Automated High Fidelity Functional Map Generation using Text Data Clustering

## names, email, organization, corresponding mail and address

# Abstract

This study presents a novel methodology for automating the generation of high-fidelity functional maps using text-based clustering of OpenStreetMap data. Functional maps, which visualize the distribution of residential, commercial, industrial, and natural zones within regions, traditionally require extensive manual effort through field surveys, data compilation, and cartographic work. While Geographic Information Systems (GIS) have streamlined this process, significant manual verification is still needed for accuracy.

We propose an automated framework that processes OpenStreetMap text data through Natural Language Processing (NLP) techniques to classify regional land use. Our methodology divides target regions into 1km² tiles and analyses the associated map text data (including building names, shop names, and point-of-interest descriptions) using the Universal Sentence Encoder (USE) for text embedding. These embeddings are then clustered using the K-means algorithm to identify distinct functional zones.

To validate our approach, we applied this framework to a 2,500 km² area within the Mumbai Metropolitan Region. The region was first manually labelled to establish ground truth data, against which we compared our automated classifications. Our results demonstrate that this methodology can effectively generate functional maps while significantly reducing manual effort. The framework's scalability makes it particularly valuable for mapping large urban areas, achieving promising accuracy, precision, recall, and F1 scores in distinguishing between residential, commercial, industrial, and natural zones.

**Keywords**: functional maps, text embedding, K-means clustering, OpenStreetMap, urban planning, automated mapping

# Introduction

# Material and Methods

## 1. OpenStreetMap Data Collection

This study utilizes OpenStreetMap (OSM) as its primary data source. OSM is a collaborative mapping platform that provides comprehensive geographic data through community contributions, functioning similarly to Wikipedia for geographic information. The platform offers detailed spatial data including:

1. Building information and classifications
2. Commercial establishments and points of interest
3. Road networks and transportation infrastructure
4. Land use designations
5. Natural features and boundaries

The data is freely accessible through the OSM API, which provides structured information in a standardized format. OSM's data quality is maintained through community verification processes, making it particularly reliable in densely populated urban areas where contributor activity is high.

## 2. Spatial Grid Generation and Region Partitioning

To systematically analyse large geographic areas, we developed a grid-based partitioning approach:

**Grid Definition:** The target region is overlaid with a uniform grid system, where each cell represents a 1km × 1km area. This granularity was chosen to:

1. Capture sufficient detail for meaningful functional analysis
2. Maintain computational efficiency

**Data Association:**

1. OSM features falling within each tile's boundaries are extracted
2. Spatial indices are created to optimize the feature-to-tile mapping process
3. Each tile accumulates all relevant text data from its contained features

This structured approach to data collection and spatial partitioning provides the foundation for subsequent text processing and clustering analyses. The uniform grid system ensures consistent spatial resolution across the study area while facilitating scalable processing of large geographic regions.

## 3. Data Preprocessing and Text Analysis

The raw data extracted from the OpenStreetMap API underwent a comprehensive filtering and preprocessing pipeline to ensure data quality and relevance. Initial filtering was performed using predefined OSM database filters to retain only pertinent geographic features, including buildings, commercial establishments, offices, transportation infrastructure, and recreational areas. This selective approach helped maintain focus on features that contribute meaningfully to functional zone classification while reducing noise in the dataset.

Following the initial feature selection, text data from each 1km² tile was aggregated to create consolidated text chunks representing the geographic characteristics of each region. The consolidation process preserved essential geographic nomenclature while maintaining the spatial relationships between features within each tile. To ensure data quality and identify potential anomalies, we conducted extensive exploratory data analysis (EDA) using statistical methods. Box plots were employed to analyse the quartile distribution of text lengths across tiles, enabling the identification of outliers and unusual patterns in the data distribution.

Regions lacking text data were subjected to additional verification processes. These void areas were systematically revalidated through cross-referencing with satellite imagery and existing land use data, leading to the identification of natural features such as mangroves and water bodies. This verification process helped distinguish between actual data gaps and legitimate natural areas, improving the overall accuracy of our classification framework.

The text preprocessing pipeline incorporated the Natural Language Toolkit (NLTK) for comprehensive text cleaning and normalization. This process included the removal of stop-words, lemmatization of terms, and standardization of text format. The lemmatization process was particularly crucial as it reduced inflected words to their base form, ensuring consistent representation of similar features across different tiles. Additionally, we analysed the binned distribution of text lengths to establish appropriate thresholds for outlier exclusion, ensuring that the final dataset maintained a balance between comprehensive coverage and data quality.

Through this systematic approach to data filtering and preprocessing, we established a robust foundation for subsequent embedding and clustering analyses. The careful attention to data quality and feature relevance during this stage significantly contributed to the effectiveness of our functional zone classification methodology.

## 4. Case Study Implementation

To validate our methodological framework, we selected the Mumbai Metropolitan Region (MMR) as our primary study area. This region presents an ideal test case due to its diverse urban landscape, encompassing a rich mixture of land use patterns across a substantial geographic area of 2,500 square kilometres.

### 4.1. Study Area Selection and Characteristics

The MMR serves as an exemplary urban testing ground for our framework due to several key characteristics. The region features a complex tapestry of land use, including high-density commercial districts, extensive residential developments, established industrial zones, and significant natural features such as the Arabian Sea coastline, creeks, and mangrove forests. This diversity provides an optimal environment for testing our classification methodology across various functional zones.

The study area was defined as a 50km × 50km square region, cantered on the metropolitan core. This delineation was carefully chosen to capture the full spectrum of urban development patterns, from the dense urban core to peripheral areas experiencing rapid transformation. The selected region also includes various stages of urban development, from historical neighbourhoods to emerging commercial corridors and industrial estates.

### 4.2. Implementation Framework

Following our established methodology, we partitioned the study area into 1km × 1km tiles, generating a dataset of 2,500 distinct spatial units. This resolution was selected to:

1. Maintain sufficient granularity for meaningful functional analysis
2. Capture local variations in land use patterns
3. Enable efficient computational processing
4. Facilitate practical validation of results

During the exploratory data analysis phase, we identified and addressed several key considerations specific to the MMR context. Tiles containing no text data were subjected to additional verification, particularly in coastal areas and regions containing large natural features. This process helped distinguish between data gaps and legitimate natural areas, enhancing the accuracy of our classification system.

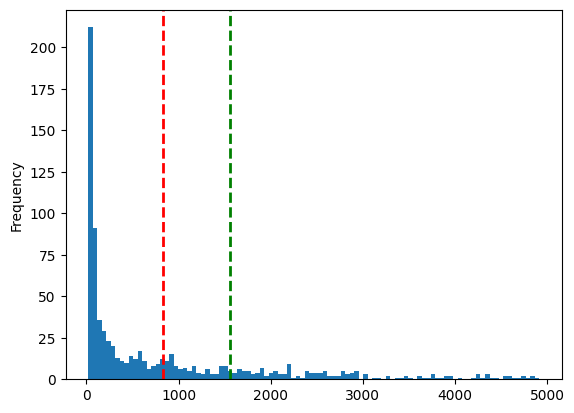
The MMR case study provided an ideal opportunity to test our framework's ability to handle complex urban environments. The region's varied development patterns, mixed land uses, and distinct natural boundaries offered appropriate challenges for validating our automated classification methodology. The results from this implementation served as the foundation for our subsequent accuracy assessments and methodology validation.

## 5. Exploratory Data Analysis

The exploratory data analysis phase revealed crucial insights about the textual characteristics and spatial distribution patterns across the Mumbai Metropolitan Region study area. Our analysis focused on understanding the distribution of text data across tiles and identifying patterns that could influence the classification process.

### Text Length Distribution Analysis

Initial analysis of text length distribution across the 2,500 tiles revealed significant variations in data density. A box plot analysis (Figure X) demonstrated that the majority of tiles (over 75%) contained between 50 and 1100 characters of pre-processed text, with a median length of approximately 192 characters and mean having 592 characters. The distribution exhibited strong positive skewness, indicating the presence of tiles with exceptionally high text content, typically corresponding to densely developed urban areas.



The quartile analysis identified several outliers, particularly in the upper range, where some tiles contained more than 1200 characters. These outliers primarily represented central business districts and major commercial zones, characterized by high concentrations of labelled buildings and points of interest. Conversely, tiles with minimal text content (below the lower quartile of 50 characters) often corresponded to natural areas or regions with limited development.

### Void Analysis and Verification

Approximately 15% of tiles contained no text data, upon further investigation, it was found that these regions mainly belonged to:

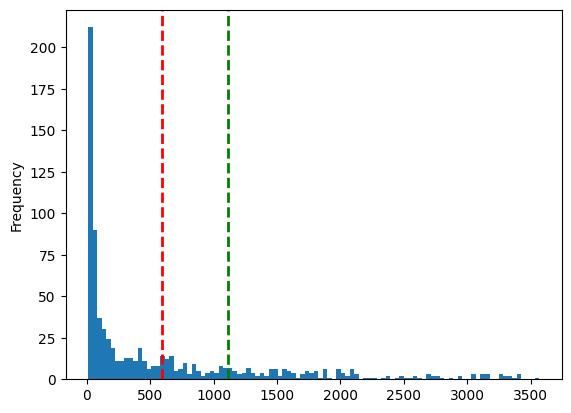
1. Water bodies (including Arabian sea and Thane/Vasai creek)
2. Protected mangrove areas
3. Undeveloped land parcels
4. Text Content Analysis

A frequency analysis of key terms and phrases across tiles provided insights into the characteristic vocabulary associated with different functional zones

1. Commercial zones showed high frequencies of terms related to retail, offices, and services
2. Residential areas were characterized by apartment complexes, housing societies, and community facilities
3. Industrial zones displayed consistent patterns of manufacturing, warehouse, and logistics-related terminology
4. Natural areas were identified through references to parks, forests, and water bodies

### Preprocessing Impact Assessment

The effect of text preprocessing steps was quantified through comparative analysis. Lemmatization reduced the unique token count by approximately 5%, while stop word removal decreased the total token count by 28%. These reductions improved the signal-to-noise ratio in the data while preserving essential semantic information for classification.



This exploratory analysis provided essential insights that informed subsequent choices in our embedding and clustering methodology, particularly in handling outliers and setting appropriate thresholds for classification.

## 6. Text Embedding Generation

Text embedding is a technique that converts words, sentences, or documents into numerical vectors (sequences of numbers) that capture their semantic meaning. These vectors allow machines to understand and compare text mathematically - similar texts will have similar vector representations. This is fundamental for many NLP applications like search, recommendation systems, and text classification.

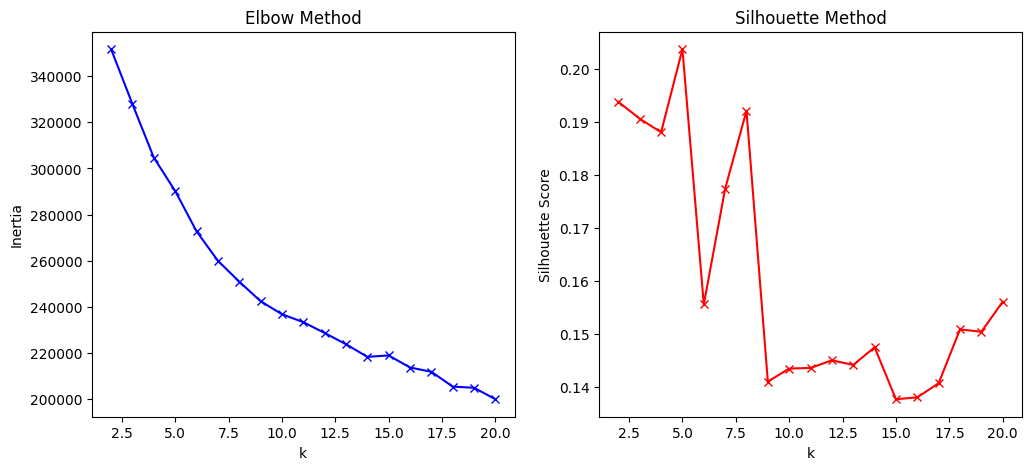
Text embeddings transform text into numerical representations to capture semantic meaning. Traditional methods like TF-IDF (Term Frequency-Inverse Document Frequency) create sparse vectors based on word frequency, but they fail to understand context. In contrast, transformer-based embeddings generate dense, context-aware representations using deep learning. Notable models include Universal Sentence Encoder (USE) by Google, which provides efficient sentence embeddings for NLP tasks, and Sentence-Transformers (all-MiniLM-L6-v2).

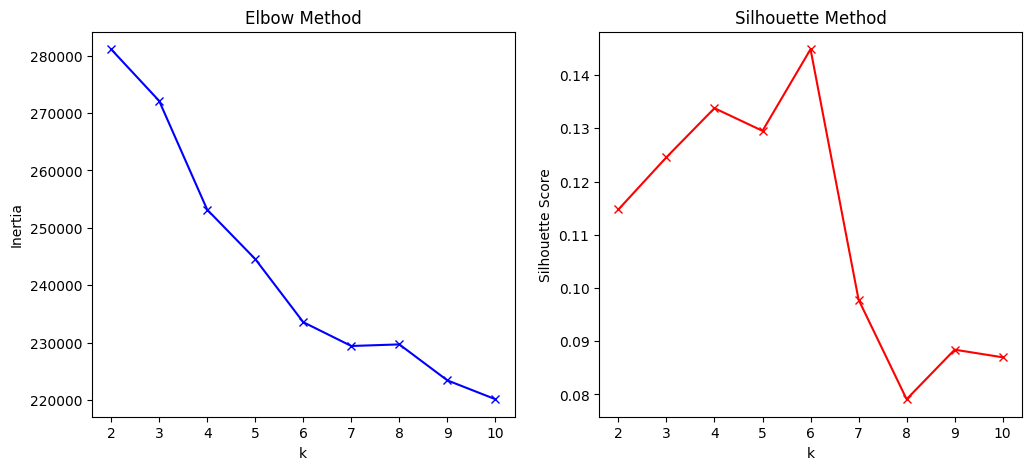
For the purpose of generating text embedding, three methods were chosen – TF-IDF, USE and sentence transformer.

## 7. Clustering

K-Means – A popular centroid-based method that partitions embeddings into M clusters by minimizing intra-cluster variance. Efficient for large datasets but assumes clusters are spherical.

The elbow point detection method was used to identify the optimal k value in K-Means clustering. The Elbow Method is a technique for determining the optimal number of clusters (K) in K-Means clustering by analysing the within-cluster sum of squares (WCSS), which measures how tightly data points are grouped within each cluster. As K increases, WCSS decreases because clusters become smaller and more refined. However, beyond a certain point, adding more clusters results in minimal improvement while increasing model complexity. By plotting WCSS against different values of K, an "elbow" shape typically appears, where the curve sharply bends before flattening out. The optimal K is chosen at this elbow point, as it represents the best balance between compact clusters and computational efficiency.





## 8. 3D Visual Representation

t-SNE (t-Distributed Stochastic Neighbour Embedding) is a dimensionality reduction technique commonly used for visualizing high-dimensional data in a lower-dimensional space (typically 2D or 3D). It preserves the local structure of data by converting pairwise similarities into probabilities, ensuring that similar points in high-dimensional space remain close in the lower-dimensional representation.

UMAP (Uniform Manifold Approximation and Projection) is a non-linear dimensionality reduction technique designed for preserving both the local and global structure of high-dimensional data while being computationally efficient. Unlike t-SNE, which focuses primarily on local neighbourhood preservation, UMAP constructs a graph-based representation of the data and optimizes a low-dimensional embedding using a probabilistic framework.

These two methods were used to visualise the embeddings and the clusters generated to evaluate the quality of the different embedding approaches.

## 9. Auto-labelling and comparison with manual coded values

The clusters generated were used to label the data points into groups. 10% samples from each cluster were manually evaluated based on their text content, to map the cluster to either comm, res, indus or nat. This method effectively generates a classification framework for unlabelled data based on clustering.

These auto generated labels were then compared with the manually coded ground truth values, to identify the accuracy and other metrics as a classification framework.

# Theory

## TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF combines local and global term importance to represent text documents as numerical vectors. For a term t in document d within corpus D:

The final TF-IDF score is:

## K-means Clustering

K-means partitions n observations into k clusters by minimizing the within-cluster sum of squares. For clusters and centroids the objective function is:

$J(C,\mu) = \sum\_{i=1}^k \sum\_{x \in C\_i} |x - \mu\_i|^2$

The algorithm iteratively:

1. Assigns points to nearest centroid: $C\_i = {x: |x-\mu\_i|^2 \leq |x-\mu\_j|^2 \space \forall j}$
2. Updates centroids: $\mu\_i = \frac{1}{|C\_i|}\sum\_{x \in C\_i} x$

## Elbow Method and Silhouette Analysis

The Elbow Method identifies optimal k by analyzing the rate of WCSS reduction:

$WCSS(k) = \sum\_{i=1}^k \sum\_{x \in C\_i} |x - \mu\_i|^2$

The optimal k is found at the "elbow" where: $\Delta WCSS(k) = WCSS(k) - WCSS(k+1)$ shows diminishing returns.

Silhouette analysis quantifies clustering quality through cohesion (a) and separation (b):

$s(i) = \frac{b(i) - a(i)}{\max{a(i), b(i)}}$

where:

* $a(i)$ = mean distance between point i and all points in its cluster
* $b(i)$ = mean distance between point i and points in nearest neighboring cluster

## t-SNE and UMAP

t-SNE converts high-dimensional affinities $p\_{ij}$ to low-dimensional similarities $q\_{ij}$:

$p\_{ij} = \frac{\exp(-|x\_i-x\_j|^2/2\sigma\_i^2)}{\sum\_{k \neq i}\exp(-|x\_i-x\_k|^2/2\sigma\_i^2)}$

$q\_{ij} = \frac{(1 + |y\_i-y\_j|^2)^{-1}}{\sum\_{k \neq l}(1 + |y\_k-y\_l|^2)^{-1}}$

Minimizing KL divergence: $KL(P|Q) = \sum\_i \sum\_j p\_{ij} \log\frac{p\_{ij}}{q\_{ij}}$

UMAP constructs a fuzzy topological representation through:

$\mu\_X(x\_i, x\_j) = \exp(-\frac{d(x\_i,x\_j) - \rho\_i}{\sigma\_i})$

where $\rho\_i$ is the distance to the nearest neighbor and $\sigma\_i$ is a local connectivity parameter. The low-dimensional representation optimizes cross entropy:

$CE = \sum\_{i,j} [\mu\_X(x\_i,x\_j)\log(\nu\_Y(y\_i,y\_j)) + (1-\mu\_X(x\_i,x\_j))\log(1-\nu\_Y(y\_i,y\_j))]$

where $\nu\_Y$ represents similarities in the low-dimensional space.

# Results

## 1. Evaluation Metrics

To assess the effectiveness of our automated functional map generation framework, we employed a comprehensive set of evaluation metrics that measure both the clustering quality and classification accuracy. The selected metrics provide complementary perspectives on the model's performance, enabling a thorough evaluation of its practical utility for urban planning applications.

1. **Confusion Matrix**

A confusion matrix is a table that visualizes the performance of a classification model by showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) - where the rows represent actual classes and columns represent predicted classes.

1. **Accuracy**

Accuracy measures the proportion of correctly classified tiles across all functional zones. While this metric provides an overall assessment of the model's performance, it was complemented with other metrics due to potential class imbalance in urban landscapes, where certain functional zones may be more prevalent than others.

1. **Precision**

Precision quantifies the proportion of correct positive predictions for each functional zone. This metric is particularly important in urban planning applications, where false positives could lead to inappropriate land-use decisions.

1. **Recall**

Recall measures the model's ability to identify all instances of a particular functional zone. This metric is crucial for ensuring comprehensive coverage of each zone type, particularly for critical areas like industrial zones where missed classifications could have significant implications.

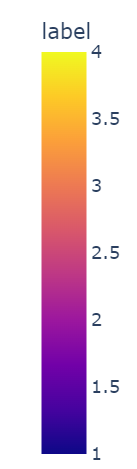
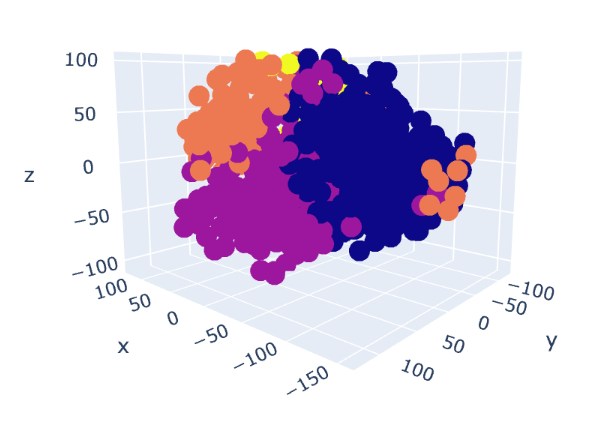
1. **F1-Score**

The F1-score provides a balanced measure of precision and recall, offering a single metric that accounts for both false positives and false negatives. This is especially relevant for our application, where both over-identification and under-identification of functional zones can impact urban planning decisions.

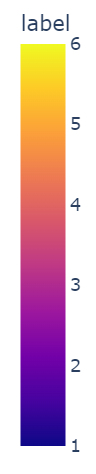
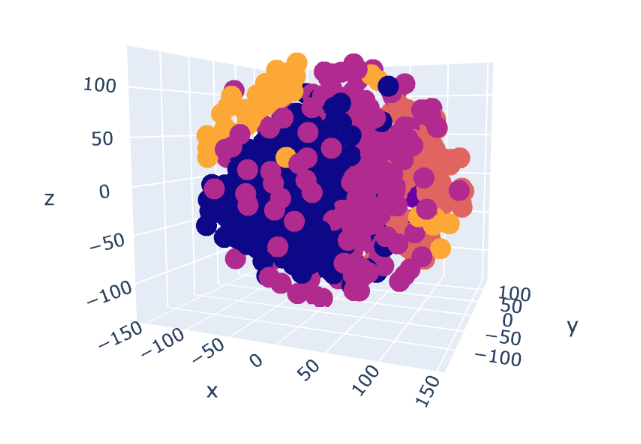
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Embedding Method | Accuracy Score | Precision | Recall | F1 Score |
| TF-IDF |  |  |  |  |
| Universal Sentence Encoder |  |  |  |  |
| Sentence Transformer |  |  |  |  |

## 2. visual cluster evaluation

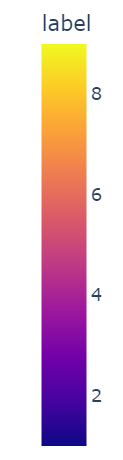
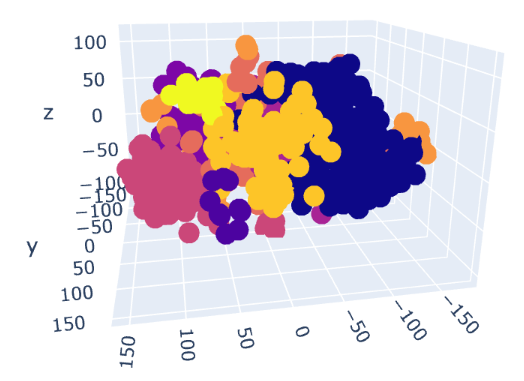
The t-SNE visualizations depict the clustering of text embeddings generated using Universal Sentence Encoder (USE) and TF-IDF features. In the USE-based projection, the clusters exhibit better-defined separations, suggesting that the semantic embeddings effectively capture contextual relationships between documents. The smooth transition between clusters indicates meaningful grouping based on text similarity, with minimal noise or overlap.



In contrast, the TF-IDF-based visualization shows a more scattered distribution, with clusters appearing less distinct and more interwoven. This suggests that TF-IDF, relying solely on word frequency statistics, struggles to capture deeper semantic connections, leading to overlapping clusters. While TF-IDF is effective for lexical similarity, the results highlight the advantage of USE embeddings in creating well-separated, semantically coherent clusters.



The t-SNE visualization for sentence-transformer-based embeddings shows well-formed clusters with noticeable separation, indicating that the model effectively captures semantic relationships. Compared to TF-IDF, sentence transformers generate dense vector representations that preserve contextual meaning, leading to more distinct groupings. While some overlap exists, the clustering pattern suggests a strong alignment with underlying similarities.



# Discussion

The t-SNE visualizations for TF-IDF, Sentence Transformers, and Universal Sentence Encoder (USE) embeddings highlight key differences in their ability to capture semantic similarity. TF-IDF relies on word frequency and lacks contextual understanding, resulting in loosely formed clusters with significant overlap. Sentence Transformers improve upon this by generating dense embeddings, capturing meaning beyond exact words, but still show some inter-cluster mixing. In contrast, USE embeddings demonstrate the best cluster separation, indicating superior contextual representation and minimal noise. For this particular use case, where clear semantic distinctions are crucial, USE embeddings provide a more structured and meaningful representation, making them the preferred choice over Sentence Transformers and TF-IDF.

# Conclusions

This study presents a novel automated approach for generating high-fidelity functional maps using text-based clustering of OpenStreetMap data. Our methodology successfully demonstrates that natural language processing techniques, particularly through the application of advanced text embeddings and clustering algorithms, can effectively classify urban spaces into distinct functional zones. The implementation in the Mumbai Metropolitan Region validates the framework's capability to process large-scale urban areas while maintaining accuracy and computational efficiency.

The key contributions of this research are multifaceted. We have developed a scalable framework for automated functional map generation using publicly available OpenStreetMap data, alongside implementing sophisticated text embedding techniques that effectively capture the semantic relationships between urban features. Through comprehensive evaluation metrics, we have validated our methodology, demonstrating its potential as a viable alternative to traditional manual mapping approaches. Furthermore, the successful application to a complex urban environment proves the framework's robustness in handling diverse land use patterns.

While our current framework shows promising results, several avenues for future research and enhancement have been identified. In particular, temporal analysis integration presents significant opportunities for advancement. This includes the development of mechanisms to track and analyse temporal changes in functional zones, implementation of time-series analysis to identify urban development patterns and trends, and the creation of predictive models for future land use changes based on historical data.

# References