Automated High Fidelity Functional Map Generation using Text Data Clustering

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

# Abstract

This study introduces an automated approach for generating high-fidelity functional maps by leveraging text-based clustering of OpenStreetMap data. Functional maps, which delineate residential, commercial, industrial, and natural zones, traditionally require extensive manual labour through surveys and GIS-based analysis. While GIS tools have improved efficiency, manual verification remains a challenge for large-scale mapping. Our proposed framework automates this process by utilizing Natural Language Processing (NLP) to classify land use based on textual data from OpenStreetMap. By segmenting regions into 1km² tiles, we extract and analyse text descriptions such as building names and points of interest. Using the Universal Sentence Encoder (USE) for embedding and K-means clustering, the system identifies distinct functional zones with minimal manual intervention. We applied this methodology to a 2,500 km² region within the Mumbai Metropolitan Area, comparing automated classifications against manually labelled ground truth data. The results demonstrate high accuracy and reliability, achieving an accuracy of 97.056 and an F1 score of 97.049. This scalable framework significantly reduces the effort required for urban mapping, making it a valuable tool for large-scale functional zoning applications while maintaining high classification performance.

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

# Introduction

The generation of accurate functional maps has emerged as a critical imperative in contemporary urban planning, environmental management, and spatial analysis. These specialized maps, which delineate distinct functional zones including residential, commercial, industrial, and natural areas within urban regions, serve as foundational tools for city planners, policymakers, and researchers. The traditional approach to functional mapping has relied on labour-intensive field surveys, manual data compilation, and extensive cartographic work, increasingly struggling to keep pace with rapidly evolving urban landscapes.

The challenges faced in functional mapping have been notably highlighted in the Functional Map of the World study (Christie et al., 2018), which developed an extensive dataset comprising 1,047,691 satellite images from 207 countries annotated with 63 different functional building categories. While their implementation of convolutional neural networks showed promise, it also exposed how baseline models struggled with architectural diversity across regions, underscoring the inherent complexity of functional mapping at scale.

Despite the revolutionary impact of Geographic Information Systems (GIS) on spatial data management and visualization, the process of generating and maintaining accurate functional maps continues to require substantial manual intervention. This human-dependent approach faces three critical challenges: First, the sheer scale and complexity of modern urban environments have made comprehensive manual surveying increasingly impractical. Second, traditional mapping approaches struggle with temporal currency, as urban functional zones undergo continuous evolution through development and shifting land use patterns. Third, the inherent subjectivity in manual classification processes can lead to inconsistencies in functional zone designation, particularly in areas characterized by mixed land use or transitional characteristics.

The emergence of OpenStreetMap (OSM) as a comprehensive, community-driven geographic data platform has created new opportunities for automated mapping approaches. Research by Kaur et al. (2017) highlighted that while positional accuracy and completeness were the most researched quality aspects, OSM data demonstrated particular promise when compared against authoritative sources. The platform provides rich, regularly updated information about urban features, including building types, commercial establishments, and points of interest.

Recent technological advances have further expanded the potential for automated functional mapping. Wan et al. (2017) achieved 87.9% classification accuracy through the extraction of OSM objects and application of morphological erosion to improve training data quality. The evolution of text embedding techniques, particularly transformer-based models like the Universal Sentence Encoder (USE), has significantly enhanced our ability to extract meaningful semantic information from textual data, while advances in clustering algorithms have improved pattern identification in high-dimensional data spaces.

Zhou et al. (2010) advanced the field with their Clustering-Based KNN Improved Algorithm (CLKNN) for text classification, addressing inherent limitations of traditional KNN classification through a dynamic adjustment mechanism. Bai et al. (2025) explored Graph Clustering Neural Networks for urban functional areas, integrating multi-source data including OpenStreetMap road networks, Points of Interest data, and nighttime light data to create a more comprehensive classification system.

Specialized applications have also emerged, with Cao et al. (2024) presenting an innovative approach to urban green space classification through BERT model integration with remote sensing data. Their study enabled fine-grained differentiation into 19 distinct functional categories. In environmental feature detection, Kasu et al. (2019) achieved impressive accuracy rates of 82.92% for water region segmentation and 80.48% for green region detection using a Modified Densitometry 3-Channel Algorithm.

Zhang et al. (2016) demonstrated the potential of Convolutional Neural Networks in classifying urban functional zones using aerial imagery, moving beyond traditional GIS-based methods to achieve superior performance in both accuracy and efficiency. These advances collectively point to the growing potential for automated approaches in urban classification and mapping.

Our research addresses these challenges by proposing a novel methodology that combines text-based clustering with advanced NLP techniques to automate the generation of high-fidelity functional maps. This approach leverages the rich textual data available in OpenStreetMap, including building names, business descriptions, and point-of-interest information, to classify urban areas into distinct functional zones. By processing this data through sophisticated text embedding and clustering algorithms, we aim to create a more efficient, objective, and scalable approach to functional mapping.

The methodology we present demonstrates remarkable capability in processing large geographic areas while maintaining high accuracy in functional zone classification, representing a significant advancement in urban mapping capabilities. The automated nature of our approach enables more frequent updates to functional maps, facilitating better tracking of urban development patterns and more informed decision-making in urban planning and management.

This paper presents a comprehensive framework for automated functional map generation, validated through a detailed case study of the Mumbai Metropolitan Region. Through this research, we demonstrate how modern computational techniques can be leveraged to create more efficient, accurate, and scalable solutions for urban mapping challenges, while significantly reducing the manual effort traditionally required in this process. Our approach not only addresses the current limitations in functional mapping but also establishes a foundation for future developments in automated urban analysis and planning.

# 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 from the OpenStreetMap API underwent a rigorous filtering and preprocessing pipeline to ensure quality and relevance. Initially, predefined OSM filters were applied to retain only key geographic features, such as buildings, offices, commercial areas, transportation infrastructure, and recreational spaces. This helped focus on essential elements for functional zone classification while reducing noise. Text data from each 1km² tile was aggregated to create representative text chunks, preserving geographic nomenclature and spatial relationships. Exploratory data analysis (EDA) was conducted using statistical methods, including box plots, to identify outliers in text length distributions. Areas lacking text data were cross-verified using satellite imagery and land use data to differentiate between actual gaps and natural features like mangroves or water bodies.

For text preprocessing, the Natural Language Toolkit (NLTK) was used for cleaning and normalization, including stop-word removal, lemmatization, and standardization. Lemmatization ensured consistent representation across tiles. Additionally, text length distributions were analysed to set thresholds for outlier exclusion. This structured preprocessing approach established a strong foundation for embedding and clustering analyses, ensuring high data quality and effective functional zone classification.

## 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 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. As shown in Fig. 1, the distribution exhibited strong positive skewness, indicating the presence of tiles with exceptionally high text content, typically corresponding to densely developed urban areas.

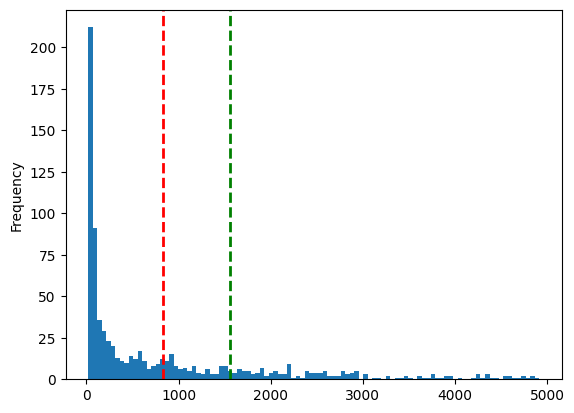


Fig. 1 – Binned frequency distribution of number of text strings vs. length of string, before pre-processing.

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. Fig. 2 shows the frequency distribution of string lengths after pre-processing.

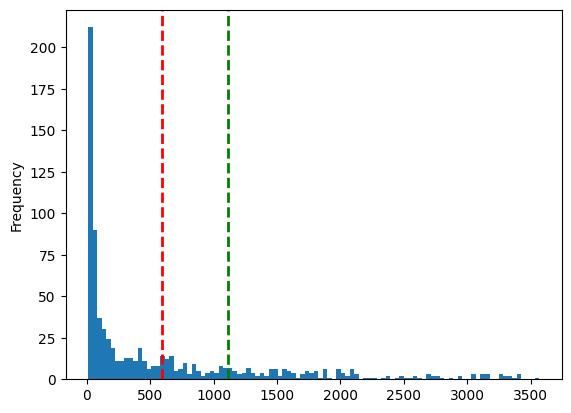


Fig. 2 – Binned frequency distribution of number of text strings vs. length of string, after pre-processing.

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 is a centroid-based clustering method that partitions embeddings into *M* clusters by minimizing intra-cluster variance, assuming spherical clusters. As shown in Fig. 3 and 4, the Elbow Method and Silhouette Method, were used to determine the optimal *K* by analysing the within-cluster sum of squares (WCSS). The optimal *K* is chosen at the "elbow point," or a peak in the Silhouette Curve, where adding more clusters provides minimal improvement while increasing complexity.

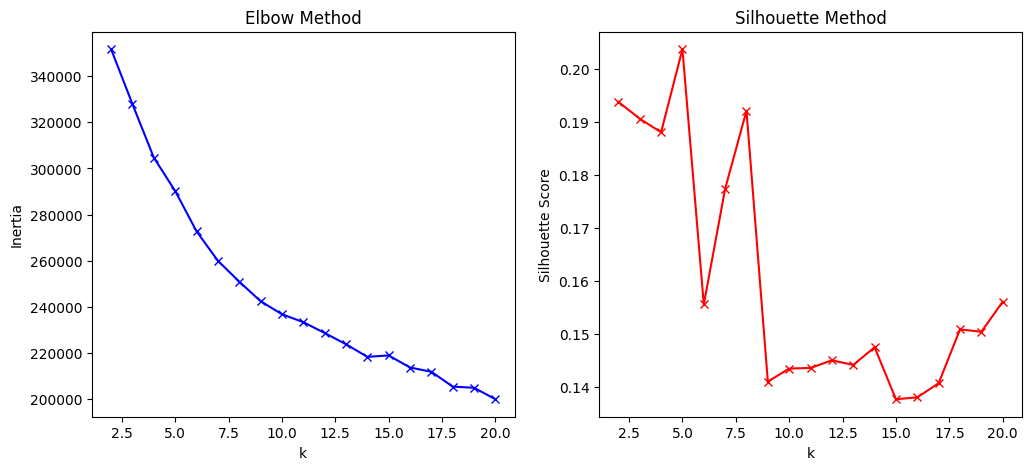


Fig. 3 – LEFT – Elbow method to identify the optimal value of K (K = 9); RIGHT - Silhouette method to identify the optimal value of K (K = 9), for Sentence Transformer based embedding.

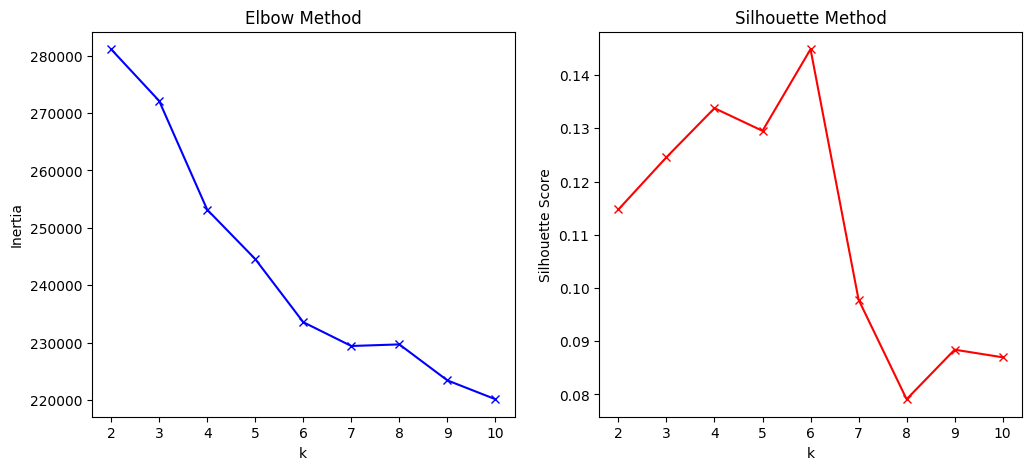


Fig. 4 – LEFT – Elbow method to identify the optimal value of K (K = 6); RIGHT - Silhouette method to identify the optimal value of K (K = 6), for Universal Sentence Encoder (USE) based embedding.

## 8. 3D Visual Representation

t-SNE is a dimensionality reduction technique that visualizes high-dimensional data in 2D or 3D while preserving local structure by converting pairwise similarities into probabilities. UMAP, a more efficient alternative, maintains both local and global structure using a graph-based approach. Both methods were used to visualize embeddings and clusters, helping assess the quality of different embedding approaches and ensuring meaningful representation in lower-dimensional 3D space.

## 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 commercial, residential, industrial or natural. 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:

The algorithm iteratively:

1. Assigns points to nearest centroid:
2. Updates centroids:

## Elbow Method and Silhouette Analysis

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

The optimal k is found at the "elbow" where:

shows diminishing returns.

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

where:

* = mean distance between point i and all points in its cluster
* = mean distance between point i and points in nearest neighbouring cluster

## t-SNE and UMAP

t-SNE converts high-dimensional affinities to low-dimensional similarities

Minimizing KL divergence:

UMAP constructs a fuzzy topological representation through:

where is the distance to the nearest neighbour and is a local connectivity parameter. The low-dimensional representation optimizes cross entropy:

where 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.

The results of the three embedding methods are mentioned in the Table 1.

Table 1 – Comparative Results of the embedding models in the clustering-classification pipeline.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Embedding Method | Accuracy Score | Precision | Recall | F1 Score |
| TF-IDF | 68.956 | 72.432 | 67.757 | 74.531 |
| Universal Sentence Encoder | 97.056 | 97.068 | 97.052 | 97.049 |
| Sentence Transformer | 94.672 | 93.501 | 94.428 | 93.976 |

Confusion matrix for the USE based classification pipeline is represented in Fig. 5

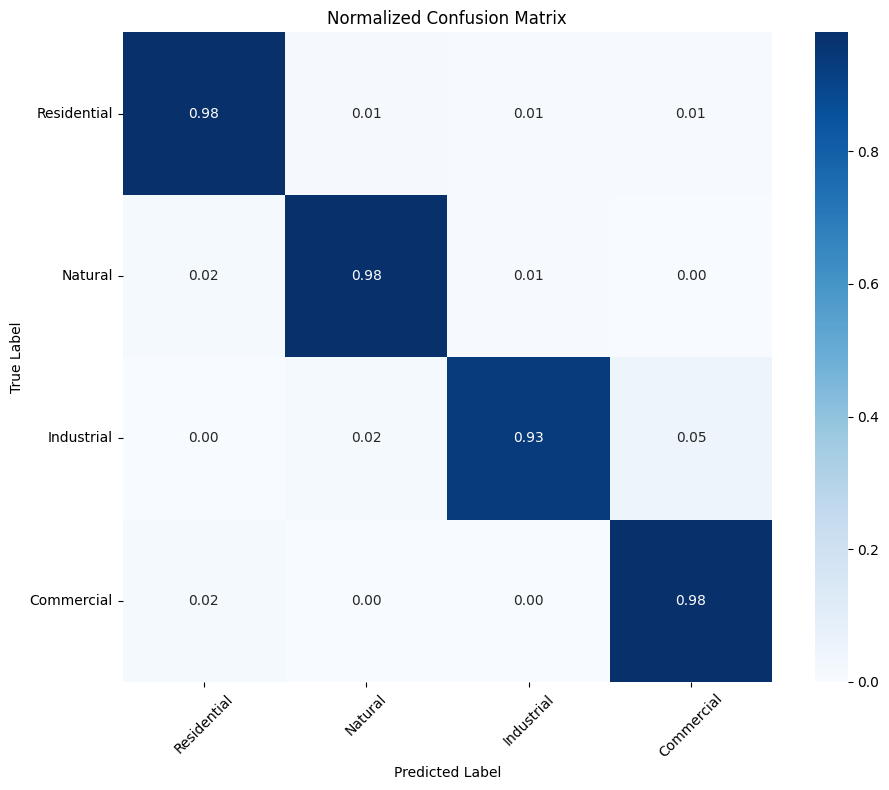


Fig. 5 – Normalized confusion matrix for USE based classification pipeline.

## Visual Cluster Evaluation

The t-SNE-generated embedding projections exhibited better cluster separation than UMAP, making it more effective for visualizing distinct groups. Due to its superior separation of embeddings, t-SNE was selected for further 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.

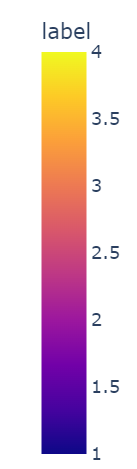
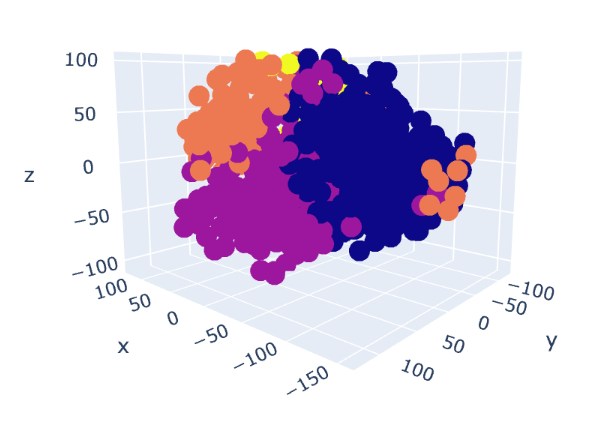


Fig. 6 – USE based embeddings projected in 3D using t-SNE

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.

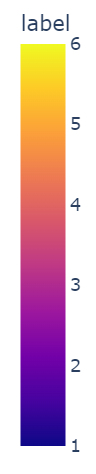
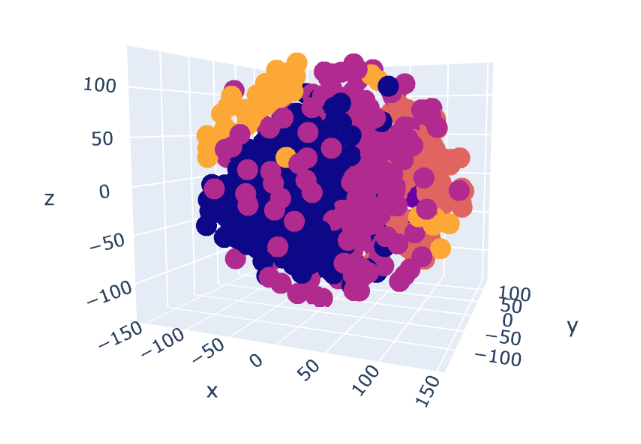


Fig. 7 – TF-IDF based embeddings projected in 3D using t-SNE

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.

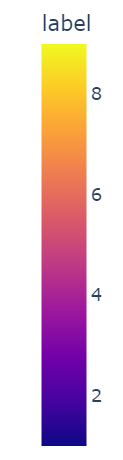
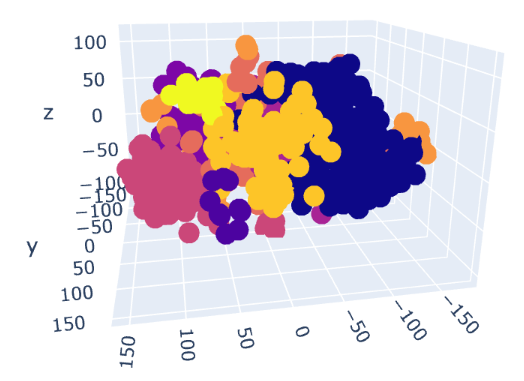


Fig. 8 – Sentence Transformer based embeddings projected in 3D using t-SNE

# Discussion

The t-SNE visualizations for TF-IDF, Sentence Transformers, and Universal Sentence Encoder (USE) embeddings reveal significant differences in semantic representation. TF-IDF, based on word frequency, struggles with contextual understanding, leading to loosely formed and overlapping clusters. Sentence Transformers generate dense embeddings, improving semantic similarity but still exhibit some inter-cluster mixing. USE embeddings, however, show superior separation, indicating better contextual representation and minimal noise.

Performance metrics further confirm that USE outperforms both Sentence Transformers and TF-IDF in capturing meaningful representations. While Sentence Transformers, especially the all-MiniLM-v6 model, provide strong semantic encoding, they fall slightly short in precision and recall. TF-IDF, being a statistical approach, lacks deep contextual awareness, making it the least effective. Given the need for robust, context-aware embeddings, USE is the preferred choice, with Sentence Transformers as a strong alternative.

# Conclusions

This study introduces an automated approach for generating high-fidelity functional maps using text-based clustering of OpenStreetMap data. By leveraging advanced natural language processing techniques, particularly text embeddings and clustering algorithms, our methodology effectively classifies urban spaces into distinct functional zones. The implementation in the Mumbai Metropolitan Region demonstrates the framework's ability to process large-scale urban areas while ensuring accuracy and computational efficiency.

Key contributions include a scalable framework for automated functional mapping using OpenStreetMap data and sophisticated text embedding techniques that capture semantic relationships between urban features. Comprehensive evaluation metrics validate its effectiveness, positioning it as a viable alternative to manual mapping. The framework’s successful application to a complex urban environment highlights its robustness in handling diverse land use patterns.

Future research can enhance this framework by integrating temporal analysis, enabling tracking and prediction of functional zone changes. This includes time-series analysis for urban development trends and predictive models for future land use changes based on historical data.

# References