# Understanding Context for Tasks and Activities

Jan R. Benetka
NTNU
benetka@pm.me

John Krumm
Microsoft Research
jckrumm@microsoft.com

Paul N. Bennett
Microsoft Research
pauben@microsoft.com

## ABSTRACT

Human activity is one of the most important pieces of context affecting an individual's information needs. Understanding the relationship between activities, time, location, and other contextual features can improve the quality of various intelligent systems, including contextual search engines, task managers, digital personal assistants, chat bots, and recommender systems.

In this work, we propose a method for extraction of an extensive set of open-vocabulary activities from social media. In particular, we derive tens of thousands of *ongoing* activities from Twitter, where people share information about their past, present, and future events and, using attached metadata, we establish spatiotemporal models of these activities at the time of posting. While public Twitter content is subject to self-censorship (not all activities are tweeted about), we compare extracted data with unbiased survey data (ATUS) and show evidence that for *activities which are tweeted about*, the underlying spatiotemporal profiles correctly capture their real distributions of activity conditioned on time and location. Next, to better understand the set of activities present in this dataset (and what role self-censorship may play), we perform a qualitative analysis to understand the activities, locations, and their temporal properties. Finally, we go on to solve predictive tasks centered on the relationship between activity and spatiotemporal context that are aimed at supporting an individual's information needs. Our predictive models, which incorporate text, personal history and temporal features, show a significant performance gain over a strong frequency-based baseline.

**ACM Reference Format:**
Jan R. Benetka, John Krumm, and Paul N. Bennett. 2019. Understanding Context for Tasks and Activities. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3295750.3298929

## 1 INTRODUCTION

It has been shown that human activity plays a major role in affecting what information needs people have. A study by Sohn et al. [34] recognized activity among four most frequent contextual triggers of information needs, which is in line with findings by other authors [9, 15]. While the majority of context-related prior work concentrates on time and location [1, 4, 36], activity as a higher-level driver of information needs remains relatively unexplored with only a handful of publications dedicated to it [3, 10, 23, 37]. In the following two examples of potential context-aware applications, which lead us to investigate the underlying predictive technology, we illustrate how including activity as context would impact a user's information access:

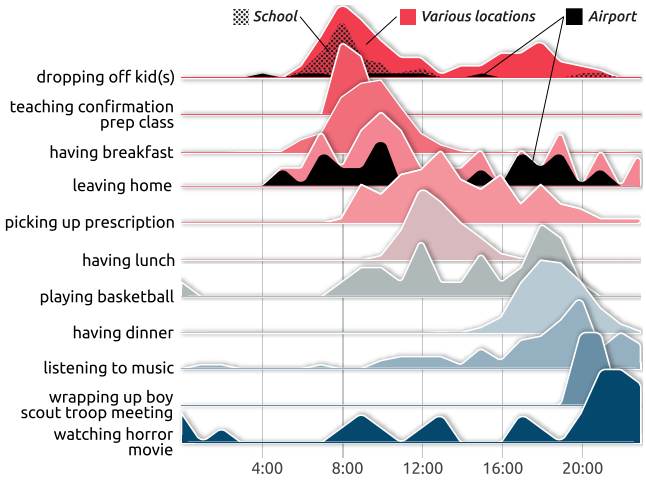



**Figure 1: Sample of human activities extracted from Twitter demonstrating the dependencies between activities, time and selected locations (school, airport, various locations).**

**a) Activity-aware reminder system.** Reminder systems allow users to set notifications for future tasks, events or activities to avoid forgetting about them. The notifications are typically invoked when a predefined date/time is met or a specified location boundary is crossed. However, Rong et al. [32] discovered that 40% of people cannot or prefer not to identify a precise time of their tasks/to-do items. This is one reason why an intelligent reminder system should not rely solely on hard-coded conditions. Instead, it should recognize activities mentioned in the reminder message (e.g., *read the article*) and proactively notify the user at any location and time when such an activity has a chance of being performed (e.g., in a *café* during the morning or at *home* in the evening).

**b) Activity recommender.** Many applications would benefit from a model that takes the user's location and time as input and returns a list of activities ranked according to the probability of being performed. Two such examples could include an automatic suggestion of a person's activity for a status update on a social network (e.g., *Enjoying coffee @ Café Lyst*) or recommendation of activities for a user's upcoming trips (e.g., *Lake Como: cycling*).

In both scenarios, as well as in similar activity-aware systems, the underlying models would be expected to capture the spatiotemporal dependencies of a wide range of open-domain activities. Ideally, the set of recognized activities should cover all activities a person might wish to be recommended or reminded about. In


*CHIIR '19, March 10–14, 2019, Glasgow, Scotland Uk*

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298929

the long-term, a predictive model between open-domain activities, locations, and time provides a simple foundation for commonsense reasoning about the world – long a goal of AI. That is, the implicit knowledge that certain activities, locations, and times correspond with each other (e.g., breakfast happens in the morning; work usually happens between 9-5 on weekdays but can be observed at other times; at a restaurant one eats but one also meets friends and celebrates milestones). We leave how to leverage these models in more general intelligence as future work and focus here solely on understanding and predicting activities in a spatiotemporal context.

However, such an endeavor requires a large dataset of human activities. This lack of a large-scale set of activities has been one of the impediments to advancing our understanding of activities and their contextual setting as well as how they might relate to tasks and information needs. In this work, we take a step toward alleviating this lack by demonstrating that activities extracted from a large-scale publicly available text source (Twitter) correspond with time and location patterns of activities reported in an independent broad survey. In contrast to surveys with predefined activity categories, textual sources can hold descriptions for tens of thousands of activities from an open vocabulary, ranging from fine-grained and very specific physical actions (e.g., "fixing a bike seat") to high-level cognitive notions (e.g., "thinking"). Prior work has proposed a number of methods for activity extraction from various textual sources [11, 13, 16, 27]. None of them, however, proceeds to jointly model the relation between activities, locations and time (e.g., "*dropping off kid(s)*" example in Fig. 1). In order to extract activities along with these two dimensions, i.e., location and time, reliable in-the-moment activity reports are necessary. ATUS[1], a survey-driven dataset with detailed records about activities of American citizens, on its own, provides rich and credible information. The taxonomy of activities it uses, however, is limited in size and would hardly cover many real scenarios. A shortage of variety is by no means a problem of microblogging platforms such as Twitter, where millions of people share updates about what they do [29].

In this work, we leverage the fact that people have become self-reporters of their own activities on social media, indirectly providing context and expressing their activities with terms that are not limited to any predefined taxonomy. We harvest Twitter for rich textual representations of tens of thousands of activities and simultaneously model their temporal and spatial dimensions by capturing timestamps and geospatial tags at the time of posting. To demonstrate the reliability of in-the-moment reports on Twitter, we are the first to report evaluation against an externally collected large-scale survey, ATUS.

We address the following research questions:

**RQ1** **What are the activities that people perform? How can we obtain an extensive set of them?** We propose a method for extraction of activities that people engage in. We demonstrate the suitability of Twitter as a self-reporting platform for extraction of reliable spatiotemporal properties. (§3)

**RQ2** **When and where do people engage in these activities?** We analyze temporal and spatial profiles of extracted activities to reveal the underlying patterns. (§4)

**RQ3** **Given an activity of a person (and possibly other context), can we predict where it is likely to happen?** We propose and evaluate a model that leverages (personalized) spatiotemporal patterns of activities in order to predict semantic location. This relates to the 'task reminder' scenario. (§5)

**RQ4** **Given a location (and possibly other context), can we predict what activities people will engage in there?** We pose an inverted problem to the one in *RQ*3, addressing the 'activity recommender' scenario. (§5)

In short, this paper makes the following novel contributions:

- Establishes an extensive set of human daily activities;
- Separates ongoing activities from future or past activities in social network posts;
- Profiles the spatial and temporal aspects of ongoing activities;
- Evaluates a sample of these activities against an external survey of activity and spatiotemporal context;
- Proposes several predictive tasks centered on activity and inspired by real-life scenarios, where using activity as additional context would make a profound difference;
- Proposes and evaluates predictive models for the aforementioned tasks.

## 2 RELATED WORK

Our aim is to profile and predict activities of all complexities while using social media as our source of evidence. Here, we review relevant prior work.

**Activity recognition**. In contrast to the traditional AI view of activity recognition using wearable/wireless sensors [8] or video [21], widespread connected mobile devices allow for collecting large amounts of contextual data about people's daily activities in a non-intrusive way. Java et al. [18] and Naaman et al. [29] have shown that users of microblogging platforms primarily post updates about their daily routines and experiences, which makes social networks by far the most scalable and diverse source of in-situ information about people's daily activities. Alternative large-scale sources of data that facilitate activity extraction [11, 27] are community-authored reviews of places (e.g., Yelp). While the reviews provide a means for mapping activities to locations, they cannot make a similar relation with time due to the ex post nature of reviews.

Methods to extract activities from textual sources range from manual curation of lists or taxonomies [19, 20, 33], through various supervised machine learning [16] or language modeling [13] techniques, to the application of natural language processing (NLP) [11, 27]. The high quality of manually crafted taxonomies [33], driven by the expertise of their curators, is counterweighted by the limited scalability of human labor involved in the process. Approaches relying on machine learning methods [16] require labeled instances of data and unsupervised language or topical models [13, 40, 41] are robust, nevertheless, they represent latent activities as a probabilistic distribution of words or concepts instead of assigning a concrete label to each activity. Finally, methods based on NLP techniques interpret the concept of activity [11, 27] or a task [12] as a verb-noun phrase pairs (e.g., *drinking coffee*), which they extract using a pattern-based paradigm. These methods are scalable and precise, however, they fail to capture expressions that are not covered by predefined patterns.

[1] American Time Use Survey (12 years, 170K participants)

**Activity classification and prediction.** A parallel line of literature focuses on activity classification, where a target set of activities already exists. Zhu et al. [42] train a model to label geo-located tweets with top-level activity classes from the ATUS hierarchy [33]. Their approach relies on crowdsourced annotations and contextual features from location-based social networks (LBSN) that include check-in time, venue name and venue category. A similar approach is adopted by Beber at al. [2] with the goal of inferring activities of moving objects based on their trajectories. In addition, the authors profile typical durations of activities, which is a feature of the work by Melià-Seguí et al. [28]. A more user-centric approach that learns activity expressions for individual users and simultaneously transfers training instances between classifiers to fight data sparsity is presented by Song et al. [35]. Weerkamp and de Rijke [38] extend the pool of microblog prediction problems with the task of activity prediction. In their pilot study, which has elements of event-extraction techniques, the authors suggest mining of likely near-future activities from the Twitter stream using a set of user-defined keywords and timeframes.

**Activity and location**. Arguably, activities that are typically performed at a location define its meaning to some degree [10] (e.g., *café* is a place where people *drink coffee*). Therefore, in some literature, activities and locations are used jointly [5] or interchangeably [17]. A common practice is to use semantic location labels as a proxy to high-level activities (e.g., *office* → *working*) [3, 23, 39]. Given the fact that detecting dominant functional places such as *home* or *work* from people's daily routines is feasible [22], such an approach is an understandable simplification. Nevertheless, a considerable number of locations can be a scene of more than one activity at the same moment [2] ("*Airport*" location in Fig. 1), and one activity can be characteristic of various locations (e.g., *"taking picture"*).

In our work, we use large volumes of tweets that are interlinked with Foursquare check-ins. This allows us to benefit from the text-to-location mapping as in [11, 27]. While Twitter users share information about past, current and future events, we only focus on messages that provide in-the-moment reports (contrary to [38] who target future activities). We extract a multitude of fine-grained activities at each location type, as opposed to [3, 39]. With respect to our data source, we use scalable NLP methods for activity extraction, and we evaluate our results against data from a handcrafted taxonomy of activities and locations (ATUS) [33]. In contrast to the activity classification literature [2, 42], we represent activities with their textual footprint, not with an abstract high-level class. We aim for establishing an extensive set of activities, answering the call by Brush et al. [7].

# 3 ACTIVITY EXTRACTION

Apart from time and location, human activity is one of the most influential cues affecting the information needs of an individual [15, 34]. While the values of time and location are known and inherently predefined (e.g., units of time, geo-coordinates, semantic labels of locations), the same cannot be said about human activities. The abstraction, ambiguity and variety of activities make it challenging to establish an exhaustive list of them. Yet, our first objective ($RQ1$) is to frame the notion of a *daily activity* and to create an extensive set of the most common activities that people perform and report on a daily basis. Let us first define the key concepts and describe our dataset before proceeding to the actual activity extraction methods:

*Definition 3.1.* **Human activity**, in the context of this work, is a real-world physical or cognitive activity of a person that she performs in space and time. The level of activity abstraction can vary from low-level (e.g., *moving*), through a more fine-grained activity expression (e.g., *riding on a bike*) to a very abstract notion (e.g, *enjoying the day*).

*Definition 3.2.* **Activity descriptor**, or simply **activity**, is a textual surface form (e.g., the actual text 'writing a report') referring to any activity that satisfies the definition of human activity.

*Definition 3.3.* **Temporal profile of an activity** is a histogram of activity observations over a specified temporal interval. In this work, we construct normalized daily and weekly profiles.

*Definition 3.4.* **Spatial profile of an activity** is a distribution capturing the normalized frequency of activity occurrences across a selected set of location categories. We work with location categories rather than individual locations to focus on patterns that generalize.

## 3.1 Data

Social networks can be seen as large crowdsourcing platforms with the potential to reveal the global picture of human activity behavior. While any social network that generates textual posts with time and location metadata could be used for our task, we found most reasons to use Twitter data: firstly, it is a widely used and well-researched microblogging platform; secondly, with the exception of our proprietary spam score weighting[2] all data is publicly accessible; finally, it has been shown that people tend to post updates that relate to their state or activity [29]. In our experiments, we operate with a sample of 6, 641, 503 geo-tagged Twitter posts that are restricted to the approximated region of the 48 contiguous states of the USA between January 1st 2016 and May 31st 2017. Further, we filter for tweets that are written in English, have a spam score below 0.5 and are cross-posted exclusively via the Foursquare application. Foursquare is a location-based social network, which is used by people to search for, retrieve details about and comment on nearby points of interest (POIs). Users can explicitly mark their visit to a POI by performing a *check-in* within the application, and many people choose to share this activity publicly via Twitter. These are the tweets that we benefit from in this work because they give us a means to pinpoint the user's precise location to the building, business or venue, and, importantly, the location's category. Locations on Foursquare are categorized in a multi-level hierarchy, where the top level has 10 categories (e.g., *Shop & Service*), the second level has 448 (e.g., *Bike Shop*).

## 3.2 Methods

Importantly, Twitter posts may express information about past, current, or future activity. We now present a simple and effective method for extraction of ongoing activities with their context.

[2]The spam score is an internally determined value that our organization applies to tweets. It is a strong function of the tweet's associated Twitter account.

*3.2.1 Preprocessing.* Twitter content is noisy, and the language is specific with heavy use of abbreviations and special characters such as emojis [14]. For that reason, we apply several preprocessing steps to minimize the noise in the data:

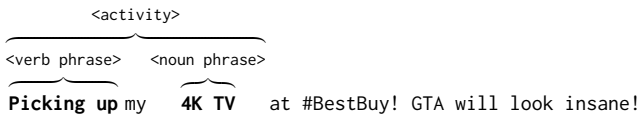
**Content cleaning** - we remove URLs, hashtags, and special characters (apostrophes, spurious white spaces, etc.).

**Foursquare pointer** - when a person checks in using the Foursquare mobile application with the sharing option switched on, the application automatically appends a string to the tweet with user's current location (e.g., `(@ Rolling Hills, CA)`). We remove these suffixes.

**Duplicate removal** - we consider two tweets to be duplicates when they are posted by the same author, have identical content (after preprocessing steps), and belong to the same Foursquare category.

**Timezone inference** - in order to calculate the local time of a tweet from its server publication time, we infer the tweet's timezone from its geo-coordinates using *timezonefinder*[3] and apply the time difference to get the local time.

*3.2.2 Activity Extraction.* Verbs or verb phrases are, by definition, sentence constituents that introduce an action (e.g., feed, go to). Nouns or noun phrases, on the other hand, typically fulfill the role of verbs' arguments (e.g., ducks, popular cafe). To extract activity descriptors from free text, we isolate the linguistically natural structure of *verb+noun* pairs (alternatively *verb phrase+noun phrase*, or combination) using a syntactic parser and a part-of-speech [6] grammar[4]. We denote the verb part as the *activity head phrase* (e.g., feed, go) and the complete pair as the *full activity descriptor* (e.g., feed ducks, go to popular café). Since we are interested in activities that take place at the time of their reporting, we only extract verbs in the present progressive tense, i.e., heuristically identified as verbs ending with '*-ing*' (e.g., feeding, going). In terms of generality, many languages contain tense markers or grammatical constructions that could be leveraged similarly (e.g., Romance or Slavic languages, Chinese, Japanese, Hindi). Future work could use tense as a weak label to bootstrap an in-progress activity descriptor extractor which could then be refined through label supervision to handle less common cases. Below, we provide an extraction example:




After extraction, articles and possessive *s*'s are removed, and the activity text is lemmatized in order to normalize the slight morphologic variations. Several activity examples are displayed in Figure 1 with the activity surface forms as labels on the y-axis. We note that a similar verb-noun extraction technique has been relied on in prior work [11, 12, 27], although, it is only our work that specializes in exclusive extraction of ongoing events.

[3]https://github.com/MrMinimal64/timezonefinder

[4]For the sake of reproducibility, the grammar is available in the Appendix (§A.3).

*3.2.3 Extraction Analysis.* To verify extraction precision, the first author conducted an evaluation by manually extracting verb+noun phrases that would be naturally understood as valid activity descriptors from 500 tweets, where activity was identified using our extraction algorithm. After comparing these results for an *exact match* with the method's output, we find that the precision is 73.6%. In the vast majority of these cases, the error does not mean that we extract activity from a tweet where there is none. The usual extraction errors are related to misspelling (e.g., *being niceâ*), not capturing the whole activity descriptor (e.g., *trying michelin* instead of *trying michelin starred sushi place*), or inclusion of extraneous text (e.g., *catching pokemon today*). False positives only account for 16% errors and are mostly caused by incorrect part of speech (POS) tagging due to limited context, which leads to the confusion of noun phrases with noun+verb phrases (e.g., *driving distance*).

*3.2.4 Semantic Location Extraction.* All tweets in our dataset come from the Foursquare mobile application and contain metadata about users' check-ins. We leverage timestamps of check-ins and their location, which, rather than by exact coordinates, are represented by the venue category (e.g., Grocery Store). The categories reflect semantic function of places [3, 39] and also increase generalizibility of our spatiotemporal profiles.

## 3.3 Results of Activity Extraction

We extracted 226,859 spatiotemporally anchored activities from which 101,869 are unique instances. The number of observations per activity follows a long-tail distribution with 82% of activities appearing only once in the whole dataset and only 2.2% of activities that are repeated ten or more times. These activities were reported by 33,116 users from which 62% performed more than one activity, though only 5% of users are associated with 25 or more activity records. The complexity and the abstraction level of extracted activities vary from very short and general (e.g., *running*) to lengthy and specific (e.g., *practicing egg drop soup delivery skill*). The most selective part of the extraction pipeline is the '*-ing*' filtering step which removes 71.5% of tweets. We would like to stress that our goal in this work is not to extract every activity mentioned in Twitter (high recall); rather we aim for precise extraction of activities that people are engaged in at the time of reporting. Also, we realize that people have individual preferences for when and where they tweet about which activities; however, in aggregate, we demonstrate in our evaluation that we collect a diverse sample that allows us to build reliable models. To aid reproducibility, the Appendix discusses the details of the extraction steps.

## 3.4 Evaluation: Comparison with ATUS

A primary research question, when deciding to work with the Twitter dataset, is whether and to which degree the microblog posts with their metadata can be trusted to reflect the true relationship between spatiotemporal context and each activity that is found there. We evaluated this by comparing the temporal profiles of certain Twitter activities with results of a large-scale survey of people's time use (ATUS) [33]. This comparison was designed to see if the Twitter profiles were relatively close to ground truth, helping to justify our deeper analysis in the remainder of the paper. ATUS (*American Time Use Survey*), a dataset which we use as our

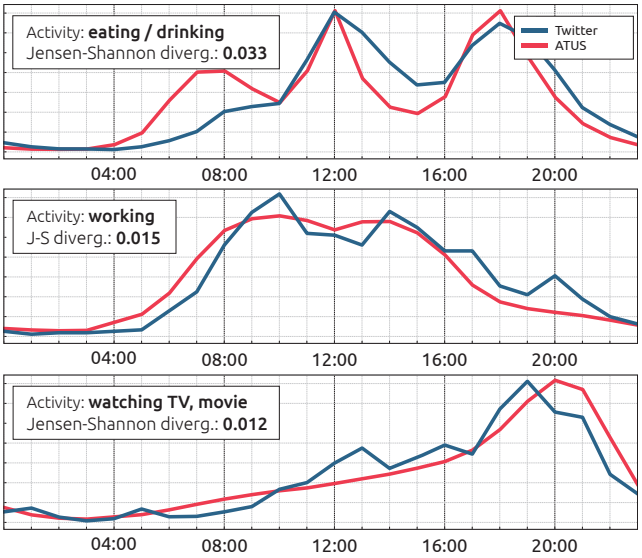



**Figure 2: Temporal profiles (normalized) of selected activities in ATUS (red) and Twitter (blue) and their divergence.**

ground truth, is an ongoing effort of the U.S. Census Bureau to collect detailed information about the ways Americans spend their time. The survey is conducted by telephone, and participants are asked to describe their day, locations they visited (26 categories), and their activities (17/105/438 categories in top 3 levels). In our study, we used data spanning from 2003 to 2015 which contain 3.35M observations from 170,842 participants.

**Evaluation method.** The difference in activity vocabularies between ATUS (limited hierarchy) and Twitter (open vocabulary) poses a challenge for direct comparison of counterpart activities. Therefore, we propose a set of evaluation approaches from the perspective of activities or locations that can be aligned manually in a straightforward way. The evaluation methods are:

**A) Quantitative (activity)** - Temporal profiles of selected activities are constructed using data from both datasets and the difference between two distributions is computed;

**B) Qualitative (activity)** - Typical activity locations in both datasets are compared and followed by discussion of differences;

**C) Quantitative (location)** - We compare temporal profiles of locations, for which all underlying activities are aggregated.

**Table 1: Activity comparison.**

| ATUS | Kahneman | Twitter (example) | $D_{JS}$ |
|---|---|---|---|
| *Socializing and communicating* | *Socializing* | *visiting friend, …* | 0.008 |
| *Television and movies* | *Watching TV* | *seeing movie, …* | 0.012 |
| *Relaxing, thinking* | *Housework* | *listening to music, …* | 0.014 |
| *Work, main job* | *Working* | *doing work, …* | 0.015 |
| *Shopping (exc. groceries/gas/food)* | *Shopping* | *doing shopping, …* | 0.016 |
| *Eating and drinking* | *Eating* | *eating lunch, …* | 0.033 |
| *Interior cleaning* | *Housework* | *doing laundry, …* | 0.041 |
| *Travel related to working* | *Commuting* | *heading to work, …* | 0.048 |
| *Physical care for children* | *Taking care of ch.* | *picking up baby, …* | 0.130 |
| *Food and drink preparation* | *Preparing food* | *making dinner, …* | 0.134 |

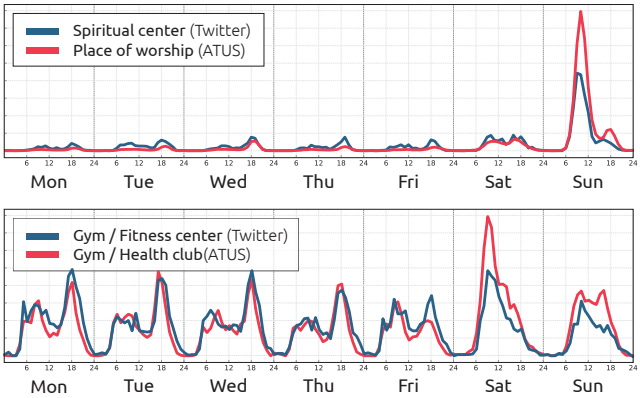



**Figure 3: Weekly activity distributions at *Spiritual center/Place of worship* and *Gym* according to Twitter and ATUS data. High correlation confirms credibility of Twitter as a source of self-reported activities.**

**A. Activity comparison (quantitative).** We compared the 15 most frequent activities [5] reported by ATUS survey participants with a study by Kahneman et al. [20] and identified 10 common activities (Table 1). In order to map ATUS activities to those from Twitter, two human assessors manually judged semantic similarity of the top 500 most frequent Twitter activities and, when possible, established a link to the corresponding ATUS activity (inter-rater agreement as Cohen's kappa: $\kappa = 0.79$). To quantify the similarity of the ATUS activities with their Twitter counterparts, we compare the temporal profiles of these activities and express the difference by calculating their Jensen-Shannon divergence ($D_{JS}$). The results indicate very high similarity for six activities ($D_{JS} \leq 0.035$), high similarity for two ($D_{JS} \approx 0.05$), and low for another two activities ($D_{JS} \geq 0.13$). The weaker correlation in the last two cases (i.e., *'Food and drink preparation'*, *'Physical care for children'*) is caused by relatively lower popularity of Twitter posts containing reference to these activities and consequent low diversity of related activities in rather limited test selection (i.e., 500 most frequent Twitter activities). For instance, all activities in the test set related to food preparation are concerned with dinner, which makes the profile skewed towards later time of the day.

With ATUS as ground truth, this comparison gives us confidence that reported activities in Twitter tend to reflect the actual timing of real activities. In Fig. 2, we depict evaluation plots of the three most frequent ATUS activities: (*'Eating and Drinking'*, *'Work, main job'*, and *'Television and movies'*), which amount for 34% of all activities in ATUS and on average occupy up to 11.4h of one's day [20].

**B. Activity comparison (qualitative).** Both activities extracted from Twitter and activities reported by participants of the survey are classified into categories of locations (e.g., spiritual center, gym). That leads us to study the typical locations of selected activities, and we find some interesting insights. The characteristic locations of majority of activities are very similar in both datasets. For instance, the activity of *'commuting'* happens in various means of transport (train, car, etc.) in both datasets; *'working'* mostly falls into the category of *respondent's workplace* in ATUS, in Twitter it is *professional*

[5] We exclude *'sleeping'* since it is impossible to tweet about it while performing it.

*& other places* (which can be considered an equivalent location). To highlight some differences, we observe that some daily activities that often take place at home according to ATUS (*'watching a movie'*, *'eating and drinking'*) are more likely to be tweeted about when performed at out-of-home places (i.e., cinema, restaurant).

**C. Location comparison (quantitative).** Apart from individual activity comparison, we construct aggregated temporal profiles for all underlying activities that happen at a location. Given that each activity has a specific temporal profile, potential differences of these aggregates in ATUS and Twitter could indicate bias of certain activities that are more or less talked about on Twitter. Despite non-compatible categorization systems of locations in ATUS and Foursquare (via Twitter), we were able to find location equivalents to 11 (out of 26) ATUS categories in Foursquare. We rendered daily and weekly temporal profiles of the location activity aggregates and in Fig. 3, on two examples, we demonstrate good alignment of extracted data with the ground truth. In Fig. 4, we use the weekly profiles to calculate the distribution divergence ($D_{JS}$) across all 11 location categories, and present the results in the form of a matrix heatmap. The low $D_{JS}$ values on the diagonal indicate positive correlation of aggregated location profiles in Twitter (y-axis) and ATUS locations (x-axis).

The evaluation, in which we compared the temporal and spatial aspects of activities from Twitter and ATUS, suggests a high correlation between the extracted activities and the ground truth in both dimensions. Based on these findings, we conclude that with respect to time and location the activities extracted from Twitter using the proposed techniques are a good representative of people's true activity behavior.

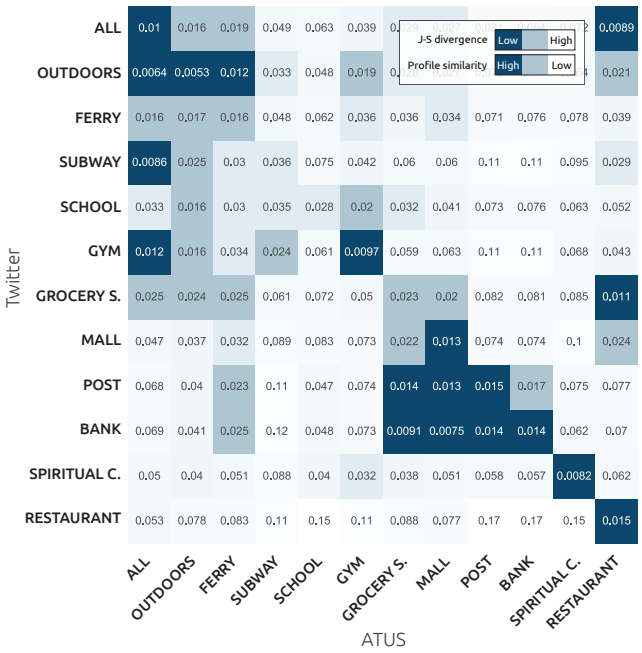



**Figure 4: Comparison of 11 counterpart ATUS and Twitter top-level location categories. The difference is expressed as Jensen-Shannon divergence between temporal profiles of location activity aggregates.**

## 4 ACTIVITY ANALYSIS

Assured that Twitter indeed provides a reliable window into activity spatiotemporal profiles, in this section, we address *RQ2* by inspecting what activities are the ones that people report on the most and where and when they tend to happen.

**What are the most common activities?** Being aware that social networks are subject to self-censorship [26], we studied which are the most common activities that people are willing to tweet about. The most common head phrase is, by a large margin, the verb *'getting.'* This phenomenon can be explained by the polymorphous nature of the verb: it is a constituent of many verbal constructions with a range of different meanings (*getting hair done*, *getting food*, *getting up*, etc.). We identified activities related to transportation (e.g., *heading home*), eating and drinking (e.g, *having lunch*) or entertainment (e.g., *celebrating birthday*) to be the most reported ones on Twitter. We list the most common activities in the appendix.

**Where do activities take place?** Spatially, activities from Twitter are unevenly distributed into location categories with strong bias towards food- or shop-related venues. An underrepresented category of locations are event venues (e.g, *conference room*). The relation between some activities and certain locations is very strong. The conditional probability of *cutting hair* being done in *Salon / Barbershop*, *having ramen* in *Asian Restaurant* or *worshiping* in *Spiritual Center* is 1.0 in all these cases. (We chose activities with 10+ observations.) On the other side of the probability spectrum we find activities such as *killing time* which is almost equally likely to happen in 120 different locations including *Beach*, *Winery*, *Zoo*, *Racetrack*, and *Bookstore*.

**When do activities happen?** Inspired by the finding of Noulas et al. [30], we inspected the aggregated temporal dynamics of users' activity patterns over the course of a day and a week (Fig. 5). We find that the most active days are Saturday and Friday, while Monday to Wednesday are almost equally quiet. We split the day into four habitual parts: morning (6:00-12:00), afternoon (12:00-18:00), evening (18:00-24:00) and night (00:00-6:00) to find that 43% of activities are reported during the afternoon, and only less than 3% reported at night. Further, when analyzing the temporal scope of various activities individually, we generate profiles of these activities to visually examine their temporal footprints, e.g., see Fig. 1. The plot depicts a sample of temporal distributions that belong to eleven activities, which are selected so that their peak hours are spread out over the timespan of a day. It should be mentioned that the underlying data originate in the US and culture dissimilarities in other regions may lead to different activity profiles.

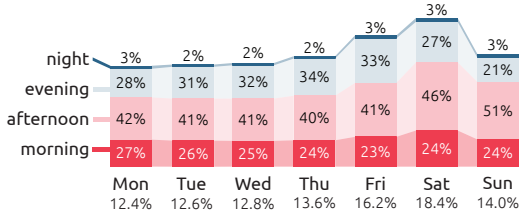



**Figure 5: Frequency of reported activities during four periods of day and on different days of the week.**

The analysis revealed interesting facts: 1) the distributions of activities and their locations are top-heavy with only a fraction of activities that re-appear frequently, 2) people often report on activities related to their journey, eating/drinking or entertainment, which were recognized as enjoyable activities according to Kahneman et al. [20], and, 3) the most active reporting periods are Friday-Saturday afternoons (which might be partially influenced by the *"Twitter social jet lag"* [24]).

# 5 PREDICTIONS

The previous section uncovered patterns of human activities. Referring back to our main task, which is to propose models that embrace human activity as a contextual feature, we go on and use the newly acquired insights in two prediction tasks, respectively:

**RQ3 Given an activity of a person (and possibly other context), can we predict where it is likely to happen?**

**RQ4 Given a location (and possibly other context), can we predict what activities people will engage in there?**

## 5.1 Methods

We cast each prediction task in this paper as a ranking problem, where a list of items $i$ is ordered by a ranking score $S_i$, which reflects its probability $P(i|c)$ of being relevant to a given context $c$.

In RQ3, the task is to return a ranked list of top- and/or second-level location categories given a full activity and other relevant context on the input. The ranking order reflects the probability of the activity taking place at the location. We denote this model as ***activity-to-location*** **(A2L)**. The problem in RQ4 is reversed, i.e., given a top- or second-level location category (and other relevant context), the goal is to rank activities according to the likelihood of them happening at the given location. We denote this model as ***location-to-activity*** **(L2A)**.

**Evaluation metrics.** We are primarily focused on precision of our predictions in this work; therefore, we use traditional binary-relevance ranking metrics: mean reciprocal rank (MRR) and precision at a cut-off position $P@k$. We evaluate results for $k = \{1, 3\}$.

**Dataset.** For the experiments, we chronologically split the Twitter sample of more than $6.5M$ tweets into an initial 90% of training data from the beginning ($D_{TR}$) and 10% of evaluation data ($D_{EV}$).

## 5.2 Baseline

An intuitive and strong baseline ranks items according to their conditional probability computed from the training dataset $D_{TR}$. In the ***activity-to-location*** task, the probability $P(l|a)$ of location $l$ is proportional to the number of cases the activity $a$ is observed at location $l$ in the training dataset $D_{TR}$. In ***location-to-activity***, activities are ranked according to the probability $P(a|l)$, which is given by the frequency of activity $a$ at given location $l$. We discovered that time, and especially the *period of day*, is a strong predictor. Therefore, we include another baseline, BL-*temp*, where the probability is also conditioned on the period of day $t$, i.e., $P(l|a,t)$ or $P(a|l,t)$.

## 5.3 Predictive Approach

We operate with a diverse set of features, which, when combined, generate a long feature vector with binary, numerical ($x \in (0,1) \subset \mathbb{R}$) or categorical values. Below, we discribe their types and, in Table 2, their usage. Ensemble models have proven to be robust and well-performing. Considering their ability to handle large feature spaces with categorical values, we opted for a random forests classifier as the machine learning algorithm of choice.

***Textual features***. We extract unigrams from activity surface forms and transform them into a feature matrix. Weights of the terms in the matrix are calculated using Tf-Idf to ensure higher values of informative terms.

***Temporal features*** are inferred from the activity timestamp (local time). We extract hour of day (0-23), part of the day as in Section 4, day of week, month, and timezone.

***Spatial / Activity prior features***. Spatial prior features capture the prior probability of top- and second-level location categories in the training dataset for a given activity $a$. The activity prior feature contains the prior probability of activities given a top- or second-level location category.

***Personal features*** encode a user's past behavior as observed in the training dataset. In *RQ*3, the behavior is represented by a feature vector that consists of binary flags indicating activities that user performed in the past. In *RQ*4, analogously, the sparse vector marks locations previously visited by the user.

**Table 2: Overview of features and their usage in *activity-to-location* (A2L) and *location-to-activity* (L2A) models.**

| Feature | Feature type | Value type | *A2L* | *L2A* |
|---|---|---|---|---|
| Activity descriptor unigrams | *Textual* | *Numerical* | ✓ | |
| Hour of day | *Temporal* | *Categorical* | ✓ | ✓ |
| Period of day | *Temporal* | *Categorical* | ✓ | ✓ |
| Day of week | *Temporal* | *Categorical* | ✓ | ✓ |
| Month of year | *Temporal* | *Categorical* | ✓ | ✓ |
| Timezone id | *Temporal* | *Categorical* | ✓ | ✓ |
| General location category prior | *Spatial prior* | *Numerical* | ✓ | |
| General activity prior | *Activity prior* | *Numerical* | | ✓ |
| User's location category prior | *Personal* | *Binary* | ✓ | |
| User's activity prior | *Personal* | *Binary* | | ✓ |

## 5.4 Experimental Setup

We compare our models with all features (*All*) against the baselines (*BL*). In addition to that, we investigate the influence of types of feature sets by applying a leave-one-out strategy. Specifically, we train with all feature sets except one: activity textual features (*w/o Text.*), temporal features (*w/o Temp.*), prior features (spatial in *RQ*3 (*w/o Spat.*), activity in *RQ*4 (*w/o Act.*)), or personal features (*w/o Pers.*). For brevity, we only display the feature set analysis for one of the datasets (H), however, the pattern is similar in all other cases.

We observed that number of records per user, activity or category (entities) follows the power-law distribution. In order to mitigate data sparsity and study its influence on prediction performance, we propose three filtering approaches. The most aggressive strategy only keeps the *head* (H) of the dataset, i.e., the 100 most frequent users, 20 categories and 100 activities. The more relaxed variants expand the data with the *body* (H+B): 1000 most frequent users, 50 categories and 500 activities; and the *tail* fraction (H+B+T) of the dataset: 5000 most frequent users, 100 categories and 1000 activities.

**Table 3: Prediction results of *activity-to-location* model.**

| Data | Model | MRR | Impr. | P@1 | Impr. | P@3 | Impr. |
|---|---|---|---|---|---|---|---|
| H | BL | 0.59 | | 0.45 | | 0.68 | |
| | All | **0.70** | **18.6%** | **0.54** | **20.4%** | **0.82** | **20.9%** |
| | w/o Text. | 0.62 | 4.8% | 0.45 | 0.4% | 0.72 | 5.9% |
| | w/o Temp. | 0.39 | -33.2% | 0.16 | -63.6% | 0.49 | -27.9% |
| | w/o Spat. | 0.61 | 3.6% | 0.44 | -2.7% | 0.71 | 4.7% |
| | w/o Pers. | 0.52 | -12.4% | 0.32 | -29.8% | 0.65 | -5.0% |
| H+B | BL | 0.47 | | 0.34 | | 0.53 | |
| | All | **0.59** | **26.2%** | **0.44** | **29.7%** | **0.66** | **23.5%** |
| H+B+T | BL | 0.43 | | 0.31 | | 0.49 | |
| | All | **0.52** | **21.3%** | **0.37** | **20.6%** | **0.57** | **16.9%** |

## 5.5 Results and Discussion

Tables 3 and 4 present prediction results of *activity-to-location* and *location-to-activity* models, respectively. Location in both cases refers to the second-level categories, which offer an order of magnitude more location categories than the top-level categories. The tables contain results for each filtering strategy ('Data' column) and compare values with the best-performing baseline (BL or BL-*temp*).

The evaluation results in Table 3 show the superiority of the ***activity-to-location*** model over the baseline by up to 26.2% in MRR for the H+B dataset. The significant improvement in MRR is accompanied by precision increase at the cut-off points of 1 and 3 by up to 29.7% and 23.5%, respectively. We see that the model benefits the most from the temporal features *(Temp.)*, suggesting that given we know which activity a person is occupied with, her location strongly depends on time. The *textual features (Text.)*, apart from enhancing the results, ensure generalizability of the model to unseen activities by leveraging prior knowledge about linguistically similar activities.

The task of predicting activity for a given location (***location-to-activity*** model) is a relatively more difficult problem, which is noticeable from the baseline (BL-*temp*) numbers in Table 4. Our technique, however, consistently outperforms the strongest baseline, achieving performance improvement of up to 32.8% in MRR and 70.6% and 39.6% in P@1 and P@3, respectively. While user's historical data *(Pers.)* have a significant impact on the predictive performance in both tasks, it is the activity prediction where knowing user's past behavior (i.e., past activities at given location) is crucial. We observed that as the datasets get sparser, the benefit of personal data grows stronger.

**Strengths.** Some human activities tend to happen at a very limited number of location types irrespective of personal preferences (*depositing check* - Bank, *waiting to board a flight* - Airport), and even simple modeling technique can capture these regularities. On the other end of the spectrum lie activities that are very general and whose entropy w.r.t. location categories is high, e.g., *'enjoying night'* that was observed 45 times in 21 distinct categories. The benefit of our models over probabilistic baselines is in their ability to handle activities whose typical location changes in time and is rather user-specific. We illustrate the temporal dependence of an

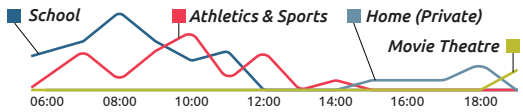



**Figure 6: Relative location probabilities changing in time for '*dropping off kid*' activity as returned by our model.**

**Table 4: Prediction results of *location-to-activity* model.**

| Data | Model | MRR | Impr. | P@1 | Impr. | P@3 | Impr. |
|---|---|---|---|---|---|---|---|
| H | BL-*temp* | 0.34 | | 0.21 | | 0.42 | |
| | **All** | **0.45** | **32.8%** | **0.36** | **70.6%** | **0.53** | **26.5%** |
| | w/o Temp. | 0.41 | 20.2% | 0.29 | 37.4% | 0.48 | 14.5% |
| | w/o Act. | 0.31 | -10.3% | 0.22 | 13.7% | 0.35 | -15.6% |
| | w/o Pers. | 0.28 | -17.0% | 0.18 | -12.8% | 0.34 | -17.8% |
| H+B | BL-*temp* | 0.25 | | 0.15 | | 0.27 | |
| | All | **0.32** | **29.2%** | **0.25** | **66.7%** | **0.37** | **39.6%** |
| H+B+T | BL-*temp* | 0.21 | | 0.13 | | 0.23 | |
| | All | **0.23** | **9.9%** | **0.18** | **41.5%** | **0.26** | **13.00%** |

activity on *'dropping off kid'* in Figure 6. We see how the probability of activity location changes over the course of a day (Thu). The probabilities depend on historical behavior of each user, and, in this particular case, our model improves the MRR by 0.51.

**Implications.** The negative correlation between prediction performance and data sparsity confirms our hypothesis that restriction of the dataset to the most frequent entities leads to more accurate predictions. The obvious explanation is that the task becomes relatively simpler, since the number of classes drops. We note that the ultimate goal is to support prediction into a rich open-set of activities and, while our model performs well there, it leaves a research opportunity for further improvements that generalize to the tail.

# 6 CONCLUSIONS AND FUTURE WORK

Human activity is clearly one of the major drivers that influence information needs of people. In this paper, we have shown that large amounts of open-domain activities are self-reported by users in their social media posts. While not all activities that people perform are tweeted about (due to self-censorship [31]), by focusing on extraction of ongoing activities present on Twitter, we are able to reliably model spatial and temporal aspects of thousands of activities that do get published. We demonstrate reliability by contrasting the extracted spatiotemporal profiles with real-life distributions captured in an independent large-scale survey of people's daily routines, ATUS. Our work was primarily motivated by two context-aware applications (i.e., reminder and recommender), which would greatly benefit from reliable activity–location prediction models. To address that, we pose two tasks: 1) prediction of locations for a given activity and 2) prediction of activities for a given location. In both of them, a proposed model outperforms a strong frequency-based baseline by a significant margin of 26.2% and 32.8% MRR improvement, in respective order.

In order to preserve variety in this initial study, we did not consider resolution of synonym activities, nor did we cluster activities into categories. Future work could try to increase the robustness of learned models by learning an embedding of activities to support synonymy. Another interesting direction for follow-up research, since we now have a good understanding of spatiotemporal patterns of various activities, would be to use them to identify web searches that are related to these activities.

To conclude, this work provides a significant step toward a probabilistic model of common-sense that enables context-aware systems to reason about the connections between location, time, and natural language descriptions of activity. Furthermore, there is an exciting opportunity for further research given both the public nature of the data source and unexplored modeling choices.

# REFERENCES


[1] Gediminas Adomavicius and Alexander Tuzhilin. 2015. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 191–226. https://doi.org/10.1145/1454008.1454068

[2] Marco Aurelio Beber, Carlos Andres Ferrero, Renato Fileto, and Vania Bogorny. 2016. Towards activity recognition in moving object trajectories from Twitter data. In *Proc. of GeoInfo*. 68–79.

[3] Jan R Benetka, Krisztian Balog, and Kjetil Nørvåg. 2017. Anticipating Information Needs Based on Check-in Activity. In *Proc. of WSDM*. https://doi.org/10.1145/3018661.3018679

[4] Paul N Bennett, Filip Radlinski, Ryen W White, and Emine Yilmaz. 2011. Inferring and using location metadata to personalize web search. In *Proc. in SIGIR*. ACM, 135–144. https://doi.org/10.1145/2009916.2009938

[5] Frank R Bentley and Crysta J Metcalf. 2008. Location and activity sharing in everyday mobile communication. In *CHI Extended Abstracts*. ACM, 2453–2462. https://doi.org/10.1145/1358628.1358702

[6] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proc. of RANLP*. Association for Computational Linguistics.

[7] AJ Bernheim Brush, John Krumm, James Scott, and T Scott Saponas. 2011. Recognizing activities from mobile sensor data: Challenges and opportunities. In *Proc. of Ubicomp*.

[8] Liming Chen, Jesse Hoey, Chris D Nugent, Diane J Cook, and Zhiwen Yu. 2012. Sensor-based activity recognition. *IEEE Trans. on SMC* 42, 6 (2012), 790–808. https://doi.org/10.1109/TSMCC.2012.2198883

[9] Karen Church and Barry Smyth. 2009. Understanding the intent behind mobile information needs. In *Proc. of IUI*. ACM, 247–256. https://doi.org/10.1145/1502650.1502686

[10] David Dearman, Timothy Sohn, and Khai N Truong. 2011. Opportunities exist: continuous discovery of places to perform activities. In *Proc. of CHI*. ACM, 2429–2438. https://doi.org/10.1145/1978942.1979297

[11] David Dearman and Khai N Truong. 2010. Identifying the activities supported by locations with community-authored content. In *Proc. of Ubicomp*. ACM, 23–32. https://doi.org/10.1145/1864349.1864354

[12] David Graus, Paul N Bennett, Ryen W White, and Eric Horvitz. 2016. Analyzing and Predicting Task Reminders. In *Proc. of UMAP*. ACM, 7–15. https://doi.org/10.1145/2930238.2930239

[13] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. 2013. Extracting Diurnal Patterns of Real World Activity from Social Media. In *Proc. of ICWSM*.

[14] Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proc. of HLT*. Association for Computational Linguistics, 368–378.

[15] Annika M Hinze, Carole Chang, and David M Nichols. 2010. Contextual queries express mobile information needs. In *Proc. of MobileHCI*. ACM, 327–336. https://doi.org/10.1145/1851600.1851658

[16] Nabil Hossain, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry Kautz. 2016. Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. *arXiv preprint arXiv:1603.03181* (2016).

[17] Giovanni Iachello, Ian Smith, Sunny Consolvo, Gregory Abowd, Jeff Hughes, James Howard, Fred Potter, James Scott, Timothy Sohn, Jeffrey Hightower, et al. 2005. Control, deception, and communication: Evaluating the deployment of a location-enhanced messaging service. *Proc. of UbiComp*, 903–903. https://doi.org/10.1007/11551201_13

[18] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2009. Why we twitter: An analysis of a microblogging community. *Advances in Web Mining and Web Usage Analysis* (2009), 118–138. https://doi.org/10.1145/1348549.1348556

[19] F Thomas Juster and Frank P Stafford. 1985. *Time, goods, and well-being*.

[20] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780. https://doi.org/10.1126/science.1103572

[21] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131. https://doi.org/10.3390/computers2020088

[22] John Krumm, Dany Rouhana, and Ming-Wei Chang. 2015. Placer++: Semantic place labels beyond the visit. In *Proc. of PerCom*. IEEE, 11–19. https://doi.org/10.1109/PERCOM.2015.7146504

[23] Michael G Lamming and William M Newman. 1992. Activity-based Information Retrieval: Technology in Support of Personal Memory. In *IFIP Congress (3)*, Vol. 14. 68–81.

[24] Eugene Leypunskiy, Emre Kıcıman, Mili Shah, Olivia J Walch, Andrey Rzhetsky, Aaron R Dinner, and Michael J Rust. 2018. Geographically Resolved Rhythms in Twitter Use Reveal Social Pressures on Daily Activity Patterns. *Current Biology* 28, 23 (2018), 3763–3775. https://doi.org/10.1016/j.cub.2018.10.016

[25] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19, 2 (1993), 313–330.

[26] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133. https://doi.org/10.1177/1461444810365313

[27] Sahisnu Mazumder, Dhaval Patel, and Sameep Mehta. 2014. Actminer: Discovering location-specific activities from community-authored reviews. In *Proc. of DaWaK*. Springer, 332–344. https://doi.org/10.1007/978-3-319-10160-6_30

[28] Joan Melià-Seguí, Rui Zhang, Eugene Bart, Bob Price, and Oliver Brdiczka. 2012. Activity duration analysis for context-aware services using foursquare check-ins. In *Proc. of Self-IoT*. ACM, 13–18. https://doi.org/10.1145/2378023.2378027

[29] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proc. of CSCW*. ACM, 189–192.

[30] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of ICWSM*.

[31] Alexandra Olteanu, Emre Kiciman, and Carlos Castillo. 2018. A Critical Review of Online Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. In *Proc. of WSDM*. 785–786. https://doi.org/10.1145/3159652.3162004

[32] Xin Rong, Adam Fourney, Robin N Brewer, Meredith Ringel Morris, and Paul N Bennett. 2017. Managing Uncertainty in Time Expressions for Virtual Assistants. In *Proc. of CHI*. ACM, 568–579. https://doi.org/10.1145/3025453.3025674

[33] Kristina J Shelley. 2005. Developing the American time use survey activity classification system. *Monthly Lab. Rev.* 128 (2005), 3.

[34] Timothy Sohn, Kevin A Li, William G Griswold, and James D Hollan. 2008. A diary study of mobile information needs. In *Proc. of CHI*. https://doi.org/10.1145/1357054.1357125

[35] Yangqiu Song, Zhengdong Lu, Cane Wing-ki Leung, and Qiang Yang. 2013. Collaborative boosting for activity classification in microblogs. In *Proc. of SIGKDD*. ACM, 482–490. https://doi.org/10.1145/2487575.2487661

[36] Jaime Teevan, Amy Karlson, Shahriyar Amini, AJ Bernheim Brush, and John Krumm. 2011. Understanding the importance of location, time, and people in mobile local search behavior. In *Proc. of MobileHCI*. ACM, 77–80. https://doi.org/10.1145/2037373.2037386

[37] Sergey Volokhin and Eugene Agichtein. 2018. Understanding Music Listening Intents During Daily Activities with Implications for Contextual Music Recommendation. In *CHIIR*. https://doi.org/10.1145/3176349.3176885

[38] Wouter Weerkamp, Maarten De Rijke, et al. 2012. Activity prediction: A twitter-based exploration. In *Proc. of TAIA*.

[39] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. 2015. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Trans. on SMC* 45, 1 (2015), 129–142. https://doi.org/10.1109/TSMC.2014.2327053

[40] Chao Zhang, Mengxiong Liu, Zhengchao Liu, Carl Yang, Luming Zhang, and Jiawei Han. 2018. Spatiotemporal Activity Modeling Under Data Scarcity: A Graph-Regularized Cross-Modal Embedding Approach. AAAI.

[41] Chao Zhang, Keyang Zhang, Quan Yuan, Fangbo Tao, Luming Zhang, Tim Hanratty, and Jiawei Han. 2017. React: Online multimodal embedding for recency-aware spatiotemporal activity modeling. In *SIGIR*. https://doi.org/10.1145/3077136.3080814

[42] Zack Zhu, Ulf Blanke, and Gerhard Tröster. 2016. Recognizing composite daily activities from crowd-labelled social media data. *Pervasive and Mobile Computing* 26 (2016), 103–120. https://doi.org/10.1016/j.pmcj.2015.10.007

# A APPENDIX

## A.1 Activity Phrases

We computed the most common activity head phrases found in Twitter, including their relative frequencies and some example noun phrases for each one, shown in Table A1. We see that each verb phrase can introduce a variety of activities, whose ambiguity is only resolved by the following noun phrase.

**Table A1: Most common activity head phrases with a sample of related common full activities.**

| | **Activity head phrase** | | **Full activity** |
|---|---|---|---|
| # | Verb phrase | Freq. | Noun phrases (most common) |
| 1. | *getting* | 11.8% | *… hair (done), stuff, lunch, food, gas, …* |
| 2. | *picking up* | 5.6% | *… dinner, lunch, thing, med, car, food, load, …* |
| 3. | *having* | 4.3% | *… lunch, dinner, fun, breakfast, time, drink, …* |
| 4. | *watching* | 3.6% | *… game, movie, TV, football, girl, fireworks, …* |
| 5. | *going* | 3.3% | *… home, to be day, to bed, to work, to gym, …* |
| 6. | *trying* | 3.3% | *… place, something, out place, spot, out spot, …* |
| 7. | *celebrating* | 2.6% | *… birthday, year, life, day, anniversary, bday, …* |
| 8. | *heading* | 2.5% | *… home, to airport, to work, to bed, downtown, …* |
| 9. | *checking* | 2.3% | *… (out) place, out spot, thing, out location, …* |
| 10. | *enjoying* | 2.2% | *… day, weather, lunch, dinner, time, view, …* |
| 11. | *taking* | 2.2% | *… care, break, advantage, tour, mom, picture, …* |
| 12. | *doing* | 1.9% | *… thing, laundry, work, shopping, something, …* |
| 13. | *making* | 1.7% | *… money, friend, round, time, dinner, progress, …* |
| 14. | *seeing* | 1.6% | *… friend, movie, show, doctor, deadpool, girl, …* |
| 15. | *grabbing* | 1.5% | *… lunch, dinner, bite, breakfast, food, coffee, …* |
| 16. | *visiting* | 1.4% | *… friend, mom, dad, family, grandma, aunt, …* |
| 17. | *dropping* | 1.4% | *… kid, off donation, off car, off load, off passenger, …* |
| 18. | *eating* | 1.4% | *… lunch, dinner, breakfast, food, pizza, supper, …* |
| 19. | *working* | 1.2% | *… today, night, hour, way, prayer, shift, …* |
| 20. | *starting* | 1.2% | *… day, week, weekend, year, morning, trip, …* |
| 21. | *playing* | 0.9% | *… game, pokemon, tune, tourist, music, golf, …* |
| 22. | *buying* | 0.9% | *… stuff, grocery, food, thing, car, book, …* |
| 23. | *opening* | 0.7% | *… day, night, ceremony, weekend, account, session, …* |
| 24. | *being* | 0.7% | *… tourist, home, adult, kid, fatty, friend, …* |
| 25. | *drinking* | 0.6% | *… beer, coffee, wine, water, tea, brew, …* |

## A.2 Overview of the Tweet-filtering Process

Filtering tweets down to activity assertions takes 10 steps, as outlined in Table A2. Our goal was to create a precise set of relevant activity tweets, but not to necessarily find all relevant tweets.

## A.3 Activity Extraction Grammar

For completeness and reproducibility purposes, we present the complete grammar we used for extraction of *verb+noun* pairs (i.e., <VPNP>) after the text was tokenized and tagged with parts of speech [6]. We refer the reader to the Penn Treebank POS tagset [25] for a full overview of the POS tags.

```
NPS: {<NNS>* <NN>* <POS>? <FW>* <NNS>* <NN>*}
JX:  {<JJ>* <JJS>* <JJR>* <,>?}
NP:  {<DT>? <CD>? <NNP>* <NNPS>* <JX>*
     (<DT> <VBG>)? <CD>? <NPS>}
VP:  {<VBG> <RP>* (<JJ> <TO>)? <TO>? <VB>*}
VPNP:{<VP> <PRP$>* <NP>}
```

For a better idea about how the extraction grammar works, we display a parse tree of an example phrase:

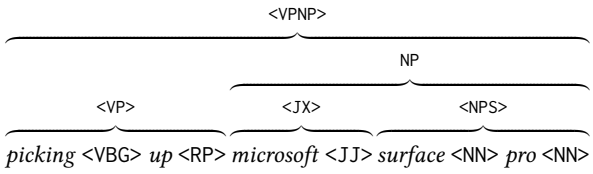



**Table A2: Filtering of tweets during the activity extraction process. Along with description of each step, the table displays a number of remaining tweets after the step is performed and the percentage of the original volume of tweets that it represents.**

| # | **Processing step** | **Number of tweets** | **Percentage** |
|---|---|---|---|
| 1. | Tweets from US region (in English) originating from Foursquare app | 6 641 503 | 100% |
| 2. | Removal of duplicate tweets (based on TweetID) | 5 845 289 | 88.0% |
| 3. | Filtering of corrupted records (due to parsing) | 5 845 108 | 88.0% |
| 4. | Filtering for tweets by users with spamScore < .5 | 5 841 176 | 87.9% |
| 5. | Filtering for tweets with spamScore < .5 | 5 788 598 | 87.2% |
| 6. | Filtering for tweets containing '-ing' | 1 044 127 | 15.7% |
| 8. | Content cleaning for tweets (removal of tweets that become empty after this step) | 544 762 | 8.2% |
| 9. | Duplicate filtering (based on User ID, Text & Category) | 518 368 | 7.8% |
| 10. | Activity extraction (filtering for tweets containing ≥ 1 activities) | 193 219 | 2.9% |