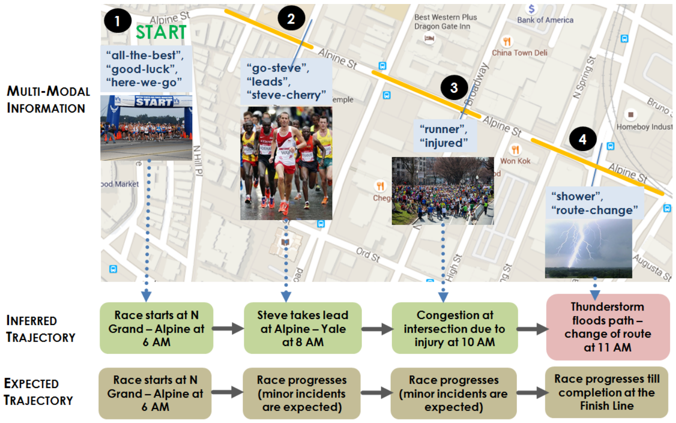
Multi-Modal Information
1
START
Alpine St
"all-the-best",
"good-luck",
"here-we-go"
START
2
"go-steve",
"leads",
"steve-cherry"
3
"runner",
"injured"
4
"shower",
"route-change"
Best Western Plus Dragon Gate Inn
Bank of America
China Town Deli
Won Kok
Homeboy Indust
N Hill Pl
Broadway
N Spring St
Ord St
High St
Augusta St
Inferred Trajectory
Race starts at N Grand – Alpine at 6 AM
Steve takes lead at Alpine – Yale at 8 AM
Congestion at intersection due to injury at 10 AM
Thunderstorm floods path – change of route at 11 AM
Expected Trajectory
Race starts at N Grand – Alpine at 6 AM
Race progresses (minor incidents are expected)
Race progresses (minor incidents are expected)
Race progresses till completion at the Finish Line

Fig. 1. Multimodal social network information for event understanding—a marathon example. The numbers 1 through 4, each represent a different stage in the race as the event progresses spatially, temporally and semantically.

Our work in this paper explores the possibility of using multi-modal fusion on such Instagram content to understand the *spatiotemporal* distribution of urban events. We specifically focus on characterizing the *when* and *where* of transient sub-events that are part of a large macro-event, rather than the macro-event itself. Figure 1 illustrates this concept: a marathon macro-event can be viewed as comprising multiple *micro-events*. At the initial stage (marked "1"), spectators and family members of the runners share images of the start line/runners with captions wishing the runners good luck. The race progresses to stage 2 where a lead runner emerges from the pack, while unexpected disruptions are caused by runner injuries at stage 3. A heavy downpour slows down the race at stage 4. We refer to this separation of stages 1 through 4 as the problem of detecting and characterizing micro-events. To focus specifically on this problem of micro-event characterization, we assume that the macro-event in question has a well-defined set of keywords/hashtags (e.g., #lamarathon for the Los Angeles marathon), and restrict our analysis to the corpus of Instagram posts whose captions contain these hashtags (and thus can be viewed as potentially related to the actual event).

Such micro-event characterization, especially if performed in near-real time, is extremely valuable for better understanding of macro-events, with applications in *anomaly determination* (e.g., inferring that the speed of runners has slowed down dramatically near a landmark, indicating likely human congestion causing the event trajectory deviate from the norm (Fig. 1)), *targeted emergency response* (e.g., dispatching law enforcement personnel to the specific street corner where vandalism seems to have broken out during a large music fair) and *causal understanding* (e.g., understanding that the sudden slowdown in pedestrian movement on a street is due to an overturned oil tanker).

Our technical objective in this paper is the development of a processing pipeline for extracting and characterizing such micro-events, from a combination of Instagram-related images and captions. A key aspect is the use of available image-analysis tools to provide semantic annotation of the image data, and use this to build a set of features defined over *both* the data and the meta-data.

Empirical analysis of Instagram posts related to 3 distinct, well-known marathon events help us to establish the following key **Research Challenges:**

- *Post Relevancy:* A significant fraction of images posted using an event's hashtag or keyword may either (a) be stock images (and thus do not really capture the real occurrences at the event) or (b) have only indirect affiliation to the event (e.g., someone posting an image of a past marathon event while watching the LA marathon on TV). It is thus important to filter out such extraneous content, prior to more careful content analysis.
- *Multi-modal Consistency:* The presence of both image and text content in a single Instagram post raises the question of consistency of whether the text labels and image content are semantically consistent–i.e., they refer to the same micro-event. For example, consider a post with the caption "good luck runners" but a picture of "having breakfast in a cafe". Our framework must thus associate event semantics with a post, based on *both* its text captions and the image.
- *Micro-event Separability:* Micro-events often occur simultaneously or with little spatiotemporal spacing. More interestingly, at least in our representative "marathon" events, the same micro-event can have significant spatio-temporal spread-e.g., different categories of runners start at different times and spectators stand dispersed along the entire route, causing the "marathon start" event to effectively span several hours before and after the actual start time (about 15 hours in our data). We will need to build improved discrimination/clustering techniques to account for such heterogeneity in the spatiotemporal span of individual micro-events.

In this paper, we present the initial version of our *Events-in-Motion* (**EiM**) framework for such micro-event extraction and characterization from image-centric social media channels. Using an Instagram data corpus captured from 3 distinct marathons in 2015 (Boston, London and LA), we make the following **Key Contributions:**

- *Pipeline for Post Relevancy, Semantic Extraction and Micro-event Detection:* To tackle the challenges mentioned earlier, we develop a framework that first performs pre-processing on the data (to eliminate irrelevant or unrelated posts), and then applies multiple data mining techniques on a broad set of metadata+ data features to identify events and their spatiotemporal boundaries.
- *Empirical Insights on Relevancy and Semantic Extraction:* We show that stock and non-relevant images can constitute around 30% of all event-related posts, and develop an image-similarity based method for extracting the likely set of original events posted from an event's location. Likewise, we also show that, in contrast to non-event related posts on topics such as "food", Instagram captions for events tend to have very little immediate semantic overlap with the semantics of the corresponding image. This observation underlines the importance of considering both caption-and-image based features for Instagram-based event analytics.
- *Empirical Evaluation of Alternative Micro-event Detection Schemes:* We formulate event detection as a problem of identifying distinct clusters over a multi-dimensional feature space, and evaluate two approaches–one that defines a vector space over all words+metadata, while the other considers LDA-based topic distribution over distinct spatio-temporal clusters. We then show that these approaches are quite promising: e.g., in the case of London Marathon in 2015, we detect the start of the race from Greenwich Park with a location error of 0.79 *km* within 30 *mins*. We also detected the winning moment of Wilson Kipsang with a location error of 2.70 *km* within 60 *mins* of his victory. The average spatial Euclidean distance error of 0.059 degrees across all micro-events is a substantial improvement over the macro-event (i.e., Earthquakes) detection error of 3.01 degrees reported in [3] using Twitter data. Overall, *EiM* achieves spatial and temporal resolutions of less than 1% and 10%, respectively, of the correponding ranges of the macro-event. Moreover, this approach also helps identify fine-grained *nano-events*–ones that have very few associated posts.

We emphasize that this work should be viewed as exploratory in nature–while our results are confined to the case of marathon events, we believe that our proposed *EiM* approach is promising and will be increasingly adopted to tackle various facets of urban event detection.

II. Dissecting an Event

We first introduce the formal notion of our micro-event characterization problem, given a corpus of posts that are associated with a specific, well-defined macro-event. Typically, the macro-event has a fairly large spatiotemporal spread–e.g., Instagram posts related to the micro-event of cheering the runners spanned a total of approx. 15 hours. Our goal is to discover and specify (as tightly as possible) the spatiotemporal range of uncertainty for an unknown number of *micro-events*–i.e., significant happenings that occurred, at different time instants, as the marathon gradually progressed from start to finish.

A. Problem Definition

Let P be the set of all posts containing a known, specific keyword that identifies a real world macro-event E. We

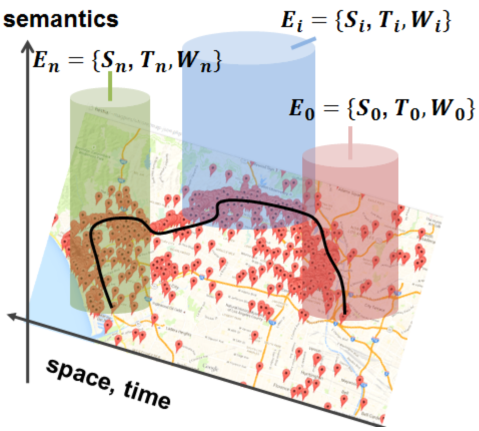
semantics
$E_i = \{S_i, T_i, W_i\}$
$E_n = \{S_n, T_n, W_n\}$
$E_0 = \{S_0, T_0, W_0\}$
space, time

Fig. 2. Space-Time-Semantic Space of an Event. Each E_i is a micro-event bounded by space S_i, time T_i and semantics W_i.

consider each post $p \in P$ as a tuple (s_p, t_p, w_p) where s_p is the latitude, longitude pair from where the content was posted, t_p is the time at which the content was posted and w_p is a bag-of-words representation of the semantics of the content. The content can be both the caption accompanying the post or a label representing the meaning of the image itself. Then E_i is a cluster (S_i, T_i, W_i) where $S_i \subseteq S$ is a set of coordinates in the universe of coordinates S, $T_i \subseteq T$ is a set of timestamps in the universe of timestamps T and $W_i \subseteq W$ is a set of words in the universe of words W that describes a meaningful, micro-event $E_i \subseteq E$, for i=0, . . . , n-1, where there are n such micro-events. We illustrate this in Figure 2.

B. Key Challenges

Leveraging crowd-sourced, multi-modal content for this problem encompasses several key challenges that we enlist below.

Spatio-temporal-semantic boundaries are not uniform: In order to describe *events in motion* as a sequence of micro-events, the boundaries of such need to be determined along the three dimensions. However, due to the varied nature of such events, this is not straightforward. For example, marathon events typically last for 2 to 5 hours over a total distance of 42 kilometers. However, an earth quake could shatter a large part of a country and the evolution of the event (from earthquake to disaster recovery to rehabilitation) could take several months. In this work, we mitigate this challenge by choosing unsupervised clustering and topic models to identify the most appropriate boundaries.

Relevancy: Social media content is noisy – not all posts shared by users, as pertaining to an event, in fact are relevant. As we describe in Section III-A, about 30% of the posts contain previously shared content such as stock photos and memes that aren't original.

Extracting Semantics of Images: Multimodal platforms are semantically richer as they provide both textual and image information about an ongoing event. We automatically extract labels representing the meaning of images using state-of-the-art computer vision techniques and show that the captions don't explain the content of the image and that it justifies the need to consider both (Section III-B).

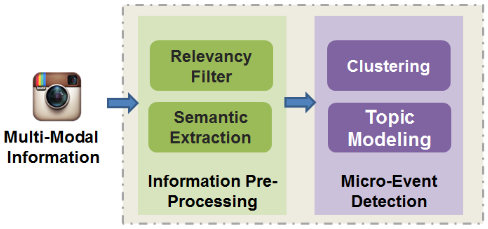

Fig. 3. High-Level Architecture of *EiM* Micro-Event Detection and Characterization.

Credibility: Of all the content shared, not all can be considered as "truthful representations" of the event. For example, a marathon enthusiast could be posting updates of progress on a race from a different part of the world altogether. The social relationships between users of the platform could also influence the credibility of a post – a user could simply reshare a friends' post without being physically there. We defer addressing this challenge as future work.

C. Our Approach

The following steps describe our overall approach in *EiM* for detecting micro-events.

Step 1: Relevancy Filter –First, of all the event-related posts, posts that contain images that are *irrelevant* or *unoriginal*, are discarded.

Step 2: Semantic Extraction –Using an existing deep learning library, the semantic labels of the incoming images are extracted which are then represented as a bag-of-words similar to captions.

Step 3: Feature Extraction –Each post is associated with metadata (coordinates in space and a time stamp) and content (bag-of-words) which form the set of all features (*STA*). For Step 4B, we consider the spatio-temporal features only (*ST*). For comparison purposes, we also consider *STL* which consists of space, time and bag-of-words from labels alone. The posts are preprocessed following the standards of tokenization, stop word removal and background words (words that are too frequent) removal.

Step 4A: Clustering –We represent each post as a word vector using the Vector Space Model with $TF-IDF$ weights in addition to the metadata dimensions. The features are scaled and normalized before clustering with either k-means (with Euclidean distance) or hierarchical clustering (with cosine similarity as affinity measure and average linkage). In Section IV, we refer to clustering using ALL features as *STA*.

Step 4B: Topic Modeling –As an alternative approach, we consider the topic distribution (using TwitterLDA [4][1]) over spatio-temporal clusters (k-means or hierarchical over *ST*) and refer to this approach as *ST+LDA*.

We illustrate this process in Figure 3.

D. Dataset Description

In this paper, we use three datasets pertaining to three different marathons events that happened in 2015 using which

[1] We use the implementation available from https://github.com/minghui/Twitter-LDA.

we evaluate our approaches. We consider ALL posts with the keywords listed in Table I. Table II lists the number of total posts collected and the percentage posts with accompanying geotags (at roughly 50%, significantly higher than that of Twitter which is typically less than 1%).

In addition, a secondary dataset consisting of a non-marathon event, and different categories of popular posts according to the taxonomy described in [5] is used to quantify the relationship between user-generated captions and auto-generated image labels. The dataset consists of 2000 randomly sampled posts from each category originating from Singapore.

Category	Observation Period	Keywords Used for Filtering
Marathon Events		
Boston	April 2015	“bostonmarathon”, “baa”
LA	March 2015	“lamarathon”
London	April 2015	“londonmarathon”, “vmlm”
Other Event		
F1 SGP	September 2015	“f1 ”,“race(s) ”,“racing ”, “formula”
Non-Events		
Food	September 2015	“food”,“yummy”, “recipe”, “delicious”
Pets	September 2015	“dog(s) ”, “cat(s) ”,“puppy ”,“doggy”
Fashion	September 2015	“fashion”, “style”,“trend”,“outfit”
Selfies	September 2015	“selfie”,“friend”,“fun”

TABLE I
SUMMARY OF THE DATASETS USED IN THIS WORK

III. EMPIRICAL OBSERVATIONS

In this section, we provide early insights into two of the key challenges we identified: establishing relevancy of a user-posted image and whether the image provides additional information orthogonal to the user-provided captions.

A. *Establishing Relevancy in Images*

We have observed user posts that carry images that are generic (e.g., memes), or were available from similar events in the past, although they may contain keywords specific to a contemporary event. It is also common for users to share/re-share images from blogs and news media; although these images may be relevant, they are less reliable as they are not the original thought/experience of the poster. To discard such images from further analyses, we propose a technique that consults a large corpus of images to identify whether a user-posted image is an already available image on the Internet by comparing its syntactic similarity against those available in the corpus. In this work, we use *Google Search*[2] as our primary corpus. We describe the steps in brief below.

1) For each user-posted image, or query image, q_k, a reverse image search is performed programmatically against the search engine. The search results page is then systematically scraped. In our implementation, we perform an X-path search on the DOM of the page returned.
2) The results returns pages that contain the image (if a match is found) and a set of visually similar images. If a page is found, then the user-posted image is deemed as a “stock” image with a $reality-score = 0$.
3) If matched pages are not found, then for the top-k visually images returned, we compute the Perceptual Hash [6] of both the query and result images and then compute the respective Hamming distances between the images. The perceptual hash is robust against simple image manipulations (e.g., rotation, scaling, borders, cropping, etc.). For a 64-bit hash, we take the $reality-score$ as $\frac{h}{64}$ where h is the minimum distance.

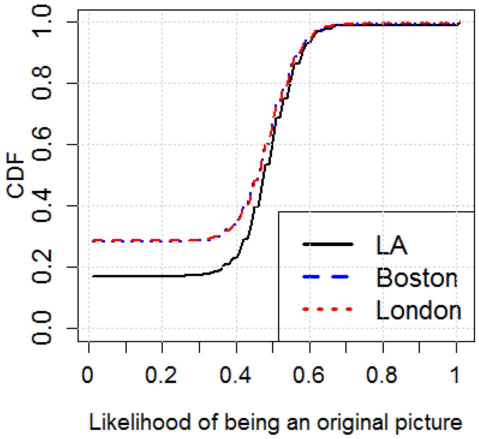

Fig. 4. CDF of $reality-score$. About 20-30% of the posts were found to have exact matches ($reality-score = 0$) against the search engine.

Dataset	# Posts	# Posts with Geotags	# Posts Detected as Original
LA	4640	2652 (57.15%)	3846 (82.88%)
Boston	7742	4176 (53.94%)	5523 (71.34%)
London	8474	4022 (47.46%)	6192 (73.07%)

TABLE II
Marathon DATASET SUMMARY.

In Table II, we list the percent posts that contained original images at a $reality-score = 0$ – i.e., exact matches. Further, in Figure 4, we plot the CDF of the $reality-score$ for each dataset. We observe that about 30% of the posts have a score equal to 0. To evaluate the effectiveness of the scoring, (1) we randomly sample 100 images from each set and (2) and two annotators manually labeled whether the images are stock or not. We report the average precision/recall values in Table III along with the κ coefficient. Overall, we observe very high precision (close to 1) and moderate to substantial agreement between the annotators. We attribute the drop in recall to the non-negligible number of false negatives. Further investigation revealed that such misclassification was caused by: (1) images with popular landmarks in the background (e.g., Big Ben during the London marathon), (2) images with specific products as the main subject (e.g., Adidas merchandise during the Boston marathon), and (3) aggregation sites which provide summary versions of Instagram posts for popular hashtags.

B. *Do captions convey what's in the image?*

Each post consists of two modes of content – text from the captions and the image content itself. In this section, we attempt to quantify the relationship between the two. In order to extract the semantic meaning of images, we use

[2] https://images.google.com/

Dataset	Recall	Precision	Cohen's κ
LA	0.906	1.000	0.535
Boston	0.769	1.000	0.682
London	0.776	0.985	0.674

TABLE III
Accuracy of Detecting Original Content in the *Marathon* Dataset with $reality - score = 0$. The loss of recall is attributed to the non-negligible number of false negatives. The κ value shows moderate to substantial agreement between the annotators.

a multimodal recurrent neural network, NeuralTalk [7][3]. It combines both object detection and the inter-object spatial relationship to generate sentence-like labels.

In understanding the relationship, we compute the following measures of similarity between the caption and the corresponding label:

Lexical Similarity (S_w): We measure this as the Jaccard similarity between the two bags-of-words in the sources: caption and label for each of the post.

Semantic Similarity (S_s): We represent each word by its word sense (i.e., synsets from WordNet [4]). Then, the semantic similarity between the two short phrases/sentences is calculated based on the path similarity between individual synsets (i.e., "based on the shortest path that connects the senses in the is-a taxonomy") and word order similarity in a sentence, as proposed in [8].

In Table IV, we tabulate the size of vocabulary (V_c for captions and V_l for labels), mean and standard deviation of lexical similarity, and the mean and standard deviation of semantic similarity, for the three marathon datasets, the non-marathon event (F1 SGP), and the four categories of popular Instagram posts. We make the following observations:

1) Lexical vs. Semantic similarity: across all datasets, we see that the semantic similarity is at least 100 times more than when considering plain word-to-word overlap.
2) Food vs. other categories: The class of *Food* shows the highest agreement between the captions and labels even thought its vocabulary is much richer (63, 736 distinct words in captions).
3) Marathon vs. non-marathon event: The F1 SGP dataset shows the lowest degree of agreement. We believe that this particular event would lack diversity in image content with most images containing the track, cars and crowds whereas the captions would be diversified.
4) However, we still note that the overall similarity between captions and labels is low ($\approx 7-12\%$) justifying the need to consider both captions and labels for accurate analyses of semantics.
5) We also note that the low similarity values could be attributed to the discrepancy in the language forms; posts may contain *casual* words with colloquial terms, abbreviations, hashtags with multiple words appended together, etc. in contrast to the formal structure of the auto-extracted labels.

Dataset	V_c	V_l	S_w	S_s
LA	14652	350	0.0017 (0.014)	**0.1135** (0.099)
Boston	21147	362	0.0015 (0.013)	**0.1233** (0.101)
London	18027	363	0.0021 (0.015)	**0.1289** (0.104)
F1 SGP	12767	299	0.0002 (0.004)	**0.0714** (0.072)
Food	63736	275	**0.0167** (0.021)	**0.1663** (0.087)
Pets	7613	302	0.0115 (0.023))	**0.1207** (0.084)
Fashion	34009	361	0.0015 (0.008)	**0.1075** (0.077)
Selfies	41897	330	0.0015 (0.009)	**0.1094** (0.082)

TABLE IV
Size of the vocabulary and lexicographic and semantic similarity between the captions and auto-generated labels for each of the dataset. 'Food" category shows the highest overlap despite its richer caption vocabulary. With an average semantic overlap of only 12% across the Marathon datasets, we use both captions and labels in EiM.

IV. Evaluation

In this section, we provide insights into the choice of parameters and our observations from evaluations using the two approaches described in Section II. In particular, we seek to understand the following:

1) Choice of clustering algorithms and number of clusters that offer the greatest clustering quality
2) The average case location and time estimation accuracy of micro-events

A. Parameter Selection

In formulating a clustering problem, the number of clusters and the quality of the resulting clusters are key concerns. To *choose* the optimal number of clusters for each of the marathon events, we varied the number of clusters k, and observed the Silhouette coefficient which is a measure of cluster quality. In our analyses, we used two fundamentally different clustering algorithms, namely, k-means and hierarchical clustering. Further, we observed these values for the different sets of features described previously (i.e., *ST*, *STL* and *STA*).

In Figure 5, we plot the Silhouette Coefficient (on the y-axis), for varying number of clusters (on the x-axis). We make the following observations: (1) across all three feature sets, hierarchical clustering outperforms k-means – this is likely due to the nature of k-means which favors spherical clusters, whereas in our case, it is not necessary that the clusters maintain this shape, and (2) the *STA* (all features) features perform the worst lending to its sparse, high-dimensional form. For the remainder of the analyses, we choose the hierarchical clustering algorithm with k set to 200 and 300 for *ST* (spatio-temporal features only) and *STA*, respectively.

B. Results

We evaluate the two approaches *STA* and *ST+LDA*, in terms of known start times of the marathon events [5], and calculated finish times based on the winners' run time [6]. We assert that the ground-truth location coordinates and timings are only approximate. For each cluster, we first find the most representative bag-of-words and then for chosen clusters (that are contain specific keywords indicating that they represent a

[3] We use the implementation available from https://github.com/karpathy/neuraltalk.

[4] http://www.nltk.org/howto/wordnet.html

[5] http://is.gd/boston15, http://is.gd/la2015, http://is.gd/london15

[6] http://is.gd/bostonwinners, http://is.gd/lawinners, http://is.gd/londonwinners

Micro-Event	Apprx. Location	Apprx. Time	Location Error in *km*		Time Error in *mins*	
			ST+LDA	*STA*	*ST+LDA*	*STA*
Boston						
Cheering	Hopkinton	8:50 - 11:15	5.22 (0.066)	13.03 (0.151)	66.93	68
Winners	Public Library	11.56 - 12:09	5.49 (0.062)	3.89 (0.047)	50.03	332.18
London						
Start of race	Greenwich Park	9:00 - 10:10	0.79 (0.027)	5.77 (0.079)	31.37	185.55
Winners	The Mall	11:43 - 12:14	2.70 (0.007)	5.44 (0.071)	57.03	56.46
LA						
Cheering	Dodger Stadium	6:30 - 6:55	3.62 (0.039)	3.88 (0.035)	93.01	99.53
Winners	Santa Monica	9:05 - 9:17	10.48 (0.094)	10.65 (0.098))	333.85	34.24

TABLE V
ERRORS IN ESTIMATING LOCATION AND TIME OF *known* MICRO-EVENTS OF THE THREE *Marathon* EVENTS. THE VALUES WITHIN BRACKETS ARE THE EUCLIDEAN DISTANCE ERRORS IN DEGREES.

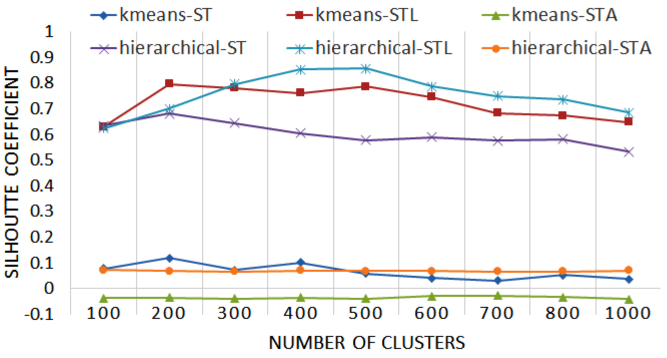

Fig. 5. Cluster Quality vs. Number of Clusters – hierarchical clustering outperforms *k*-means across all three feature sets. A common *k* value of 300 is chosen for the remainder of the analyses.

start or win micro-event of the race), we compute the location and time estimation errors based on the ground-truth.

*1) **Representative Bag-of-Words of Clusters**:* In the case of *ST+LDA*, each cluster consists of several posts each assigned to a specific topic. Each such topic has a pre-computed distribution (called proportions) over the universe of words. Hence, to arrive at the most dominant bag of words for that cluster, we choose the top-*k* words with the highest cumulative proportion within that cluster. If $topic(i)$ is the topic of post i, and $topic(i)$ has proportion $p_{i,j}$ over word w_j, then the cumulative weight of the word in that cluster is $\sum_i p_{i,j}$. Then, the dominant words of that cluster are those with the highest such sum.

In the case of *STA*, each post in each cluster has a vector form. Hence, we take the centroid of this cluster as the most representative vector for that cluster, and choose the top-*k* words with the highest weights as the dominant bag of words. In both cases, we choose $k = 10$.

*2) **Error estimation**:* We manually selected clusters that semantically match the ground-truth micro-events (based on generic keywords such as "elite", "cheer", etc. for race starts and "finishline", "finish", etc. for race ends), and computed location and time errors as the haversine distance and absolute distance from the cluster centroid(s), respectively.

In Table V, we tabulate the observed errors for the cheering/start and win/finish micro-events (for which it was possible to infer approximate ground-truth). In Table VII, we provide top keywords retrieved by the two approaches for some interesting micro-events. Overall, we achieve a Euclidean distance error of 0.059 degrees across all micro-events with a substantial improvement over the 3.01 degree error reported in [3] using Twitter data. We make the following observations:

Cheering along the route: The cheering micro-events are the most common and largest clusters observed in our datasets (e.g., in the Boston set, 35 clusters with an average size of 94 with the term "luck"). As such, they are also the least discriminative. For instance, spectators along the entire route of the marathon post content indicative of cheering or wishing the runners good luck. Also, the wishing and cheering from the anticipating crowd starts well before the stipulated start time and continues as the race progresses. Hence, we see high variability in both space and time.

In particular, we observed the spatio-temporal variation of the terms "luck" (top keyword for cheering) and "desisa" (a top keyword for winning) and fit a Gaussian models along the three dimensions latitude, longitude and time after removing outliers. In Table VI, we list the standard deviations seen in latitude, longitude and time and observe the estimation errors with respect to this variability. For example, *ST+LDA* has a best case location error of $\approx 5.22km$ and time error of $\approx 66.93mins$ in the case of the Boston Marathon. However, compared to the variability, this translates to only 0.15% - 0.35% in location error. Interestingly, in the case of the London marathon, discriminative keywords (e.g., "Greenwich" and "charity" – the race had three separate start points and some charity runs started at specific points) results in a dramatic drop in location errors down to $\approx 1-6km$. The time performance also improves – the improvement is more significant in *ST+LDA*.

Similarly, in the case of finishes, spectators use generic keywords such as "finish line" from the beginning of the race and well after the winning time (which is our ground-truth), and not necessarily towards the end or when a winner emerges. This is apparent in the case of the LA marathon where the best case location error is $\approx 11km$. As noted earlier, with the use of discriminative keywords indicating the actual winner (e.g., "Lelisa", "Desisa"), we see that the error reduces significantly ($\approx 5km$ for both Boston and London).

Highly Discriminative Sub-Events: We observe a number of *rare* clusters that represent interesting micro-events. Due to the lack of reliable ground truth, we are unable to provide error estimates. Inspirational posts of the first ever woman runner in the Boston Marathon (in 1967), Kathrine Switzer, was detected along with keywords suggesting her audacity ("fearless", "adversity") and women empowerment. Another interesting cluster emerged as Tatyana McFadden emerged as winner in the wheelchair runners category. Rebekah Gregory, a survivor from the 2013 bombings, completed the race on prosthetic legs which was also picked up as a small cluster.

Late Finish: Maickel Melamed completed the Boston Marathon in 20 hours, well after the spectators and organizers had dispersed. This story first emerged as a news article [7], and not surprisingly, the location estimates and time estimates are far from ideal. We observe a average error of 1162 *km* and a standard deviation of 732.28 *km*.

We conclude that with the presence of discriminative keywords, the two approaches are able to detect micro-events with less than 5km error and an hour delay, on average. We also realize the need for solutions to the orthogonal problems of automatically (1) identifying which words represent which micro-event (micro-event classification) and (2) identifying the most likely cluster for each micro-event (in the presence of multiple clusters). We defer this for future work.

Micro-Event	**Keyword**	SD_{lat}	SD_{lon}	SD_{time}	E_{lat}	E_{lon}	E_{time}
Cheering	"luck"	7.81	17.12	875.95	0.15%	0.35%	7.64%
Winning	"desisa"	5.47	13.02	802.98	0.17%	0.50%	6.23%

TABLE VI
SPACE AND TIME CHARACTERISTICS OF KEYWORDS FROM THE BOSTON MARATHON. THE STANDARD DEVIATIONS OF LATITUDE AND LONGITUDE ARE IN DEGREES AND OF TIME IS IN MINS. *SD*-STANDARD DEVIATION, *lat*-LATITUDE, *lon*-LONGITUDE, *E*-ERROR AS PERCENTAGE OF STANDARD DEVIATION.

V. Discussion and Future Work

Current Limitations and Future Work: In detecting non-stock images, we observed that most errors resulted from misclassification of original content as stock content due to the presence of "landmark" objects or backgrounds. To improve the performance (reduce the misclassification of event images as stock photos), we intend to investigate the use of time-distance based filtering – i.e., if two pictures were posted close together in time, although sharing the same background, due to their proximity in time, they could both be in fact original. Another possible approach is to introduce additional background metadata (e.g., whether the day in concern was cloudy or sunny) in the classification process. Moreover, in Section III-B, we show that the vocabulary of the auto-generated labels is much smaller than the size of the captions. This is an inherent limitation of the training corpus of images and sentences used in extracting the labels – hence, a larger scale training with diverse images and multiple human annotators could help in significantly improving the richness of the labels.

Given our focus on establishing some baseline measures in this paper, we have intentionally limited ourselves to relatively simple clustering and topic modeling techniques. In future, we plan to expand on the topic model (to include both space and time as additional variables) which will then allow us to estimate the geotags of untagged posts (which is about 50% of the total number of posts). By using such spatiotemporal distance features, we can then include the untagged posts in the processing pipeline, instead of simply discarding them. We anticipate that this will let us recursively fine-tune the model for better performance. Further, in dealing with very high dimensional data, we also intend to consider dimensionality reduction and covariance between word dimensions to understand its effect on performance.

Open Problems: In our current work, we attempted to tackle the challenge of establishing credibility in user posts using a outlier detection approach. We identified local outliers of every micro-event cluster using the Local Outlier Factor [9] algorithm. However, the initial results did not show promise. In our discussions of semantic similarity between captions and labels, although we assert that the two are *different*, we did not consider whether the two statements *corroborate* each other or are in *conflict*. This remains an interesting question on whether such contradictions *hint* at a lack of credibility in the post.

In Section IV, we introduce two orthogonal problems. In evaluating the location and time accuracy, first, we manually identified which bag-of-words represents which micro-event (e.g., cheering vs. winning). Here, the development of domain-specific ontologies and knowledge representations can automate the process by classifying or labeling the micro-event based on the bag-of-words. Second, we manually chose the best cluster amongst the set of clusters that identifies a micro-event (i.e., one that minimizes the errors). This can be formulated as a path estimation problem to identify the best cluster for each micro-event to find the set of clusters that maximizes the overall probability of the event trajectory.

VI. Related Work

Event Detection and Tracking using Social Media: Event detection from user-generated social media content (e.g., Twitter feeds) is a widely studied topic. Events are detected primarily by identifying changes in the frequency distribution of usage of hashtags or keywords (i.e., volume of usage) using a combination of machine learning techniques [3]. In Walther [1], spatially localized events such as house fires or parties are detected by first setting up spatial filters to identify clusters of tweets and then using a combination of topic and semantic analyses to identify events. In contrast, Twitcident [10] focuses on monitoring events; here, it is assumed that events to be monitored are known a priori and Twitter data are then mined to monitor those events. Some work, such as [11], focused on the classification and tracking of such events. While many solutions rely on pre-computed vocabularies for such classification, practical systems should develop and update their vocabulary autonomously, which remains an important research challenge.

Spatio-Temporal Topic Discovery: Previous work in text analysis research have demonstrated the use of techniques such as Space-Scan-Statistics ([12]) and topic models involving space, time, user attributes and semantics[13] in uncovering spatio-temporal topics. Their focus was mainly on *detecting* large-scale events.

Complementary to the above work, we envision to dissecting such events to uncover micro-events or event stages to understand the evolution of the event.

[7] http://is.gd/maickelarticle

Micro-Event Type	Keywords from STA	Keywords from ST+LDA
Starts	"elite", "cheering", "elitemen", "nearlythere"	"elite", "cheering", "big", "day", "amazing", "luck"
Finishes	"finish", "finishline", "congrats", "personalbest", "desisa"	"finish", "finishline", "lelisa", "desisa", "rotich"
Last runner	"maickelmelamed", "venezuela", "muscular", "dystrophy"	"maickelmelamed", "venezuela", "boylston", "spirit", "story"
First woman runner	"kathrine", "switzer", "adversity", "women"	"switzer", "women", "kathrine", "fearless"
Wheelchair runner	"tatyana", "womens", "wheelchair", "mcfadden"	Not picked up as a dominant word/topic

TABLE VII
OBSERVATIONS FOR DIFFERENT TYPES OF MICRO-EVENTS FROM THE BOSTON MARATHON.

Multimedia in Social Sensing: Although there has been a flurry of research in leveraging crowdsourced textual content for a variety of applications, the same does not hold true for multimodal platforms such as Instagram. Early works in exploiting Instagram have primarily focused on two areas: (1) characterization of posts and workload and (2) empirical analyses on cultural and travel aspects due to the higher availability of geo-coded information as opposed to its text-based counterparts (e.g., Twitter). In [14], the authors characterize user posting behavior, spatially and temporally, over a large dataset of 2.3 million images, and demonstrate the use of such characterization by building a tool for recommending regions of interest. While [5] looks at developing a taxanomy of the user posts, [15] tries to answer key questions related to what makes a post more popular. While [14] employs only the associated metadata, [5], [15] analyze the image content similar to our work. Our work differs in the fact that we combine both metadata and multimodal semantics to understand urban events and their evolution.

In this work, we expand on our initial vision [16] for a multimodal, social signal processing framework.

VII. Conclusion

In this paper, we have introduced and discussed the problem of characterizing micro-events (i.e., transient occurrences that occur within a larger event, and that are highly localized to specific neighborhoods and time intervals) via social sensing of image-sharing social networks–specifically, Instagram. Our proposed *EiM* framework includes novelties at both the data preprocessing and subsequent event detection stage. During data preprocessing, image similarity based measures are used to weed out irrelevant or non-live photos, while tools from image semantic labeling are used to provide a richer feature set for Instagram posts (quite distinct from the captions used in such posts). Subsequently, we use both a generative model and a statistical clustering based approach to identify such micro-events. Results show that the generative model, *ST+LDA*, is able to bound events with spatial errors that are less than 1-2% of the reported location tags and less than 10% of the reported time interval. While more work is clearly needed to improve the localization of such events (perhaps by considering semantics-based *inter-event relationships*), our work establishes the importance of including *semantics of image content* in using Instagram feeds as a social sensor channel for event characterization.

VIII. Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory, under agreement number FA2386-141-0002, and partially supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre @ Singapore Funding Initiative. Kasthuri Jayarajah is supported by an A*STAR Graduate Scholarship. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References

[1] M. Walther and M. Kaisser, "Geo-spatial event detection in the twitter stream," in *Proc. of ECIR'13*.

[2] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models:\# twitter trends detection topic model online." in *COLING*, 2012, pp. 1519–1534.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. of WWW'10*.

[4] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, *ECIR 2011*, ch. Comparing Twitter and Traditional Media Using Topic Models.

[5] Y. Hu, L. Manikonda, S. Kambhampati *et al.*, "What we instagram: A first analysis of instagram photo content and user types." in *Proc. of ICWSM*, 2014.

[6] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," *Masters' Thesis, Fraunhofer Institute for Secure Information Technology.*

[7] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[8] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. on Knowl. and Data Eng.*

[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2, 2000, pp. 93–104.

[10] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams," in *Proc. of HT'12*.

[11] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proc. of SIGKDD'12*.

[12] 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 SIGKDD'13*.

[13] T. Cheng and T. Wicks, "Event detection using twitter: A spatio-temporal approach," in *PLoS ONE, 9(6), e97807.*

[14] T. H. Silva, P. O. S. V. d. Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "A picture of instagram is worth more than a thousand words: Workload characterization and application," in *Proc. of DCOSS'13*.

[15] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *Proc. of CHI'14*.

[16] K. Jayarajah, S. Yao, R. Mutharaju, A. Misra, G. D. Mel, J. Skipper, T. Abdelzaher, and M. Kolodny, "Social signal processing for real-time situational understanding: A vision and approach," in *Proc. of MASS'15*. IEEE, 2015, pp. 627–632.