

Analysis of Points of Interests Recommended for Leisure Walk Descriptions

Ehsan Hamzei ✉ 

University of Melbourne, Parkville, Australia

Thi Minh Hoai Bui ✉ 

University of Melbourne, Parkville, Australia

Martin Tomko ✉ 

University of Melbourne, Parkville, Australia

Stephan Winter ✉ 

University of Melbourne, Parkville, Australia

Abstract

Leisure walking is a physical activity where locomotion through a natural or even urban environment is the goal in itself, e.g., in pursuit of health and wellbeing. In contrast to destination-oriented walks that are focused on navigation efficiency (i.e., shortest or simplest walk from source to destination), leisure walks emphasize experiencing the environment, engaging in activities, and discovering places that may be off route, or intermediate destinations en-route, summarily called points of interest (POIs). POIs are key for recommending leisure walks, yet a detailed analysis of POIs in the context of leisure walking is missing in the literature. This study extracts and annotates POIs of leisure walking recommendations available in *WalkingMaps.com.au*, creating an annotated dataset to address this research gap and provide a first analysis of leisure walking descriptions. We classify POIs using the verbal description provided in the dataset, match them with data available in OpenStreetMap (OSM), and compare the POIs with nearby alternatives in OSM. Our analysis reveals thematic and spatial patterns in POI selection, offering a machine learning approach to model POI choices for leisure walks. We further evaluate the availability of rich data in OSM for future automated leisure walking recommendation. This study contributes to automated systems for recommending leisure walks, tailoring suggestions based on available information in the spatial open data, and presents an annotated dataset to facilitate future research in this field.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Location based services

Keywords and phrases leisure walks, points of interest, places, platial information

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.5

Supplementary Material *Software (Source code)*: <https://github.com/hamzeiehsan/leisure-walking-analysis> [8], archived at `swb:1:dir:cddb6d133e212246c2e458e4ea46f1358cd27927`

1 Introduction

Leisure walking is a physical activity where locomotion through a natural or even urban environment is the goal in itself, e.g., in pursuit of health and wellbeing. In contrast to the everyday walks, which are *destination-oriented* (where the path is merely a means of reaching the destination), leisure walks are *path-oriented* with a focus on the experience of the environment along the path [20]. In other words, in leisure walks the points of interest experienced at distance (off-route) or visited en-route as intermediate destinations, are as important or maybe even more important than the final destination, which may be identical with the start location.



© Ehsan Hamzei, Thi Minh Hoai Bui, Martin Tomko, and Stephan Winter;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 5; pp. 5:1–5:16



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Destination-oriented walks prioritize efficiency, enabling walk recommendations using optimal path algorithms (e.g., shortest or simplest paths) and spatial data typically stored in graph structures. Such walks can be externalized into verbal descriptions focused on efficient and effective navigation, with survey or route perspectives as the narrative strategy [25]. In contrast, leisure walk recommendations require platial information to suggest walks that focus on experiencing the environment [22, 23]. Leisure walk descriptions also demand a more intricate approach, moving beyond navigation, to offer rich descriptions of where to stop, what to see, and what activities to do, alongside narratives about places and their social and historical significance [23, 24].

Previous studies analyzing walk recommendations and descriptions provide insights into how people choose and communicate landmarks for navigation purposes [19, 10]. These studies are either limited to a specific walk [20] or a specific type of environment (e.g., natural landscape) [19], or they focus solely on route instructions [10] rather than the characteristics of points of interest (POIs) that fundamentally shape the experience of leisure walking. A detailed analysis of how people communicate descriptions of leisure walks and how they select relevant POIs is thus missing in the literature. This study aims to find insights and patterns of the characteristics of the chosen POIs in a leisure walking corpus.

The study addresses gaps in leisure walking recommendation research, particularly the lack of a dedicated dataset for POIs in the context of leisure walking. By collecting data from the *WalkingMaps* website¹, a sharing platform for leisure walking experiences in Victoria, Australia, we create an annotated dataset covering both leisure walk descriptions and POIs. To investigate why POIs are relevant for leisure walks, we classify the chosen POIs in the dataset and discuss their similarities and differences compared to nearby available POIs in OSM. Additionally, we conduct an evaluation of OSM data for leisure walking recommendations. Using a semi-automatic matching approach, we assess the availability of the selected POIs in OSM, contributing insights into the feasibility of automated leisure walk recommendations using open data.

This study hypothesizes that the thematic and spatial characteristics of POIs within a geographic extent can be used to identify patterns that describe which POIs are relevant for a leisure walking experience. To investigate this hypothesis, the following research questions must be addressed:

- What types of POIs are selected in leisure walking recommendations?
- To what extent is rich thematic and spatial information for the recommended POIs available in OSM?
- How can a machine learning model imitate the POI selection process for a given geographic area?

In short, the contributions of this study are:

- A leisure walk recommendation dataset, enriched with recommended POIs matched to OSM objects;
- A classification of the POIs in the dataset;
- A preliminary analysis of the availability of rich data in OSM for such POIs;
- A baseline prediction model for choosing relevant POIs for a leisure walk given a geographic extent.

¹ <https://walkingmaps.com.au/>

2 Related Works

POIs are defined as locations or objects that cartographers add to maps using cartographic symbols or labels to communicate relevant places [16]. Alternatively, POIs can be described as specific locations that individuals might find interesting or useful [9, 3]. Both definitions emphasize the importance of *relevance and interest* to the individual seeking POI recommendations. Consequently, POI identification is context-dependent and inherently subjective. For instance, POIs recommended to tourists may differ from those relevant to residents of a neighborhood, and the concept of POIs in urban analytics varies from that in mobility studies. Some research has narrowed the scope of POI to systematic definitions of specific types of POIs. For example, natural POIs have been defined in tourism and conservation management contexts using a structured rubric, with applicability demonstrated through examples from *OSM*, *iNaturalist*², and *Scenic-or-Not*³ data sources [9].

Categorizing POIs has been a major focus in POI-related research, using topic modeling in a semi-supervised manner to identify meaningful taxonomies for describing POIs in specific datasets (e.g., [5, 21, 6, 11]). These categorizations often result in POI classes relevant to specific domains, frequently focused on urban spaces due to the availability of rich datasets [21, 5, 11]. The Latent Dirichlet Allocation (LDA) topic modeling method is commonly used to identify dominant POI classes in textual datasets such as Foursquare⁴ and Yelp⁵ [11]. Typical POI types identified in these studies include *restaurants, cafes, and bars, shops and malls, public spaces such as parks and squares, museums, and religious and historic sites*. Additional classes depend on the dataset and method; for instance, [5] identified beach-related categories (e.g., beaches, piers, surf spots) due to their geographic focus, while [6] included businesses, transportation facilities, and government, health, and education-related POIs. However, these studies are predominantly focused on urban areas and do not specifically address a leisure walk context, where POIs are locations along a route, sometimes at the cost of few more steps from the walk, to spend a short time to view, visit, explore and interact [17].

Other common areas in POI-related research include predicting POIs based on previously selected POIs along a path [3, 27], efficiently storing and retrieving POIs [15], and evaluating the availability and quality of POI datasets [26]. These studies often provide general-purpose solutions for POI research and do not specifically address the unique challenges associated with domains such as leisure walks. For instance, predicting POIs for leisure walks involves distinct challenges: a walk may have a specific theme (e.g., visiting historic landmarks or bird habitats in rural areas), influencing what should be considered relevant as a POI. Additionally, POIs for leisure walks are not necessarily tourist attractions or places selected solely for navigational purposes but may include lesser-known places and objects discovered through personal experiences, often appealing especially to local residents [17]. Regarding data quality and availability, POIs associated with leisure walks often include a diverse set of less prominent places and objects, which means general-purpose studies may not adequately represent the specific data quality and availability conditions for these POIs.

This study addresses these research gaps by collecting and annotating a dataset specifically relevant to leisure walking POI research. We classify the collected POIs using a topic modeling approach and analyze the availability and quality of leisure walk POIs in OSM, examining

² <https://www.inaturalist.org/>

³ <https://scenicornot.datasciencelab.co.uk/>

⁴ <https://location.foursquare.com/products/places>

⁵ <https://www.yelp.com/dataset>

how the identified classes relate to data quality aspects. Finally, we present a baseline machine learning model to imitate human selectivity in choosing relevant POIs for leisure walks and discuss the challenges involved in developing such predictive models.

3 Leisure Walking Dataset

WalkingMaps is a publicly available service provided by *Victoria Walks Inc.*, a non-profit organization dedicated to promoting walkable communities in Australia. The platform allows users to explore and share leisure walks. Each recommended leisure walk includes several types of information: (1) the *walk*, represented as a linear geometry; (2) *POIs*, each represented as a point with a verbal description and optionally an image; and (3) a *verbal description* of the walk, highlighting the experience and providing navigational instructions. We utilized web scraping techniques to extract this information from the *WalkingMaps* website, resulting in a dataset of 386 leisure walks and 4392 POIs⁶. Detailed statistics derived from the verbal descriptions of the walks and POIs are presented in Table 1. The table shows that POI descriptions average about 23 words, sufficient to describe the POI and convey the rationale for the POI recommendation.

■ **Table 1** Statistics of collected textual data from *WalkingMaps*.

Item	Min	Median	Mean	Max
Walk description (word count)	7	130	181	540
Walk description (character count)	43	764	1062	3052
POI description (word count)	2	23	22	116
POI description (character count)	11	127	130	296

To further enrich the collected data, we developed a semi-automated approach to match the POIs with OSM objects. Using the Nominatim Geocoding API provided by OSM, we found that only 14.16% of the POI descriptions could be automatically matched to OSM data. This low geocoding rate is attributed to the rich verbal descriptions included in the POI data, well beyond simple names and feature types, making it challenging for the Geocoding API to interpret and locate the corresponding objects in OSM.

To achieve more accurate matching between POIs in the *WalkingMaps* dataset and OSM, we designed a simple annotation interface. This interface displays the verbal description of a POI, its location on the OSM map, and the ten OSM objects with the highest matching scores to the POI description, facilitating the interactive identification of relevant OSM objects. The matching score is calculated using the cosine similarity of embedding vectors derived from the POI verbal description and concatenated OSM key values for nearby OSM objects. These embeddings are generated using sentence transformers [18] with the *msmarco-distilbert-dot-v5* model. We selected this model, trained on the MS MARCO dataset [1], because its characteristics align closely with our problem of matching POI descriptions to OSM thematic representations. In the MS MARCO dataset, queries consist of keyword-based prompts (e.g., “largest river”) paired with natural language answers (e.g., “The Nile River is the longest river in the world at 4,132 miles”), which aligns to our task of mapping less structured, keyword-like OSM representations (e.g., “nature beach”) to rich textual descriptions of *WalkingMaps* POIs (e.g., “Byron Bay Main Beach is a walker’s paradise. Flat, hard sand and plenty of things to look at.”).

⁶ This dataset has been collected in February 2023.

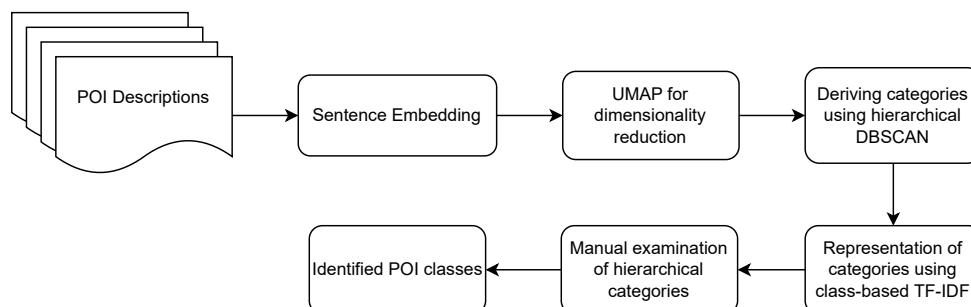
Using the annotation interface, annotators can select matches from the top ten suggestions or explore the map to identify other relevant objects in OSM. When a match is found, the annotator records the OSM identifier(s) alongside the corresponding POI in the dataset. Three scenarios were observed during the annotation process: (1) no match, when the POI is missing in OSM; (2) single match, when only one object matches the POI description; and (3) multiple matches, when several objects align with the POI description. The third scenario occurs when a POI description refers to an aggregate of similar objects (e.g., “there are a couple of ponds in the garden”) or a collection of spatial objects of different types (e.g., “a great picnic area and playground, complete with multiple BBQs, toilets, and plenty of play equipment”).

In total, 2385 POIs were matched with 3119 OSM objects. Due to multiple matches, the number of matched OSM objects exceeds the number of POIs in the leisure walking dataset. Among the 2385 matched POIs, 2022 are single matches, while 363 are multiple matches, associated with 1097 OSM objects. The remaining 2007 POIs could not be matched with any OSM objects at the time of annotation (i.e., May-June 2024).

To study the subjectivity involved in matching POIs to OSM objects, 5% of the dataset was independently annotated by another annotator. The results showed an inter-annotator agreement of 81.2%, with 407 out of 497 POIs in complete agreement between the two annotators. When considering partial agreements cases, where multiple matches for a POI resulted in overlapping lists of matches between the two annotators, the agreement increased to 88.1%, covering 438 out of 497 POIs.

4 POI Classification

In the first step, we use topic modeling to categorize POIs based on their verbal descriptions (see Figure 1). To ensure the topics are derived from the intrinsic characteristics of the POIs (i.e., what the POIs are) rather than their geographic locations (i.e., where the POIs are located), we pre-process the text to remove names of common places such as suburbs, local government areas, towns, and cities. For topic modeling, we apply the BERTopic workflow [7], which involves transforming descriptions into sentence embeddings, reducing the dimensionality of the embeddings using Uniform Manifold Approximation and Projection (UMAP) [14], clustering the reduced vectors with the Hierarchical DBSCAN method [2] to minimize noise, and creating bag-of-words representations for each cluster. The clusters are then characterized by top keywords identified using topic class-based Term Frequency-Inverse Document Frequency (TF-IDF). This workflow results in a detailed, hierarchically organized list of topics, each represented by a set of keywords.



■ **Figure 1** Classification workflow.

To identify classes based on the identified topics, we manually examine each topic within the hierarchy. We verify whether the top-10 keywords for each topic represent a particular type of places (i.e., well-formed topics). We then interpret the hierarchy of these well-formed topics to label and classify them. The proposed classification of POIs, based on the 75 identified topics, includes four classes further divided into ten sub-classes (209 POIs belonging to five malformed topics are not classified, and labeled as unknown). The four classes are:

- **Nature-related POIs:** This class includes 22 identified topics related to nature, describing natural features such as mountains, bays, beaches, and waterfalls, as well as habitats for flora and fauna (e.g., “the native grass trees are particularly striking”). In the dataset, 1478 POIs are classified as nature-related, 745 as natural features and 733 as habitats for flora and fauna. Out of 386 leisure walks, 338 include at least one nature-related POI, and 125 walks have the majority of POIs (at least 50%) belonging to this class. This indicates that nature-related POIs are the dominant class in terms of both the number of POIs and the number of leisure walks that include these POIs. The definition of nature-related POIs extends beyond the previous definition by [9] and encompasses places or regions with fuzzy boundaries that describe, for example, habitats for flora and fauna.
- **Activity-related POIs:** These POIs are described in terms of their functional roles and what activities they offer, which is the primary reason for their recommendation. For example, “gather your friends and start training for the WNBL or NBL by shooting some hoops at Braybrook Park” (the (women’s) national basketball league). The four major subclasses of this category are places that support sport-related activities (212 POIs), aquatic activities (227 POIs), picnic and camping (482 POIs), and food and beverages (167 POIs). This class includes 1088 POIs in the dataset, with 315 leisure walks having at least one activity-related POI and 63 walks having the majority of their POIs belonging to this class.
- **Society-related POIs:** These POIs are man-made features and objects that people found interesting and recommended for investigation during leisure walks in the neighborhood. For example, “established by Aboriginal artist Lin Onus, this was a social and political meeting place during the 1960s for young people influenced by the Black Power Movement.” This class includes two major subclasses: (1) human-made landmarks (e.g., hospitals, hotels, and law courts), and (2) places and objects with artistic, historical, and cultural significance (e.g., native people establishments, historic gold mines, and murals). This class includes 1070 POIs in our dataset, with 376 human-made landmarks and 694 POIs with artistic, historical, and cultural significance. In total, 256 leisure walks include POIs of this class, with 64 walks having the majority of their POIs belonging to this class.
- **Transport-related POIs:** This class includes two major subclasses: (1) trails, paths, streets, canals, and bridges that are described as POIs due to the atmosphere and experiences they afford (e.g., “the Boardwalk is fantastic and makes for easy walking”), and (2) transport-related facilities (i.e., lines and stations) that aid in planning commuting to and from the leisure walk area (e.g., “start at Ringwood Railway Station fronting Maroondah Highway”). This class includes a total of 547 POIs, with 472 related to trails, paths, and canals, and 75 describing lines and stations. In total, 238 leisure walks include POIs of this class, with twelve walks having the majority of their POIs belonging to this class.

5 POIs and Data Quality of OSM

As discussed in Section 3, we matched the POIs from the WalkingMaps dataset to OSM objects wherever possible. Here, we present detailed results on the matching status (no match, single match, multiple matches) of the POI classes and subclasses. We further discuss how these classes and subclasses are described using OSM tags (i.e., key/value pairs).

Table 2 shows the matching status of POIs based on their class and subclass. The results reveal that most unmatched POIs belong to the nature-related class, specifically to habitats of flora and fauna (552 POIs, roughly 75%). One reason for this missing data in OSM is the complexity of mapping habitat regions due to their fuzzy and possibly unknown boundaries. Table 2 shows that activity-related POIs are mostly found in OSM, but with a considerable number of POIs in this category matched to multiple OSM objects. This is because their POI descriptions may include several features (e.g., multiple shops in a neighborhood, or BBQ facilities and seats in a park) or a single feature conceptualized as multiple entities in OSM (e.g., a sports complex with multiple buildings). Society-related and transport-related POIs are also mostly found in OSM as well. The unmatched POIs for these classes are due to (1) places that no longer exist (e.g., an former shop that has changed, no longer exists, or was demolished), (2) places or features that have not been mapped, mostly trails and small artistic objects (e.g., sculptures and murals), and (3) a difficulty in identification due to a lack of thematic information for OSM objects related to these classes.

■ **Table 2** Matching status for each class and subclass.

Class	Subclass	Matching status
Nature-related	flora and fauna	no match (552), single match (152), multiple matches (29)
Nature-related	natural landmarks	no match (322), single match (357), multiple matches (66)
Activity-related	picnic and camping	no match (158), single match (257), multiple matches (67)
Activity-related	aquatic	no match (80), single match (112), multiple matches (35)
Activity-related	sport	no match (76), single match (108), multiple matches (28)
Activity-related	food and beverage	no match (54), single match (105), multiple matches (8)
Society-related	human-made landmarks	no match (122), single match (238), multiple matches (16)
Society-related	art, history and culture	no match (351), single match (310), multiple matches (33)
Transport-related	trails, paths and canals	no match (165), single match (258), multiple matches (49)
Transport-related	lines and stations	no match (27), single match (44), multiple matches (4)

Table 3 shows the top OSM tags for each class and subclass of the POIs matched to the OSM database. The results indicate that for the nature-related class, the dominant keys are natural, leisure, foot, and highway, providing information about the type of places

and their accessibility. Activity-related POIs are mainly described using the leisure key with values such as playground, pitch, and park, with common tags such as the building tag also being popular. Society-related POIs are described primarily using the building tag, along with specific tags for places of worship and the tourism key. Other popular tags for society-related POIs, not listed in the table, include `tourism:[museum, hotel]`, `amenity:[school, post_office]`, and `historic:memorial`, highlighting their cultural and historical significance. The transport-related class includes popular tags describing types, such as `bridge:yes`, `highway:cycleway` for the trails, paths, and canals subclass, and `railway:station`, `train:yes` for lines and stations. As shown in the table, the number of frequent tags, except for common tags like `building:yes`, is much smaller than the number of actual matched POIs for each class and subclass. This discrepancy is due to either the POIs belonging to diverse categories or the OSM records lacking sufficient key-value information to describe them (i.e., 1078 OSM objects in this experiment only had one tag, either name or type, without any other thematic information available).

■ **Table 3** Popular OSM key/values for each class and subclass.

Class	Subclass	Most frequent OSM tags (count)
Nature-related	flora and fauna	leisure:park (30), natural:water (29), highway:footway (19)
Nature-related	natural landmarks	foot:yes (102), highway:cycleway (102), highway:path (70)
Activity-related	picnic and camping	leisure:playground (98), access:yes (58), leisure:park (43)
Activity-related	aquatic	man_made:pier (19), building:yes (13), leisure:slipway (11)
Activity-related	sport	leisure:pitch (63), leisure:park (21), building:yes (18)
Activity-related	food and beverage	building:yes (21), amenity:restaurant (19), amenity:cafe (18)
Society-related	human-made landmarks	building:yes (56), religion:christian (23), amenity:place_of_worship (22)
Society-related	art, history and culture	building:yes (57), addr:state:VIC (41), tourism:artwork (31)
Transport-related	trails, paths and canals	layer:1 (92), bridge:yes (90), highway:cycleway (67)
Transport-related	lines and stations	railway:station (15), train:yes (11), railway:miniature (10)

The relationship between the identified classes and subclasses with OSM tags (key-value pairs) can be numerically described using Cramér's V categorical association [4]. Table 4 shows the measured associations between OSM keys, OSM values, and OSM key-value pairs (i.e., tags) with classes, subclasses, and topics. It indicates that key-value pairs have a high association value with classes (0.73) and subclasses (0.70). With both key and value available, we can predict the corresponding class and subclass. This number is lower for individual keys and values, especially for keys, which are often general – e.g., the *leisure* key can have values such as park, fishing, garden, or pitch, describing POIs belonging to different classes and subclasses in our classification.

■ **Table 4** Cramér’s V categorical association between class/subclass/topic.

	OSM key	OSM value	OSM key-value
Class	0.43	0.68	0.73
Subclass	0.34	0.67	0.70
Topic	0.20	0.61	0.63

To demonstrate the impact of low-quality thematic information on classifying POIs using OSM tags, we trained two baseline predictive models. The first model uses a Ridge Classification approach with TF-IDF vectors of all concatenated key-value tags available in OSM as input features. The second model is a Gradient Boosting Classification model trained with the presence of popular key values as binary features (defined by a hard threshold of at least 25 counts, resulting in 472 keys). Both classifiers were tested using 10-fold cross-validation, showing similar results in predicting POI classes using OSM thematic information (see Table 5). Removing OSM matches that only have a name and no further descriptive tags significantly improves the accuracy of both models from about 0.54 to 0.72. This indicates that even with few available tags, we can identify POI types; however, a significant number of matched POIs do not have such descriptive tags available. The *SHapley Additive exPlanations* (SHAP) values [12] of the Ridge model and the feature importance measures of the Gradient Boosting model highlight the importance of features such as *amenity*, *playground*, *viewpoint*, *natural*, *waterfall*, *building*, *garden*, and *office* in the thematic descriptions to predict the correct POI classes.

■ **Table 5** Accuracy, weighted precision, recall and F-score values of the baseline predictive models.

	Accuracy	Precision	Recall	F-score
Ridge Classifier	0.54	0.6	0.54	0.53
Gradient Boosting Classifier	0.55	0.58	0.55	0.53

6 POI Selection

In this section, we analyze the selectivity involved in choosing a POI for a leisure walk and test a baseline machine learning (ML) method to predict whether a POI candidate is suitable for a leisure walk (i.e., a binary classifier). We already have the selected POIs in the dataset (the matched POIs to the OSM database with at least one tag other than *name*, i.e., 2367 OSM objects). To define alternative POI choices not recommended for leisure walks, we selected all other OSM objects within 200 meters of the actual leisure walk POIs (extending the buffer zone to 1000 meters if no alternative POI candidate was found) and filtered them to spatial objects containing one or more of the following OSM keys (with any value) in their tags: *amenity*, *shop*, *railway*, *bridge*, *club*, *building*, *historic*, *tourism*, *place*, *waterway*, *landuse*, *leisure*, *natural*, *office*, *boundary*, *highway*, *man_made*. These keys were selected based on the top frequent tags observed in OSM objects matched with the leisure walk POIs.

The list of POI candidates that are not recommended by people within their leisure walk descriptions contains 183906 spatial objects with at least one tag other than *name*. The number of not recommended POIs is more than 77 times larger compared to the recommended POIs in our dataset, highlighting the selectivity and challenge of POI selection. The random

baseline theoretically results in 1.3% accuracy, meaning only one out of 77 suggestions is a POI recommended by people for leisure walks. Another challenge is related to the sparsity of the data; some not-recommended POIs could be relevant for (other) leisure walks, but due to the relatively small number of recommended walks in the dataset, their relevance is unknown and not being captured.

To test how a baseline ML model can imitate human selectivity in recommending a POI for a leisure walk, we trained a binary classifier model and select the best performing model using 10-fold cross-validation. This classifier includes an ensemble of two classifiers focused on the available thematic and spatial features of the OSM objects, respectively. Both classifiers are trained with over-sampled data using the SMOTE technique to better model this highly imbalanced dataset, and tested with an imbalanced unseen dataset. The test dataset includes 10% of the whole dataset, randomly selected and stratified based on the binary outputs to ensure it includes both positive and negative cases.

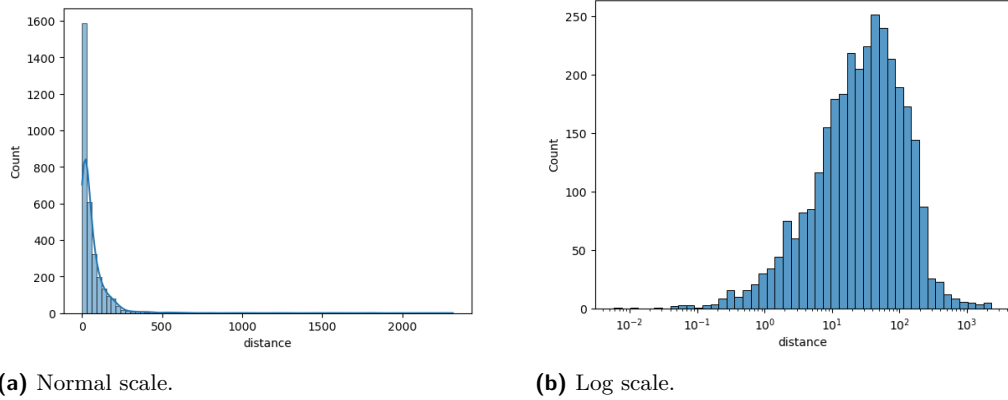
The thematic classifier is trained using the Ridge Classification method with features being TF-IDF vectors of textual descriptions generated by concatenating OSM key-value pairs. Table 6 shows the confusion matrix of the test dataset for predicting whether a candidate is a relevant POI in the leisure walk context. This classifier has an overall accuracy of 0.90, precision of 0.10, recall of 0.82, and ROC AUC of 0.86. These results indicate that while the model performs much better than a random classifier and has a high recall, it still struggles with precise predictions. In other words, out of 10 predicted POIs, only one was actually recommended by people, while most of the recommended POIs are predicted correctly (i.e., 82.3% recall). The low precision is primarily attributed to the subjectivity inherent in POI selection and not necessarily that suggested POIs are irrelevant to the context of leisure walk. Although several POIs may be relevant and useful for providing a leisure walk experience, only a few are suggested by users, influenced by their personal preferences and experiences. However, a much larger and more diverse dataset is required to minimize the impact of subjectivity in POI selection task.

■ **Table 6** Confusion matrix predicting relevant POIs using thematic information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	16597	1794
POI (actual)	42	195

The spatial classifier is built based on the Euclidean distance from POI candidates belonging to the leisure walk, the area of POI candidates (0 for point-based and linear geometries), and their length (0 for point-based geometries). The histogram of distances from the recommended POIs to their leisure walks is shown in Figure 3. This figure demonstrates that the thresholds used for creating the list of the not-recommended POI candidates are within realistic distance ranges for the actual recommended POIs. Thus, we are not creating a single metric that makes the prediction obvious or trivial for the model. These three spatial features are used to train a Gradient Boosting classifier, and the confusion matrix of the test data is shown in Table 7. The results show slightly worse performance compared to the thematic classifier but are still much better than a random classifier. The accuracy is 0.82, precision is 0.06, recall is 0.84, and ROC AUC is 0.83.

When we ensemble the thematic and spatial classifiers, we observe the complementary role of these features through the improvement in prediction precision. The ensemble model predicts a POI candidate as relevant for leisure walks if both thematic and spatial classifiers



■ **Figure 3** Euclidean distance between the leisure walk path and matched OSM objects.

■ **Table 7** Confusion matrix for predicting recommended POIs using spatial information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	15067	3324
POI (actual)	39	198

predict it as relevant. The results of this ensemble model are shown in Table 8. This model achieves an accuracy of 0.98, precision of 0.34, recall of 0.69, and ROC AUC of 0.84. The significant improvement in precision, from one correct prediction out of ten predicted POIs for the thematic model to one correct prediction out of three predictions in the ensemble model, demonstrates the complementary role of thematic and spatial features in predicting the relevance of a POI candidate for leisure walks using OSM data.

■ **Table 8** Confusion matrix for predicting recommended POIs using spatial and thematic information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	18072	319
POI (actual)	74	163

7 Discussion and Conclusion

In this paper, we introduced a new dataset collected from the *WalkingMaps* website, which includes verbal descriptions of leisure walks, geometric representations of the walks, and a set of POIs with their point-based representations and verbal descriptions for each leisure walk. We further enriched the POIs by matching them with OSM objects using a semi-automated approach and classified each POI using a topic modeling based on their verbal descriptions. Our proposed classification includes four top-level classes: nature-related, activity-related, society-related, and transport-related POIs. The classification further breaks down into ten subclasses: habitats of flora and fauna, natural landmarks, places that offer food/beverages or sport/aquatic activities, places related to picnic and camping, human-made landmarks, places with artistic, historical, or cultural significance, transport-related places such as trails, paths, and canals, and finally lines and stations.

Next, we discussed the availability of OSM data for the POIs recommended in leisure walks. We observed that only 14.16% of the descriptions can be geocoded automatically. Even with manual inspection to find data in OSM, only 54.3% of POIs can be found in OSM database. During the annotation process, we noticed that several POI descriptions need to be matched with multiple OSM objects, either because these descriptions describe multiple objects at once (e.g., BBQ and playground in a park) or because the single object described is a complex entity modeled with multiple OSM objects (e.g., sports complexes that include several buildings). We noticed that several POI descriptions included sensory details (e.g., visual, auditory, olfactory, or tactile), which are highly relevant and useful in the context of leisure walking but are often ephemeral and context-dependent. These rich, human-centric descriptions were challenging to match with OSM records, as OSM often provides only rudimentary thematic information compared to the richness of these narratives. [13] conducted a detailed analysis of landscape and place descriptions incorporating sensory information; however, in the context of leisure walks, further research is needed to analyse these complex POI descriptions.

We further discussed the relationship between data availability and the identified classes, noting that roughly 75% of the POIs related to habitats of flora and fauna are missing in the OSM database. Activity-related POIs have more matches to multiple OSM objects compared to others, as their descriptions often include multiple activities offered in an area (i.e., multiple objects described in a description) or places with multiple buildings or compartments, modeled as multiple objects in OSM. Most activity and society-related POIs can be found in OSM, with missing ones mainly related to places that have changed use or no longer exist (e.g., a cafe or restaurant described in a leisure walk that now changed to another entity). The main reason for transport-related POIs not being matched to OSM objects is that the POI descriptions only describe part of a mapped path or trail, while OSM captures the whole path or trail. These parts of the paths in these cases are often important due to the views they offer or the atmosphere and vibe of walking in there (e.g., *“follow the gravel path down towards the falls for a spectacular view”*).

Finally, we study whether we can automate the POI selection for leisure walks, given the walk area, geometry, and a set of POI candidates (both the recommended POIs in the dataset and a large set of POI candidates from the OSM database). We trained an ensemble model that utilizes thematic information from OSM objects (all available tags) and their spatial features (i.e., distance to walk, area, and length). The results show the complementary role of spatial and thematic features, with the ensemble model significantly outperforming individual spatial and thematic models, improving from one correct guess out of 10 suggestions to one correct guess out of three suggestions. We also highlight the challenges of POI selection tasks due to subjectivity in the process, data sparsity (i.e., 2367 matched POIs), and highly imbalanced train/test datasets (one recommended POI for 77 not-recommended POI candidates).

This study can be further extended, by refining the methodology for classifying POIs for leisure walks and by further verifying our findings against leisure walk POIs from other datasets and geographies. Our focus in this paper is to provide baseline methods for selecting POIs and demonstrate these on the Leisure Walks dataset, to discuss its coverage and limitations. Using more sophisticated machine learning methods, we may expect improvements in the POI selection task. This task can be reformulated in other ways, such as predicting new POIs given a path and previous POIs. While beyond the immediate scope of this study, the dataset presented here can be used to develop and test such approaches. The data also have potential applications outside the leisure walking

context, such as studying and improving the automatic identification of OSM objects based on verbal descriptions, beyond the usual geocoding task, as here we consider multiple matches for a single description. The task here also differs conceptually from geocoding since the point-based location is already available and provided in the dataset. Instead of finding the place/object's location, we aim to find the OSM object(s) (or spatial entities in any other spatial database) related to what is verbally and geometrically described by a person. As described in the paper, the results of such matching may yield zero, one, or multiple objects depending on data availability and how objects are conceptualized within the database (e.g., OSM). For example, a sports complex may be represented as multiple areal features in OSM, yet described as a single entity in the verbal description.

In leisure walk descriptions research, it is essential to differentiate between two types of POIs: those where one can visit, investigate, and engage in activities, and those that serve as locations to view other objects of interest (e.g., a view from a hill to a distant mountain). Leisure walking, and the inclusion of POIs in walk descriptions demands further conceptual research to investigate the role of such POIs and why they are recommended by the authors of descriptions. This includes examining the purpose of a POI: is it recommended to aid navigation, for its functional role, or as a focus of attention for interest and enjoyment during the walk.

References

- 1 Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated Machine Reading COmprehension Dataset, 2018. [arXiv:1611.09268](#).
- 2 Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- 3 Buru Chang, Yonggyu Park, Donghyeon Park, Seongsoon Kim, and Jaewoo Kang. Content-aware hierarchical point-of-interest embedding model for successive POI recommendation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18*, pages 3301–3307. AAAI Press, 2018. doi:10.24963/IJCAI.2018/458.
- 4 Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.
- 5 Song Gao, Krzysztof Janowicz, and Helen Couclelis. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467, 2017. doi:10.1111/TGIS.12289.
- 6 Yunfan Gao, Yun Xiong, Siqi Wang, and Haofen Wang. Geobert: pre-training geospatial representation learning on point-of-interest. *Applied Sciences*, 12(24):12942, 2022.
- 7 Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022. doi:10.48550/arXiv.2203.05794.
- 8 Ehsan Hamzei, Thi Minh Hoai Bui, Martin Tomko, and Stephan Winter. hamzeiehsan/leisure-walking-analysis. Software, swbId: swb:1:dir:cddb6d133e212246c2e458e4ea46f1358cd27927 (visited on 2025-07-30). URL: <https://github.com/hamzeiehsan/leisure-walking-analysis>, doi:10.4230/artifacts.24214.
- 9 Charlie Hewitt, SD Sabbata, A Ballatore, S Cavazzi, and Nicholas Tate. Defining natural points of interest. In *29th Annual GIS Research UK Conference (GISRUK)*. Cardiff, Wales, UK (Online), 2021.
- 10 Olga Koblet and Ross S. Purves. From online texts to landscape character assessment: Collecting and analysing first-person landscape perception computationally. *Landscape and Urban Planning*, 197:103757, 2020. doi:10.1016/j.landurbplan.2020.103757.

- 11 Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. *Cities*, 99:102610, 2020. doi:10.1016/j.cities.2020.102610.
- 12 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- 13 Nora Fagerholm Manuel F. Baer, Flurina Wartmann and Ross S. Purves. Extracting sensory experiences and cultural ecosystem services from actively crowdsourced descriptions of everyday lived landscapes. *Ecosystems and People*, 20(1):2331761, 2024. doi:10.1080/26395916.2024.2331761.
- 14 L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018. arXiv:1802.03426.
- 15 Grant McKenzie and Krzysztof Janowicz. OpenPOI: An open place of interest platform (Short paper). In *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.GISCIENCE.2018.47.
- 16 Achilleas Psyllidis, Song Gao, Yingjie Hu, Eun-Kyeong Kim, Grant McKenzie, Ross Purves, May Yuan, and Clio Andris. Points of Interest (POI): A commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2(1):20, 2022.
- 17 Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 116–125, 2014. doi:10.1145/2631775.2631799.
- 18 Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019. doi:10.18653/V1/D19-1410.
- 19 Tiina Sarjakoski, Pyry Kettunen, Hanna-Marika Halkosaari, Mari Laakso, Mikko Rönneberg, Hanna Stigmar, and Tapani Sarjakoski. Landmarks and a hiking ontology to support wayfinding in a national park during different seasons. In Martin Raubal, David M. Mark, and Andrew U. Frank, editors, *Cognitive and Linguistic Aspects of Geographic Space: New Perspectives on Geographic Information Research*, pages 99–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-34359-9_6.
- 20 Nancy Stevenson and Helen Farrell. Taking a hike: exploring leisure walkers embodied experiences. *Social & Cultural Geography*, 19(4):429–447, 2018. doi:10.1080/14649365.2017.1280615.
- 21 Demi van Weerdenburg, Simon Scheider, Benjamin Adams, Bas Spierings, and Egbert van der Zee. Where to go and what to do: Extracting leisure activity potentials from web data on urban space. *Computers, Environment and Urban Systems*, 73:143–156, 2019. doi:10.1016/J.COMPENVURBSYS.2018.09.005.
- 22 James Williams, Stefano Cavazzi, James Pinchin, Adrian Hazzard, Gary Priestnall, and Andrea Ballatore. Context for leisure walking routes: A vision for a spatial-platial approach. In *Spatial Data Science Symposium 2022 Short Paper Proceedings*, 2022. doi:10.25436/E20W2J.
- 23 James Williams, James Pinchin, Adrian Hazzard, Gary Priestnall, Stefano Cavazzi, and Andrea Ballatore. Emerging platial narratives and themes from a leisure walking study. In *Proceedings of the 4th International Symposium on Platial Information Science (PLATIAL’23)*, pages 23–28, 2023.
- 24 James Williams, James Pinchin, Adrian Hazzard, Gary Priestnall, Stefano Cavazzi, and Andrea Ballatore. Walkgis: Exploring platial analysis of leisure walks via linked video narratives. In *31st Annual Geographical Information Science Research UK Conference (GISRUK)*, 2023.

- 25 Stephan Winter, Ehsan Hamzei, Nico Van de Weghe, and Kristien Ooms. A graph representation for verbal indoor route descriptions. In Sarah Creem-Regehr, Johannes Schöning, and Alexander Klippel, editors, *Spatial Cognition XI*, pages 77–91, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-319-96385-3_6.
- 26 Lih Wei Yeow, Raymond Low, Yu Xiang Tan, and Lynette Cheah. Point-of-Interest (POI) Data Validation Methods: An Urban Case Study. *ISPRS International Journal of Geo-Information*, 10(11), 2021. doi:10.3390/ijgi10110735.
- 27 Kangzhi Zhao, Yong Zhang, Hongzhi Yin, Jin Wang, Kai Zheng, Xiaofang Zhou, and Chunxiao Xing. Discovering subsequence patterns for next POI recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3216–3222, 2021.

A Frequent Words by POI Class

■ **Table 9** Most frequent OSM key/values for each class and subclass.

Class	Most frequent OSM tags (count)
Nature-related	view (281), tree (224), bird (183), lake (182), river (165), beach (161), creek (154), path (148), see (130), track (116)
Activity-related	park (307), playground (208), picnic (164), area (151), BBQ (120), spot (102), river (101), great (95), place (86), club (83)
Society-related	garden (147), building (143), built (124), centre (117), art (114), church (103), street (81), memorial (78), community (77), school (77)
Transport-related	bridge (232), track (157), path (107), station (89), walking (83), railway (66), start (57), street (51), boardwalk (49), park (47)