Human Activity Recognition Using Social Media Data

Zack Zhu, Ulf Blanke, Alberto Calatroni, Gerhard Tröster
Wearable Computing Lab, ETH Zurich
Zurich, Switzerland

{zack.zhu, ulf.blanke, alberto.calatroni, troester}@ife.ee.ethz.ch

ABSTRACT

Human activity recognition is a core component of context-aware, ubiquitous computing systems. Traditionally, this task is accomplished by analyzing signals of wearable motion sensors. While such signals can effectively distinguish various low-level activities (e.g. walking or standing), two issues exist: First, high-level activities (e.g. watching movies or attending lectures) are difficult to distinguish from motion data alone. Second, 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 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. Applying an L1-regularized Linear SVM on features derived from textual content, semantic location, and time, we are able to infer 10 meaningful classes of daily life activities with a mean accuracy of up to 83.9%. Our work illustrates a promising first step towards comprehensive, high-level activity recognition using free, crowd-generated, social media data.

Keywords

Web Mining; Activity Recognition; Crowd Sensing

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

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, include accelerometers, gyroscopes, and sometimes magnetometers, which acquire the user's motion and thereby his physical movements. Using such sensors, activities that have been investigated range from low-level ones (standing, sitting, or walking) [12] to physical activity (cycling or working out) [1, 14] to higher level routines (having dinner or commuting) [2, 8] that consist of numerous sub-activities (preparing food, eating, clearing the table).

With the advent of the smartphone and the 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 followers. Messages are usually short: just 140 characters with optional image attachment in the case of Twitter. According to Twitter's official blog [25, 24], 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 [9], 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 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" [17]. Second, "daily chatters" share content multiple times a day [9]. 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 [16]. 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 extra 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 believe that data collected by social

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

MUM '13, December 02 - 05 2013, Luleå, Sweden

ACM 978-1-4503-2648-3/13/12$15.00.
http://dx.doi.org/10.1145/2541831.2541852

media platforms can be directly fed into activity- or context-aware systems. We believe that artefacts from user-social platform interaction can be understood as a reflection of the user's daily activities.

However, since users are not systematically submitting their life activities onto social platforms, the data content we access is unstructured, ambiguous or can contain manifold activities. 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. In addition, users are not constrained beforehand to specific daily situations or activities for self-reporting. This opens the question as to how to define a common scheme for *activity*. In this work, we make use of a standardized activity taxonomy from the American Time-Use Survey (ATUS) [22]. 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 [19, 3]. We select this taxonomy for its relevancy, comprehensive coverage, and also its overlap with other activity surveys from healthcare: such as the social rhythm metric or activities of daily living [15, 11].

In this paper, we investigate the potential of harvesting and extracting publicly self-reported activities through social media. Utilizing 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 labeling of noisy social media posts to identify human activities?
- Can we automatically estimate the activity of a user from social media posts?

Towards these two questions, we make the following contributions:

- We present an architecture for gathering and labeling 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 labeling 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 behavior.
- Finally, we construct an activity recognition model capable of recognizing 10 types of activities plus a null-class with an overall accuracy of up to 83.9%. We analyze three sets of features (textual content, geo-mapped location semantics, and time) to quantify their respective importance in inferring different classes of activities.

In Section 2, we first review existing work in the area of activity-related research using social media. Then, in Section 3, we present our system architecture and the crowdsourced labeling 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 the crowd-labeled dataset. We present quantitative results evaluating our approach in Section 6 and discuss our findings in Section 7. Finally, we conclude and provide an outlook for our work in Section 8.

2. RELATED WORKS

Previous work have investigated the use of "freely-available" information to augment the performance of AR systems without additional instrumentation. For example, temporal features have been leveraged by Ye et. al. [27, 28] 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 [13]. Although useful, these studies present experiments conducted with 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 [21] utilize 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, were exploited to aid context-aware systems. Patridge et. al. [19] leveraged the American Time-Use Survey data to learn mappings between location semantics and activities. This work is extended by [3] 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 (e.g. via Foursquare to obtain venue type).

One rationale of our approach to use large-scale social data comes from [16], 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. [18], communication data from Gmail, Facebook, and Twitter are analyzed to extract daily behavior patterns. Based on an assumed correlation to daily life behavior, 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 [5] shows that it is possible to utilize Yelp reviews to identify potential activities. As opposed to defining activity classes *a priori*, they parse for nearby verb-noun pairs to extract activity descriptions (e.g. buy book, appreciate art) as they are encountered in the textual data. 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 [4]. Similar to [19], [5, 4] generates a model of *potential* activities for various venues. It does not attempt to infer the *current* activity that the user is engaged in. In addition, only 14 venues from four venue categories are examined. In our approach, we assume every microblog instance captures an activity, which may or may not be explicitly described by a verb-noun pair. Using the textual features in addi-

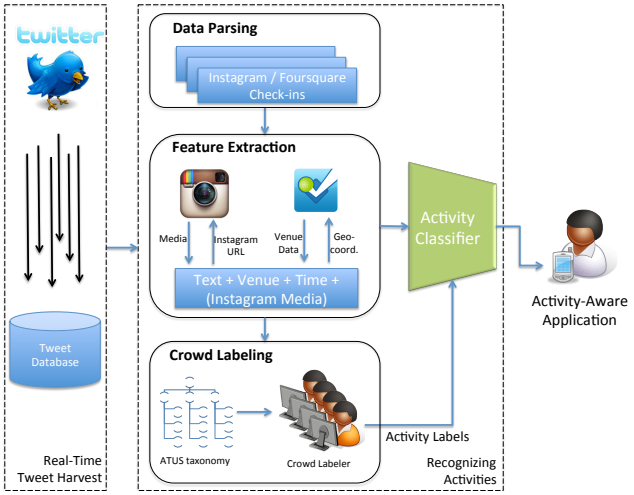

Figure 1: System architecture for gathering and modeling activities from in-situ, self-report social platform data.

tion to location and time, we construct a broader feature set, which is then used by a machine learning algorithm to learn the mapping between these signals and activity classes.

3. GATHERING SELF-REPORTS FROM SOCIAL MEDIA DATA

In Figure 1, we present our system architecture. In the following two subsections, we will discuss the data collection and parsing process, how we extract additional contextual information to build features, and finally how the data is labeled by crowd-workers. Then, in Section 5, we will discuss our activity recognition model.

3.1 Collecting Social Media Data

To gather the wealth of crowd generated self-reports, we use Twitter's streaming API[2] 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 Figure 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. 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 must filter out content irrelevant to what one is doing at the moment. Fortunately, as a general communication platform, Twitter is linked to by location-based services such as Foursquare and Instagram. There, users use their mobile devices to "check-in" to a geo-location while posting a text and/or photo as self-reports signaling their current activities. With a linked Twitter account, content generated for Foursquare or Instagram can be simultaneously Tweeted with a reference link. From these Tweets, one is able to identify the in-situ activity with much more clarity since additional context information (e.g. at a library) and even photographs that

[2]https://dev.twitter.com/docs/streaming-apis

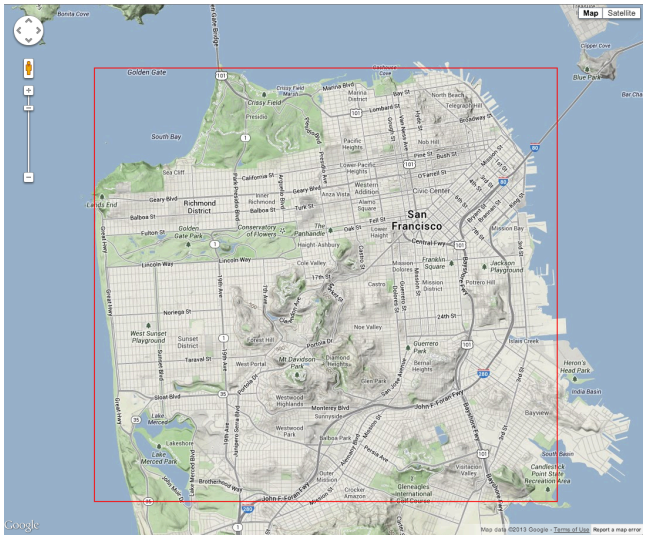

Figure 2: Map of geographical area in San Francisco from which self-report instances were gathered. The bounding box has coordinates of (37.7099, -122.5137) in the lower corner and (37.8101, -122.3785) for the upper corner.

capture the moment (Instagram posts) can be collected.

In this study, we filter for Foursquare and Instagram Tweets in our Twitter dataset as a rough filter for activity-related self-reports. From the Feature Extraction block of Figure 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[3] by looking up the geo-coordinates of 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, our model is able to recognize activities with just the text contained in the tweet; nevertheless, the additional pieces of information allow us to build contextual information that clarify the ground-truth labeling process and improve classification performance in our activity recognition system.

3.2 Labeling for Self-Reported Activities

As mentioned previously, 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 [29], the inherent bias in social media data is skewed and underrepresents 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 [29] *Socializing, Relaxing, & Leisure; Eating & Drinking; Sports, Exercise, & Recreation; Consumer Purchases; Work-Related; Education; Traveling; Professional Services; Household Activities; Personal Care*. In addition, we include

[3]https://developer.foursquare.com/docs/venues/search

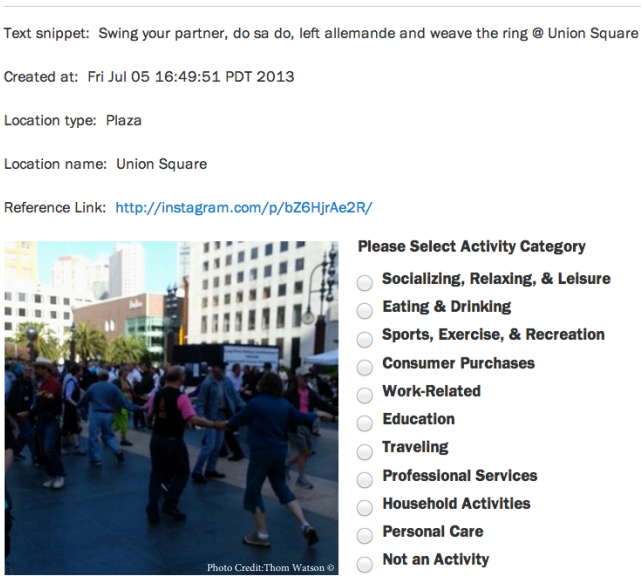

Figure 3: Screenshot of a task instance, in which a crowd labeler is asked to provide the ground truth to the user's current activity. Upon mouseover, example activities for each category would appear to guide the decision.

a null class in case an activity is not clearly apparent from the instance information.

Using these categories, we crowdsource the task of manually inferring activities from instances by using the CrowdFlower[4] platform. Although Amazon Mechanical Turk (AMT) is a popular platform for crowdsourcing labeling tasks, we select CrowdFlower for two reasons: first, it distributes our tasks on tens of crowdsourcing platforms, including AMT. Of the 579 workers who elected to complete our tasks, the largest contributing platform is InstaGC with 175 workers (∼30%). From Prodege and Neodev, we received 144 and 106 workers, respectively. On the other hand, only 24 workers were from AMT. Therefore, we believe CrowdFlower allows us to reach a larger pool of workers who are prompt in responding to our tasks. Second, AMT currently requires the task purchaser to have a U.S.-based credit card to be able to pay for labor. This is relatively inconvenient for international researchers.

In Figure 3, an example instance is shown to demonstrate what a crowd-worker sees. We deploy tasks to online workers from Canada and the United States with the following instructions:

> *Given a tweet describing an Instagram upload, categorize the* ***main activity*** *that the user is trying to capture at that moment.* ***Please mouseover first (press alt while mouseovering if necessary) to familiarize yourself with activity examples of each category.***
>
> *Please base your selection 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.*

For the 6004 tasks we post on CrowdFlower, a budget of $550 USD is applied and the tasks are finished within 48 hours of posting. To assure labeling quality, CrowdFlower collects "gold" labels from

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

Label Category	Instances	Proportion
Socializing, Relaxing, & Leisure	2627	43.75%
Eating & Drinking	1442	24.02%
Sports, Exercise, & Recreation	669	11.14%
Work-Related	390	6.50%
Consumer Purchases	316	5.26%
Not an Activity	227	3.78%
Traveling	188	3.13%
Education	65	1.08%
Professional Services	52	0.87%
Personal Care	17	0.28%
Household Activities	11	0.18%

Table 1: Distribution of labels in dataset as generated by crowd workers.

experimenters in order to assess the qualification and seriousness of crowd workers. We manually labeled 170 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 4 "gold" units correct before proceeding to non-"gold" units. Second, based on labeling 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. To be robust to noisy labels, we tune the task redundancy to 3, which implies every task will be rated by at least 3 workers. However, CrowdFlower utilizes the trust score to increase this number automatically if initial workers who complete this task have low trust scores.

4. CROWD-GENERATED ACTIVITIES DATA

In this section, we analyze the 6004 Tweeted instances of Foursquare "check-ins" and Instagrams that received activity category labels. In Table 2, we provide some sample instances to make intuitive the appearance of our raw data. In the rest of this section, we demonstrate that, although unevenly distributed and affected by labeling noise, our dataset captures a wide variety of activities and is coherent with common understandings of people's daily activities.

4.1 Activity Label Distribution

From Table 1, we find that the majority of data (>78%) is contributed by the top 3 categories: "Socializing, Relaxing, & Leisure", "Eating & Drinking", and "Sports, Exercise, & Recreation". This is not surprising since leisurely experiences are easily shared as opposed to demanding activities, such as work-related ones. Although there is significant bias in the data we collected, we notice that even the least populated categories contain some instances. Therefore, we believe prolonged data collection will alleviate the lack of data issue for the lesser represented categories.

4.2 Ambiguity of Daily Activities

The short and unstructured nature of Tweets is problematic for labeling in two ways: First, there may not be enough information contained within the text to indicate the engaged activity. For example, the Tweet *"Winner of my San Francisco Gibraltar Challenge: Four Barrel. (close second: Blue Bottle) #coffee"* from an Instagram does not provide sufficient indication as to the user's activity. To decrease the labeling ambiguity, we supplement the text with contextual information via venue name (Four Barrel Coffee) and type (coffee shop). With the added information, it becomes

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

Table 2: Table illustrating sample data instances harvested from Tweeted "check-ins" in the San Francisco metropolitan region.

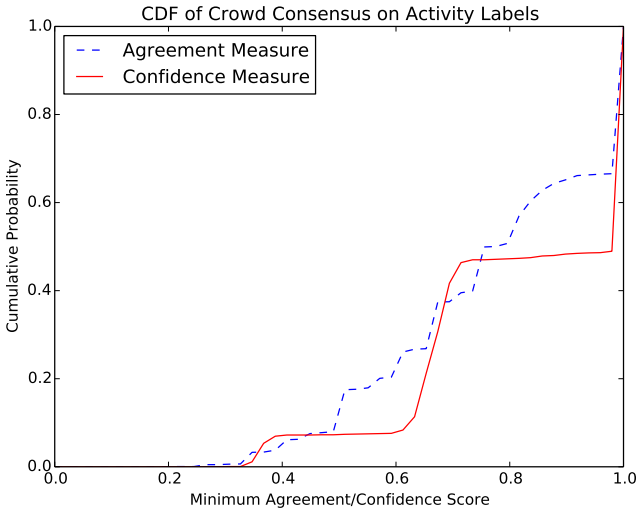

Figure 4: Empirical CDF of crowd consensus on activity category labels. The random variable X denotes the minimum score for agreement or confidence.

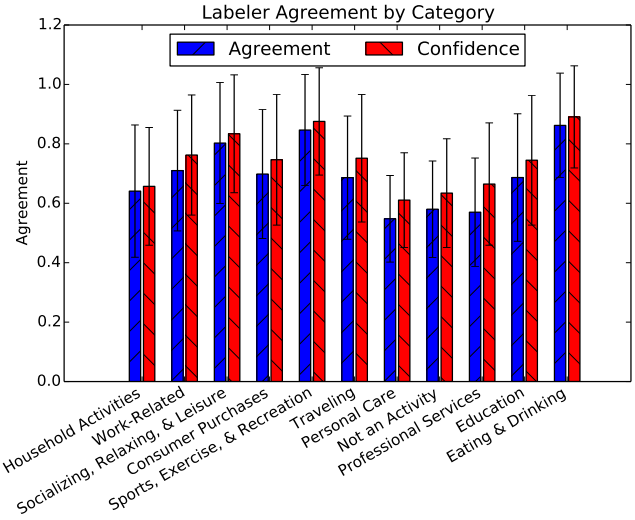

Figure 5: Agreement and CrowdFlower confidence scores of different activity categories as a fraction between 0 and 1.

more likely that the user is engaged in a work-related activity. In addition, the labeler can view the attached photo that captures the immediate activity as well as photo comments via the original Instagram URL. Second, some activities can be interpreted to fit multiple activity categories. For example, an activity where friends eat or drink together can be considered an "Eating & Drinking" activity as well as "Socializing, Relaxing, & Leisure". In our experiment, we ask labelers to judge the main activity type and conduct majority voting to reach consensus. In future work, we intend to allow activities to receive multiple labels and train our activity recognition model via multi-label learning.

To quantify the labeling noise in our dataset, we use two metrics provided by CrowdFlower to measure consensus: agreement and confidence. Agreement is simply the number of labels in the majority category over all labels assigned. According to CrowdFlower[5], the confidence score is based on agreement weighted by worker trust score, although the original function not given. Both scores range from 0 to 1.

In Figure 4, we plot the empirical cumulative distribution function (CDF) of minimum agreement and confidence scores. From the figure, we see that a larger portion of our data received quite high confidence scores from CrowdFlower (almost 50% of instances received a confidence score close to 1.0). In terms of label agreement, about 50% received an agreement between 0.8 and 1.0. In Figure 5, we show the mean agreement and confidence scores for each activity category with their standard deviation (as error bars). We see that leisurely activities (Socializing, Relaxing & Leisure; Eating & Drinking) are the least ambiguous (0.84 & 0.86) while Professional Services, Personal Care, and Not an Activity (0.54-0.58) are the most ambiguous to label. Upon deeper examination of the data, we notice that many Professional Service activities (e.g. getting a haircut or nail treatment) were mistakenly labeled as Personal Care. Although the strict definition by the ATUS taxonomy (also reflected in our example activities) is to label these as Professional Services (since a service is rendered by a professional), such activities may be more intuitive as Personal Care to relatively indiligent

[5]https://crowdflower.com/blog/stopworrying/

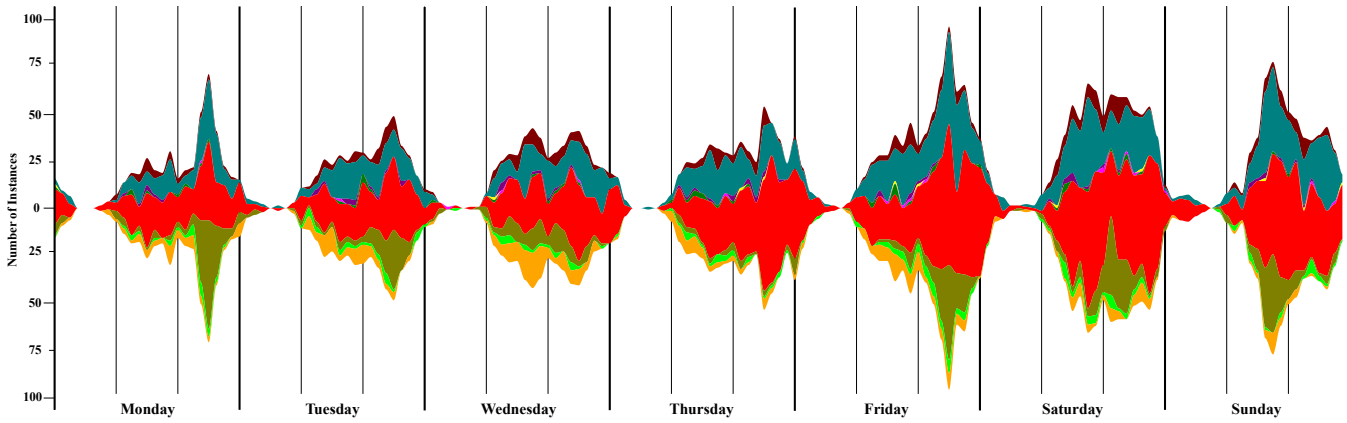

Figure 6: Weekly pulse of various activities in the San Francisco area. The varying heights of the wave depict variations in the number of instances for that time slot. The colored layers depicting activity categories are in order from top to bottom: Consumer Purchases (Maroon), Eating & Drinking (Teal), Education (Purple), Household Activities (Yellow), Personal Care (Fuchsia), Professional Services (Green), Socializing, Relaxing, & Leisure (Red), Sports, Exercise, & Recreation (Olive), Traveling (Lime), Work-Related (Orange).

labelers. As for the Not an Activity category, it is not surprising the agreement is low because activities can be inferred for any instance depending on how many assumptions a labeler makes. In Section 6, we will more closely examine how the labeling ambiguity affects our model's ability to infer activity categories.

4.3 Weekly Pulse 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 natural environment.

In Figure 6, we plot a wave graph to show the 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.

As expected, the quantity of activity reports are much higher, per day, on weekends (Friday, Saturday, Sunday) than weekdays (weekend average: $\sim$1050 instances/day *vs.* weekday average: $\sim$714 instances/day). For each day of the week, a clear increase and decrease in activity levels can be observed to mark a 24-hour 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 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 modeling the mapping of these signals to ground truth activity categories.

As the main form of our data is unstructured, 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.

In the work of Dearman and Truong [5], part-of-speech tagging is used to extract explicit verb-noun pairs (e.g. purchase-book) to identify activities from crowd-generated text. However, we believe some phrases (e.g. at the park, it's a beautiful day for recreational activities) implicitly signal the user's activity without using verb-noun pairs. By using a 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.

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 labeling, we currently do not explore computer vision techniques to derive features for our model. To fuse the multiple sources of features, we simply concatenate our feature matrices. We leave tuning and other feature engineering to future work. We derive the following sets of features:

1. **Tweet Text:** By extracting unigrams and bigrams, we obtain 52,918 n-gram features from the tweeted text of each instance. We use Tf-Idf scaling to construct our textual feature

matrix. The feature weight is calculated as follows, where $tf_{t,d}$ is the 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:

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

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 of 267 features (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. From the venue names, we derive 6,318 n-gram features. 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. We chunk the time data by hour of the week and binarize to construct 167 features.

Given the large number of features from textual features, our feature matrix is large-scale, sparse, and high-dimensional. From text mining literature [6, 10], 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 [20] 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:

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

where $\boldsymbol{\omega}$ is the feature weight vector and $\|\cdot\|_1$ denotes the 1-norm [6]. As a result, implicit feature selection will take 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 $\boldsymbol{\omega}$ [7]. This allows us to compare the usefulness of the various feature sets. In the next section, we evaluate the results of our approach and provide some insight to which features are predictive of people's activities.

6. EXPERIMENTAL VALIDATION

We begin this section by introducing classification results using only data instances that received a agreement score of 1 (complete agreement from all labelers). With a classification of 83.9%, we show promise for building activity recognition models from social media data. Given the inherent ambiguity in activity definitions, we provide a sensitivity analysis revealing that classification performance degrades to ∼70% when the data is unfiltered for labeling agreement. Finally, we analyze the whole dataset and discuss the varying importance of the three feature sets.

Label Category	Instances	Precision	Recall	F1-Score
Socializing, Relaxing, & Leisure	1155	0.83	0.88	0.86
Eating & Drinking	792	0.84	0.85	0.85
Sports, Exercise, & Recreation	353	0.96	0.91	0.93
Work-Related	95	0.58	0.47	0.52
Consumer Purchases	79	0.74	0.58	0.65
Traveling	39	0.69	0.56	0.62
Education	13	0.78	0.54	0.64
Not an Activity	10	0.25	0.10	0.14
Professional Services	4	0.00	0.00	0.00
Household Activities	2	0.00	0.00	0.00
Personal Care	0	N/A	N/A	N/A
Average / Total	2542	0.83	0.84	0.84

Table 3: L1-Regularized Linear SVM classfication performance. The table shows the score of the average test-fold when 2542 instances are divided using five-fold cross-validation. These instances received full agreement amongst labelers (agreement score = 1.0). This multi-class classification task (11-classes) achieved an overall classification accuracy of 83.9%

6.1 Activity Classification Performance of Non-Ambiguous Data

Using the L1-Regularized Linear SVM described in the previous section, we conduct 5-fold cross-validation on 2542 labeled instances that have full labeling agreement and present the results in Table 3. We note that there is no data for the Personal Care category as all instances received agreement scores of less than 1. At this stage, we use the complete feature set, although feature selection is conducted implicitly due to the L1-regularization.

We obtain an overall mean testing accuracy of 83.9% and F1-score of 84%. Although this is relatively high, we note that many categories with a low number of instances (e.g. Household Activities) could not be correctly inferred. It is interesting to note that although lower number of training instances indeed negatively impact classification performance, classification performance can still be better for some classes that have less data. For example, the F1-score for the Education class is higher than both the classes of Traveling and Work-Related, which contain more data. This speaks to the relative suitability of our features in inferring some classes of activities over others. Since Education activities are specific to particular types of places (e.g. schools, library) and often described with specific words (e.g. read, study, etc.), it is understandable it achieves relatively higher performance despite less training data.

Since the performances are based on data with full labeling agreement, it is expected that the bottom four classes (Not an Activity, Professional Services, Household Activities, Personal Care) in terms of agreement (see Figure 5) are represented with only minimal data for classifier training. To address this issue in the future, we intend to increase our dataset size so that sufficient instances remain in the least populated classes even after filtering for high agreement data.

6.2 Impact of Activity Ambiguity on Activity Classification Performance

To study the influence of activity ambiguity on classification performance, we sub-select instances with various levels of agreement and train the same classifier to obtain corresponding classification accuracies. The result is plotted in Figure 7 where the x-axis mea-

sures the agreement score. We plot the classification accuracy as a solid line using the left y-axis and a dashed line indicating quantity of data sub-selected, which is specified with a second y-axis on the right. Even with unfiltered data (minimum agreement score of 0), system performance is still reasonable with 70.2% of accuracy. With maximum agreement, system performance increases to 83.9%. Interestingly, the classifier trained with data containing perfect agreement does not give the best performance (peak performance of 84.6% achieved via data with agreement of 0.8). Since our data quantity is limited, this could be due to the label quality vs. quantity trade-off selected by the minimum agreement score.

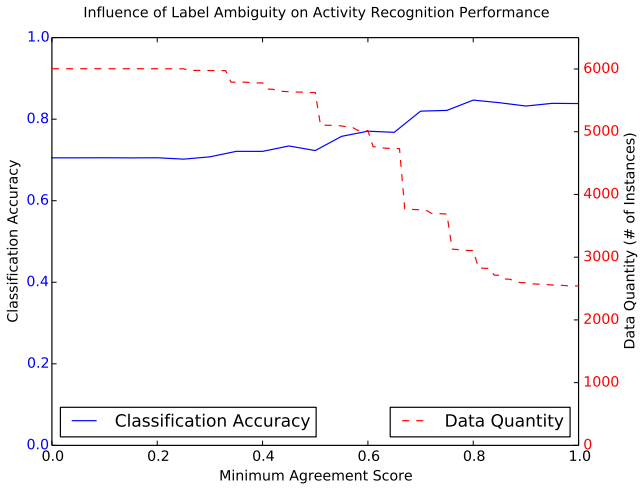

Figure 7: Sub-selecting data with various agreement scores (plotted on the x-axis) and classification accuracy (plotted on the y-axis), we notice classification accuracy is heavily influenced by the noise in data labels.

Calculating the Pearson correlation between agreement score and classification accuracy, we note a high level of correlation ($\rho = 0.94$, $p < 0.001$). Therefore, we believe the classification accuracy may be improved further by gathering additional data and increasing the number of labelers per instance. This would increase high-agreement data quantity and potentially increase agreement via a larger majority, respectively.

6.3 Feature Analysis for Activity Recognition Performance

One of the key questions that arises is the importance of each feature set (text, location, and time) in terms of their contribution to classification performance. This is significant since, practically, we may not always have access to location-based services (privacy or power consumption concerns) or venue type information. Then, our feature set would only include time and n-gram features.

In Figure 8, we visualize the F1-scores of the various activity classes for comparison. The differences between the three feature sets plotted is small (mostly 0.0-0.03) within each category while significant differences exist between categories (in accordance with Table 3). Therefore, the lack of a geo-lookup service will not significantly decrease activity recognition performance, as long as the content generated by the user 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 42.7%. For all categories, using all features result in the best performance.

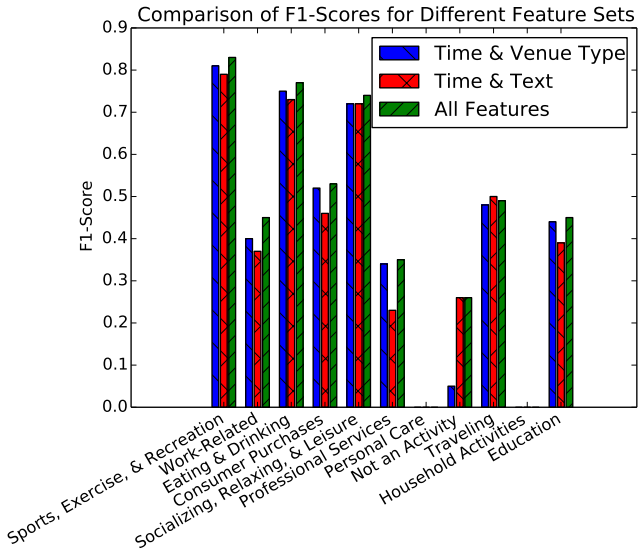

Figure 8: A comparison of F1-Scores for classifiers trained with all features, time and text features, and time and venue type features.

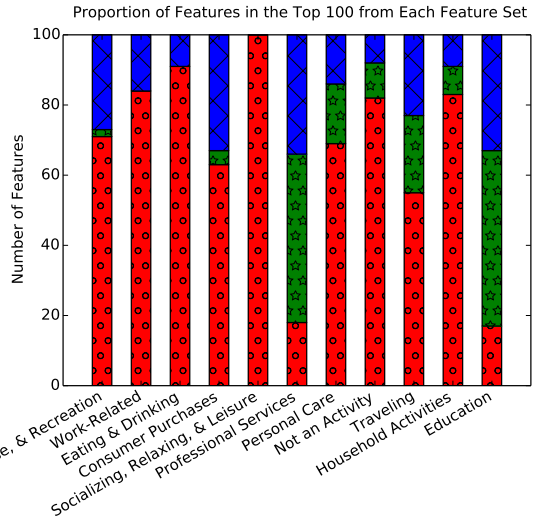

Figure 9: Proportion of Top-100 important features from each feature set. The feature sets stacked are textual (red, circles), time (green, stars), and venue type (blue, cross-hatch).

The overall accuracy for using all data (6004 instances), regardless of labeling agreement, is 68.0%, 68.1%, and 70.5% for Text + Time, Venue Type + Time, and All Features, respectively. This supports the intuition 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 important for making fine-grained adjustments. For illustrative purposes, we post the top 10 most indicative n-grams in Table 4, which are quite intuitive in categories with higher accuracy.

By sorting the magnitude of coefficients for all features, we can plot the composition of the top 100 most important features in terms of their feature type (text, time, venue type). From Figure 9, we notice textual features (represented by red bars) dominate the top-

Label Category	Vocabulary
Socializing, Relaxing, & Leisure	chowder, justin, thi view, dinner, merc, page, hiltonhotel hilton, firework, hood, card
Eating & Drinking	chowder, meet, fresh, hey, coffe, breakfast, fri, food, meal, box
Sports, Exercise, & Recreation	yoga, gg, hous air, golf, brief, girl club, room, netvib hq, run, garlic fri
Work-Related	radio, work, docusign, spin, beauti sf, confer, workin, team, bring, mobil
Consumer Purchases	shirt, jest, sale, bookstor, shop, thi photo, fabric, bought, cinta, cloth
Not an Activity	refresh, 111, exhbit, squar hiltonhotel, hiltonhotel hilton, ymca chinatown, spg, cutleri, blackheart, alexand wang
Traveling	foggi today, commut, train, 49, final san, shuttl, sfmta muni, gate bridg, presidio gate, bu
Education	scienc calacademi, depart educ, public librari, academi, hackbright, art build, ccsf, pic, school, jewish museum
Professional Services	pierc, dmv, campu, mortgag san, glass pro, appl store, barbershop, rafa, bodi, chase bank
Personal Care	franci, beauti skin, tick, brogan, skin care, flag footbal, flair spot, flair, flagship origin, flagship
Household Activities	zynga, flake truffl, flake crust, flake, flair spot, flair, flagship origin, flagship, flag ve, flag state

Table 4: Top 10 most indicative n-gram features for each activity category.

100 list for most categories. It is interesting to note that textual features are not descriptive of activities in Professional Services and Education while time and venue type features represent most of the top-100. This is reasonable since a large variety of text can be used to describe activities in these categories (as opposed to just "food-related" words for Eating & Drinking), which renders textual features relatively less useful. In addition, activities from these two categories tend to be specific to time (business hours) and places (schools, banks, etc.), which explains why time and venue category features are heavily represented in the top-100.

7. DISCUSSION

From Table 4, we see highly indicative vocabulary as n-grams that are relatively specific to San Francisco (e.g. chowder, yoga, foggi today). Although this may indicate 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 can select the appropriate city-scale model for activity recognition.

In an effort to clearly demonstrate the usefulness of social media data, we have not yet incorporated any traditional sensor data (e.g. accelerometer data) known to be useful for activity recognition. From Figure 8, we note the imbalance of predictability across various activity classes. We believe this can be improved by fusing traditional sensor data. However, our approach as is would already have useful applications. For example, understanding what types of activities are possible where, especially types well-classified by our approach, can be used to specify what one can *do* via key n-gram descriptors. 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. This fine-grained information would provide valuable insight for tourists who may be unfamiliar with the unique function of a local venue. For muti-purpose venue owners, our approach for activity recognition would provide valuable data on how people are utilizing their facilities. Potentially, insights can be generated to indicate the composition of activities (e.g. proportion of people shopping, eating and drinking, leisurely activities at a shopping center).

Even though we show the highly indicative 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. 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). Furthermore, 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 Traveling activity class in our label space) would naturally suffer. One way to remedy this would be to consider sequences of GPS readings to recognize large-area locomotion as in [13]. In addition, we believe another "free" data source, population-wide time use data, could also complement the activity recognition chain. 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. We believe this can be improved by taking advantage of population-scale time use data for activity prediction as attempted in [19, 3]. Ultimately, we believe 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 geographical availability is very appealing, the data is unstructured and can be ambiguous, even for human labelers. 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 daily activities with an overall accuracy of 83.9% if unambiguous labels are used for training. From our analysis of feature importance, we see that textual features perform almost as well as features capturing location semantics. However, when combining all feature sets, we achieve the best performance. We believe while both textual features and venue semantics are relatively indicative, together they allow more fine-tuning of the model to better represent subtleties in the data.

In future work, we intend to modify our crowd-labeling effort to allow multi-labeling, where an instance can receive multiple labels as to capture the inherent ambiguity of some activity instances. We will then formulate a multi-label learning problem [23] to train classifiers to output a probability distribution over the label space. We also intend to realize a mobile system to conduct real-time activity recognition. From batch learning results, we note the feature density is only 0.6% with the L1-regularized SVM. Therefore, we believe it is possible to store a classifier onboard the smartphone and conduct real-time classification of in-situ measurements as check-ins and/or textual signals are detected. To update our classifier, we would collect corrections from users when the classified activity is manually rejected. With these instances, we can conduct server-side, batch retraining and deployment of updated classifiers when internet connectivity is detected. We will deploy field trials and assess the usefulness and limitations of our system.

9. ACKNOWLEDGEMENT

We thank the anonymous reviewers for their valuable suggestions and comments. This work is partially supported by the Hasler Foundation through the SmartDAYS project.

10. REFERENCES

[1] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, pages 1–17, April 2004.

[2] U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *4rd International Symposium on Location- and Context-Awareness (LoCA)*, 2009.

[3] M. Borazio and 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*, pages 108–115. ACM, 2013.

[4] D. Dearman, T. Sohn, and 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, pages 2429–2438, New York, NY, USA, 2011. ACM.

[5] D. Dearman and 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, pages 23–32, New York, NY, USA, 2010. ACM.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[8] T. Huynh, U. Blanke, and B. Schiele. Scalable recognition of daily activities with wearable sensors. In *3rd International Workshop on Location- and Context- Awareness (LoCA 2007)*, page 50–67, september 2007.

[9] A. Java, X. Song, T. Finin, and 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*, pages 56–65. ACM, 2007.

[10] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML98*, 1998.

[11] S. Katz, T. Downs, H. Cash, and R. Grotz. Progress in development of the index of ADL. *The Gerontologist*, 10(1 Part 1):20, 1970.

[12] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, and 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*, page 27. ACM, 2011.

[13] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *Int. J. Rob. Res.*, 26(1):119–134, Jan. 2007.

[14] D. Minnen, T. Starner, I. Essa, and C. Isbell. Discovering characteristic actions from on-body sensor data. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers (ISWC)*, 2006.

[15] T. H. Monk, E. Frank, J. M. Potts, and D. J. Kupfer. A simple way to measure daily lifestyle regularity. In *European Sleep Research Society*, 2002.

[16] M. Naaman, J. Boase, and 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*, pages 189–192. ACM, 2010.

[17] A. Oulasvirta, E. Lehtonen, E. Kurvinen, and M. Raento. Making the ordinary visible in microblogs. *Personal and ubiquitous computing*, 14(3):237–249, 2010.

[18] R. Pan, M. Ochi, and Y. Matsuo. Discovering behavior patterns from social data for managing personal life. In *2013 AAAI Spring Symposium Series*, 2013.

[19] K. Partridge and P. Golle. On using existing time-use study data for ubiquitous computing applications. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 144–153. ACM, 2008.

[20] 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, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[21] M. Perkowitz, M. Philipose, K. Fishkin, and D. J. Patterson. Mining models of human activities from the web. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 573–582, New York, NY, USA, 2004. ACM.

[22] K. J. Shelley. Developing the american time use survey activity classification system. *Monthly Lab. Rev.*, 128:3, 2005.

[23] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[24] Twitter. 200 million tweets per day. https://blog.twitter.com/2011/200-million-tweets-day, June 2011.

[25] Twitter. Twitter turns six. https://blog.twitter.com/2012/twitter-turns-six, Mar. 2012.

[26] K. Van Laerhoven, D. Kilian, and B. Schiele. Using rhythm awareness in long-term activity recognition. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 63–66. IEEE, 2008.

[27] J. Ye, A. K. Clear, L. Coyle, and S. Dobson. On using temporal features to create more accurate human-activity classifiers. In *Artificial Intelligence and Cognitive Science*, pages 273–282. Springer, 2010.

[28] J. Ye, L. Coyle, S. Dobson, and P. Nixon. Using situation lattices in sensor analysis. In *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, pages 1–11. IEEE, 2009.

[29] Z. Zhu, U. Blanke, A. Calatroni, and G. Tröster. Prior knowledge of human activities from social data. In *Proceedings of the 17th International Symposium on Wearable Computers (ISWC '13)*, 2013.