AGILE: GIScience Series, 6, 6, 2025. https://doi.org/10.5194/agile-giss-6-6-2025
Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025.
Eds.: Auriol Degbelo, Serena Coetzee, Carsten Keßler, Monika Sester, Sabine Timpf, Lars Bernard
This contribution underwent peer review based on a full paper submission.

Exploring Urban Place Function through User-Generated Textual Content

Shahreen Muntaha Nawfee[1], Stef De Sabbata[1], and Nicholas J. Tate[1]

[1] School of Geography, Geology and the Environment, University of Leicester, Leicester, LE1 7RH, UK

Correspondence: Shahreen Muntaha Nawfee (smn18@leicester.ac.uk)

Abstract. Understanding the functions played by different places as part of the urban fabric is essential for the sustainable development of cities. The increase in the amount and variety of people's digital footprint, combined with the recent emergence of Large Language Models (LLMs), holds the potential to derive novel and useful insights into the function of urban places from unstructured textual content. This paper presents a reproducible and robust methodological framework to extract place functional characteristics from user-generated textual content. In particular, we analyse tags of Points of Interest (POIs) from OpenStreetMap (OSM) and summaries of geo-tagged articles from Wikipedia through spatial clustering and topic modelling, exploring classic and current approaches in Natural Language Processing (NLP). We evaluate and compare the performance of different LLMs, including Mistral and Llama. Our results show that the framework provides a good insight into place functionality for our case study.

Submission Type: Analysis

Keywords. Natural Language Processing (NLP), Urban Place function, Large Language Models (LLMs), OpenStreetMap (OSM), and Wikipedia

1 Background

More than half of the world's population lives in urban areas, and the number is projected to continue rising (World Bank, 2023). Such unprecedented and, in some cases, unplanned rates of urbanisation challenge the sustainable planning and management of urban resources. In this context, one of the United Nations' Sustainable Development Goals focuses on the urban environment by proposing the sustainable development of cities (United Nations, 2024). Reaching that goal requires a better understanding of different dimensions of urban environments, including the thematic and functional characteristics of urban places and their spatiotemporal distribution, which influences the socio-economic development of the cities.

Place is a fundamental concept in Urban Studies and Geographic Information Science (Goodchild, 2011). A large amount of work in the literature has been devoted to describing places in cities (Bahrehdar et al., 2020), analysing their structures (Calafiore et al., 2021), categorising their types (Adams & McKenzie, 2013; Hu et al., 2021; Tang & Painho, 2023), extracting their thematic (Adams & Janowicz, 2015) and functional characteristics (Gao et al., 2017; Niu & Silva, 2021). Places are produced and shaped by human activities and experiences in space, which give meaning to particular locations (Cresswell, 2004). Place functional characteristics are derived from human activities that shape them and reflect the socio-economic patterns of human activity, which determine how physical settings (buildings and other facilities) are arranged in urban areas (Niu & Silva, 2021). The functional characteristics are a combination of place affordances, events, and key descriptors (Oh & Yao, 2021) that relate to the activities places can offer, occurring individually or in a planned and organized manner.

Understanding place function in an urban context is necessary because it can benefit urban planners, policymakers, and researchers by enabling them to gain knowledge about urban areas (Hu et al., 2021) and develop better plans and strategies for sustainable development (Niu & Silva, 2021). Moreover, such human-centric approaches, supported by place-based research, can provide a better understanding of urban issues, such as the inefficient use of land resources and unplanned growth.

Representations of place can be varied, and they are communicated through different mediums. With the increased digitalisation of everyday life, the human activities that shape place functions are captured as digital footprints (De Sabbata et al., 2023), including User-Generated Content (UGC) in digital platforms, creating layers of information that become an integral part of places they describe (Ballatore & De Sabbata, 2019). The ability to capture these activities predominantly in an urban context (e.g., urban functional zones, geodemography, gentrification) makes UGC particularly suitable for studying urban places. Compared to traditional data sources, some UGC platforms can provide information at a higher frequency (Jain et al., 2021) and granular level on a spatiotemporal scale (Gao et al., 2017).

Over the last decade, a growing body of literature has explored how UGC sources can be used to understand urban processes (Bahrehdar et al., 2020; Chapple et al., 2021; Chen et al., 2018; Comber et al., 2024; Du et al., 2021; Gao et al., 2017; Hu et al., 2015; Purves & Hollenstein, 2010; Yao et al., 2016; Zhang et al., 2018). The textual content in UGC platforms is a reservoir of how people experience places. The recent advances in Natural Language Processing (NLP) can potentially render the analysis of text-based UGC more effective. NLP encompasses various statistical and artificial intelligence approaches to analyze human natural language (Albalawi et al., 2020), including keyword extraction, translation, question answering, and topic modelling. Moreover, the advent of Large Language Models (LLMs) can unlock more information about places and our everyday experience with them (De Sabbata et al., 2023).

While previous research has advanced our understanding of urban place function, significant gaps remain unaddressed. First, although Points of Interest (POIs) locations have been examined to determine place type, their tags have been explored less to determine place functions. Second, recent advances in NLP have not been extensively explored for extracting place functions from textual UGC. Third, a distinct methodological framework for urban place function determination is lacking. A robust, reproducible framework can demonstrate how the available datasets and methods can be best utilised to extract urban function information. Considering these gaps, in this study, we propose and showcase a framework exploring textual UGC through both classic and current NLP approaches to enhance the understanding of urban place function.

To that purpose, we analyse POI data from OpenStreetMap (OSM) and summaries of geotagged articles from Wikipedia as sources of complementary information on geographic entities in the same place. The practical benefits of using POIs include a bottom-up approach to delineating urban functions, which identifies multiple functions offered at the exact location. OSM combines information about the built environment and land use at fine detail by including mundane objects such as benches and trees. At the same time, Wikipedia focuses on renowned buildings, streets and institutions, providing more in-depth information about their history and context. In this context, we assume that combining these two sources might allow us to understand better the places they contribute to define. We employ various topic modelling approaches, including Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), and Large Language Models (LLMs), as well as a topic modelling approach based on the Bidirectional Encoder Representations from Transformers (BERTopic), to leverage their capabilities in understanding urban functions.

The remainder of this article is organised as follows. In Section 2, we provide an outline of the methodology, including details on the dataset used, the methods employed, and the case study area where the method is tested. In this section, the methodological framework that we establish is also presented. In Section 3, we present the results and analysis of applying the methodological framework to the case study area. Finally, we conclude the article with section 4, discussing the findings and providing future research directions.

2 Methodology

2.1 Datasets

This study employs two datasets: OSM POIs and Wikipedia article summaries. The OSM dataset provides very short textual data in the form of tags, which essentially represent the category of what is located at that position. Wikipedia articles can vary quite substantially in their length and structure. Our analysis is limited to the summaries of geo-located Wikipedia pages, which are more homogenous while still providing key information about the represented object.

OSM is one of the most widely studied and used crowdsourced mapping projects, with the most active registered users among the web-based crowdsourced mapping platforms. OSM has become a prominent data source for several studies in Geographic Information

Science and Quantitative Geography (Jin, 2020; Fathi et al., 2020). We extracted OSM POIs (6304 objects for the case study area) from the ohsome API [1]. The data collection is initiated by retrieving OSM data (including the geometry and tags of elements) for each UK Census Output Area (OA)[2] within the selected case study area. All the data were collected for the same date, so they are comparable across different regions.

Wikipedia is an online-based crowdsourced encyclopaedia that hosts information on various topics. The pages are composed of voluntary contributions from users who aim to produce articles about places, people, events, and other topics. The Wikipedia Articles' information was collected from the Wikimedia data dump[3] and the page titles were used to retrieve the summary texts for 259 articles of our case study area using Wikipedia-API[4].

2.2 Case study area

As our case study area, we chose the Royal Borough of Kensington and Chelsea in Greater London, which is centrally located west of the City of London and north of the River Thames. The borough is quite diverse, including areas classified in the top most deprived decile of the UK Indices of Multiple Deprivation[5], especially in its northern part, and wealthy areas, particularly in its south-eastern section, which hosts several restaurants, renowned museums, boutiques and antique shops.

2.3 Methods

In this project, we conceptualise urban place at a mesoscale as spatial clusters of POIs. These are then assigned a function label based on the topic identified by applying topic modelling to the tags or text summaries attached to the POIs composing the cluster. We interpret the topics generated by the topic modelling approaches for each urban place (i.e., a cluster of POIs) as their functional signatures, labelling them depending on the topic-term distribution for each topic. Our study includes four main components. First, we utilize the OSM and Wikipedia datasets separately, creating urban places as spatial clusters of POIs, which serve as input for topic modelling. Second, we conduct various Topic modelling approaches and label places with their functions. Third, we conduct a qualitative and quantitative assessments of the results. Finally, we select the best-performing approach to analyse the two datasets combined for place function extraction. The methodological flowchart represented in Fig. 1 illustrates the components of our proposed methodology

2.3.1 Spatial Clustering

We began our analysis by performing spatial clustering of OSM POIs and Wikipedia articles separately by applying a density-based clustering, the DBSCAN (Density-based Spatial Clustering of Applications with Noise) method (Ester et al., 1996). DBSCAB was chosen here for several reasons. First, the DBSCAN method can identify spatially distributed clusters of diverse shapes and sizes within a dataset. Such clusters are likely to mirror urban forms and functions known to distribute irregularly without spatial smoothness (Arribas-Bel & Fleischmann, 2022). Second, using DBSCAN enables us to create places where geographic objects are close to one another within a specified distance rather than requiring a predefined number of clusters. This is important in our analysis as the number of functions in the study area is unknown; instead, we aim to find the varying number of functions that our datasets represent. Third, choosing density-based clustering works well with emergent places, which are captured as dense content areas in our datasets, rather than binding our analysis to predefined spatial units (e.g., census boundaries or grid-based distribution).

DBSCAN requires the choice of two parameters to create spatial clusters: the minimum number of points that comprise a cluster and epsilon, i.e., the maximum distance from a point to neighbouring points to consider it part of the same cluster. We experimented with different numbers of minimum points to create clusters, balancing the cluster size and the number of objects classified as noise so that the noise is minimal and as many POIs as possible are part of a place. The distance selection process was adopted from Rahmah and Sitanggang (2016). We measured the average distance of each point to its nearest neighbours in the vertical axis and plotted it against the number of points sorted at each distance in the horizontal axis. The point of curvature is chosen as the optimum distance because the maximum points are sorted into clusters at this distance. We experimented with a range of

[1] https://heigit.org/big-spatial-data-analytics-en/ohsome/ Data for 15th October 2021; Accessed on: 30th October 2021. This dataset is licensed under the Open Data Commons Open Database License (ODbL) by the OpenStreetMap Foundation (OSMF) (© OpenStreetMap contributors).

[2] https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/2011censusgeographies

[3] Data dumps - Meta https://wikimedia.org , Accessed on January 2021

[4] Wikipedia-API - https://www.mediawiki.org/wiki/API:Main_page, Accessed on: January 2022

[5] https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019, Accessed on January 2025

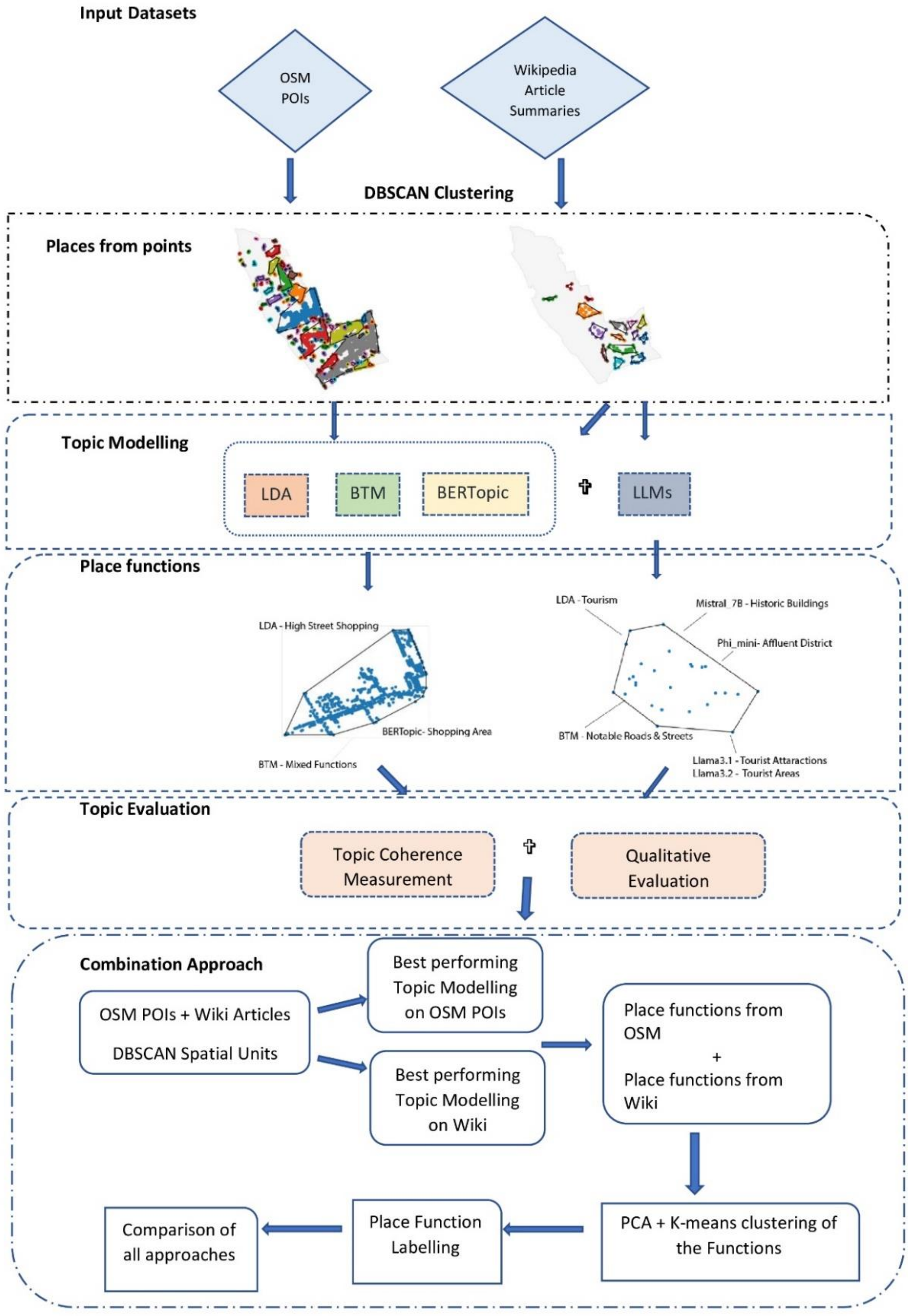

Figure 1. Methodological Flowchart. The figure contains public sector information licensed under the Open Government Licence v2.0. Also contains: National Statistics data © Crown copyright and database right 2025 and Ordnance Survey data © Crown copyright and database right 2025, retrieved through the London Datastore

distance values below or above the point of curvature to choose the optimum epsilon that will also reduce noise. We then execute DBSCAN using the related function from the Scikit learn Python library to create the spatial clusters. For DBSCAN clustering, we chose a distance of 60 m and 4 nearest neighbours for OSM POIs, a minimum of 5 nearest neighbours, and a 200 m distance for Wikipedia articles.

Following DBSCAN clustering, the next step was to draw polygons representing the resultant clusters. The aim of the process was purely for visualisation purposes so that the urban functions represented at each spatial cluster could be mapped. The convex hull shapes were chosen to create these regions where lines join the outermost points in each cluster. The method was implemented using the SciPy package. The noises were removed before applying the regionalisation process.

2.3.2 Topic Modelling

Topic modelling is an NLP task that involves extracting thematic scopes from textual data by analyzing how words are structured within a document and linking documents that share similar content. This study examines a set of topic models designed to capture the functional signatures of places from OpenStreetMap (OSM) and Wikipedia content. Over the last decade, the LDA method proposed by Blei et al. (2003) has been effectively used to identify places of similar thematic (Adams & McKenzie, 2013; Oh & Yao, 2021) and functional signatures (Gao et al., 2017; Jain et al., 2021; Lin et al., 2025), from a corpus of documents. LDA employs a bag-of-words approach and word frequency in the corpus to determine the probability distribution of the documents over the latent topics, assigning each document a unique probability score for each topic it belongs to.

We consider each DBSCAN spatial cluster (our urban places) as a document for input into the LDA model. The OSM POI tags in the clusters are considered words that compose the sentences of the document. The OSM POI tags are typically one to two words, with no specific order. We create a dictionary and a corpus using the Gensim package (Rehurek and Sojka, 2011). The dictionary refers to creating an index of each unique word, and the corpus represents the distribution of words. We iteratively ran a topic coherence measure to determine the coherence value for varying topics ranging from two to twenty. Then, we ran the LDA process with 10 topics, as this yielded the highest coherence score. We also ran the LDA model on the pre-processed Wikipedia article summary texts to obtain six topics. We then analysed the resultant topic-term distribution, which shows the probability of occurrence of each term (the OSM POI tags or words from Wikipedia summaries) in a topic. We considered the top ten highest probability terms to annotate a place function for that topic manually.

Our second tested method was BTM introduced by Yan et al. (2013), using the BTM package to create the model (Wijffels, 2019). The BTM emerges from the difficulty of analysing short texts without rich contexts, which are prevalent in online media, such as social media statuses and brief messages. It generates topics by modelling bi-terms, which are word-pair co-occurrence patterns, over the whole corpus. The number of topics is kept similar to that one selected for LDA (10 for OSM and 6 for Wikipedia) to maintain consistency and comparability at later stages. Place functions were similarly labelled for BTM.

Third, we tested the BERTopic framework (Grootendorst, 2022), a neural topic modelling approach composed of the following steps. First, we used the Sentence Transformers (SBERT) model (Reimers & Gurevych, 2019) “all-MiniLM-L6-v2” to generate sentence embeddings, creating dense vectors of 384 dimensions from the input sentences – i.e., the sequences of OSM tags and the Wikipedia page summaries. Second, we reduced the dimensionality of the embeddings using the Uniform Manifold Approximation and Projection (UMAP) approach (McInnes et al., 2018) and clustered the resulting low-dimensional representations using Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) model (Campello et al., 2013). Finally, the thematic clusters were described and named using keywords extracted using the classic tf-idf approach. Each step requires parameter selection, which differs by the input dataset. For UMAP, we selected cosine distance, 5 nearest neighbours for OSM, and 2 for Wikipedia. For HDBSCAN, we selected 4 and 2 as the minimum cluster size and 3 and 2 as the sample size for OSM and Wikipedia, respectively. The choice of the parameter here is based on the size of the datasets, with OSM having a higher number of nearest neighbours and minimum cluster size as it has more objects than Wikipedia.

We evaluated the performance of LDA, BTM, and BERTopic using topic coherence. This widely used measure considers the pairwise semantic similarity of the top words in a topic and helps determine how interpretable the topics are (Athukorala & Mohotti, 2022). We calculated Cv (Röder et al., 2015), UMass (Mimno et al., 2011), and UCI (Newman et al., 2010) as metrics of topic coherence for LDA and BERTopic using Gensim. These metrics are automated topic coherence metrics for intrinsically evaluating topics resulting from topic modelling. They mostly consider the frequency of co-occurrence of word pairs in the top-N word list of each topic. Following Grootendorst (2022), we used the Countvectorizer to create the corpus and dictionary for the Coherence Model for BERTopic.

2.3.3 Large Language Models

We further leveraged different-sized LLMs[6] for place functions and keyword generation with explanations from the Wikipedia summaries. We omitted the OSM POIs at this stage as the bag of words from OSM tags will not be useful for topic label extraction compared to the article summaries from Wikipedia. We adopted the KeyLLM function of the Keybert library introduced by Grootendorst (2023)[7]. Although primarily developed for keyword extraction, it can be leveraged to generate place functions.

For this part of the analysis, we tested the tiny model 'Phi-3-mini-128k-instruct' from Microsoft (Abdin et al., 2024), two small models, Mistral-7B (Jiang et al., 2023) and Llama3.1-8B, and the more recent Llama3.2-3B (Grattafiori, 2024), which sits somewhere in between. Medium and large-sized models (usually ranging in the 70B and 400B+ categories, respectively) have been excluded from this study under the assumption that their computationally expensive text-generation abilities would not be necessary for the task.

The LLMs were deployed locally using the pipeline function from the Hugging Face transformers library[8]. The prompts were structured based on the templates shown in Table 1. Each prompt includes two parts: the System prompt and the User prompt. The system prompt clearly and precisely states the role of the LLMs, providing a clear direction regarding the expected result. The user prompt states the specific task given to the particular model (e.g., a specific urban place). We adopted a 'one-shot' approach, providing an example of the expected output when an input document is given. It depicts the format of the resultant text once the model is run on the text datasets.

2.3.4 Combined dataset analysis

We combined the OSM and Wikipedia data, assuming it would generate a broader range of functions by extracting the uniqueness and different perspectives these two datasets capture about place. We begin the combination method by applying the DBSCAN clustering method, using a minimum of four samples and an epsilon value of 50 m to create clusters.

Once a set of places combining OSM POIs and Wikipedia geolocated pages is established, we created a table summarising the urban functions linked to these new combined places. Each combined place aggregates urban functions originally associated with its constituent OSM-based places and Wikipedia-based places. For each previously computed OSM-based place whose POIs are part of a new combined place, the OSM-based urban function is counted for the new combined place, and the same procedure is followed for the Wikipedia-based places. Therefore, a new table is generated, including one row for each new combined place and one column for each OSM-based and Wikipedia-based urban function. The value of each cell expresses how common an urban function is in each new combined place. represented by the frequency of each functional topic present within it.

The resultant merged dataset has high dimensions. Hence, we performed a Principal Component Analysis (Hotelling, 1933) to reduce its dimensionality. Finally, we performed *k-means* clustering on the merged dataset to group places with similar functional topic distributions. Before running *k-means* clustering, we experimented with the 'within-cluster sum of squared distances' for a range of cluster sizes, varying from 2 to 20. We obtained the optimal *k*-value of 12, thus creating 12 k-means clusters. Lastly, we manually annotated the place function label for each k-means cluster.

Finally, we qualitatively evaluated the place functions generated by the classic topic models (LDA and BTM), neural topic model (BERTopic), LLMs, and our combination approach.

3 Results

3.1 Spatial Clustering

In Kensington and Chelsea, the spatial clustering process identified 100 urban places (i.e., clusters) based on the OSM POIs, leaving 367 POIs classified as noise. The analysis based on the geolocated Wikipedia articles identified 16 urban places, leaving 95 articles as noise. Finally, combining both datasets, the spatial clustering

[6] Hugging Face Model Hub for LLM models – Available at - https://huggingface.co/models. Accessed on October 2024

[7] KeyLLM - https://maartengr.github.io/KeyBERT/guides/keyllm.html

[8] Available at - https://huggingface.co/docs/transformers/en/main_classes/pipelines

Table 1. The Prompt for text generation from a Wikipedia Article summary of the Kensington and Chelsea borough using the Mistral-7B model

System Prompt
"""
<s> [INST] <<SYS>>
You are a helpful assistant capable of generating keywords and topic labels for documents.
Your task is to:
1. Generate a topic label that best describes the topic of the document based on its content place type and place function.
2. Return the topic label starting with '<l>' and ending in '</l>'
3. Give reasons for choosing why the topic label best describes the document, start by saying, 'Let's think step by step'.
4. In the explanation of the reasons give a list of keywords that best describes the reasons for choosing the keywords. Give the list of keywords separated by commas starting with '<k>' and ending in '</k>'
5. Exclude place names like London, England, Kensington and Chelsea as the Topic label and as keywords.
6. Exclude any person's name as a topic label and keywords
[/INST] <</SYS>>
"""
Example Prompt
"""
<s>[INST]
I have the following document:
'Bao is a Korean restaurant in Chelsea, London. From 2000 to 2010, it was awarded a Michelin star, making it, along with the Gyozo in London, the first Korean restaurant to be awarded stars. 'Kenzi is a hotel in the centre of Kensington borough in London. It is an old hotel in the area and has been popular to tourist for a long time'. 'The Tate is a modern art museum in Kensington borough. It has a display of numerous contemporary art. 'The Glory is an indoor children's park. It acts as a centre of attraction for many in the area. The ticket price to enter is £5 per person.'
Give me the topic label that best describes the place type and function mentioned in these documents.
[/INST]
The topic label for the document is <l> 'Tourist Area' </l> Let's think step by step, the documents are about place types visited by tourists, for example, the keywords for the documents are, <k> Restaurants, Hotels, Museums </k>.
</s>"""
Keyword Prompt
"""[INST]
I have the following document:
- [DOCUMENT]
Give me the topic label that best describes the place type and function mentioned in these documents.
[/INST]
"""

process identified 139 urban places, leaving 433 objects as noise. The choice of distance and minimum points to form a DBSCAN cluster in each case has allowed to capture the varying number of places occupying the case study area, keeping the number of noises to a minimum.

3.2 Topic Modelling

The LDA method struggles to adequately distinguish the functions of different urban places when applied to OSM due to the frequent occurrence of certain OSM POIs (e.g., trees, benches, restaurants, bicycle parking) throughout the borough. Such a distribution renders the method ineffective when working with the sequences of tags we constructed from the set of POIs in each urban place. Since the top two to five most probable terms comprising each function are similar, we also considered the top sixth to tenth terms, which appear unique to each function representing place.

The BTM captures a few distinct functions from the OSM POI tags. The models show the occurrence of Local Shopping Streets and High-Street shops. The model also identifies the presence of functions unique to the case study area, i.e., a Shopping Street of antique elements and Diplomatic Missions.

Twelve functions are returned by running BERTopic on OSM POIs with the selected parameters. Similar to the results obtained from LDA modelling, a few POI tags, such as bench, bicycle parking, and tress, are repeatedly found in most functions. As a result, the functions can't be easily distinguished.

On the other hand, the LDA output in Wikipedia summaries is also unsatisfactory. It shows mostly place names, such as London, Chelsea, and Kensington, as the most frequent terms for each function. These occurrences of place names instead of place types make it difficult to identify distinct place functions. Among the six topics generated, the only two distinguishable functions are the presence of Notable Roads, Royal Houses, Garden, and Tourism.

Running BTM on Wikipedia article summaries yields better results than LDA, which returns six distinct functions. BTM reveals that an underlying theme emerges, relating to the presence of notable roads and streets and tourist attractions such as museums and exhibitions alongside diplomatic missions, which is also correctly identified as a distinct characteristic. Other functions focus on areas with educational institutions. The BERTopic modelling returns distinguishable functions to describe the article summaries for Wikipedia articles. The functions include Renowned Roads and Streets, Museums and Exhibitions, Eateries and Boutique shopping.

Table 2 presents the topic coherence measures to identify the best Topic Model approach. For both Cv and UCI metrics, the higher value represents better coherence. For UMass with negative results, the values closer to zero equate to higher coherence. From the results presented in Table 2, we can interpret that between LDA and BERTopic, the latter shows the best scores across almost all the metrics (except for UCI on Wikipedia) for both the OSM and the Wikipedia datasets. The analysis shows that for the LDA and BERTopic models, the performance across all the metrics (except for UCI) is higher for the Wikipedia article summaries. Thus, in our later analysis stage, we consider BERTopic a suitable method to extract urban functions from the Wikipedia dataset. A similar quantitative evaluation of topic coherence could not be achieved for the BTM model. However, a qualitative comparison of function labels between LDA, BERTopic, and BTM for OSM POIs shows that the BTM method returns more interpretable functions. The shorter texts of the OSM tags probably enable BTM to produce better results.

Table 2. Topic Coherence Metrics for different Topic Models

Topic Model	***OSM POIs***			***Wikipedia***		
	C_V	***UMass***	***UCI***	***C_V***	***UMass***	***UCI***
LDA	0.184	- 0.881	- 4.258	0.338	- 0.454	- 1.402
BERTopic	0.498	- 0.518	- 3.229	0.636	-0.357	-3.517

3.3 Place Function Labels from Large Language Models

Fig. 2 illustrates the results of extracting functional labels through LLMs (Mistral, Llama, and Ph-mini models). The labels in the Figure demonstrate that the Mistral-7B model identifies the study area as comprising Arts & Cultural Hub, Historic Buildings and Neighbourhoods, Upscale Residential and Diplomatic areas, Shopping and Dining District, and others. The function labels obtained from the Llama 3.1-8B and Llama 3.2-3B (Fig. 2) models are very similar. Both the Llama models capture a diverse range of urban functions in the study area, although the Llama 3.1-8B model identifies more functions than the smaller Llama 3.2-3B model. Similar to the Mistral-7B model, they both capture the presence of residential areas, diplomatic missions, shopping districts, royal residences, and tourist destinations. The larger Llama3.1-8B model successfully identifies commercial areas, leisure and cultural districts, and the shopping and entertainment district for which the Kensington and Chelsea borough is renowned.

The Phi-mini model, on the other hand, identifies the case study area as a mixture of urban place functions (Fig. 2), which includes areas of cultural significance with landmarks of historical importance and heritage. The presence of residential areas as both royal and high-end residences and diplomatic missions is also captured. Identifying such a diverse range of topics makes the model no less effective than the larger-sized models mentioned above.

3.4 Integration of methods and datasets

For this part of the analysis, we focused on the 139 urban places identified by combining the OSM and Wikipedia datasets, applying the best-performing approach to each component separately. Therefore, we applied BTM to the OSM component of each of the 139 combined urban places, identifying 7 functions from 10 topics, and BERTopic to the Wikipedia component, which also returned 7 functions from 10 topics. Figures 3 and 4 illustrate the topic-term distribution of BTM on OSM POIs and BERTopic on Wikipedia article summaries, respectively.

The combination of the BTM and BERTopic model is illustrated through a heatmap in Figure 5. The heatmap represents the frequency of occurrence of each function in each *k-means* cluster. A value of 0 represents absence of the function in the cluster, whereas 1 mean the function is most frequent in that cluster. From the heatmap we manually annotated each k-means cluster with a function label based on the frequency of the BTM and BERTopic functions present in it. Representing functions through the combination of methods effectively extracts a variety of

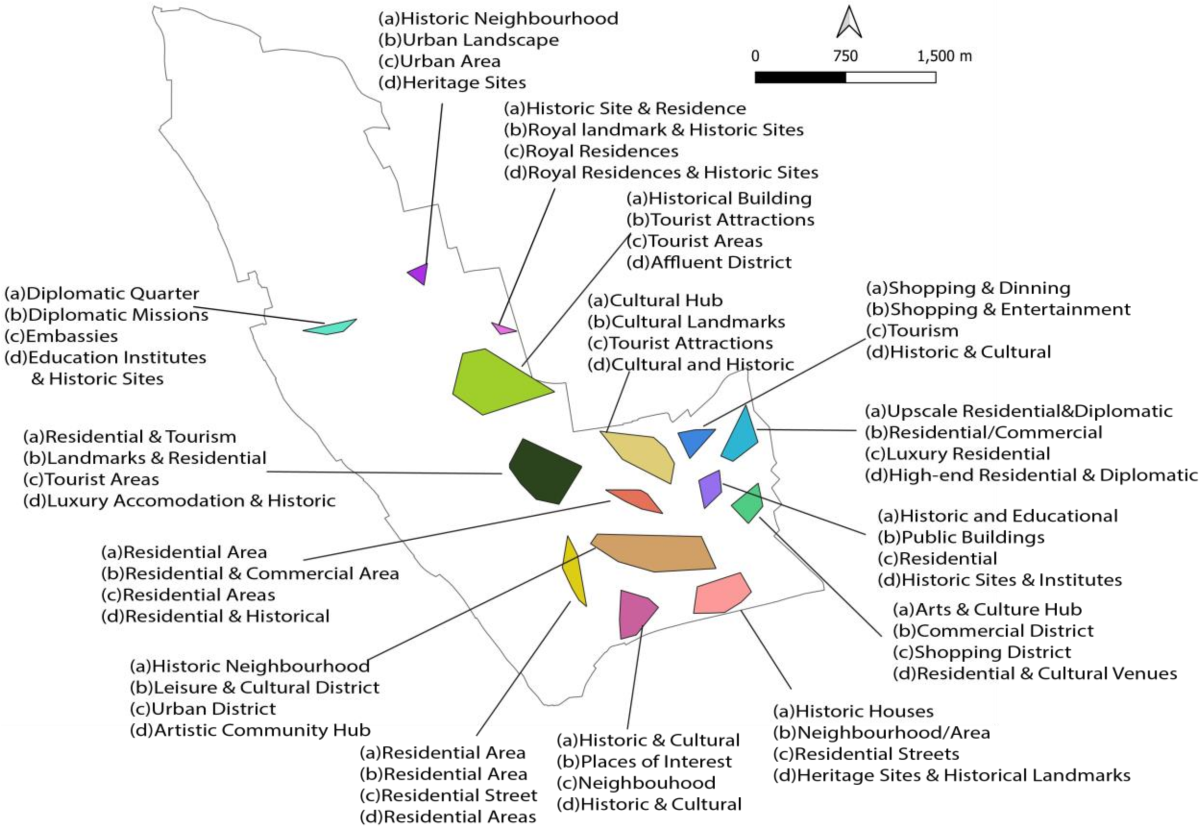

Figure 2. Function Labels from the Large Language Models (a) Mistral-7B, (b) Llama3.1-8B, (c) Llama3.2-3B, and (d) Phi-mini on Wikipedia Article Summaries for Kensington and Chelsea. The figure contains public sector information licensed under the Open Government Licence v2.0. Also contains: National Statistics data © Crown copyright and database right 2025 and Ordnance Survey data © Crown copyright and database right 2025, retrieved through the London Datastore

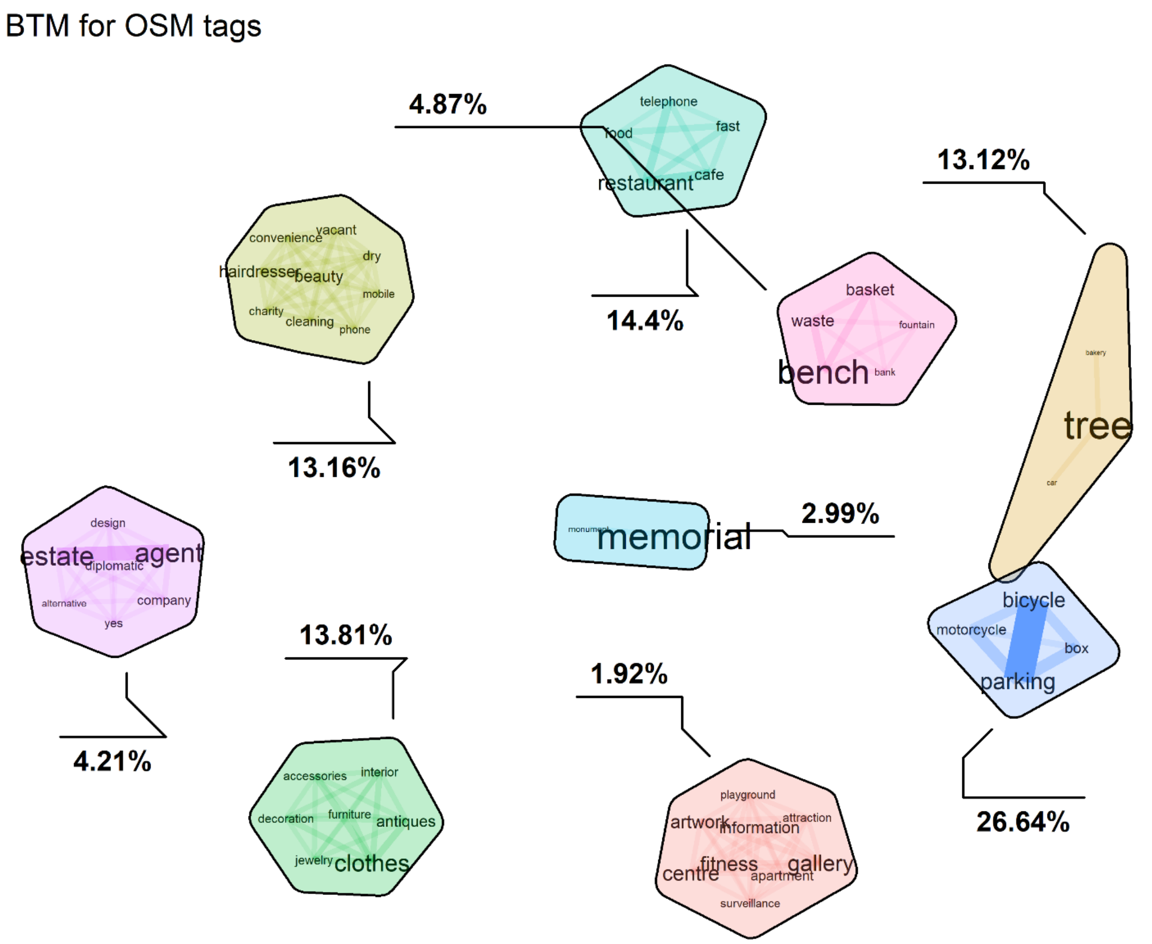

Figure 3. The topic-term distribution of topics generated by the Biterm Topic Model on the OSM POI tags

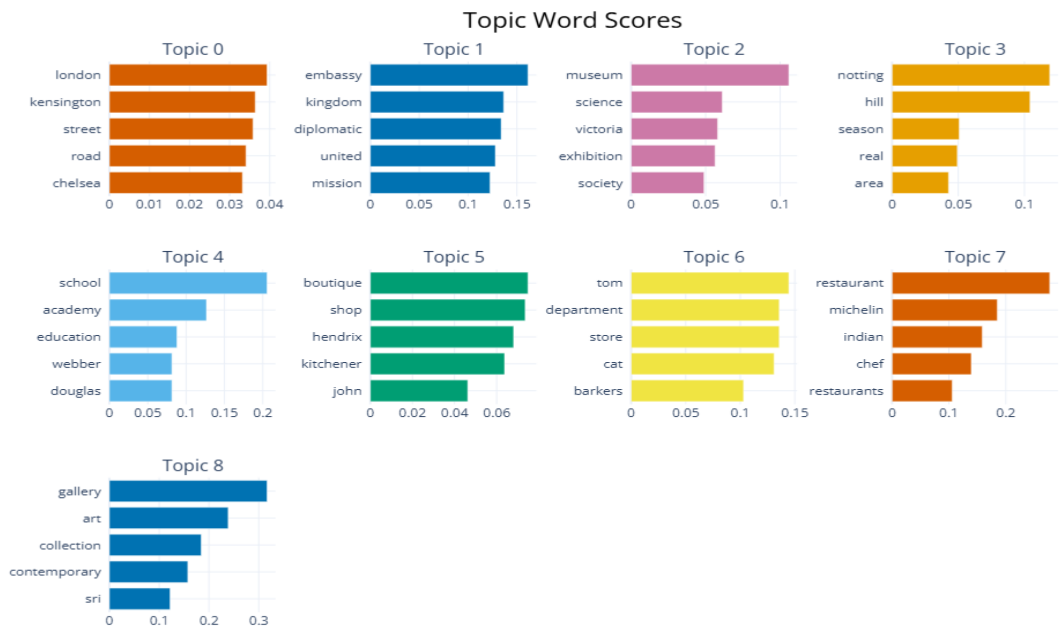

Figure 4. The topic-term distribution per topics from BERTopic Model on Wikipedia Article Summaries

functions. The spatial distribution of these functions, as represented in Figure 6, identifies the Boutiques and Fashion Street and the presence of a Tourist area in the Southern part of the study area. Urban parks and Green spaces are observed in the central part of the study area. A Diplomatic Zone, and a congregation of Educational institutes and Healthcare facilities are also noticeable. Most other places distributed throughout the study area have a mixed functional signature.

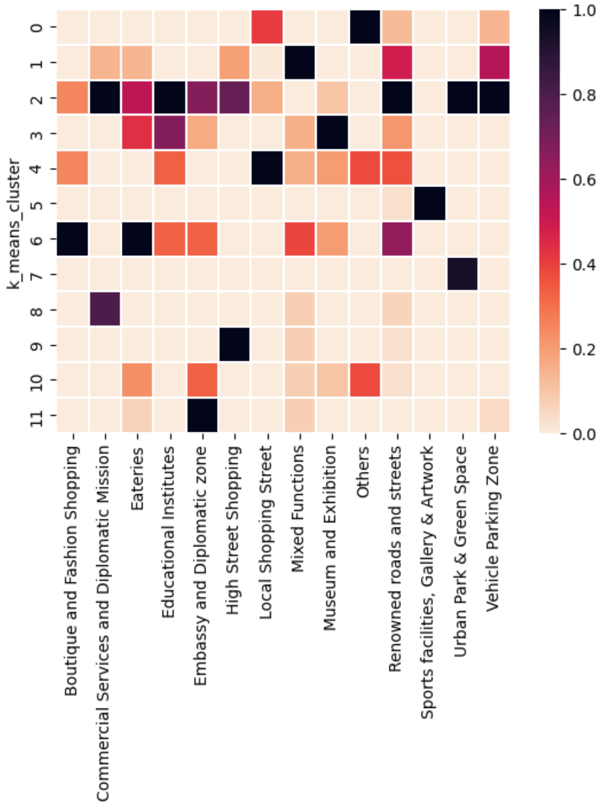

Figure 5. Heatmap of *k-means* clustering of integrated Biterm Topic Model on integration from OpenStreetMap POIs and BERTopic on Wikipedia Article Summaries

Each Topic modelling approach we have examined so far broadly captures the major urban functions prevailing in the study area. The function labels provide an insight into the places that occupy our case study area. However, upon closer examination of the models, we can identify differences in how each model categorizes urban functions for the same place. In this context, we compared the approaches qualitatively by examining the results for a specific place (a single DBSCAN spatial cluster).

We chose a place in between South Kensington and Chelsea, where multiple museums, galleries, hotels, and restaurants exist. Our analysis reveals that this location is designated as a Tourism function by both LDA and BTM on Wikipedia article summarises (Table 3). Our combination approach also identifies the place's function as a tourism site, owing to numerous museums, exhibition centres, restaurants, and cafés. Among the LLMs, the Mistral-7B, Llama 3.1 and the Phi-mini model identify the cluster as Cultural Hub/Landmark/Institutions. Only Llama3.2 labels it as Tourist Attractions. Thus, most models we experimented with concur with a similar label for the place function.

Table 3. The Topic Models and LLMs' output for a particular place in Kensington and Chelsea borough

Model Name	**Topic Label**
LDA	Tourism
BTM	Tourism
BERTopic	Eateries
Integration of OSM BTM & Wiki BERTopic	Tourism
Mistral-7B	Cultural Hub
Llama-3.1	Cultural Landmarks
Llama-3.2	Tourist Attractions
Phi-mini-128K	Cultural Institutions & Historical Preservation

4. Discussion & Conclusion

Our study provided a methodological framework for studying urban place functions from text-based UGC utilising the advances in NLP methods. Our work stemmed from the idea that the textual content in UGC contains more depth of information than just a place's location, which can be efficiently extracted with the robust methodology that we suggest here. Our study's contribution is discussed below.

First, we have shown how we can combine outputs of topic modelling on two different datasets (OSM and Wikipedia) to extract information on urban place function, which is of immense importance in urban sustainable development. The OSM dataset captures all the objects that comprise an area, thus providing functional signatures concerning everyday activities. On the other hand, Wikipedia presents articles on renowned landmarks and events, thereby capturing more notable places and functions within the area. Our integration method demonstrates that we can gain further insight into a place function by integrating functions related to everyday activities from OSM POIs with functions referring to remarkable events and places from Wikipedia

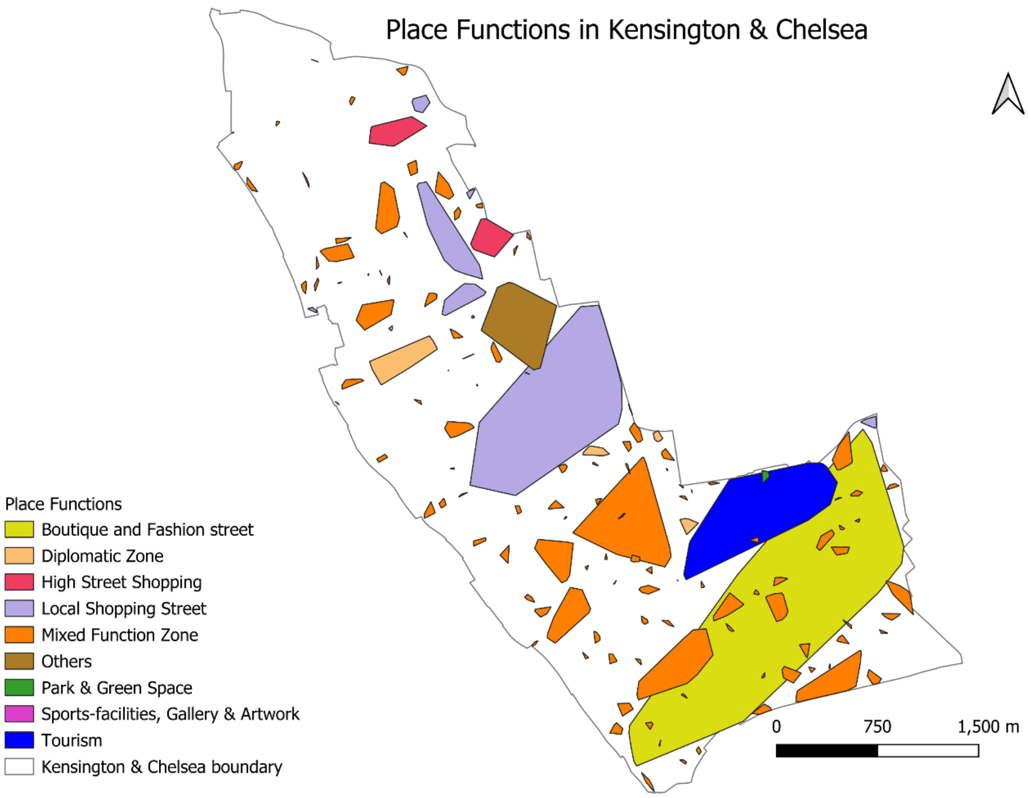

Figure 6. The spatial distribution of Urban Place functions identified by integration of OpenStreetMap BTM topics and Wikipedia BERTopic Model. The figure contains public sector information licensed under the Open Government Licence v2.0. Also contains: National Statistics data © Crown copyright and database right 2025 and Ordnance Survey data © Crown copyright and database right 2025, retrieved through the London Datastore

article summaries. The method also shows that combining datasets improves the coverage of place function labelling. In contrast, the Wikipedia article summaries only highlight the southern part of the case study area.

Second, we have shown how to effectively leverage the existing NLP techniques for extracting place-based information from unstructured texts. The efficiency of the NLP techniques varies with the datasets. We show how integrating the best-performing methods for each dataset can gain further insight into place functions from the texts. Combining BTM for short texts from OSM and BERTopic for longer texts from Wikipedia summaries has returned useful place function information.

Third, the LLM-based function generation returns meaningful outputs. However, the functions returned by our integration method are of equal quality to those of the LLMs. In contrast, the computational power required for running the LLM might hinder its widespread usage. However, we can utilise the power of the LLMS to reduce the manual work in labelling a function from the most probable terms.

In future work, we plan to compare our findings with existing indexes, such as those used in Great Britain (Arribas-Bel & Fleischmann, 2022) and land use maps of the study area. We also plan to apply the framework to different case study areas to see how it captures the urban functional variations between them. Finally, we aim to examine the framework's scalability by applying it to a larger-scale area.

Data and Software Availability

Data are collected from the sources cited in the footnotes in Section 2.1 of the paper. The codes for the data collection process and their analysis are made available at https://github.com/smnawfee/Exploring-Urban-Place-Function-through-User-Generated-Textual-Content.git

Acknowledgements

The authors would like to express gratitude for a College of Science and Engineering studentship from the University of Leicester funding the first author's Ph.D. They would also like to acknowledge the ALICE High Performance Computing facility at the University of Leicester for conducting the LLM-based analysis. The contributors of the OpenStreetMap and Wikipedia platforms are thanked for the data as are the Agile Scientific reviewers for valuable feedback in preparing this paper.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., et al., Phi-3 Technical Report: A Highly Capable Language Model Locally On Your Phone, arXiv,. https://doi.org/10.48550/arXiv.2404.14219, 2024.

Adams, B., and Janowicz, K.: Thematic signatures for cleansing and enriching place-related linked data. International Journal of Geographical Information Science, 29(4), 556-579. https://doi.org/10.1080/13658816.2014.989855, 2015.

Adams, B., and McKenzie, G.: Inferring thematic places from spatially referenced natural language descriptions, in: Crowdsourcing Geographic Knowledge edited by Sui, D., Elwood, S., and Goodchild, M., pp. 201-221). https://doi.org/10.1007/978-94-007-4587-2_12, 2020.

Arribas-Bel, D., and Fleischmann, M.: Spatial signatures - understanding (urban) spaces through form and function. Habitat International, 128, 102641. https://doi.org/10.1016/j.habitatint.2022.102641, 2022.

Athukorala, S., and Mohotti, W.: An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter. Soc Netw Anal Min, 12(1), 89. https://doi.org/10.1007/s13278-022-00898-5, 2022.

Bahrehdar, A. R., Adams, B., and Purves, R. S.: Streets of london: using Flickr and OpenStreetMap to build an interactive image of the city. Computers, Environment and Urban Systems, 84. https://doi.org/10.1016/j.compenvurbsys.2020.101524, 2020.

Ballatore, A., and De Sabbata, S.. Los Angeles as a digital place: The geographies of user-generated content. Transactions in GIS, 24, 888-902. https://doi.org/10.1111/tgis.12600, 2019.

Blei, D. M., Ng, A. Y., and Jordan, M. L.: Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022, 2003.

Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., & Singleton, A.: A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. Computers, Environment and Urban Systems, 85. https://doi.org/10.1016/j.compenvurbsys.2020.101539, 2021.

Campello, R. J. G. B., Moulavi, D., and Sander, J.: Density-based clustering based on hierarchical density estimates, Advances in Knowledge Discovery and Data Mining. PAKDD 2013. https://doi.org/10.1007/978-3-642-37456-2_14, 2013.

Chapple, K., Poorthuis, A., Zook, M., and Phillips, E.: Monitoring streets through tweets: Using user-generated geographic information to predict gentrification and displacement. Environment and Planning B: Urban Analytics and City Science, 49(2), 704-721. https://doi.org/10.1177/23998083211025309, 2021.

Chen, M., Arribas-Bel, D., and Singleton, A.: Understanding the dynamics of urban areas of interest through volunteered geographic information. Journal of Geographical Systems, 21(1), 89-109. https://doi.org/10.1007/s10109-018-0284-3, 2018.

Comber, A., Kieu, M., Bui, Q.-T., and Malleson, N.: Using social media data to identify neighbourhood change. AGILE: GIScience Series, 5, 1-6. https://doi.org/10.5194/agile-giss-5-20-2024, 2024.

Cresswell, T.: Place a short introduction. Blackwell Publishing Ltd, 2004.

De Sabbata, S., Bennett, K., and Gardner, Z. Towards a study of everyday geographic information: Bringing the everyday into view. Environment and Planning B: Urban Analytics and City Science. https://doi.org/10.1177/23998083231217606, 2023.

Du, Z., Sui, H., and Wang, J.: A novel semantic recognition framework of urban functional zones supporting urban land structure analytics based on open-source data. Transactions in GIS, 25(3), 1460-1484. https://doi.org/10.1111/tgis.12736, 2021.

Gao, S., Janowicz, K., and Couclelis, H.: Extracting urban functional regions from points of interest and human activities on location-based social networks. Transactions in GIS, 21(3), 446-467. https://doi.org/10.1111/tgis.12289, 2017.

Goodchild, M.: Formalizing Place in Geographic Information Systems, in: Communities, Neighborhoods, and Health Expanding the Boundaries of Place, edited by Burton, L. M.,. Kemp, S. P., Leung, M., Matthews, S. A., and Takeuchi D. T., pp. 21-34. Springer. https://doi.org/10.1007/978-1-4419-7482-2, 2011

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al., The Llama 3 herd of models, arXiv,. https://doi.org/10.48550/ARXIV.2407.21783, 2024

Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arxiv.2203.05794, 2022.

Hu, S., Xu, Y., Wu, L., Wu, X., Wang, R., Zhang, Z., Lu, R., and Mao, W.: A framework to detect and understand thematic places of a city using geospatial data. Cities, 109. https://doi.org/10.1016/j.cities.2020.103012, 2021.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., and Prasad, S.: Extracting and understanding urban areas of interest using geotagged photos. Computers, Environment and Urban Systems, 54, 240-254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001, 2015.

Jain, S., Proserpio, D., Quattrone, G., and Quercia, D.: Nowcasting Gentrification using Airbnb data. Proceedings of the ACM on human-computer interaction, 5(CSCW1), 1-21. https://doi.org/10.1145/3449112, 2021.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand,F., et al., Mistral 7B. arXiv, https://doi.org/10.48550/arXiv.2310.06825, 2023.

Lin, X., Yang, T., and Law, S.: From points to patterns: An explorative POI network study on urban functional distribution. Computers, Environment and Urban Systems, 117. https://doi.org/10.1016/j.compenvurbsys.2024.102246, 2025.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A. Ammar Ahmad, A., Bach, N.,Bahree, A., et al., Phi-3 technical report: a highly capable language model locally on your phone. arXiv,. https://doi.org/10.48550/arXiv.2404.14219, 2024

McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arxiv.1802.03426, 2018.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A.:. Optimizing Semantic Coherence in Topic Models. Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. July 27–31, 2011, 2011.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T.: Automatic Evaluation of Topic Coherence Annual Conference of the North American Chapter of the ACL, 2010.

Niu, H., and Silva, E. A.: Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. Computers, Environment and Urban systems, *88*. https://doi.org/10.1016/j.compenvurbsys.2021.101651, 2021.

Oh, D., and Yao, X. A.: Assessing place type similarities based on functional signatures extracted from social media data. ISPRS International Journal of Geo-Information, 10(9). https://doi.org/10.3390/ijgi10090626, 2021.

Purves, R., and Hollenstein, L. Exploring place through user-generated content: Using Flickr to describe city cores. Journal of Spatial Information Science (1), https://doi.org/10.5311/josis.2010.1.3, 2010.

Rahmah, N., and Sitanggang, I. S.: Determination of Optimal Epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra. IOP Conference Series: Earth and Environmental Science, 31. https://doi.org/10.1088/1755-1315/31/1/012012, 2016.

Reimers, N., and Gurevych, I., Sentence-BERT: sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084. Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.48550/arXiv.1908.10084, 2019.

Röder, M., Both, A., and Hinneburg, A.: Exploring the space of topic coherence measures, Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015.

Tang, V., and Painho, M.: Content-location relationships: a framework to explore correlations between space-based and place-based user-generated content. International Journal of Geographical Information Science, 37(8), 1840-1871. https://doi.org/10.1080/13658816.2023.2213869, 2023.

United Nations: Development of Economic and Social Affairs. Sustainable Development. Available at: https://sdgs.un.org/goal Accessed on 9th April, 2025, 2024.

World Bank Group: Urban Development. Available at: https://www.worldbank.org/en/topic/urbandevelopment/overview, Accessed on 9th April, 2025, 2023.

Wijffels, J.: BTM: Biterm Topic Models for Short Text. R package version 0.3.7. https://cran.r-project.org/web/packages/BTM., 2019.

Yan, X., Guo, J., Lan, Y., and Cheng, X.:. A biterm topic model for short texts, 22nd International Conference on World Wide Web, New York, NY, USA., 2013.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., and Mai, K.: Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. International Journal of Geographical Information Science, 31(4), 825-848. https://doi.org/10.1080/13658816.2016.1244608, 2016.

Zhang, X., Li, W., Zhang, F., Liu, R., and Du, Z.: Identifying urban functional zones using public bicycle rental records and point-of-interest data. ISPRS International Journal of Geo-Information, 7(12). https://doi.org/10.3390/ijgi7120459, 2018.