Journal of Ambient Intelligence and Humanized Computing manuscript No.
(will be inserted by the editor)

Where am I? Predicting User Location Semantics from Engagement with Smartphone Notifications

Andreas Komninos · Ioulia Simou · Antonis-Elton Frengkou · N. Gkorgkolis · John Garofalakis

Received: date / Accepted: date

Abstract Location semantics are important for the delivery of context-aware ubiquitous services to users, such as the contextually-relevant handling of interruptions on mobile devices. For such purposes, user coordinates can be used to query global venue databases, to get back the likely venue (and its categories) where the user is located. This potentially compromises user privacy, allowing service providers to track users. We analyse data from a longitudinal study of 44 participants (university students and staff in Patras, Greece), including notification handling, device state and location information. Using semantic labels from the Google Places API as ground truth, we demonstrate that it is possible to semantically label a user's location based on their notification handling behaviour, even when location coordinates are obfuscated so as not to precisely match known venue locations. On the other hand, the reliability of this ground truth is questioned through a crowdsourcing exercise. We demonstrate that Places API data can only be reliably used for some venue categories, and recommend best practices for using such data to establish ground truth in location context aware services.

Keywords Interruption management · Mobile notifications · Semantic location labelling · Location Services.

1 Introduction

As users of mobile devices roam through urban environments, a wealth of data can be collected from their devices about their current whereabouts and activities. While it is relatively easy to obtain the location of a user, within a given accuracy estimate (e.g. through GPS, connection to Wi-Fi or 4G networks), a harder task is to assign *semantics* to the user's location. The typical method of resolving this,

A. Komninos · J. Garofalakis
University of Patras, Rio 26504, Greece

A. Komninos · I. Simou · A. E. Frengkou · N. Gkorgkolis · J. Garofalakis
Computer Technology Institute & Press "Diophantos", Rio 26504, Greece
E-mail: akomninos, simo, frengkou, gkorgkolis, garofala@ceid.upatras.gr

is by comparing the user's coordinates against a database of known locations, and there are several commercial services that offer this type of information (e.g. Google Places API). Therefore, given a user's location coordinates, it is relatively easy to obtain the venue and venue type that a user might currently be at, and therefore to infer their current activity (e.g., they are at Cinema X, and thus quite likely watching a movie). More formally, from positioning data (coordinates), one could infer various abstractions of the location semantics (e.g. the venue name, the venue type, the venue's function, the purpose of visitation, etc.). Naturally, it's not always useful, or necessary, to obtain a complete picture of all semantic knowledge about a location, in order to offer a ubiquitous service. In fact, respect for the user's privacy requires that only the knowledge which is necessary to deliver a service should be obtained, deduced or inferred by a provider.

To demonstrate, let's consider the example of offering contextually relevant notification handling to users. Currently, users are left on their own in terms of how they might manage notifications under different contexts (Auda et al, 2018). However, automatic notification management can offer opportunities for a better and more socially aware mobile use experience (Anderson et al, 2018). Taking the cinema example mentioned earlier, a device could automatically suppress incoming notifications which are not relevant at the current location, as per Saikia and She (2017), or automatically set the device ringer mode to silent for the duration of the user's stay at that location.

There are several confounding factors to being able to achieve this goal. First, user location coordinates might not be available, or accurate enough to provide a reasonable estimate of venue (e.g. the user might be indoors, or the user might be connected to a sparse 4G network only). Even more, for services such as this to work, the user's location needs to be sent to a remote server, potentially compromising user privacy. Finally, it's not really necessary for the service to know exactly *which cinema* the user is at - only the fact that the user is located at *a cinema* is enough for the service to fulfil its purpose.

As discussed in existing literature, users receive a significant volume of notifications during the day, from on-device events (e.g. network availability, battery status) and external services (e.g. instant messaging), which can reach several hundreds (Visuri et al, 2019). These events can become opportune moments for assessing the user's location. The user behaviour in handling these notification events can vary significantly across time (Komninos et al, 2018), and we can assume that the behavioural choices are influenced by the location context and semantics as well, even though there is no previous literature to investigate this. For example, while watching a movie at the cinema, the user might take longer to notice an incoming notification since their device will probably be set to "silent mode" and tucked away, or even if they do, they might chose to ignore it until the show is over.

This paper is an extended version of our previous publication at AmI2019 (Komninos et al, 2019). In that paper, and also presented here in Sections 3 and 4, we explore the use of notification handling behaviour and device state information, as an additional source of information for overcoming problems with user coordinate availability and accuracy. Using supervised machine learning algorithms on a dataset of notification and location samples from several users, we predict user location semantics and demonstrate that notification handling behaviour can overcome the problem of location accuracy. Additional contributions in this pa-

per, on top of our previously reported findings, are presented in Section 5. They constitute additional work that address a major limitation of our previous publication, namely the reliance on Google Places API as a source for location semantic labels. Using a crowdsourcing technique, we find that a large number of venues are incorrectly labelled by the API. As a result, we are able to significantly improve the reliability of our machine-learning approach, using the crowdsourced semantic labels.

2 Related work

Discovering location semantics is the research effort directed towards assigning categorical labels (e.g. "Home", "School", "Shop") to venues represented in a dataset with at least a set of coordinates (latitude, longitude) and optionally a given name (e.g. "Mike's cafe"). Location semantics are important for a range of location based services, such as point-of-interest (POI) search and recommendation. Commercial applications such as Google Maps, Foursquare and Tripadvisor maintain large databases of POIs, relying largely on users adding and/or modifying these. One issue with this approach is that represented venues are not always correctly semantically labelled by the users, and also the reliance on user effort means that many real-world POIs may be often left out of the service. Previous research has frequently focused on the automatic semantic labelling of locations, with a variety of means. An overview of related work, including datasets used, classifier types, feature types and resulting performance is shown in Table 1. Researchers have examined features based on data "fingerprints" left by users, such as user behaviour (e.g. check-in locations and temporal patterns), linguistic behaviour (tweet content), relationship to other users also present at a location, which are easy to mine from publicly available datasets. Others have supplemented these with additional hardware-based features from users' mobiles, such as application use, calling and texting behaviour, battery status etc. These constitute a more significant invasion of privacy and are hard to collect at a large scale for research use.

There are some common themes in the previous literature, which can be identified. First, where multiple classifiers have been used (e.g. decision trees, SVMs, random forests), the results do not seem to vary significantly. Most often, it is the type and number of features introduced to the model which have the most impact. Secondly, a larger number of categories makes the likelihood of misclassifications higher. Another issue is that in most papers, there is a significant class imbalance in the datasets used. This is somewhat problematic since in most reviewed works, the measure of accuracy is used, which is heavily influenced by the prevalence of certain categories (Akosa, 2017). Hence, comparisons with the performance of these previous approaches is done with some hesitation.

To the best of our knowledge, the use of notification handling behaviour as a feature for semantic place labelling has not been investigated in the past. Hence the goal of our paper is to explore how this information can be used for the task of semantic place labelling. We also attempt classification at a more fine-grained level (24 categories). Further, rather than taking the root-level categories from category hierarchies, or focusing only on the top-N most frequent categories in the dataset, as done in many papers, we adopt a more methodical approach to deriving the final categories to use.

Paper	Data	Model	Features	Target	Performance
Celik and Incel (2018)	Own dataset	RF	Time, network, human activity, app use, system	Place category (10 types)	Accuracy, 71%
Falcone et al (2014)	Twitter	J48, Decision Table, Multilayered Perceptron, Bayesian Network, K* and LogitBoost	Spatiotemporal visitation statistics, tweet statistics	Place category (8 types)	Accuracy, 67%
Gu et al (2016)	Foursquare	Social friendship, trust model and check-in model	Foursquare check-ins and social network relationships	Home label	Accuracy, up to 92%
He et al (2016)	Foursquare	Model based on spatiotemporal and textual features	Spatiotemporal check-in statistics, user ratings, user comments	Place category (unspecified number)	Accuracy, 63%
Huang et al (2012)	Nokia MDC	multilevel classification models (e.g., SVM, J48, RF, LMT, PART, SMO, SipleLogistic)	Location visitation and spatial statistics, App usage and Device statistics, Communication statistics, Network statistics	Place category (10 types)	Accuracy (REPTree) 66%.
Kinsella et al (2011)	Twitter	Language models	Tweet content and coordinates	User country, state, town, zipcode	Accuracy Country (76%), State (45%), Town (32%), Zipcode (15%).
Krumm and Rouhana (2013)	ATUS and PSRC datasets	Boosted decision trees	User statistics, visitation statistics, temporal context	Place category (ATUS: 14 types, PSRC: 13 types)	Accuracy, ATUS: 73%, PSRC: 74%
Leppäkoski et al (2017)	Nokia MDC, Microsoft	NB, DT, Bagged Tree, NN, KNN, SVM, LogReg,	Time, network, human activity, app use, system, call log	Place category (10 and 3 types)	Accuracy, 69% (10 types), 89% (3 types)
Mahmud et al (2012)	Twitter	ensemble of statistical and heuristic classifiers	Tweet frequency & content	User city, state, timezone	Recall City (58%), State (66%), Timezone (78%)
Wu et al (2017)	Nokia MDC, ATUS	Naïve Bayes, RF, J48	Time, app ise, call log, system, media, network features	Place category (9 types)	Accuracy, 74%
Yang et al (2016)	Foursquare	KNN classifier (SketchMinMaxWeighted)	Foursquare check-ins	Place category (9 types)	Accuracy, ≈65%.
Ye et al (2011)	Foursquare	SVM	Check-in statistics, relatedness of venues	Place category (199 types)	Precision ≈80%
Zhu et al (2013)	Nokia MDC	LogReg, SVM, GBT, RF	Time, acceleration, network, app use, call log, system and media features	Place category (10 types)	Accuracy, 75%

Table 1: Overview of past work in predicting location semantics

3 Study methodology

3.1 Apparatus and participants

We developed a UI-less notification logging application for Android devices, which runs unobtrusively on the device as a background service. We collected features about incoming notifications and the user's device state at the time of issue. For this, we used the Android `NotificationListener` service, which allows our application to be informed with the details of every notification, as soon as it is issued by any app or the operating system. The `NotificationListener` service returns a `Notification` object which contains all the necessary information, especially about programmed modality, in a consistent way across all API versions. We also employed a range of other Android APIs to capture device context from hardware states (e.g. `PowerManager, DisplayManager`).

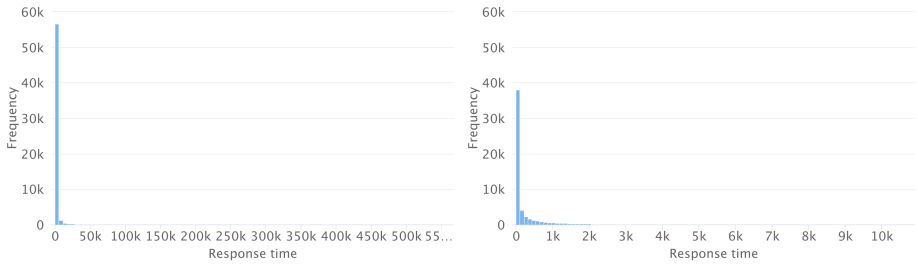
We also exploited the Google Places API to retrieve details about the user's presumed location at the time of notification issue. This API requests the user's location coordinates, and returns a list of likely places where the user is located, along with a confidence level. We logged the place which had the highest confidence value. The data features collected are discussed in detail in section 3.3. All data was uploaded to a remote server at frequent intervals during the day, provided the user had wi-fi connectivity.

A call for participation was issued to undergraduate students at our local university. The application was installed on their device, a consent form was signed and participants were instructed that they could quit the study at any time. The study automatically ended after 3 months of use. They were requested to leave location services enabled on their device for the duration of the study, although we did not enforce this condition. In total, 44 participants took part in the study (26 female). From this set of participants, we excluded several participants who participated for fewer than 10 days and who provided fewer than 50 notification log entries, resulting in a subset of 31 participants. Participants provided data that spanned an average of 30.87 days (sd=16.15, min=13, max=84).

3.2 Dataset preparation

In total we collected 204,074 notifications from the users. In the dataset, we noticed that a significant number of notifications (38,400) were issued by the system and immediately dismissed. This phenomenon was observed for all users, although for some users the proportion of such notifications was unusually large. We are not certain why this happens. Further investigation of the package name showed that some system applications might be issuing such notifications (perhaps as a means of interprocess communication), although it might be the case that a user is also manually quickly dismissing some notifications (within the resolution of 1 second). We decided to exclude such notifications from the dataset. Further, we removed from the dataset all notifications for which the "flag" feature values indicated that they were ongoing events and not user-dismissable (e.g. an ongoing phonecall or download). These notifications are automatically dismissed by the system and hence offer no value to our research goal. From the remaining notifications, a significant number did not contain location information, since the user's location

services might have been switched off at the time, or the service might not have been available. We also excluded these from the dataset. After these exclusions, the dataset contained 59,221 user-dismissed notifications with location details.



(a) Entire dataset (b) Response time < 3h

Fig. 1: Distribution of response time (in seconds) to notifications (100 bins)

Examining the pruned dataset, we observed that the average response time to notifications is 1,366.93s (sd=11,255.82), with a maximum response time of 562,302s. A histogram of response time to notifications shows a power-law distribution (Fig. 1). Based on this observation, we limited the dataset to only notifications that were attended to within 3 hours of issue, resulting in 57,737 notifications (97.5% of the original dataset). As can be seen, even after culling the dataset further, the distribution of response times to notifications maintains a power-law shape. This finding is consistent with previous works such as (Komninos et al, 2018).

3.3 Dataset features

To address the problem at hand, we used raw and synthetic features obtained from the user's device. To begin, the raw data features collected from users are shown in Table 2. Something to note here is that while all devices support sound and vibration for notifications, not all devices incorporate the status LED. All except two of our participants had phones incorporating a status LED, hence we maintain this feature.

From these raw features we synthesized a further set of features, to create the final dataset to be used for prediction, as shown in Table 3. First, we used the current device ringer mode and programmed notification modalities (custom or default) to determine the true modalities used to deliver the notification, as per (Komninos et al, 2018). An illustration of how the combination of the raw features for modality and ringer mode result into the synthetic modality features is shown in Table 4. Further, a place can belong to multiple categories. These are reported in a non-ordered list by Google, ostensibly therefore the order of appearance shows the prevalence of a category type (e.g. "Bar, Restaurant, Cafe" shows that a place is primarily of type "Bar", but also functions as a restaurant and cafe). We therefore extract the primary category of a venue. In doing so, we observed that many places included the vague category "Point of Interest".

Table 2: Raw data features collected

Notification Details	
Time posted	Unix timestamp of notification issue
Time dismissed	Unix timestamp of notification dismissal
Package name	Application that created the notification
Sound	Whether the notification was programmed to issue a custom sound alert
LED	Whether the notification was programmed to issue a custom status LED blink pattern
Vibration	Whether the notification was programmed to issue a custom vibration pattern
DefaultSound	Whether the notification was programmed to use the default sound alert
DefaultLED	Whether the notification was programmed to use the default status LED blink pattern
DefaultVibration	Whether the notification was programmed to use the default vibration pattern
Priority	The notification priority category
Notification flags	Additional information about the notification
Device state	
Ringer mode	The current device ringer mode (Normal, Vibrate, Silent)
Idle state	Whether the device is in an idle state
Interactive state	Whether the device is in a state ready to interact with the user (screen on, processor awake)
Lockscreen notifications allowed	Whether notifications are visible from the user's lock screen
Location Details	
Place name	Name of the most likely current place
Place categories	The categories assigned to the most likely current place
Confidence	Confidence of reporting the most likely current place
Latitude	Decimal coordinates of the most likely current place
Longitude	Decimal coordinates of the most likely current place

Hence, where this was the primary category, it was replaced by the immediately subsequent category type.

Another note here relates to Google's list of categories, where 127 different categories are listed. Predicting on 127 category classes is possible, but presents an unnecessary complexity to the problem, as many venue categories are quite similar in nature and it can be expected that a user will exhibit similar behavioural patterns in these. For example, "Church" and "Mosque" are both places of worship, where devices are typically kept on silent, and users do not readily engage in notification handling. We therefore attempted to group the individual categories into larger sets, as per Table 5. Ultimately, we assigned to each place the super-category to which it belongs, based on its primary category type. Ultimately, we assigned to each place the super-category to which it belongs, based on its primary category type. An exception to this were the "Miscellaneous" and "Entertainment areas" categories, since for these the user behaviour might be quite different depending on conditions (e.g. a user probably can't notice a notification in a night club as easily as in a cafe), hence for these we used the primary categories ungrouped. As a result, we find that the user notifications were issued at 24 distinct place categories and distributed unevenly (Table 5, non-grouped primary categories capitalised). Finally, it's important to note that the location coordinates collected by our app,

Table 3: Final feature set

Notification Details		
Response time	Time dismissed - time posted	Synthetic
Hour issued	Hour of day at notification issue [0-23]	Synthetic
Day of week issued	Day of week at notification issue [1-7]	Synthetic
Had Sound	Whether the notification was issued with a sound	Synthetic
Had LED	Whether the notification was issued with a LED blinking pattern	Synthetic
Had Vibration	Whether the notification was issued with a vibration pattern	Synthetic
Priority	The notification priority category	Raw
Device state		
Idle state	Whether the device is in an idle state	Raw
Interactive state	Whether the device is in a state ready to interact with the user (screen on, processor awake)	Raw
Lockscreen notifications allowed	Whether notifications are visible from the user's lock screen	Raw
Location Details		
Place category	The primary place category	Synthetic
Latitude	Decimal coordinates of the most likely current place	Raw
Longitude	Decimal coordinates of the most likely current place	Raw

Table 4: An example of synthesis of the true modality feature values, based on raw feature values at the time of notification issue. The example assumes that the app developer specified that a notification should be issued with a sound clip, vibration pattern and LED blink pattern, using programmer-specified or system default options for each.

Raw Feature Values Example			
	Sound	Vibration	LED
Programmed Modality	1	1	1
Synthetic Feature Values based on Ringer Mode			
	Had_Sound	Had_Vibration	Had_LED
Ringer Mode "Normal"	1	1	1
Ringer Mode "Vibrate"	0	1	1
Ringer Mode "Silent"	0	0	1

are not the user's actual coordinates, but the coordinates of the venue that is the user's most likely current place, as reported back by Google's API. We do not store the user's actual location coordinates for privacy reasons.

As can be seen in Fig. 2a, users receive a varying amount of notifications throughout the day. The distribution is similar to that reported in previous literature, such as (Celik and Incel, 2018). More importantly, we note that the diurnal distribution varies distinctly across categories, as exemplified in Fig. 2b. This is an expected result, since different venue types exhibit different diurnal visitation patterns (Falcone et al, 2014). Further, we note the distribution of response times to various notifications on a hourly basis (Fig. 3a). The pattern is similar to the findings in (Komninos et al, 2018), showing the distinct user behaviour in handling notifications throughout the day. Distinct response time averages are also noted across the categories (Fig. 3b shows three category examples). Furthermore, while it could be intuitively assumed that a "Normal" ringer mode might lead to shorter reactions to notifications, we note that the mean response time is not drastically

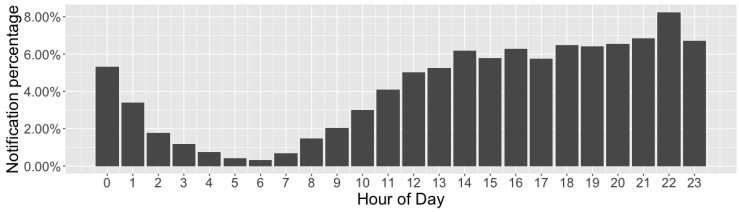
Table 5: Grouped place categories

Category group	Categories	Samples
Accomodation	Campground, Lodging, Room, Rv Park	1,350
Address	Administrative Area Level 1, Administrative Area Level 2, Administrative Area Level 3, Country, Geocode, Locality, Political, Post Box, Postal Code, Postal Code Prefix, Postal Town, Street Address, Sublocality, Sublocality Level 1, Sublocality Level 2, Sublocality Level 3, Sublocality Level 4, Sublocality Level 5, Synthetic Geocode	86
Civil Services	City Hall, Courthouse, Embassy, Fire Station, Local Government Office, Police, Post Office	89
Contractors	Electrician, General Contractor, Moving Company, Painter, Plumber, Roofing Contractor	76
Education	Library, School, University	11,996
Entertainment Areas	Amusement Park, Aquarium, Bar, Bowling Alley, Cafe, Casino, Gym, Movie Theater, Museum, Night Club, Restaurant, Stadium, Zoo	11,157
Financial Services	Bank, Atm, Finance	93
Healthcare	Dentist, Doctor, Health, Hospital, Physiotherapist	617
Miscellaneous	Establishment, Floor, Other, Point Of Interest, Premise, Subpremise	18,347
Outdoor Areas	Colloquial Area, Natural Feature, Neighborhood, Park, Parking, Route	516
Personal Care	Beauty Salon, Hair Care, Spa	1,104
Place Of Worship	Cemetery, Church, Hindu Temple, Mosque, Place Of Worship, Synagogue	758
Professional Services	Lawyer, Accounting, Car Dealer, Car Rental, Car Repair, Car Wash, Funeral Home, Insurance Agency, Laundry, Locksmith, Real Estate Agency, Storage, Travel Agency, Veterinary Care	659
Public Transport	Airport, Bus Station, Intersection, Subway Station, Taxi Stand, Train Station, Transit Station	580
Shopping	Art Gallery, Bakery, Bicycle Store, Book Store, Clothing Store, Convenience Store, Department Store, Electronics Store, Florist, Food, Furniture Store, Gas Station, Grocery Or Supermarket, Hardware Store, Home Goods Store, Jewelry Store, Liquor Store, Meal Delivery, Meal Takeaway, Movie Rental, Pet Store, Pharmacy, Shoe Store, Shopping Mall, Store	10,309

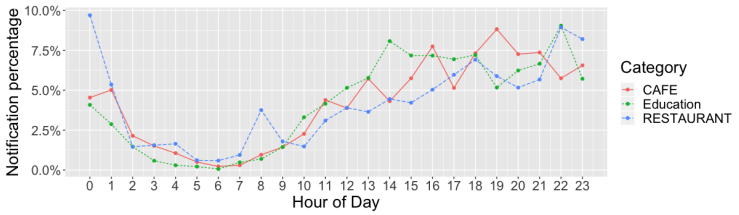
different (Fig. 4a). The examples in Fig. 4b demonstrate that attentiveness to the device is not simply dependent on device ringer mode, but is mediated by other factors, such as time of day, current user activity and/or social norms. For example, while response times are largely similar at *cafes* with any ringer mode, we can see that it is much longer with the ringer mode on *Silent* when the users are at an *education* venue. The placement of the device on *Silent* and the long response time demonstrate high engagement with the task context at hand (learning).

4 Study 1: Predicting user location category

In this section, we attempt to predict the category of the user's current location based on notification handling features in Table 3. This is, in essence, a multino-

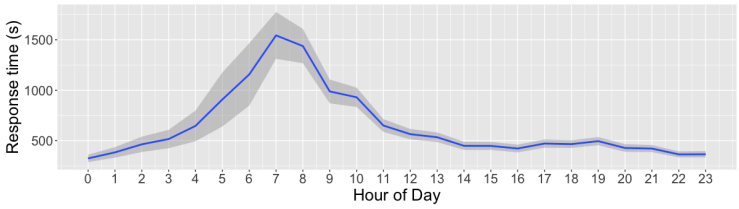


(a) All notifications

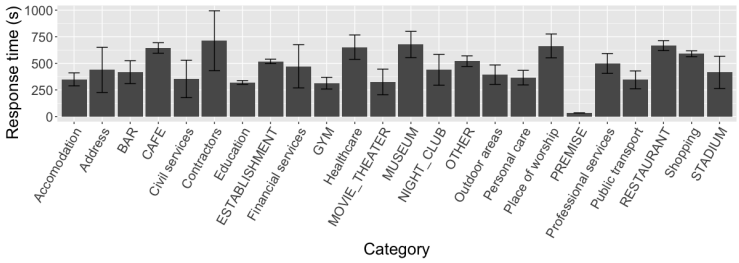


(b) Example categories

Fig. 2: Diurnal distribution of notifications



(a) Diurnal, all notifications (error band at 95%c.i.)



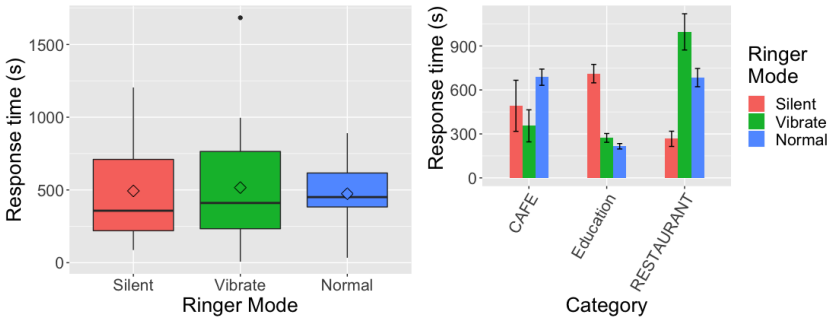
(b) Per place category (error bars at 95%c.i.)

Fig. 3: Distribution of response time to notifications

mial classification task using *Place Category* as the target. Implementation of the analysis process was done with the RapidMinder software platform.

4.1 Classifier and parameter selection

We used decision tree classifiers, since they have been shown to demonstrate comparable performance to other methods (Falcone et al, 2014). To tune the classifier



(a) Per ringer mode, all categories (b) Example categories

Fig. 4: Average response time to notifications per ringer mode

hyperparameters, we employed RapidMiner's evolutionary parameter tuning process on a small hold-out dataset. The final parameters used for the decision tree are Maximal depth:23, Minimal gain:0.013, Minimal leaf size:2, Minimal split size:4. Throughout the analysis reported in the following sections, we used a 10-fold cross-validation approach. We note that there is an imbalance in the frequency of location categories (Table 5), hence for performance we adopt the F-score (macro-averaged), which is more appropriate for imbalanced datasets (He and Ma, 2013), compared to the accuracy measure usually encountered in previous literature.

4.2 Decision tree modelling performance

As a starting point, we apply the decision tree classification algorithm to the entire dataset. To clarify the process further, the classifier is fed with all features as shown in Table 3, and returns the predicted place category. We assume the user's location is the same as each venue's reported coordinates. Therefore, given the user's notification handling behaviour, their location, and the device state, we attempt to predict the type of venue that they are currently at. Overall, we obtained a macro F-score μ. 88.96%, σ=11.05%). Examining the results, we wondered whether the broader categories "Miscellaneous" and "Entertainment areas" categories might be best split up, since for these the user behaviour might be quite different depending on conditions (e.g. a user probably can't notice a notification in a night club as easily as in a cafe), hence for the rest of the analysis, we used these two categories ungrouped.

As seen in Fig. 5, the classification performance remains quite good for most categories (F-score macro μ. 82.9%, σ=12.6%). During analysis, we noted that there is some discrepancy in the confidence reported for the most likely current user place, across the place categories (Fig. 6). For this reason, we decided to repeat the analysis in multiple steps, each time limiting the dataset to contain only notifications reported where the most likely current user place was reported above a certain confidence threshold $T \in [0, 0.1, ..0.9]$. The results are shown in Fig. 7. We note that the average F-score is not majorly affected by the reduction of the dataset, however the best nominal performance is achieved when considering

venues reported with a confidence threshold $T \geq 0.7$ (μ=84.6%, σ=13.51%, dataset size = 13,558 entries).

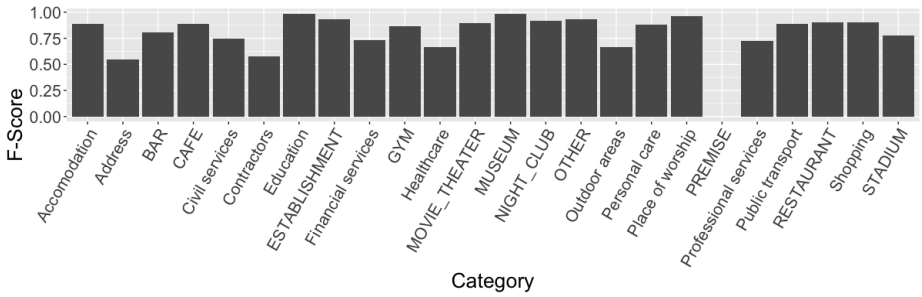


Fig. 5: Average F-score using decision trees, all notifications

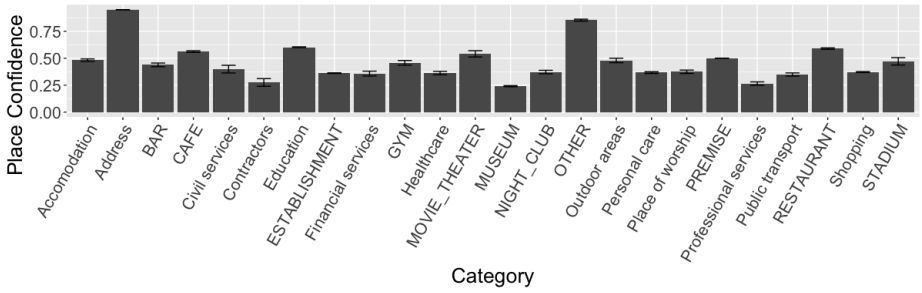


Fig. 6: Average confidence of most likely user place, all notifications, error band at 95%c.i.

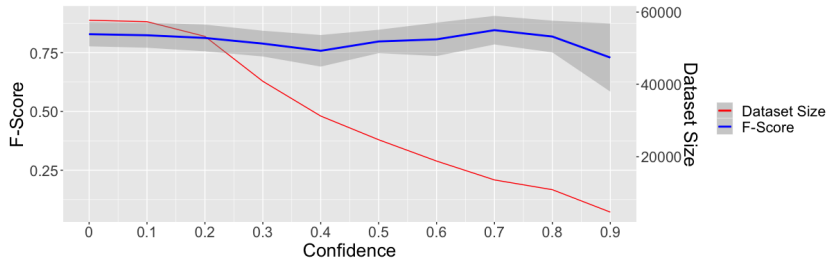


Fig. 7: Average F-score using decision trees (error bars at 95%c.i.)

4.3 Modelling with inaccurate user coordinates

In the preceding analysis, we assumed that a user's current coordinates are the same as those corresponding to places reported by Google's API. Of course, it would be rare that the user's actual coordinates would be precisely the same as those that match a specific venue, especially for venues that cover a large area (e.g. outdoor parks). To overcome this limitation, we proceeded to modify the user's coordinates by adding random noise to the known place coordinates (latitude and longitude). This noise was applied to each coordinate component individually, following a Gaussian distribution with a standard deviation set by us. The noise standard deviation was calculated using the formula $n \times 10^{-x}$ and was applied to each coordinate component (latitude and longitude), therefore the resulting random coordinates would fall within a certain circular range of a specific venue. An example of how this process generates the random user coordinates within a gaussian distance distribution of a specific venue is shown in Table 6. Distance is calculated using the Haversine formula.

Table 6: Sample random coordinate range generation

Noise σ	**Lat**	**Lng**	**Dist. at 1σ (m)**
0 (Place coords.)	38.2836678	21.7889705	0
1.0×10^{-6}	38.28370608	21.78899229	4.7
2.0×10^{-6}	38.28374437	21.78901408	9.3
3.0×10^{-6}	38.28378265	21.78903587	14.0
4.0×10^{-6}	38.28382093	21.78905766	18.6
5.0×10^{-6}	38.28385922	21.78907944	23.3

To assess the effect of imprecise user coordinates, we repeated the analysis for each value of $n \in [1, 2, .., 9]$, limiting the dataset to locations with a confidence threshold $T \geq 0.7$, since this achieved the best nominal performance in the preceding analysis. As can be seen in Fig. 8, the algorithm remains quite robust when adding noise to the decimal coordinates with a $\sigma \leq 9 \times 10^{-6}$ ($\approx$42m), after which, performance begins to deteriorate.

At this point, it becomes interesting to observe which categories suffer the heaviest penalty then the user coordinates are further away from the actual place coordinates. Taking the largest noise σ distance (233.1m), we note that the categories Place of worship, Outdoor areas, Professional services, Stadium and Civil services take the worst hit between -35.46% and -55.45% reduction of their F-score, compared to the smallest σ (4.7m). On the other hand, some categories like Shopping and Cafe only take a small penalty (-7.56% and -7.20%) respectively. The explanation for this possibly rests in the spatial clustering of these venue types (e.g. see Fig. 9). In Fig. 9, we see that cafes are mostly clustered together, hence we may not be able to accurately guess *exactly which* cafe a user is at, but we can be quite certain that they might be at *some* cafe, as long as their location and notification response behaviour is proximal to that captured at a nearby cafes. Although this might suggest that spatial distribution may have a significant effect on the accuracy of the classifier, it must be borne in mind that this is a very extreme scenario. Most users' location data is obtained via A-GPS, which, in an urban environment, has been shown to have an accuracy of about 9m (Zand-

bergen, 2009). In any case, we might expect similar results to be generalisable to many similar-sized cities, since it has been shown that a representation of cities as m-dimensional vectors based on their venue categories can uncover the similarities between them (Preoţiuc-Pietro et al, 2013).

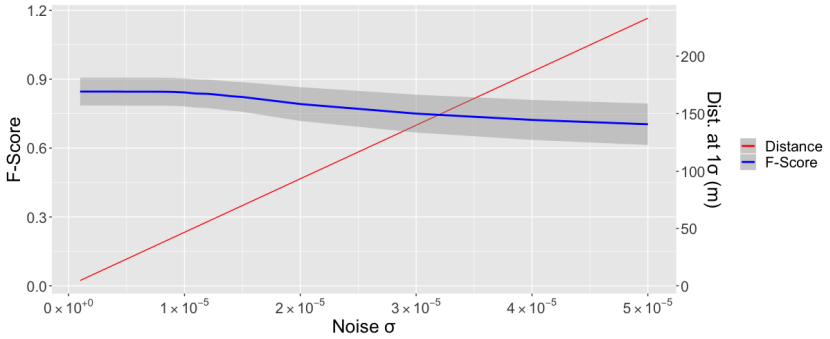
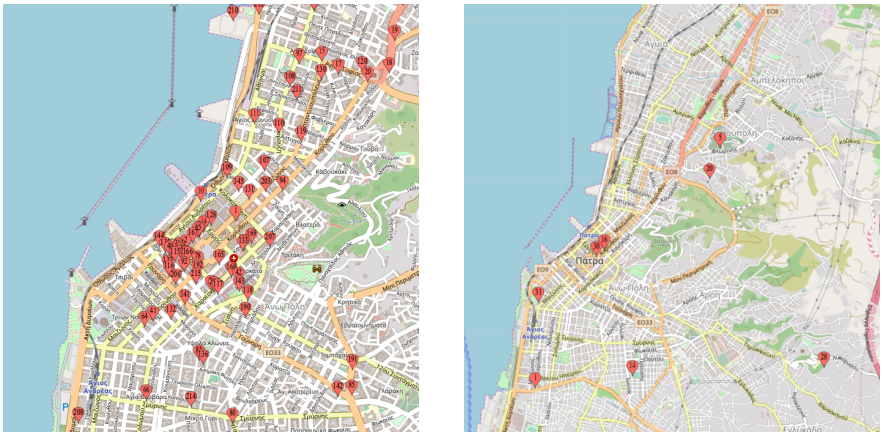


Fig. 8: Average F-score using decision trees, under random coordinate input noise (error band at 95%c.i.)



(a) Cafe category (b) Places of worship category

Fig. 9: Spatial distribution of places in our dataset

4.4 Effect of user location coordinates

In the preceding analysis, one of the input features is the user's location. This feature is certainly obtainable from the user, but its availability depends on whether

a user has enabled positioning on their device, their surroundings (indoors or outdoors) and connectivity (wi-fi, 4G, off). So far we have demonstrated that guessing the user's current location type is possible based on their notification behaviour, device state and geographic location, even if the latter is not precisely correspondent to a known place. For the next step, we wanted to experiment without taking user position coordinates into account. The same process as in the previous analysis was repeated, limiting the dataset iteratively to contain notifications at locations above a confidence threshold T. As shown in Fig. 10 the results are much worse than in our previous analysis, showing that the prediction model depends heavily on the knowledge of the user's coordinates, even though these do not necessarily need to correspond with great precision to the true location's coordinates. The reduction of the dataset size has no major impact on the performance of the classification.

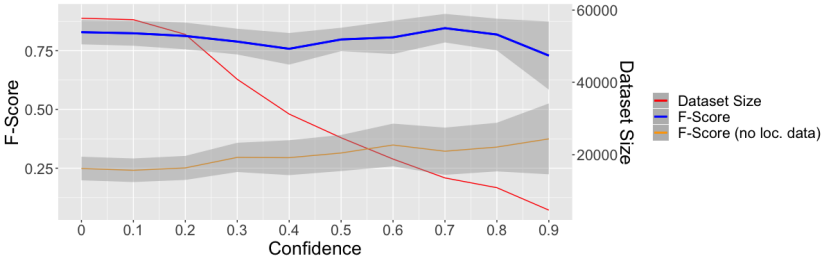


Fig. 10: Average F-score using decision trees (error bands at 95%c.i.)

5 Study 2: Obtaining reliable ground truth

In Study 1, we used the location semantic labels as reported by the Google Places API, in order to train our machine learning algorithm and validate its results. This means that we treated the Places API results as the "ground truth" for the entire study. However, this assumption does not necessarily hold. Previous work by Hochmair et al. (Hochmair et al, 2018) has highlighted the lack of actual "ground truth" POI datasets, and that various POI databases (e.g. by Google, OSM, Facebook, Yelp and others) provide varying degrees of quality in terms of coverage, position and classification accuracy. In this work, the Google dataset is found to be one of the most reliable, even though the researchers assessed the dataset quality for a single European city only (Salzburg). Since the population target from which we collected data refers to another country (Greece), we cannot be certain regarding the quality of Google Places data for the regions covered by our participants. Therefore, in Study 2, we attempted to establish a more reliable ground truth for locations covered in our dataset, and to repeat the analysis as in Study 1, this time using these more reliable location semantics to train and validate the machine-learning algorithms.

5.1 Region of interest

Since we recruited students from our university, located in Patras, Greece, as can be expected, the majority of the notification data were gathered at POIs located in that city, although a number of notifications were collected in nearby cities or even far away countries, due to participant mobility through the experiment. Our original dataset contained 2,210 unique POIs and though the vast majority were located in Greece, it also included POIs in some in three countries (Bulgaria, Cyprus and Russia). To narrow down the problem, we chose to focus on POIs located in Patras and, more specifically, excluded from the dataset any notifications and associated locations outside a bounding box that includes the city center and its surrounding neighbourhoods, as well as the university campus that is located approximately 8km from the city center. Furthermore, we removed POIs that had fewer than 5 notifications, in order to focus on locations that were systematically visited, and therefore would not artificially dilute the quality of available data. As a result of this pruning, the resulting dataset contained 419 unique POIs (Fig. 11).

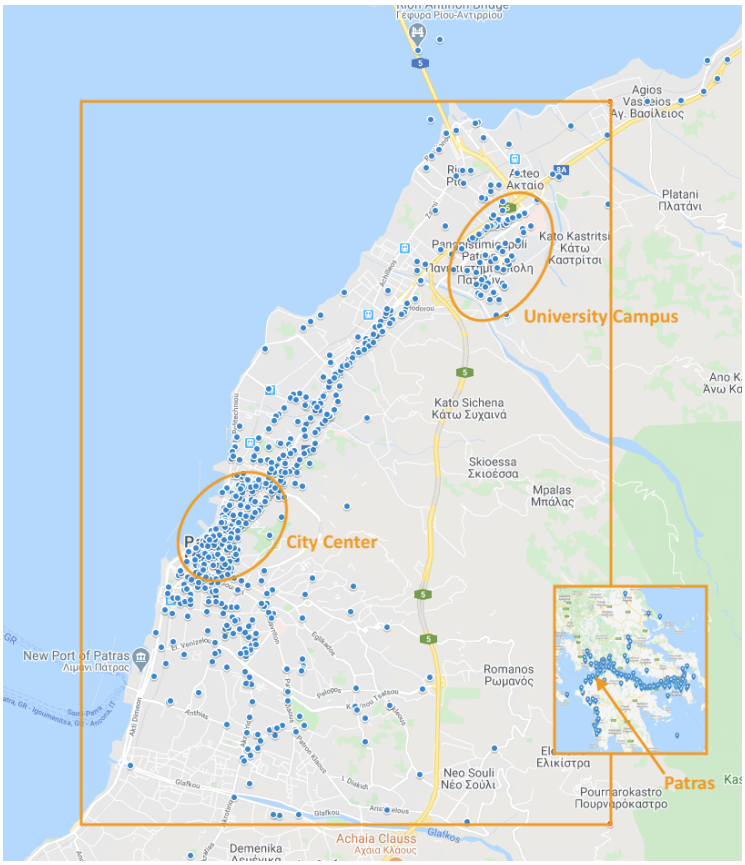


Fig. 11: Venues in the target area (Patras, Greece) present in our dataset. The spread of venues in a wider geographic area is shown in the inset.

5.2 Crowdsourcing the ground truth

In order to obtain a more reliable ground truth for the classification of these 419 POIs, we decided to engage in a crowdsourcing experiment. To this end, we developed a simple web application in responsive HTML5 (Fig. 12), which queried users about the semantic classification of 20 semi-randomly chosen POIs from the pool of 419. Our goal was to obtain multiple user classifications for each venue, so that we could determine the best label for each POI by selecting the its most frequently selected label. Since a completely random selection could result in some POIs gathering many more responses than others, we prioritised the selection of POIs which had received the fewest responses, so that they would be more likely to be chosen to be presented to a user.

In the web app, we showed participants the POI name (Fig. 12-A), its location on the map (Fig. 12-B), and presented them with some options. At first, a participant had to indicate whether they were familiar with that POI (Fig. 12-C). If they were not, the rest of the options were disabled and the participant could move to the next one. Otherwise, we asked participants to select the Category Group (cf. Table 5) to which they believe this POI belonged to (Fig. 12-D), giving them a free choice between the 14 category groups, and adding "Other" as a further option. Additionally, we asked participants to indicate the believed primary (Fig. 12-E) and secondary category of the POI (Fig. 12-F), however, for these, the options were limited to the categories reported by the Places API only, (or "Other"). Participants could also select the level of confidence for each of these choices on a scale between 1 (not confident at all) and 5 (very confident), using the sliders (Fig. 12-D*;E*;F*).

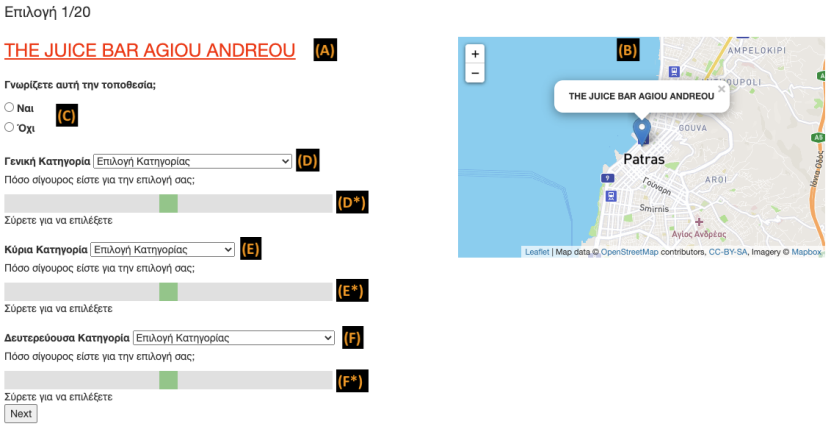


Fig. 12: Crowdsourcing web app UI. The various UI elements are marked in orange font with a black background

We publicised the crowdsourcing app to a range of Facebook groups relating to students and residents in Patras, over a period of 2 weeks. Overall, we received valid responses (completed questionnaires) from 133 participants (male: 62; female: 71; other: 0). Basic demographics were collected by allowing participants to

select between value ranges for age and years living in Patras. Participant ages varied, with the majority being relatively young adults, as can be expected due to recruitment from social media (18-25: 57; 26-30: 38; 31-40: 20; 40+: 18). The majority of participants lived in Patras for quite a number of years (0-3 years: 12; 4-9 years: 33; 10-15 years: 9; 16+ years: 79) and hence can be considered to be reasonably familiar with the city. We also asked them about their education level, with the majority being university or masters degree holders (high-school graduate: 42, university graduate: 55; masters graduate: 31; PhD graduate: 5).

We also captured the time it took participants to complete the questionnaire, in order to ensure that they were spending at least some time to reflect on their choices and thus to provide valid responses, and not simply clicking through the presented POIs. On average, participants spent 11m58s to complete the questions ($\sigma = 6m10s, min = 3m15s; max = 37m08s$). As such we consider all responses to be valid and included them in the ensuing analysis.

5.3 Crowdsourcing results

Due to an unexpected logging error, 10 venues ended up being excluded from the dataset, hence we present results for the remaining 409 POIs. From the 133 participants, only 5 indicated that they did not know any of the 20 places shown to them. For the rest of the participants, on average, they indicated being familiar with 11 of the 20 POIs shown to them on average ($\sigma = 3.454, min = 3, max = 20$). For these POIs, the average confidence in reporting the general category was quite high ($\mu = 4.580, \sigma = 0.459, min = 3, max = 5$). As expected, not every POI gathered the same amount of responses, owing to the semi-random selection of POIs for presentation, and the varying familiarity of participants with the POIs presented to them. To determine the final category group of each POI, we selected the majority choice as reported by participants. In cases where the choices were tied, we selected one at random. As can be seen in Figure 13, all POIs received at least one response, with the majority of POIs receiving up to 5 responses.

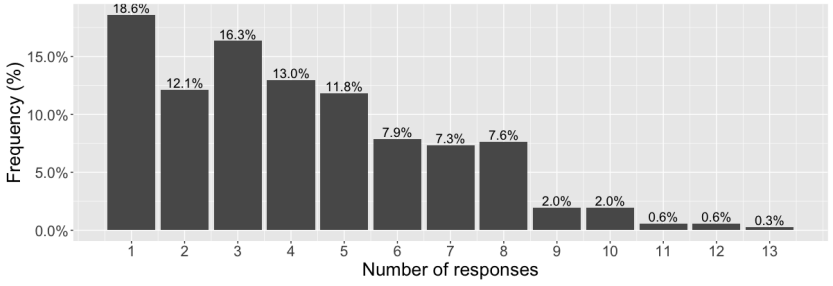


Fig. 13: Distribution of crowdsourced response frequency for POIs

Overall, we found 204 POIs (49.89%) where the participant classification differed from the original classification derived from the Places API results. This is a significant finding that indicates that the quality of the data offered by Google

for our region of interest is not as high as expected. In Table 7 we present an outline of the issues encountered in this analysis. As can be seen, the largest problem appears in the Miscellaneous, Professional Services and Shopping categories. Other categories such as Contractors show a large proportion of mismatches but are generally under-represented in our dataset. Notably, the Miscellaneous category is probably the place where the crowdsourcing exercise offers the most value, since it has allowed for a more specific classification of POIs instead of this generic description.

Table 7: Mismatch between Places API-derived and crowdsourcing-derived group categories

Category Group	**Matches with Places API**	**Mismatches with Places API**	**Mismatch %**
Contractors	0	2	100.00%
Miscellaneous	3	102	97.14%
Outdoor areas	2	3	60.00%
Professional services	7	10	58.82%
Personal care	2	2	50.00%
Shopping	59	53	47.32%
Civil services	3	1	25.00%
Entertainment areas	99	28	22.05%
Education	14	3	17.65%
Accomodation	9	0	0.00%
Financial services	5	0	0.00%
Healthcare	7	0	0.00%
Place of worship	2	0	0.00%

5.4 Predictions using crowdsourced POI labels

Following these results, we continued to repeat the analysis in Sections 4.2 and 4.3, this time using the crowdsourced category group labels instead of those derived from the Places API. The algorithm used and the parameters are the same as in described in Section 4.1. As a result of limiting our dataset to notifications received in the region of Patras, the dataset used includes 30,240 notifications (51.06% of the original dataset as reported in Section 3.2).

As a note, before the results are reported, it should be stated that the predictive algorithm trains and predicts on the data it is given. Hence, when training and testing with Places-API derived labels, it will attempt to predict what the Places API would return for each case. Conversely, when training and testing with crowdsourced labels it will attempt to predict what our participants would return for each case. Therefore to compare the classification performance directly between these two cases is not appropriate. Instead, to obtain a better idea about the effect of the discrepancy of classifications between real users and the Places API, we can train the algorithm using data received from the Places API, and attempt to predict on the actual ground truth, as reported by real users. This effectively becomes equivalent to using one dataset to train an algorithm, and performing tests on an entirely different dataset, a technique common in machine learning literature. We can expect here that the performance should drop considerably, since

we already know that users have a different opinion on the proper classification of a POI compared to the Places API (49.89% of labels differed). Next, we report results for all these cases, without obfuscating the coordinates.

Using the crowdsourced labels for training and testing, we achieved a macro average F-score of 92.06% (σ = 2.70%). Using the Places API-derived labels for training and testing, the performance is slightly increased to a macro average F-score of 95.75% (σ = 2.43%). To compare the effect of training and predicting with crowdsourced labels, we performed again a k-fold cross validation (k=10), but this time, we used the Places API-derived categories as training labels, and attempted to predict the POI categories in each fold, examining the predicted labels against the crowdsourced labels (i.e. the ground truth). It's worth noting here that the k-fold splits are stratified based on the distribution of the Places API label, which is used for the training set. As expected, the macro average F-score achieved dropped to 44.39% (σ = 1.92%, excluding categories for which the F-score is undefined, i.e. no correct predictions at all), representing a considerable departure from the scores achieved using the crowdsourced labels, or the Places API-derived labels for both training and testing. As can be seen in Fig. 14, while F-score performance is comparable in several categories, there exist several categories for which the performance is down to zero.

Since we noted that there exist several categories for which the mismatch between Places API and crowdsourced labels is large (Table 7), we attempted the same predictive process, this time removing the cases belonging to the categories with large discrepancies from the dataset (Miscellaneous, Outdoor areas, Personal care, Professional services, Shopping). The rationale here is that these mis-aligned labels could be overly affecting the result. Removing these cases (and associated labels) maintains a level of mismatch, but at more reasonable levels. As a result, the predictive performance increased to $\mu = 71.62\%, \sigma = 2.09\%$.

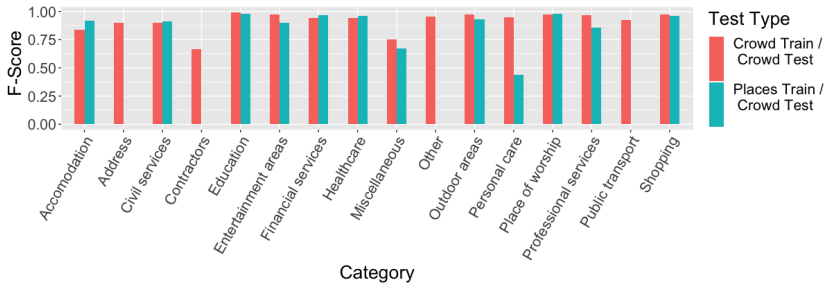


Fig. 14: F-scores per category during training and testing with crowdsourced labels only, vs. training with Places API-derived labels and testing with crowdsourced labels (10-fold cross validation).

Finally, we repeat the analysis using the noise addition process (Fig. 15), obtaining results for: a) training and testing with Places API labels only (red line); b) training and testing with crowdsourced labels only (blue line); c) training with Places API labels and predicting on crowdsourced labels (grey line), and; d) removing high-mismatch categories before training with Places API labels and predicting

on crowdsourced labels (yellow line). Here, we note that adding coordinate noise has a much less pronounced detrimental effect in all cases of using training/test label settings, compared to our original analysis. This is explainable since the composition and spatial distribution of POIs in this reduced dataset is different to the original (entire) dataset.

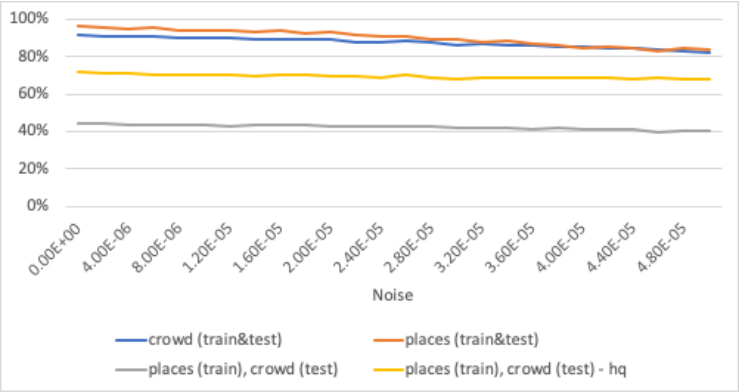


Fig. 15: Effect of coordinate noise addition on F-score performance.

6 Discussion

In this paper, we examined the use of notification handling behaviour as a cue for semantically labelling the user's current location. We found that, when paired with location coordinates, the resulting models can yield useful results with high classification accuracy. Such models can be pre-trained on the cloud and then stored and ran locally on the user's device, as part of an application or service framework, without the need for an internet connection. Further, we demonstrated that such models are robust to small deviations of user coordinates from the actual place coordinates, thereby allowing for positioning errors, or even, the obfuscation of precise user coordinates, in order to maintain privacy.

In Study 1, reported in our original paper (Komninos et al, 2019), we assumed that Google's labelling of the place categories could be used as the ground truth. As with other studies that leverage social network data, e.g. Falcone et al (2014), the algorithms are tuned to predict the ground truth as reported by the location identification services, therefore introducing an inherent element of inaccuracy. In this extended paper, we addressed the issue of reliable ground truth by obtaining semantic labels through crowdsourcing. We found that for the region of interest we focused on, there was significant discrepancy between the labels reported by Google, and the labels reported by city residents. As a result, we note that the training of algorithms using labels provided by the Places API yields unacceptably bad results, and therefore should highlight the need for better consideration of data quality prior to use in such predictive tools and services. On the positive side, we demonstrate that the technique we used can still provide excellent results,

when trained on accurate data. Performance might be improved through better hyperparameter optimisation (we kept the same for both studies) or choice of different classification algorithms (e.g. SVM, neural networks).

Therefore, for future studies, we recommend that, where possible, relevant labelling information should be crowdsourced from local experts, or at the very least, cross-validated against other datasets (e.g. Facebook, Foursquare), if and where available. In our study, we were able to obtain crowdsourced labels with relative ease, since the scale of the covered area is not very large. Scaling this approach to a planetary scale would be unrealistic. However, since we were able to obtain reasonably good results by excluding the categories where high levels of mismatch were identified, we could recommend that as a practical approach to cover much larger geographical areas (e.g. a country), it could be enough to obtain a small sample of labels through crowdsourcing for the whole area, and to limit predictions for those categories only where a reasonable level of matching is found.

A further underlying assumption in our analysis is that the user is currently positioned and has a certain non-trivial stay time at the location where the notification was received. This is likely true for most cases - users spend more time stationary at various places, than being mobile. However, further work here could include filtering of notification events during transit times, which in our case could not be done (since we did not keep GPS logs for privacy).

Finally, as we note different behaviours across venue categories, it would be of value to learn the reasons leading to these variances in user behaviour. However, this would be the subject of a further qualitative study. The generalisability of the findings presented here is limited to the body of the participants (students), hence the varying distribution of sample across categories. The models can be improved by mining information from other populations, to build up the number of samples across as many categories as possible. Personalised models depending on user type can then also be applied to better improve classification performance.

7 Data availability

The data used in this paper are openly accessible at https://github.com/komis1/ami2019-notifications

Conflict of interest

The authors declare that they have no conflict of interest.

References

Akosa JS (2017) Predictive accuracy: A misleading performance measure for highly imbalanced data. In: SAS Global Forum 2017 Conference

Anderson C, Hübener I, Seipp AK, Ohly S, David K, Pejovic V (2018) A Survey of Attention Management Systems in Ubiquitous Computing Environments. ACM Interactions on Mobile and Wearable Ubiquitous Technology 2(2):58:1–58:27, DOI 10.1145/3214261

Auda J, Weber D, Voit A, Schneegass S (2018) Understanding User Preferences Towards Rule-based Notification Deferral. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, CHI EA '18, pp LBW584:1–LBW584:6, DOI 10.1145/3170427.3188688

Celik SC, Incel OD (2018) Semantic place prediction from crowd-sensed mobile phone data. Journal of Ambient Intelligence and Humanized Computing 9(6):2109–2124, DOI 10.1007/s12652-017-0549-6

Falcone D, Mascolo C, Comito C, Talia D, Crowcroft J (2014) What is this place? Inferring place categories through user patterns identification in geo-tagged tweets. In: 6th International Conference on Mobile Computing, Applications and Services, pp 10–19, DOI 10.4108/icst.mobicase.2014.257683

Gu Y, Yao Y, Liu W, Song J (2016) We Know Where You Are: Home Location Identification in Location-Based Social Networks. In: 2016 25th International Conference on Computer Communication and Networks (ICCCN), pp 1–9, DOI 10.1109/ICCCN.2016.7568598

He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons

He T, Yin H, Chen Z, Zhou X, Sadiq S, Luo B (2016) A Spatial-Temporal Topic Model for the Semantic Annotation of POIs in LBSNs. ACM Trans Intell Syst Technol 8(1):12:1–12:24, DOI 10.1145/2905373

Hochmair HH, Juhász L, Cvetojevic S (2018) Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. In: Kiefer P, Huang H, Van de Weghe N, Raubal M (eds) Progress in Location Based Services 2018, Springer International Publishing, Cham, Lecture Notes in Geoinformation and Cartography, pp 293–313, DOI 10.1007/978-3-319-71470-7_15

Huang CM, Ying JJC, Tseng V (2012) Mining users' behaviors and environments for semantic place prediction. In: Mobile Data Challenge Workshop, citation Key: Huang2012MiningUB

Kinsella S, Murdock V, O'Hare N (2011) "i'm eating a sandwich in glasgow": modeling locations with tweets. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, Association for Computing Machinery, SMUC '11, p 61–68, DOI 10.1145/2065023.2065039, URL https://doi.org/10.1145/2065023.2065039

Komninos A, Frengkou E, Garofalakis J (2018) Predicting User Responsiveness to Smartphone Notifications for Edge Computing. In: Kameas A, Stathis K (eds) Ambient Intelligence, Springer International Publishing, Lecture Notes in Computer Science, pp 3–19

Komninos A, Simou I, Frengkou E, Garofalakis J (2019) Discovering user location semantics using mobile notification handling behaviour. In: Chatzigiannakis I, De Ruyter B, Mavrommati I (eds) Ambient Intelligence, Springer International Publishing, Lecture Notes in Computer Science, p 219–234, DOI 10.1007/978-3-030-34255-5_15

Krumm J, Rouhana D (2013) Placer: Semantic Place Labels from Diary Data. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, UbiComp '13, pp 163–172, DOI 10.1145/2493432.2493504

Leppäkoski H, Rivero-Rodriguez A, Rautalin S, Muñoz Martínez D, Käppi J, Ali-Löytty S, Piché R (2017) Semantic Labeling of User Location Context Based on Phone Usage Features. Mobile Information Systems DOI 10.1155/2017/3876906,

URL https://www.hindawi.com/journals/misy/2017/3876906/

Mahmud J, Nichols J, Drews C (2012) Where is this tweet from? inferring home locations of twitter users. In: ICWSM, citation Key: Mahmud2012WhereIT

Preoţiuc-Pietro D, Cranshaw J, Yano T (2013) Exploring venue-based city-to-city similarity measures. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Association for Computing Machinery, New York, NY, USA, UrbComp '13, DOI 10.1145/2505821.2505832, URL https://doi.org/10.1145/2505821.2505832

Saikia P, She J (2017) Effective Mobile Notification Recommendation Using Social Nature of Locations. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), pp 1265–1270, DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.203

Visuri A, van Berkel N, Okoshi T, Goncalves J, Kostakos V (2019) Understanding smartphone notifications' user interactions and content importance. International Journal of Human-Computer Studies 128:72–85, DOI 10.1016/j.ijhcs.2019.03.001

Wu X, Chen L, Lv M, Han M, Chen G (2017) Cost-Sensitive Semi-Supervised Personalized Semantic Place Label Recognition Using Multi-Context Data. ACM Interactions on Mobile and Wearable Ubiquitous Technology 1(3):116:1–116:14, DOI 10.1145/3131903

Yang D, Li B, Cudré-Mauroux P (2016) Poisketch: semantic place labeling over user activity streams. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, IJCAI'16, p 2697–2703

Ye M, Shou D, Lee WC, Yin P, Janowicz K (2011) On the semantic annotation of places in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Association for Computing Machinery, KDD '11, p 520–528, DOI 10.1145/2020408.2020491, URL https://doi.org/10.1145/2020408.2020491

Zandbergen PA (2009) Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. Transactions in GIS 13(s1):5–25, DOI 10.1111/j.1467-9671.2009.01152.x

Zhu Y, Zhong E, Lu Z, Yang Q (2013) Feature engineering for semantic place prediction. Pervasive and Mobile Computing 9(6):772–783, DOI 10.1016/j.pmcj.2013.07.004