Pervasive and Mobile Computing 26 (2016) 103–120

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Recognizing composite daily activities from crowd-labelled social media data

Zack Zhu *, Ulf Blanke, Gerhard Tröster

Wearable Computing Lab, ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Available online 13 October 2015

Keywords:
Web mining
Activity recognition
Crowd sensing

ABSTRACT

Human activity recognition is a core component of context-aware, ubiquitous computing systems. Traditionally, this task is accomplished by analysing signals of wearable motion sensors. While successful for low-level activities (e.g. walking or standing), high-level activities (e.g. watching movies or attending lectures) are difficult to distinguish from motion data alone. Furthermore, instrumentation of complex body sensor network at population scale is impractical. In this work, we take an alternative approach of leveraging rich, dynamic, and crowd-generated self-report data from social media platforms as the basis for in-situ activity recognition. By treating the user as the "sensor", we make use of implicit signals emitted from natural use of mobile smartphones, in the form of textual content, semantic location, and time. Tackling both the task of recognizing a main activity (multi-class classification) and recognizing all applicable activity categories (multi-label tagging) from one instance, we are able to obtain mean accuracies of more than 75%. We conduct a thorough analysis and interpret of our model to illustrate a promising first step towards comprehensive, high-level activity recognition using instrumentation-free, crowdsourced, social media data.

1. Introduction

Human activity recognition (AR) provides the basis for developing context-aware services and applications. Novel applications are recently surfacing to provide just-in-time information. A well known example is the commercial product Google Now,[1] which learns the daily routine of the user to provide relevant information like local weather or driving directions.

Intuitively, two dominant signals for such context-aware services are location and time, which can already provide rough estimates to infer simple and non-specific activity routines like "working" or "staying at home". To detect more fine-grained activities, Inertial Measurement Units (IMUs) are popularly employed. Using such sensor packages, often worn by the user, accelerometers, gyroscopes, and sometimes magnetometers acquire the user's motion and thereby his physical movements. In this way, researchers have been investigating the recognition of various activities, ranging from low-level ones (standing, sitting, or walking) [1] to physical activity (cycling or working out) [2,3] to higher level routines (having dinner or commuting) [4,5] that consist of numerous sub-activities (preparing food, eating, clearing the table).

With the advent of the smartphone and availability of mobile Internet access, users can stay connected with their friends at any time and express themselves and their current situation via status updates or image uploads. Known as "microblogging", users write short on-the-spot updates about their life and publish these to their social circles or interested

* Corresponding author.
E-mail address: zack.zhu@ife.ee.ethz.ch (Z. Zhu).

[1] Google Now: http://www.google.com/landing/now/.

http://dx.doi.org/10.1016/j.pmcj.2015.10.007
1574-1192/

followers. Messages are usually short: just 140 characters with optional image attachment in the case of Twitter. According to Twitter's official blog [6,7], Twitter users were generating 340M tweets daily in 2012 with 140M active users, compared to 200M Tweets daily in 2011, 65M in 2010 and 2M in 2009. Therefore, it is to be expected that a large fraction of users posts regularly about their routine life experiences. Investigated by [8], typical content ranges from daily life experiences to special interests and news. In this work, we explore a novel path to conduct activity recognition. Instead of collecting evidence from instrumented sensors, we "probe" users indirectly by picking up implicit signals from their natural mobile phone usage as they post to social media platforms.

As smartphones essentially enable any-time use of social media platforms, relevant properties emerge for collecting evidence about the user's activity. First, content is shared in real-time and focuses on experiences that "happen right now" [9]. Second, "daily chatters" share content multiple times a day [8]. Third, and most importantly, it has been shown that the majority of users focus on themselves, rather than on, for example, sharing plain information or opinions [10]. Moreover, social media usage is widespread geographically and has become a natural part of people's daily lives, much of it taking place on smartphones. As a consequence, an abundance of data revealing a user's activities is generated *implicitly* by the user. Through social media platforms that record such data, we can obtain rich signals for activity recognition without any additional instrumentation. This data is spontaneously-generated and naturally occurring, thereby providing in-the-wild sensing without the restrictions of laboratory environments. Our goal is not to incentivize users to post explicitly about his activities or to post in higher quantities. Instead, we argue that data collected by social media platforms can be directly fed into activity- or context-aware systems. As such, artefacts from user-social platform interaction can be understood as a reflection of the user's daily activities.

However, because users are not constrained beforehand to systematically post specific daily situations or activities for self-reporting, this opens the question as to how to define a common scheme for *activity*. Here, we make use of a standardized activity taxonomy from the American Time-Use Survey (ATUS) [11]. It is defined by the Bureau of Labor Statistics in the United States for investigating time-use of the American population. The taxonomy describes a comprehensive, multi-tier hierarchy of typical activities people perform in everyday life. It has been investigated for ubiquitous computing systems as well in the past [12,13]. We select this taxonomy for its relevancy, comprehensive coverage, and also its overlap with other activity surveys from healthcare [14,15].

Social media instances are short and abrupt in nature, leading to potentially ambiguous and/or manifold interpretations of activity classes. Take the common example of dining out with friends, should it be classified, or even labelled, as "Eating & Drinking" or "Socializing, Relaxing, & Leisure" (categories of [11])? Therefore, it is a challenging task, not only for machine learning techniques, but even for humans to agree on a single activity when examining the expressed content afterwards. We contrast this work with our earlier investigation [16] by allowing multi-labelling and learning of composite activities in hopes of alleviating such ambiguity to comprehensively capture activity implications from social media.

In this paper, we investigate the potential of harvesting and extracting publicly self-reported activities through social media. Using text mining and machine learning techniques, we build statistical models to map user signals to activity classes. The key research questions we aim to answer are, therefore:

- Is it feasible to crowdsource labelling of noisy social media posts to identify human activities?
- Can we automatically estimate the main activity and simultaneously occurring, manifold activities of a user from social media posts?

Towards these two questions, we make the following contributions:

- We present an architecture for gathering and labelling activity reports. In an attempt to comprehensively cover the variety of possible human activities, we rely on social media platforms for large-scale gathering of data and crowdsourcing engines for labelling of data.
- Although noisy and unstructured, we characterize our dataset to reveal the rich variety of activities contained within it and the potential of such data to reflect collective human behaviour.
- Finally, we construct an activity recognition model capable of recognizing the main activity category as well as all applicable categories (for manifold activities) from 10 activity classes, with accuracies of 76% and 75%, respectively. We provide a thorough evaluation and model interpretation comparing the use of single-labelled and multi-labelled training data. We find that, while both approaches perform similarly well for main activity classification, multi-labelled training data is necessary to successfully capture the manifold activities implied within social media instances.

In Section 2, we first review existing work in the area of instrumentation-free approaches for human activity recognition. Then, in Section 3, we present our system architecture and the crowdsourced labelling task. In Section 4, we discuss the collected dataset and challenges that arise from harvesting social media data for activity inference. We describe our model in Section 5 for automatic activity recognition based on a crowd-labelled dataset. We present quantitative results evaluating our approach in Section 6 and discuss the limitations of our approach as well as potential solutions to address these limitations in Section 7. Finally, we conclude and provide an outlook for our work in Section 8.

2. Related work

Since the pioneering work of Bao and Intille [17], activity recognition (AR) research has evolved significantly due to the increasing ubiquity of commercially-available mobile and wearable devices. A summative review by Lane et al. [18]

in 2010 highlights existing achievements, challenges, and future directions for the role of AR research in the context of mobile sensing. From the AR community, recent research tackle challenges still preventing AR as a widely deployed service. Significant topics include reducing energy consumption on mobile devices [19,20], personalization of activity models [21], and detection of group activities [22,23]. While such issues are critical for conventional AR approaches leveraging physical mobile sensors, our work complements these developments by leveraging another significant aspect, the user. By treating the user as the sensor, we conduct AR from informative self-reports as they are emitted naturalistically. For the rest of this section, we discuss existing literature on two fronts relevant to our work: the use of instrumentation-free signals for AR and crowdsourcing techniques for obtaining training labels.

2.1. Crowd-generated signals for activity recognition

There exists previous work that investigates the use of "freely-available" information to augment the performance of AR systems without additional instrumentation. For example, Ye et al. [24,25] leverage temporal features to increase activity classification performance and [26] achieve significant performance gains by augmenting sensor-based features with a temporal rhythm model of the user's daily activities. In addition to time, routinely visited locations such as home, work, or a school can indicate pursued activities such as leisure, working, or picking up someone [27]. However, burdensome data acquisition efforts limit these studies to a small number of users and simple activities, inhibiting a general application to a multitude of users.

Towards large-scale data usage, earlier work by [28] utilizes query results of Google to build models for activities of daily living. Despite the use of web-scale knowledge, primitive activities (e.g. brushing teeth) were addressed while the detection of activity routines were not investigated. Recently, time-use survey data, collected by government organizations through telephone interviews, are being exploited to aid context-aware systems. Partridge and Golle [12] leverage the American Time-Use Survey data to learn mappings between location semantics and activities. This work is extended by [13] in 2013, where the German Time-Use Survey is compared. Although such data is well-annotated and incorporates input from thousands of subjects, there is significant cost for governmental organizations to conduct such surveys regularly. Therefore, coverage is limited to certain parts of the world. Furthermore, an additional step to obtain "semantics" of a user's current location is required for activity inference. In other words, the user's absolute location (geo-coordinates) would need to be converted to a relative location (i.e. a reverse geocoding step to obtain venue type).

One rationale to substantiate the use of large-scale social data comes from [10], where they illustrate the different types of content that is generated on Twitter. In their study, 41% of the content can be categorized as "Me Now" and is the leading category. This shows that self-reporting is readily available from such data. Moreover, the user-base is geographically widespread and increasingly encompassing of various strata of society. Although a relatively new idea, others have already started exploring the use of social media data to understand human activities. In the recent work of Pan et al. [29], communication data from Gmail, Facebook, and Twitter are analysed to extract daily behaviour patterns. Based on an assumed correlation to daily life behaviour, they infer the well-being, in turn, the sentiment of a person. We differentiate our work by focusing on inferring user activity from online content as opposed to user emotions. The work of Dearman and Truong [30] shows that it is possible to utilize Yelp reviews to identify potential activities. They validate their system with human reviewers to compare precision and recall of potential activities extracted. In later work, this method is applied to construct a mobile system to provide guidance for exploration of urban spaces [31]. Essentially, their approach relies on part-of-speech tagging to extract explicit verb–noun pairs (e.g. purchase-book) to identify activities from crowd-generated text. However, the disadvantage is that some phrases (e.g. at the park, it is a beautiful day for recreational activities) implicitly signal the user's activity without using indicative verb–noun pairs. On the other hand, a sentence may contain multiple nouns or verbs that create an ambiguous scenario for automated algorithms to make the correct matching. Therefore, the nuances of language, particularly in the case of unstructured web texts, pose significant challenges and ambiguities for rule-based approaches to conduct activity extraction. By using a supervised machine learning approach, we can learn patterns of concurrence between certain grams and activity labels. For example, we may discover significant statistical relationships associating the grams "recreational" and "beautiful" with leisurely activities, even though both of these are tagged as adjectives.

We differentiate our work from [12,30,31] in another fundamental way. As they generate a model of *potential* activities for various venues, we illustrate our approach for inferring the *current* activity of a user. By assuming every microblog instance captures an activity, our machine learning approach draws from textual features as well as other facets of context, like location and time. Through this multi-faceted view of the context, our employment of supervised machine learning is capable of learning the mapping between naturalistic self-reports to in-situ activity classes.

2.2. Crowdsourced labels for activity recognition

Given the free-form, incomplete, and noisy nature of social media text, providing sufficiently labelled training instances for machine learning algorithms is an essential challenge. A recent approach employed by researchers is to crowdsource a component that is difficult for computers, but easy for humans (e.g. data labelling) [32,33]. In addition, crowdsourcing requires gathering input from a large number of contributors via the web or another information access platform, such as SMS [34].

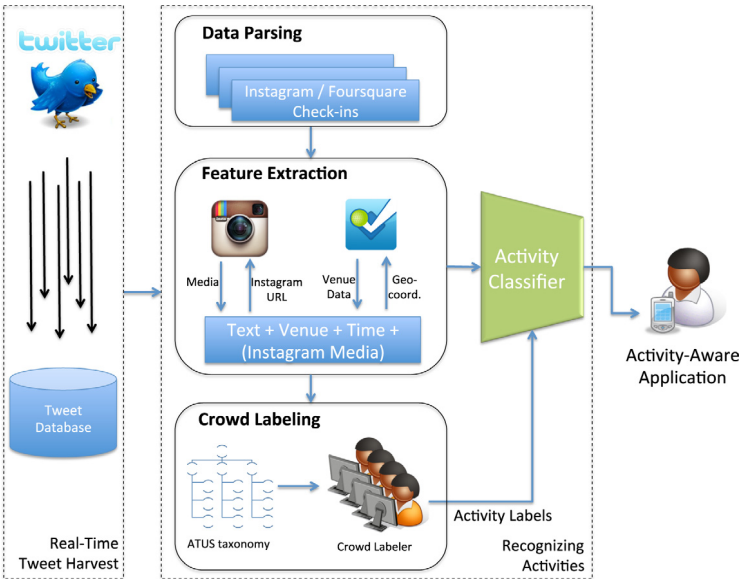

Fig. 1. System architecture for gathering and modelling activities from in-situ, self-report social platform data.

To conduct crowdsourcing, a number of platforms exist. In paid crowdsourcing, Amazon Mechanical Turk[2] is one of the earliest and leading platforms where workers from around the world work for monetary incentives. Interestingly, experimental platforms for crowdsourcing are emerging to ensure work quality and enabling of developing regions [35,36]. Furthermore, researchers have also assessed the utility of crowdsourcing for various tasks and provide guidelines for properly motivating workers [37,38]. Given the findings of previous researchers, we deem crowdsourcing as an appropriate venue for labelling social media instances, particularly due to the low difficulty of such tasks and scalability provided by crowdsourcing platforms. In our experimentation, we utilize the paid crowdsourcing Crowd Flower,[3] as it aggregates a large number of workers from multiple platforms for scalability, provides effective quality assurance measures, and extensive reporting features for analysis.

3. Gathering self-reports from social media data

In Fig. 1, we present our system architecture. In the following two subsections, we will first discuss the data collection and feature generation processes. Then, we illustrate the labelling of social media instances by crowdsourcing.

3.1. Collecting social media data

To gather crowd-generated self-reports, we use Twitter's streaming API[4] to gather real-time Tweets as they are posted on Twitter. Although our standard developer account receives only a very small fraction of the total volume of Tweets, we collected 157,257 instances between July 4, 2013 and July 30, 2013. Our collection is limited to English tweets in the San Francisco metropolitan area as indicated in the bounding box shown in Fig. 2.

Even though Twitter's original purpose is to serve as a venue for concise self-expressions, which is indeed its main purpose currently, it has evolved to become a general purpose platform for online communication. As such, Twitter is linked to by various other social media platforms, such as Foursquare and Instagram, so that users posting to these platforms can also automatically replicate their post on Twitter.

Sometimes, opinions or general thoughts are posted for sharing. Other times, entire conversations of multiple parties are Tweeted through the ReTweet function as replies are generated. For our purpose of understanding people's in-situ activities, we filter out content irrelevant to what one is doing at the moment. As a coarse selection, we filter for tweets "checking-in" to a geo-location while posting a text and/or photo. From these tweets, one is able to identify the in-situ activity with much more clarity via additional context information (e.g. venue location) and potentially moment-capturing photographs.

From the Feature Extraction block of Fig. 1, additional context is fetched to augment the original Tweets. From Foursquare, we obtain venue information, such as venue category, using the Venue Search API[5] by looking up the geo-coordinates of

2 https://www.mturk.com/.
3 http://www.crowdflower.com/.
4 https://dev.twitter.com/docs/streaming-apis.
5 https://developer.foursquare.com/docs/venues/search.

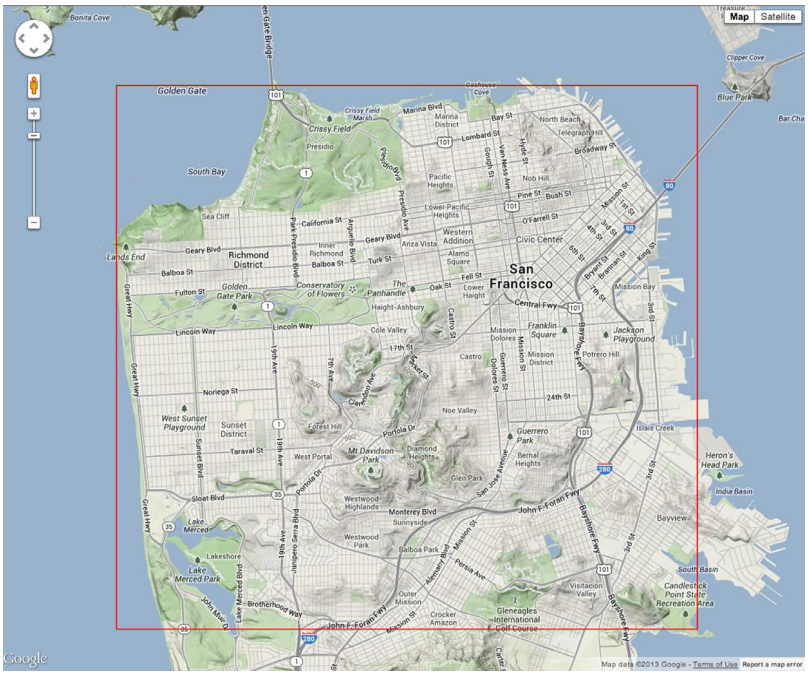

Fig. 2. Map of geographical area in San Francisco from which self-report instances were gathered. The bounding box has coordinates of (37.7099, −122.5137) in the lower corner and (37.8101, −122.3785) for the upper corner.

Tweets. From Instagram, we obtain corresponding "check-in" photos if the Tweet instance contains an Instagram reference link. After augmentations from these two platforms, our Tweet instances contain the text, venue name, venue category, local post time, reference link to the original Foursquare or Instagram page, and corresponding photo for Instagram Tweets. As we will show later, to a certain extent, our model is able to recognize activities with just the text contained in the tweet; nevertheless, additional pieces of contextual information clarify the crowdsourced labelling process and improve classification performance in our activity recognition system.

3.2. Labelling for self-reported activities

We treat all information from an augmented instance as in-situ signals depicting the abstract notion of an "activity" that the user is engaged in and expressing. Using the ATUS taxonomy, we structure the infinitely possible number of high-level daily activities into major activity categories. As shown by [39], the inherent bias in social media data is skewed and under-represents activity categories such as "Caring for household members" or "Religious and spiritual activities". We find a similar pattern in our dataset and adapt the ATUS taxonomy to include the same activity categories as in [39]: *Socializing, Relaxing, & Leisure; Eating & Drinking; Sports, Exercise, & Recreation; Consumer Purchases; Work-Related; Education; Commuting; Professional Services; Household Activities; Personal Care*.

Using these categories, we crowdsource the task of manually labelling activities through the CrowdFlower[6] platform. To assure labelling quality, CrowdFlower collects "gold" labels from experimenters in order to assess the qualification and seriousness of crowd workers. We manually labelled 167 instances ourselves to provide coverage in all categories. The "gold" labels are used in two ways: First, crowd workers are required to obtain at least 10 "gold" units correct before proceeding to standard units. Second, based on labelling accuracy of "gold" units, CrowdFlower calculates a trust score between 0 and 1 to weigh the contribution of each worker on the final, aggregated label result. CrowdFlower adjusts worker trust based on randomly inserted "gold" instances to continuously monitor worker trust. In our tasks, workers are removed if their trust score dips below 0.7. To be further robust to noisy labels, we tune the task redundancy to 10, which implies every task will be labelled by at least 10 trusted workers.

Due to the short, unstructured nature, and varying usage purpose of social media posts, there exist ambiguity when labelling categories of activities. For example, an instance with the self-report text of *"Training so hard at the gym today"* can be easily categorized as "Sports, Exercise, & Recreation". On the other hand, posting *"San Francisco rules! Tonight we are in Oakland at Eli's, spread the word!"* can be quite ambiguous as to whether the activity is "Work-Related" or "Socializing, Relaxing, & Leisure". To alleviate such ambiguity, we leverage the aforementioned context augmentation to

6 http://www.crowdflower.com.

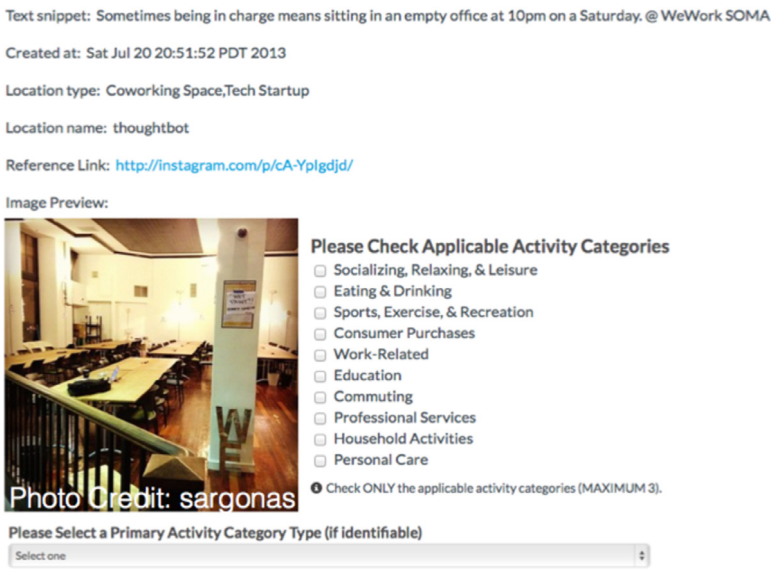

Fig. 3. Screenshot of a task instance, in which a crowd labeller is asked to provide the ground truth to the user's current activity. Upon mouseover, example activities for each category would appear to guide the decision.

provide additional cues for labellers. In addition, the original post's URL is provided so the labeller can access additional information about the instance, such as comments.

Another source of ambiguity occurs when labelling instances with a single activity category. Often, instances may contain manifold activities, such as the instance *"Let's start our foodie weekend with dim sum and friends!"* upon checking-in at a Dim Sum Restaurant. Given the labelling task of identifying a single activity category as in our previous work [16], crowdsourced labels may be divided between "Eating & Drinking" and "Socializing, Relaxing, & Leisure", thereby introducing noise into the classification process. In this work, we extend our labelling task by allowing for multi-activity tagging of instances to have a maximum of three applicable activity categories per instance. Effectively, this transforms our modelling approach to address a multi-label classification problem, allowing us to decrease labelling ambiguity and capture the potentially manifold nature of self-reports.

In Fig. 3, an example instance is shown to demonstrate what a crowd-worker sees. For each instance, two tasks are posed: check up to three applicable activity categories and specify a main activity from a drop-down field that best describes the moment. We deploy tasks to online workers from English-speaking countries (Canada, United States, United Kingdom, Australia, and New Zealand) with the following instructions:

> *Given a tweet describing an Instagram upload, identify the categories of activities that the user is trying to capture at that moment.* ***Please mouseover first (press alt while mouseover if necessary) to familiarize yourself with activity examples of each category.*** *If multiple activities are selected, identify one main primary activity in the drop-down menu. Please base your selection also on the context (geographical, time, venue type, photo if available, and text snippet) provided. Aside from the information we provide, feel free to click the link to check out the original post.*

4. Crowd-generated activities data

The 157,257 collected tweets from the San Francisco area are filtered by examining whether the embedded URLs contain the domain names "4sq.com" or "instagram.com". The resultant set contains 43,835 instances of geo-tagged tweets representing Foursquare "check-ins" or Instagram posts. Due to limited labelling funding, we uniformly sub-select from these instances to obtain 6024 samples to post on the crowdsourcing platform, CrowdFlower.

In total, we paid out approximately $500 USD to obtain all labels. Dividing up all instances into 4044 tasks of Instagram instances and 1813 Foursquare instances, we initially launched the Instagram tasks for $0.05 USD per 10 judgements (8 regular + 2 "gold" tasks). Within 92 h, all Instagram tasks were labelled satisfactorily after receiving 56,978 responses, of which 40,850 were trusted judgements according on-going quality monitoring by comparing against the "gold" labels. In terms of users, the 621 out 3217 unique workers were rejected from the experiment due to poor performance. To maximize labelling quantity, we lowered the payment to $0.03 USD for 10 judgements for the Foursquare instances. The 1813 Foursquare tasks were finished after 181 h, during which, 18,160 trusted judgements and 1760 untrusted judgements were entered. The number of disqualified workers was 37 out of 363 total contributors.

In Table 1, we provide some sample instances of the raw data and the label distribution from both the single-labelling and multi-labelling tasks. It can be seen that while some instances belong distinctly into one category (e.g. rows 1 and 6), multiple categories are equally applicable for others (e.g. row 5).

Table 1
Table illustrating sample data instances harvested from Tweeted "check-ins" in the San Francisco metropolitan region.

Row	Posting Time	Location Name	Location Type	Main Activity	Multi-Label Distribution
1	Fri Jul 05 11:35:28 PDT 2013	The Tipsy Pig	Gastropub	Eating & Drinking	Eating & Drinking (64%); Socializing, Relaxing, Leisure (36%)
	Text: *"Lunch at the Pig. Pulled Pork Sandwich."*				
2	Sat Jul 06 07:39:00 PDT 2013	YetiZen Innovation Lab	Tech Startup	Work-Related	Work-Related (58%); Socializing, Relaxing, Leisure (42%)
	Text: *"It's going to be a very Zen-like #Hackathon #AngelHack #HumanApi"*				
3	Wed Jul 10 16:45:10 PDT 2013	Don Rafa's Cyclery	Bike Shop	Professional Services	Professional Services (50%); Consumer Purchases (25%); Sports, Exercise, & Recreation (25%)
	Text: *"Best bike repair shop in the city. Getting the gears tuned up. "*				
4	Fri Jul 12 13:54:18 PDT 2013	San Francisco Caltrain Station	Train Station	Commuting	Commuting (62%); Sports, Exercise, & Recreation (31%); Socializing, Relaxing, & Leisure (8%)
	Text: *"Off work early so headed home and then the gym"*				
5	Sat Jul 13 10:31:58 PDT 2013	City View Restaurant	Dim Sum Restaurant	Eating and drinking	Eating & Drinking (56%); Socializing, Relaxing, & Leisure (39%); Consumer Purchases (6%)
	Text: *"Let's start our foodie weekend with dim sum and friends!"*				
6	Sat Jul 20 13:02:10 PDT 2013	Yerba Buena Gardens	Park	Socializing, Relaxing, & Leisure	Socializing, Relaxing, & Leisure (90%); Sports, Exercise, & Recreation (10%)
	Text: *"Merola opera in the gardens. Really sunny day downtown."*				

Table 2
Distribution of labels in dataset as generated by crowd workers.

Label category	Single label proportion	Multi label proportion
Socializing, relaxing, & Leisure	44.33% (2661)	27.44% (5831)
Eating & drinking	26.45% (1588)	15.17% (3223)
Sports, exercise, & recreation	12.21% (733)	11.15% (2370)
Work-related	6.70% (402)	13.18% (2801)
Consumer purchases	5.05% (303)	13.58% (2886)
Commuting	2.85% (171)	6.93% (1472)
Education	1.27% (76)	4.49% (955)
Professional services	0.85% (51)	4.08% (867)
Personal care	0.18% (11)	2.07% (440)
Household activities	0.12% (7)	1.91% (405)

4.1. Activity label distribution

Table 2 shows the distribution of labels received from CrowdFlower. We find the majority of instances (>80%) are labelled with "Socializing, Relaxing, & Leisure", "Eating & Drinking", and "Sports, Exercise, & Recreation" as the main activity category. This is not surprising since leisurely experiences are readily shared on social media platforms as opposed to work-related ones. Although there is significant skew in the categories, we notice that even the least populated categories contain some instances. Therefore, prolonged data collection would alleviate the lack of data for the lesser represented activity categories.

Comparing single-labelling (main activity category) with multi-labelling (all applicable activity categories), the skew is less significant when multiple activity categories can be assigned to a single instance. This demonstrates the potential of social media posts to capture manifold activities simultaneously. By allowing multi-labelling, we prevent the loss of applicable labels in secondary or tertiary activity categories.

To assess the extent in which instances are multi-labelled, we calculate the standard metrics of label cardinality [40] defined as:

$$LC = \frac{\sum_{i=1}^{|D|} |S_i|}{|D|} \tag{1}$$

where $|D|$ is the number of instances in the dataset and $|S_i|$ is the number of labels tagged for the instance i. For our dataset, we find an LC value of 3.54 on the raw labels we receive. This means, on average, each instance is labelled with approximately 3–4 activity labels from 10 crowd labellers.

4.2. Labelling ambiguity of daily activities

Labelling ambiguity can arise from inappropriate labelling noise as well as inherent ambiguity of self-report instances. For the task of identifying the main activity category, we quantify the level of labelling ambiguity by examining the agreeability of labellers. This is defined as the number of labels in the most common category over number of all labels obtained. Similarly for multi-labelling of activities, agreement is defined as the count of the *label set* with the highest frequency over all unique *label sets* obtained.

In Fig. 4, we plot the empirical cumulative distribution function (ECDF) of agreement levels. The stepping pattern noticeable in the ECDF functions is an effect of having 10 labellers per instance. For labelling the main activity category, we see that a larger portion of our data received quite high agreement, where approximately 50% of the instances receive an agreement score between 0.8 and 1.0. On the other hand, multi-labelling of all applicable categories resulted in less agreement, where 50% of the instances receive an agreement score between 0.6 and 1.0. This shows that, while a main activity may be quite salient, potential secondary and tertiary interpretation is more subjective and up for labeller interpretation.

4.3. Temporal distribution of activities

Although our data contain bias due to the nature of what people share on social media platforms, it is reasonable to assume that we capture a realistic representation for activity categories with significantly more instances (e.g. Socializing, Relaxing, & Leisure; Eating & Drinking; Sports, Exercise, & Recreation). For these categories, we note that the temporal patterns are aligned with common expectations of what people do in their routine day-to-day lives.

In Fig. 5, we plot a wave graph to show the main activity category distribution and variation in activity levels against time. On the horizontal axis, we plot time as the weekly hour (0th–167th). On the vertical axis, the number of instances is depicted (see scale for absolute numbers). As a result, we notice the "pulse"-like pattern for the days of the week. It is interesting to note some visually salient patterns from the figure.

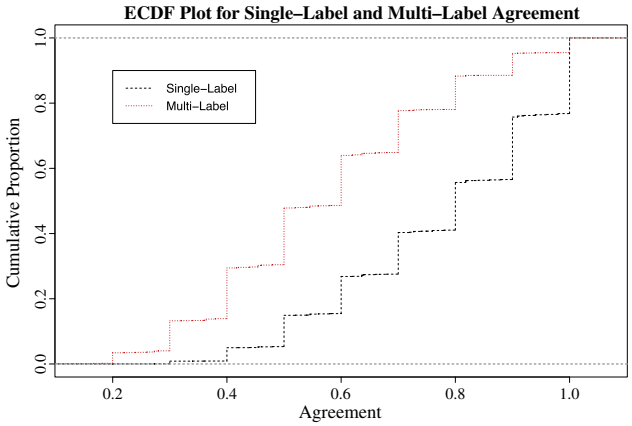

Fig. 4. Empirical CDF of crowd consensus on activity category labels for single-labelling of main activity category and multi-labelling of all applicable activity categories.

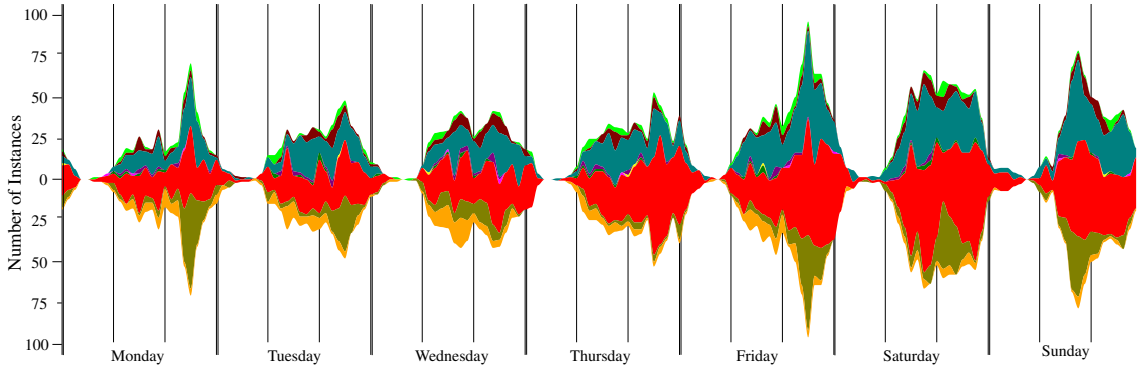

Fig. 5. Weekly pulse of various activities in the San Francisco area. The varying heights of the wave depict variations in the number of instances for that time slot. The coloured layers depicting activity categories are in order from top to bottom: Commuting (Lime), Consumer Purchases (Maroon), Eating & Drinking (Teal), Education (Purple), Household Activities (Yellow), Personal Care (Fuchsia), Professional Services (Green), Socializing, Relaxing, & Leisure (Red), Sports, Exercise, & Recreation (Olive), Work-Related (Orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As expected, the quantity of activity reports are much higher, per day, on weekends (Friday, Saturday, Sunday) than weekdays (weekend average: ∼1050 instances/day *vs.* weekday average: ∼714 instances/day). For each day of the week, a clear increase and decrease in activity levels can be observed to mark a 24-h interval. However, weekend days sustain activity levels much longer into the evening than weekdays. In terms of when different activities take place, we notice social activities (red) spread more or less evenly (relative to overall activity quantity) throughout the week, although increasing significantly over the weekend, as expected. On the other hand, sports-related activities are more time-specific in their occurrence. Namely, they tend to take place Monday/Friday/Saturday evenings, and Sunday morning. This is understandable since people tend to have more time to exercise over the weekend; hitting the gym Monday evening as a result of guilt from a lazy weekend is also commonplace.

5. Constructing activity models from social media data

From the above harvesting methodology and data characterization, we show that it is feasible to capture a rich, diverse, and abundant collection of implicit signals for inferring human activities. In this section, we present our method for modelling the mapping of these signals to ground truth activity categories.

5.1. Feature construction

As the main form of our data is unstructured text, we take a classical text mining approach using n-gram features. Specifically, we extract unigrams and bigrams from text snippets, where each gram serves as one feature in our model. We follow standard text processing techniques of stemming and stop-word removal to reduce feature dimensionality. For example, the phrase "the cats are home" would generate three features: the unigrams "cat" and "home", as well as the bigram "cat home". The words "the" and "are" are removed while "cats" is stemmed to remove the inflectional suffix.

As mentioned, we also augment the context for which the activity occurs by leveraging venue type, name, and occurrence time. Even though Instagram photos are used in the labelling, we currently do not explore computer vision techniques to derive graphical features for our model. To fuse multiple sources of features we do derive, we simply concatenate our feature matrices. We derive the following sets of features:

1. **Tweet Text:** After extracting unigrams and bigrams we use Tf-Idf scaling to construct our textual feature matrix. The feature weight is calculated as follows, where $tf_{t,d}$ is the frequency of n-gram t in tweet d, $|D|$ is the total number of tweets in the corpus D, and df_t is the number of times the term t appears in all documents:

$$wf_{t,d} = \begin{cases} \dfrac{1 + \log tf_{t,d}}{(|D|/df_t) + 1} & tf_{t,d} > 0 \\ 0 & \text{else.} \end{cases} \tag{2}$$

2. **Venue Semantics:** Since each instance is geo-referenced to a Foursquare venue, we extract the semantic category of the venue (e.g. synagogue, Mexican restaurant). We binarize these categories to construct an indicator matrix. Venues with multiple venue tags are indicated with all tags.
3. **Venue Name:** Similar to how we extract features from Tweet Text, we also extract n-gram features from the name of the venue in which the activity happens. These features could be indicative in some cases since venue owners typically name their establishment according to the main activity provided (e.g. China Garden Restaurant for eating).
4. **Posting Time:** Associated with each time is also the posting time. From Fig. 5, our visualization of activity distribution over time shows several intuitive characteristics, described in 4.3. While weekdays and weekends illustrate clearly different distribution and volume of activities, activity patterns also differ within weekdays (e.g. Monday evening) and weekend days (e.g. Friday night *vs*. Sunday night). We chunk the time data by hour of the week to provide the classifier with fine-grained delineation of the data. The hour of week is then binarized to construct an indicator vector for each instance.

5.2. Classifying for activity categories

5.2.1. Classification tasks

We investigate two classification tasks in this work: multi-class classification of the main activity category represented in each social media post and multi-label classification of activity tags. The motivation of the former is to precisely identify the key activity as in our earlier work [16]. Our motivation for the latter is to more comprehensively capture the potential of a social media post to contain manifold activities.

Two paradigms of multi-label classification exist: algorithm exploitation (e.g. intrinsic multi-output nature of decision trees [41] and ensemble methods [42]) and problem transformation. To permit the general applicability of any binary classification algorithm, we elect to use the standard binary relevance problem transformation approach (BR) [40]. In this approach, multi-labelled data are assigned as members of all applicable classes. Then, $|L|$ binary classifiers are trained to distinguish whether an element belongs to the lth class, where L is the set of all possible classes. The approach is similar to the *One-vs-Rest* strategy (OvR) for multi-class classification, with the difference that each instance can be predicted as members of multiple classes.

5.2.2. Classification algorithm

Given the large number of features from textual features, our feature matrix is large-scale, sparse, and high-dimensional. From text mining literature [43,44], such problems have benefited from the use of fast and highly scalable Support Vector Machine (SVM) algorithms. Therefore, we apply the Linear SVM package from the Scikit-Learn library [45] to learn the mapping between our feature space and the multi-class label space. We select L1-regularization with squared hinge loss and keep the default parameters of the package.

Two key benefits of L1-penalized Linear SVM are: implicit feature selection and feature importance ranking. By using L1 regularization (with squared hinge loss), the objective function is effectively:

$$\min_{\boldsymbol{\omega}} \| \boldsymbol{\omega} \|_1 + C \sum_{i=1}^{D} (\max(0, 1 - y_i \boldsymbol{\omega}^T \boldsymbol{x}_i))^2 \tag{3}$$

where $\boldsymbol{\omega}$ is the feature weight vector and $\| \cdot \|_1$ denotes the 1-norm [43]. As a result, implicit feature selection takes place as some coefficients in $\boldsymbol{\omega}$ are forced to 0. Storing only the remaining features achieves a lightweight activity classifier, suitable for real-time mobile use cases. Since our kernel function is linear and all our features are in the range of [0, 1], feature importance is directly indicated by the magnitude of the coefficient element in $\boldsymbol{\omega}$ [46]. This allows us to compare the usefulness of the various feature sets.

5.2.3. Preprocessing crowd-generated labels and post-processing model output

Aggregating the multi-labelled activity categories from crowd-workers, we can calculate the number of labels for each activity category over number of labels received in total for an instance i, summing to 1. However, this distribution needs to be converted into a set of binary labels for classification. To achieve this, we use a simple thresholding factor θ to binarize the assignment of an instance to each activity category, as follows:

$$S_i = \begin{cases} 1, & p_l \geq \theta \\ 0, & \text{else} \end{cases} \quad \forall\, l \in L(i) \tag{4}$$

where p_l is the proportion of labels in class l over all labels received for the instance i. In cases where no p_l passes the threshold, S_i reduces to the single class that received the maximum number of votes. Intuitively, a low threshold θ allows more activity categories per instances but may be susceptible to labelling noise. A higher θ utilizes majority voting to eliminate false-positive labelling noise but reduces label cardinality and negates the effect of per instance, multi-labelling.

In addition to using multi-labelled data to conduct multi-label classification and single-labelled data to conduct multi-class classification, it is also possible to apply post-processing to use multi-labelled data to predict the main activity category and single-labelled data to predict all applicable activity categories:

Single-labelled data for multi-label classification: The use of the OvR scheme with SVM for multi-class classification outputs distance of an instance to each of the $|L|$ decision functions. We post-process this vector of to extract all classes for which an instance is on the positive side of the decision function. These classes are used as the multi-label prediction.
Multi-labelled data for single-label classification: Given the distances of an instance to all decision boundaries, we simply post-process to select the class with the maximum positive distance to output as the final classification result.

In Section 6, we will compare the performances of both these data types for the two classification tasks.

6. Experimental validation

In this section, we first evaluate our approach in answering two key questions through the classification tasks presented in Section 5: how accurate can we infer the *main* activity category from an in-situ, social media self-report? Also, how accurate can we infer all applicable activity categories? For the first question, we evaluate our model with the standard accuracy and F1-score metrics. To evaluate multi-label classification for the latter question, we calculate the average Jaccard similarity coefficient (corresponding to multi-class accuracy) [40] and an adapted multi-label F1-score similar to [47] as follows:

$$Accuracy_{multi} = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|S_i \cap Y_i|}{|S_i \cup Y_i|} \tag{5}$$

$$F1_{multi} = \frac{1}{|L|} \sum_{j=1}^{|L|} \left(\omega_j * \frac{2 * p_j * r_j}{p_j + r_j} \right) \tag{6}$$

where Y_i is the predicted set of activity categories for instance i while S_i is the ground truth activity set. The accuracy measure for multi-label classification compares a predicted *set* against the ground truth *set*. The Jaccard similarity index is appropriate here because its numerator gauges the amount of overlap in S_i and Y_i while the denominator term would penalize irrelevant activities not found in S_i. For the F1-score shown in Eq. (6), the class precision and recall of class j are depicted as p_j and r_j, respectively, while w_j weighs each class by the proportion of data in class j to account for class imbalance in the dataset. This is the same formula for how we calculate the F1-score for multi-class classification of the main activity category since the scores are also calculated per-class.

In the latter parts of this section, we examine the details of per-class classification and quantify performance in scenarios without location or textual features. Finally, we illustrate the interpretability of our model through data-driven vocabulary identification to characterize each activity category.

6.1. Inferring main activity category

Using the aforementioned L1-Regularized Linear SVM described, we conduct standard 10-fold cross-validation. We report the average testing accuracy to demonstrate the predictive ability of our approach for inferring the main activity category from social media instances.

To study the influence of labelling ambiguity on classification performance, we sub-select instances with various levels of labelling agreement and report the prediction approach in Fig. 6. On the secondary y-axis on the right, we plot the number of instances remaining after agreement filtering. For comparison purposes, we also plot the accuracy and F1-score of main activity prediction using multi-labelled data with post-processing (as described in Section 5). We set the binary thresholding factor θ to 0.3 here and in the remaining analysis, unless otherwise specified.

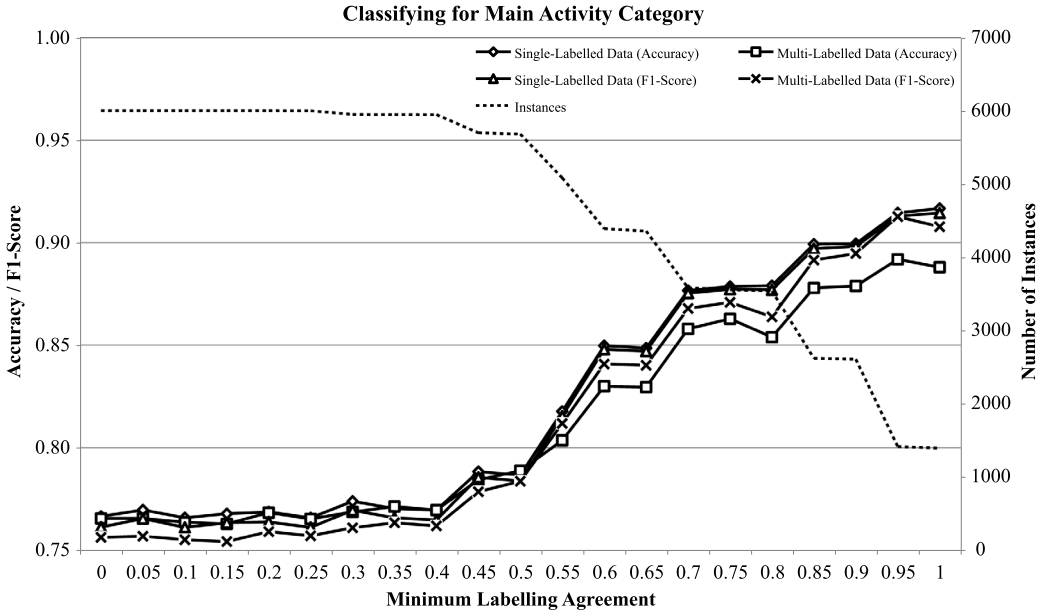

Fig. 6. Sub-selecting data with various agreement scores (x-axis), prediction performance (left y-axis) increases significantly when data contains less labelling ambiguity. However, the amount of data also decreases, thereby eliminating the possibility to predict some lesser populated activity classes.

Comparing the use of single-labelled and multi-labelled data, we notice negligible difference in model performance. Specifically, the maximum difference in accuracy is 2.87% when selecting data with perfect agreement while a difference of 0.1% is observed using the complete dataset. This implies the use of multi-labelled data causes negligible distraction for the model to select the main activity category amongst all predicted activities.

Looking across the x-axis, we notice the prediction accuracy ranges between 77% and 92% as data is filtered based on level of labelling agreement, from no filtering to selecting only those with complete labelling agreement, respectively. By filtering with respect to labelling agreement, we remove the contribution of inadequate workers and spamming effects. However, this process also removes inherently ambiguous instances difficult for both humans and machine learning algorithms to classify, leading to higher model performance. Interestingly, we achieved accuracies between 70% and 84% in our earlier work [16] using the same approach except only 3 labellers per instance. In comparison, this increase in number of labellers to 10 per instance uniformly lifts up model performance by about 15%. Therefore, it is clearly beneficial to obtain a larger number of opinions when crowdsourcing for activity labelling, due to the ability of majority voting to diminish noise.

6.2. Recognizing manifold activity categories

Examining how well we can recognize multi-labelled activity categories from single instances, we plot mean testing accuracy and F1-Score in Fig. 7. We find the maximum accuracy of 74.9% when $\theta = 0.15$. From the figure, it is apparent that, with the post-processing to convert to multi-label output, the single-labelled model is not able to deliver similar performances as multi-labelled model in comprehensively capturing all the applicable activities categories. Only as we increase θ, effectively diminishing the label cardinality, do we find single-labelled model approach the performance of multi-labelled model. Hence, we verify that single-labelled training data provides insufficient information to comprehensively capture the manifold nature of social media instances.

Tuning of θ effectively makes the trade-off between labelling noise and label cardinality for the multi-labelled model. We notice that initial values of $\theta < 0.15$ increase performance for multi-labelled model due to the muting of labelling noise through the label binarization process. Beyond $\theta = 0.15$, higher binarization thresholds may artificially penalize the manifold nature of some instances, leading to slightly lower performance when multiple categories are predicted for one instance.

Another aspect we investigate is the ability of single-labelled and multi-labelled models to capture co-occurrence of activities, or the manifold nature of instances. For this, we use a $|L| \times |L|$ matrix to illustrate the co-occurrence likelihood of each activity category with all other activity categories. To derive the co-occurrence likelihood, we remove the diagonal entries and normalize row-wise such that each row sums to 1. We construct such matrices for the ground truth labels and predicted outputs from both the single-labelled and multi-labelled models. In Fig. 8, we plot the differential heatmaps when the co-occurrence matrix of each model is subtracted from the ground truth co-occurrence matrix. Patches with grey colouring indicate correctly predicted co-occurrences. Green patches reveal where the model under-predicted co-occurrences while red patches reveal an over-prediction. Juxtaposing the two differential heatmaps, we show that multi-labelled model outperforms single-labelled model with noticeably more grey patches as well as diminished colouring

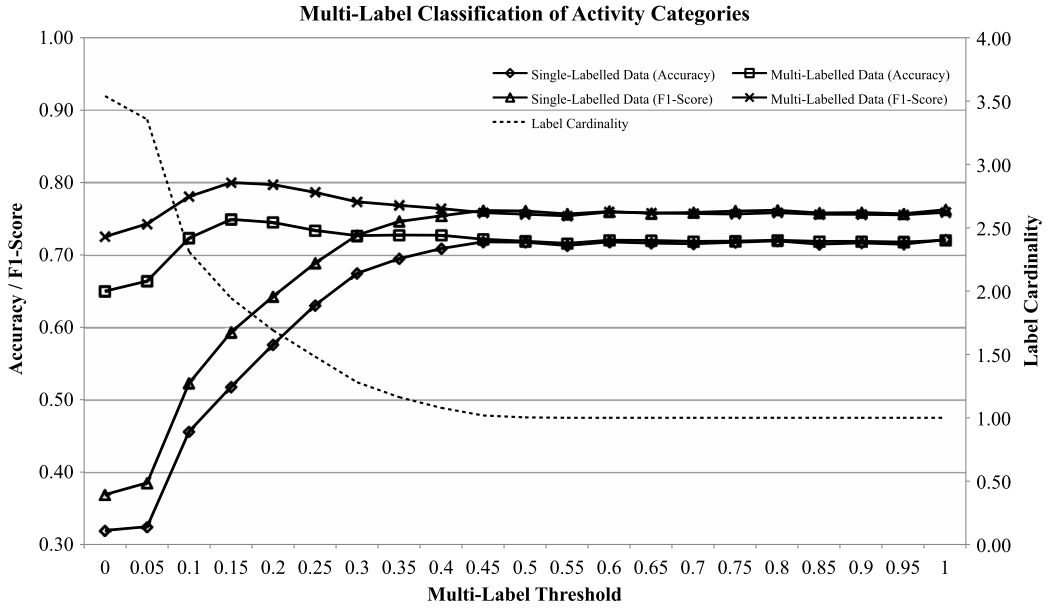

Fig. 7. Sub-selecting data with various agreement scores (*x*-axis) and predictive ability (left *y*-axis), we notice classification performance is heavily influenced by the labelling agreement in data labels.

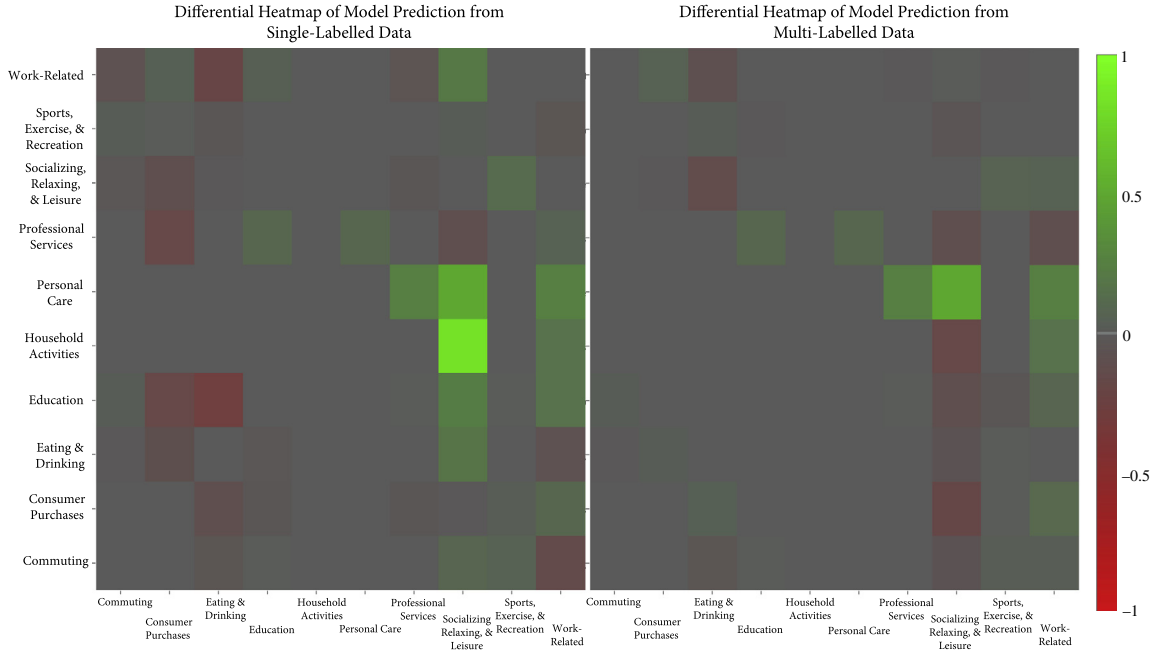

Fig. 8. Differential heatmaps showing the prediction errors of models trained with single-labelled data and multi-labelled data. Green patches reveal where the model missed co-occurrences of labels while red patches show an over-prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where mistakes do still occur. Understandably, single-labelled model suffers from under-prediction of activity co-occurrence (shown in green) while multi-labelled model may over-predict (shown in red).

6.3. Feature analysis for activity recognition performance

Testing with both types of input data (single and multi-labelled) for main activity prediction, we show the mean testing accuracies in Table 3 when using all features (textual n-grams, venue categories and name, and time) as well as when venue or text features are removed. Not surprisingly, using all features deliver the highest performance. This supports the intuition

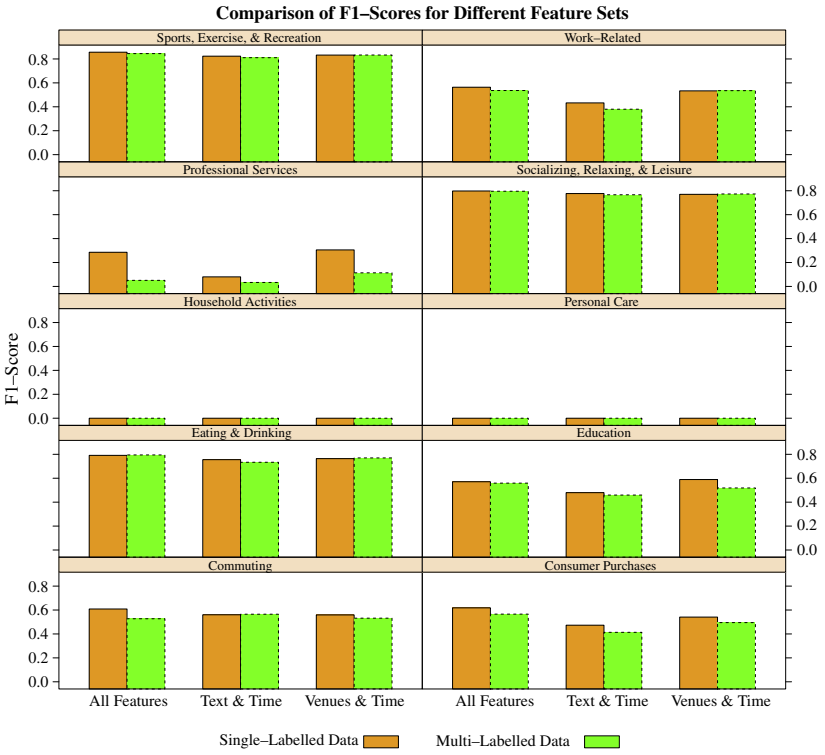

Fig. 9. A comparison of F1-Scores for classifiers trained with all features, time and text features, and time and venue type features.

Table 3
Mean testing accuracy of single and multi-labelled data with different sets of features for main activity prediction.

	Full feature set	Text & time	Venue & time
Single-labelled data	76.81%	73.43%	73.71%
Multi-labelled data	76.39%	72.13%	74.15%

that, although the type of venue constrains what activities are possible, venues of the same category do not necessarily offer the same activities. As a result, textual features are able to make fine-grained adjustments on top of coarse discrimination by venue features, delivering additional gains in prediction accuracy.

We notice negligible difference between the use of textual features and venue features. This is significant since, practically, we may not always have access to location-based services or venue type information, due to privacy or power consumption concerns. Then, our feature set would only include time and n-gram features. Therefore, we expect the lack of a geo-lookup service to not significantly decrease activity recognition performance, as long as the textual content generated is of similar nature to that shared onto online social networks. Although time-based features are free, the overall accuracy with just those features is poor at 43%.

In Fig. 9, we visualize the F1-scores achieved with the three different feature combinations. Since each F1-score measures the ability to correctly infer the main activity category, we facet over the 10 activity categories. Within each sub-plot, we compare the performance of the single-labelled data with multi-labelled data and notice little differences (mostly between 0.0 and 0.03), which agrees with our earlier illustration of overall F1-score in Fig. 7. However, comparing the performance between activity categories, we notice some large performance differences between activity categories. Corresponding activity-specific performance with quantity of instances for that category, we notice the less reported activities on social media platforms suffer in performance (e.g. Household Activities, Personal Care, Professional Services). We discuss potential solutions to overcome the data bias of leveraging social media platforms in the discussion section.

6.4. Data-driven vocabulary of activities

By sorting the magnitude of feature coefficients of a fitted model, $\boldsymbol{\omega}$, we can extract the most distinguishing features for each activity category. In Table 4, we extract up to 25 features with the highest positive feature coefficients for each class. While the model is clearly prone to noise, the mix of venue categories (in bold) and textual n-grams (in quotes) are quite intuitive. This is especially true for activity categories with higher F1-scores (e.g. Sports, Exercise, & Recreation). By

Table 4
Top 25 most indicative text (in quotes) and venue features (in bold) for each activity category.

Category	Top Vocabulary (Single-Labelled Data)	Top Vocabulary (Multi-Labelled Data)
Commuting	"commut", "toll", "car", "station -", "way home", "muni", "80", "termin", **College Auditorium**, **Subway**, **Train Station**, **Platform**, "(j," "ride",**Taxi**,"portal", "home", "drive", "ferri", "train", "mission district", "alamo squar", "octavia", "j,"	"drive", "commut", "car", "bridg toll", "cross", "muni", "home", "way home", **Subway**, **Train Station**, **Platform**, "ride", **Bus Line**, **Boat or Ferry**, **Bus Station**, "ferri", **Taxi**
Consumer Purchases	"jest", "buy", ". size", "store", "bought", "shop", "bag", "collect", "shirt", "fabric", "sale", "$", "box", "new", "shoe", **Racetrack**, "need", **Department Store**,"farm", "choos.", "color", **Video Store**, **Hardware Store**, "light bookstor", **Hot Spring**	"jest", "sale", "buy", "store", "collect", "box", "shop", "color", "shoe", "new", "fabric", "shirt", "$", "bought", ". size"
Eating & Drinking	"breakfast", "chees", "lunch", "beer", "food", "breweri", "chowder", "blossom", "eat", "rice", "mushroom", "dinner", "toast", "sandwich", "delici", "ghirardelli squar", "tomato", "coffe", "soup", "dessert", "fri", "meal", "#food", "truck", "water"	"breakfast", "lunch", "beer", "drink", "snack", "eat", "food", "fri", "cocktail", "chees", "dinner", "fresh", "coffe", "pork", "chowder", "#beer", "sandwich", "#food", "chocol", "toast", "bbq", "bar", "tast"
Education	"colleg", "scienc", "@calacademi (san", **College Administrative Building**, **College Classroom**, **Community College**, **College Lab**, **College Academic Building**, "univers", "school", **Library**, **College Arts Building**, "learn", **Medical School**, **General College & University**, **University**, **History Museum**, **Trade School**, **Music School**, **Temple**, **Student Center**, "art", "institut", "@twitter", **Dentist's Office**	"scienc", **College Classroom**, "colleg", **Library**, **Community College**, **College Lab**, **College Academic Building**, **College Administrative Building**, **College Arts Building**, **General College & University**, **University**, **High School**, "institut", **Trade School**, **History Museum**, **Medical School**, "univers", "school"
Household Activities	**Outdoors & Recreation**, **Laundry Service**, **Comedy Club**, **Light Rail**, **Residential Building (Apartment / Condo)**, **Pet Store**, **Home (private)**	**Outdoors & Recreation**, **Laundry Service**, **Residential Building (Apartment / Condo)**, **Comedy Club**, **Home (private)**
Personal Care	". week", **Emergency Room**, **Hostel**, **Assisted Living**, **Nail Salon**, **City**, **Fast Food Restaurant**, **Salon / Barbershop**, "nail", **Other Great Outdoors**, **Dessert Shop**, **Park**	**Emergency Room**, **Hostel**, **Nail Salon**, **Assisted Living**, **City**, **Salon / Barbershop**, **Dessert Shop**, **Park**
Professional Services	**Courthouse**, **Government Building**, "(@", **Post Office**, **Salon / Barbershop**, "check", "bank america", **Nail Salon**, **Military Base**, "xxi", **Rental Car Location**, **Bank**, **Tattoo Parlor**, **Resort**, **Automotive Shop**, **Hospital**, **Doctor's Office**, "gener hospit", **Cosmetics Shop**, **Speakeasy**	"(@", **Courthouse**, **Government Building**, **Post Office**, "bank america", **Salon / Barbershop**, **Nail Salon**, **Hospital**, **Rental Car Location**, **Tattoo Parlor**, **Doctor's Office**, **Automotive Shop**, **Bank**, **Cosmetics Shop**, **Laundry Service**
Socializing, Relaxing, & Leisure	"good time", "area", "mr.", "churchil", "background", "parti", "rememb", "picnic presidio grid", "music", "babi", "night", "daughter", "idol", "pictur", "birthday", "downtown", "wash cafe", "ladi", "#baybridg", "#art", "human", "photo", "uncl", "rainbow", "scene"	"hang", "parti", "boat", "pictur", "casa", "birthday", "babi", "girl", "spend", "photo", "peopl", "kid", "celebr", "friend", "night", "music"
Sports, Exercise, & Recreation	"golf", "yoga", "mile", "trail", "lake merc", "bike", "& t", "giant", "(@ sport", "sport", "ride", "south beach", "aid walk", "t", "run", "center northern california", "beat", "workout", "gate park", **Gym / Fitness Center**, **Basketball Stadium**, **Climbing Gym**, **Tennis Court**, **Football Stadium**, "rang"	"yoga", "golf", "giant", "aid walk", "mile", "run", "& t", "hous air", "workout", "sport", "cup", "bike", "t", "ride", **Climbing Gym**, **Gym / Fitness Center**, "@ lake", **Gym**, **Tennis Court**
Work-Related	"inman", "confer", "work", "interview", "21a", "cowork", "job", "talk", "techshop", "offic", ". readi", "autofuss", "media", "shoot", "meet", "#herb", "power", "releas", "tomorrow night", "#san", "featur", "mobilebeat", "lab", "hq", "studio"	"work", "inman", "confer", "job", "tomorrow night", "interview", "offic", "power", "@ fort mason", "releas", "custom", "desk", "team", "talk", "intern", "come", "cowork", "meet", "hq", "help", "excit", "finish"

interpreting the model, it is interesting to notice that the activities which are constrained to specific locations (e.g. school buildings for Education or banks for obtaining Professional Services) have more venue features in their top 25. On the other hand, activities less constrained by venue (e.g. Socializing, Relaxing, & Leisure and Eating & Drinking) are more distinguishable by textual n-grams.

7. Limitations & discussion

Even though we show the highly predictive nature of leveraging implicit signals from people's natural interaction with their smartphones, it is ultimately only one source of information. As such, its information content biases towards certain activities regularly reported on social media platforms. From Fig. 9, we note the imbalance of predictability across different activity classes. To comprehensively capture the wide variety of high-level daily activities, other information sources (e.g. physical sensor data) should be fused into the final activity recognition chain. This would be especially helpful for activity recognition in situations where the general population is unlikely to engage with social media platforms (e.g. work-related activities). For scalability, commercially available devices such as wrist-based smart-watches and mobile phones would be useful to collect and predict such activities.

However, our approach as is would already have useful applications. Similar to [30], the use of crowd-generated textual features allow us to make fine-grained activity discoveries. This fine-grained information would provide valuable insight for tourists who may be unfamiliar with the extraordinary function of a local venue. For example, our approach could reveal a bar is popular for its darts tournaments or a university terrace offers a panoramic view of the city. One way to enable such a feature would be to allow free-form text queries into our model and generate heatmaps representing likelihoods of keywords. In addition, by defining a comprehensive set of activity classes *a priori* and conducting classification of unstructured signals into manifold activity distributions, we can also assess the activity composition of a venue in an aggregated and structured manner. For example, we would be able to provide multi-purpose venue owners (e.g. shopping centres) with structured information on how people are utilizing their facilities as a composition of activity classes (e.g. proportion of people shopping, eating and drinking, and conducting leisurely activities). Furthermore, the standardization gained from structured classification of activities would enable direct semantic comparison between different urban spaces in terms of actual usage. This information would be quite informative to urban planners when designing and/or monitoring the interaction of residents and urban spaces.

From Table 4, we see some highly indicative vocabulary as n-grams that are relatively specific to San Francisco (e.g. "giant", "mission district", "lake merc"). Although this indicates overfitting of our model to one geographical region, it would be straightforward to obtain separate supervised models for other metropolitan areas (e.g. New York). Then, a simple GPS lookup or user-input can select the appropriate city-scale model for activity recognition.

Finally, we have relied heavily on tweets deemed as location-specific "check-ins". Although this can be indicative of stationary activities, spatially transient activities (e.g. the Commuting) would naturally suffer. One way to remedy this would be to consider sequences of GPS readings to recognize the method of locomotion as in [27]. In addition, we believe another "free" data source that could also complement the activity recognition chain is population-wide time-use data. Although we have incorporated time as a feature and show it is not entirely indicative, we have only trained the time feature from a limited pool of social media users. This can be improved by taking advantage of population-scale time-use data for activity prediction as attempted in [12,13]. Ultimately, the combination of traditional sensor data, time use surveys, and social media data, altogether would deliver the best performance.

8. Conclusion and future work

In this paper, we argue for the feasibility and benefit of using social media data as an approach to conduct automatic in-situ activity inference. By regarding the user as our most informative "sensor", we implicitly and naturalistically obtain rich and comprehensive signals without any instrumentation effort. Although the abundance and widespread geographic availability is very appealing, the data is unstructured and can be ambiguous, even for human labellers. Furthermore, our data characterization shows significant skew in social media data for covering various types of daily activities.

Despite these challenges, we successfully show that machine learning techniques can be employed to infer the main activity category as well as the manifold activity categories embedded in one social media instance. Comparing single-labelled and multi-labelled training data, we find that multi-label-trained models perform similarly as single-label-trained models for main activity recognition, with average testing accuracy of 76%. However, the multi-label-trained model is able to outperform the single-label-trained model by more than 20% (at 75% accuracy) when recognizing multiple activities embedded in a single instance. Examining per-class recognition performance, we notice high correspondences with number of instances available.

To address the limitation of data imbalance from social media platforms, we intend to complement our approach with other activity recognition approaches, such as the use of wearable devices and population-scale, time-use and travel survey data. Also, we intend to realize a mobile system able to conduct real-time activity recognition. As demonstrated by Kim et al., smartphone-based location gathering and in-field classification for activities is highly feasible and scalable for a large number of users [48,49]. We believe our approach of classifying for activities based on user-generated text should

be feasible, partially due to the light-weight nature of the L1-regularized SVM used. With such trials, we look forward to in-field validation of our approach.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions and comments. This work is partially supported by the Hasler Foundation through SmartDAYS (Project No. 11078).

References

[1] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, N. Nishio, Hasc challenge: gathering large scale human activity corpus for the real-world activity understandings, in: Proceedings of the 2nd Augmented Human International Conference, ACM, 2011, p. 27.
[2] L. Bao, S.S. Intille, Activity recognition from user-annotated acceleration data, in: Proceedings of the 2nd International Conference on Pervasive Computing, 2004, pp. 1–17. http://dx.doi.org/10.1007/b96922.
[3] D. Minnen, T. Starner, I. Essa, C. Isbell, Discovering characteristic actions from on-body sensor data, in: Proceedings of the 10th IEEE International Symposium on Wearable Computers, ISWC, 2006.
[4] U. Blanke, B. Schiele, Daily routine recognition through activity spotting, in: 4rd International Symposium on Location- and Context-Awareness, LoCA, 2009.
[5] T. Huynh, U. Blanke, B. Schiele, Scalable recognition of daily activities with wearable sensors, in: 3rd International Workshop on Location- and Context-Awareness, LoCA 2007, 2007, p. 50–67.
[6] Twitter, Twitter turns six, March. 2012. https://blog.twitter.com/2012/twitter-turns-six.
[7] Twitter, 200 million tweets per day, June. 2011. https://blog.twitter.com/2011/200-million-tweets-day.
[8] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, 2007, pp. 56–65.
[9] A. Oulasvirta, E. Lehtonen, E. Kurvinen, M. Raento, Making the ordinary visible in microblogs, Pers. Ubiquitous Comput. 14 (3) (2010) 237–249.
[10] M. Naaman, J. Boase, C.-H. Lai, Is it really about me?: message content in social awareness streams, in: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, ACM, 2010, pp. 189–192.
[11] K.J. Shelley, Developing the american time use survey activity classification system, Monthly Lab. Rev. 128 (2005) 3.
[12] K. Partridge, P. Golle, On using existing time-use study data for ubiquitous computing applications, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, 2008, pp. 144–153.
[13] M. Borazio, K. Van Laerhoven, Improving activity recognition without sensor data: a comparison study of time use surveys, in: Proceedings of the 4th Augmented Human International Conference, ACM, 2013, pp. 108–115.
[14] T.H. Monk, E. Frank, J.M. Potts, D.J. Kupfer, A simple way to measure daily lifestyle regularity, in: European Sleep Research Society, 2002.
[15] S. Katz, T. Downs, H. Cash, R. Grotz, Progress in development of the index of ADL, The Gerontologist 10 (1 Part 1) (1970) 20.
[16] Z. Zhu, U. Blanke, A. Calatroni, G. Tröster, Human activity recognition using social media data, in: Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia, 2013.
[17] L. Bao, S. Intille, Activity recognition from user-annotated acceleration data, in: A. Ferscha, F. Mattern (Eds.), Pervasive Computing, in: Lecture Notes in Computer Science, vol. 3001, Springer, Berlin, Heidelberg, 2004, pp. 1–17. http://dx.doi.org/10.1007/978-3-540-24646-6_1.
[18] N.D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A.T. Campbell, A survey of mobile phone sensing, Comm. Mag. IEEE 48 (9) (2010) 140–150.
[19] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, K. Aberer, Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach, in: Proceedings of the 2012 16th Annual International Symposium on Wearable Computers, ISWC, ISWC '12, IEEE Computer Society, Washington, DC, USA, 2012, pp. 17–24. http://dx.doi.org/10.1109/ISWC.2012.23.
[20] G. Metri, W. Shi, M. Brockmeyer, A. Agrawal, Batteryextender: An adaptive user-guided tool for power management of mobile devices, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, ACM, New York, NY, USA, 2014, pp. 33–43. http://doi.acm.org/10.1145/2632048.2632082, http://dx.doi.org/10.1145/2632048.2632082.
[21] A. Reiss, D. Stricker, Personalized mobile physical activity recognition, in: Proceedings of the 2013 International Symposium on Wearable Computers, ISWC '13, ACM, New York, NY, USA, 2013, pp. 25–28. http://doi.acm.org/10.1145/2493988.2494349, http://dx.doi.org/10.1145/2493988.2494349.
[22] T. Hirano, T. Maekawa, A hybrid unsupervised/supervised model for group activity recognition, in: Proceedings of the 2013 International Symposium on Wearable Computers, ISWC '13, ACM, New York, NY, USA, 2013, pp. 21–24. http://doi.acm.org/10.1145/2493988.2494348, http://dx.doi.org/10.1145/2493988.2494348.
[23] D. Gordon, M. Scholz, M. Beigl, Group activity recognition using belief propagation for wearable devices, in: Proceedings of the 2014 ACM International Symposium on Wearable Computers, ISWC '14, ACM, New York, NY, USA, 2014, pp. 3–10. http://doi.acm.org/10.1145/2634317.2634329, http://dx.doi.org/10.1145/2634317.2634329.
[24] J. Ye, A.K. Clear, L. Coyle, S. Dobson, On using temporal features to create more accurate human-activity classifiers, in: Artificial Intelligence and Cognitive Science, Springer, 2010, pp. 273–282.
[25] J. Ye, L. Coyle, S. Dobson, P. Nixon, Using situation lattices in sensor analysis, in: Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on, IEEE, 2009, pp. 1–11.
[26] K. Van Laerhoven, D. Kilian, B. Schiele, Using rhythm awareness in long-term activity recognition, in: Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on, IEEE, 2008, pp. 63–66.
[27] L. Liao, D. Fox, H. Kautz, Extracting places and activities from gps traces using hierarchical conditional random fields, Int. J. Robot. Res. 26 (1) (2007) 119–134. http://dx.doi.org/10.1177/0278364907073775.
[28] M. Perkowitz, M. Philipose, K. Fishkin, D.J. Patterson, Mining models of human activities from the web, in: Proceedings of the 13th international conference on World Wide Web, WWW '04, ACM, New York, NY, USA, 2004, pp. 573–582. http://doi.acm.org/10.1145/988672.988750, http://dx.doi.org/10.1145/988672.988750.
[29] R. Pan, M. Ochi, Y. Matsuo, Discovering behavior patterns from social data for managing personal life, in: 2013 AAAI Spring Symposium Series, 2013.
[30] D. Dearman, K.N. Truong, Identifying the activities supported by locations with community-authored content, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp '10, ACM, New York, NY, USA, 2010, pp. 23–32. http://doi.acm.org/10.1145/1864349.1864354, http://dx.doi.org/10.1145/1864349.1864354.
[31] D. Dearman, T. Sohn, K.N. Truong, Opportunities exist: continuous discovery of places to perform activities, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM, New York, NY, USA, 2011, pp. 2429–2438. http://doi.acm.org/10.1145/1978942.1979297, http://dx.doi.org/10.1145/1978942.1979297.
[32] J. Goncalves, V. Kostakos, S. Hosio, E. Karapanos, O. Lyra, Inclucity: Using contextual cues to raise awareness on environmental accessibility, in: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '13, ACM, New York, NY, USA, 2013, pp. 17:1–17:8. http://doi.acm.org/10.1145/2513383.2517030, http://dx.doi.org/10.1145/2513383.2517030.
[33] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, M. Dredze, Annotating named entities in twitter data with crowdsourcing, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 80–88. URL http://dl.acm.org/citation.cfm?id=1866696.1866709.

[34] N. Eagle, txteagle: Mobile crowdsourcing, in: N. Aykin (Ed.), Internationalization, Design and Global Development, in: Lecture Notes in Computer Science, vol. 5623, Springer, Berlin, Heidelberg, 2009, pp. 447–456. http://dx.doi.org/10.1007/978-3-642-02767-3_50.
[35] A. Kulkarni, P. Gutheim, P. Narula, D. Rolnitzky, T. Parikh, B. Hartmann, Mobileworks: Designing for quality in a managed crowdsourcing architecture, IEEE Internet Comput. 16 (5) (2012) 28–35. http://doi.ieeecomputersociety.org/10.1109/MIC.2012.72.
[36] A. Gupta, W. Thies, E. Cutrell, R. Balakrishnan, mclerk: Enabling mobile crowdsourcing in developing regions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, ACM, New York, NY, USA, 2012, pp. 1843–1852. http://doi.acm.org/10.1145/2207676.2208320, http://dx.doi.org/10.1145/2207676.2208320.
[37] A. Kittur, E.H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, ACM, New York, NY, USA, 2008, pp. 453–456. http://doi.acm.org/10.1145/1357054.1357127, http://dx.doi.org/10.1145/1357054.1357127.
[38] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, M. Vukovic, An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets, 2011. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778.
[39] Z. Zhu, U. Blanke, A. Calatroni, G. Tröster, Prior knowledge of human activities from social data, in: Proceedings of the 17th International Symposium on Wearable Computers, ISWC '13, 2013.
[40] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, Int. J. Data warehous. Min. (IJDWM) 3 (3) (2007) 1–13.
[41] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, Decision trees for hierarchical multi-label classification, Mach. Learn. 73 (2) (2008) 185–214. http://dx.doi.org/10.1007/s10994-008-5077-3.
[42] R.E. Schapire, Y. Singer, Boostexter: A boosting-based systemfor text categorization, Mach. Learn. 39 (2-3) (2000) 135–168. http://dx.doi.org/10.1023/A:1007649029923.
[43] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, J. Mach. Learn. Res.
[44] T. Joachims, Text categorization with suport vector machines: Learning with many relevant features, in: ECML98, 1998.
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res.
[46] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[47] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, in: W. Buntine, M. Grobelnik, D. Mladeni, J. Shawe-Taylor (Eds.), Machine Learning and Knowledge Discovery in Databases, in: Lecture Notes in Computer Science, vol. 5782, Springer, Berlin, Heidelberg, 2009, pp. 254–269. http://dx.doi.org/10.1007/978-3-642-04174-7_17.
[48] Y. Kim, F. Pereira, F. Zhao, A. Ghorpade, P. Zegras, M. Ben-Akiva, Activity recognition for a smartphone based travel survey based on cross-user history data, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 432–437. http://dx.doi.org/10.1109/ICPR.2014.83.
[49] Y. Kim, F.C. Pereira, F. Zhao, A. Ghorpade, P.C. Zegras, M.E. Ben-Akiva, Activity recognition for a smartphone and web based travel survey, CoRR abs/1502.03634. http://arxiv.org/abs/1502.03634.