**Automated High Fidelity Functional Map Generation using Text Data Clustering**

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

**Abstract**

This study presents an automated approach for generating high-fidelity functional maps using text-based clustering of OpenStreetMap data. We address the challenges of traditional manual mapping by implementing a Natural Language Processing framework that classifies urban land use from textual descriptions. Our methodology segments regions into 1km² tiles, extracts text data (building names, points of interest), applies Universal Sentence Encoder embeddings, and uses K-means clustering to identify functional zones. Testing on a 2,500 km² region in Mumbai yielded 97.056% accuracy and 97.049 F1 score when compared to manually labelled ground truth. The framework significantly reduces manual effort while maintaining high performance across residential, commercial, industrial, and natural zone classifications. 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 creation of accurate functional maps has become essential in modern urban planning, environmental management, and spatial analysis. These specialized maps categorize urban regions into distinct zones—residential, commercial, industrial, and natural areas—serving as crucial tools for city planners, policymakers, and researchers. Traditional mapping approaches, relying on labour-intensive field surveys and manual data compilation, are increasingly unable to keep pace with rapidly evolving urban landscapes.

The Functional Map of the World study (Christie et al., 2018) highlighted these challenges, developing a dataset of over 1 million satellite images from 207 countries with 63 functional building categories. While their convolutional neural network approach showed promise, baseline models struggled with architectural diversity across regions, revealing the complexity of large-scale functional mapping.

Despite Geographic Information Systems (GIS) revolutionizing spatial data management, generating and maintaining accurate functional maps still requires substantial manual intervention. This human-dependent approach faces three critical challenges: the impracticality of manually surveying increasingly complex urban environments; difficulty maintaining temporal currency as functional zones continuously evolve; and inconsistencies from subjective classification, particularly in mixed-use or transitional areas.

OpenStreetMap (OSM) has emerged as a comprehensive, community-driven platform creating new opportunities for automated mapping. Research by Kaur et al. (2017) showed OSM data demonstrates particular promise compared to authoritative sources, providing rich, regularly updated information about urban features, including building types and points of interest.

Recent technological advances have expanded automated functional mapping potential. Wan et al. (2017) achieved 87.9% classification accuracy by extracting OSM objects and applying morphological erosion. The evolution of text embedding techniques, particularly transformer-based models like the Universal Sentence Encoder, has enhanced semantic information extraction, while improved clustering algorithms better identify patterns in high-dimensional data.

Zhou et al. (2010) advanced the field with their Clustering-Based KNN Improved Algorithm, addressing limitations of traditional KNN classification. Bai et al. (2025) explored Graph Clustering Neural Networks integrating multi-source data for comprehensive classification. Specialized applications include Cao et al.'s (2024) approach to green space classification using BERT models with remote sensing data, and Kasu et al.'s (2019) impressive accuracy rates for environmental feature detection using a Modified Densitometry 3-Channel Algorithm. Zhang et al. (2016) demonstrated Convolutional Neural Networks' potential in classifying urban functional zones using aerial imagery.

Our research addresses these challenges by proposing a novel methodology combining text-based clustering with advanced NLP techniques to automate high-fidelity functional map generation. This approach leverages OpenStreetMap's rich textual data—building names, business descriptions, and point-of-interest information—to classify urban areas through sophisticated text embedding and clustering algorithms, creating a more efficient, objective, and scalable approach.

Our methodology demonstrates remarkable capability in processing large geographic areas while maintaining high classification accuracy, representing a significant advancement in urban mapping. This paper presents a comprehensive framework for automated functional map generation, validated through a detailed case study of the Mumbai Metropolitan Region, demonstrating how modern computational techniques can create more efficient, accurate, and scalable solutions while significantly reducing manual effort.

**Material and Methods**

**1. OpenStreetMap Data Collection**

This study relies primarily on OpenStreetMap (OSM), a collaborative mapping platform that provides extensive spatial data through community contributions. OSM offers detailed building information and classifications, commercial establishments and points of interest, road networks and transportation infrastructure, as well as land use designations and natural features. The data is freely available through the OSM API in a standardized format. Data quality is maintained through community verification processes, making OSM particularly reliable in densely populated urban areas where contributor activity is highest. This comprehensive geographic database serves as the foundation for our spatial analysis methodology.

**2. Spatial Grid Generation and Region Partitioning**

To systematically analyse large geographic areas, we developed a grid-based partitioning approach. The target region is overlaid with a uniform grid system, where each cell represents a 1km × 1km area. This granularity captures sufficient detail for meaningful functional analysis while maintaining computational efficiency. OSM features falling within each tile's boundaries are extracted, with spatial indices created to optimize the feature-to-tile mapping process. 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, ensuring consistent spatial resolution across the study area while facilitating scalable processing of large geographic regions.

**3. Data Preprocessing and Text Analysis**

The OpenStreetMap data underwent systematic filtering and preprocessing to ensure quality and relevance. Predefined filters retained only key geographic features—buildings, offices, commercial areas, transportation infrastructure, and recreational spaces—reducing noise while preserving classification-relevant elements. Text data from each 1km² tile was aggregated into representative chunks, maintaining geographic nomenclature and spatial relationships. Statistical methods including box plots identified outliers in text length distributions, while data-sparse areas were cross-verified using satellite imagery to distinguish between actual gaps and natural features.

Text preprocessing employed NLTK for cleaning and normalization, including stop-word removal and lemmatization to ensure consistent representation. Text length distributions were analysed to establish outlier exclusion thresholds. This structured approach provided a robust foundation for subsequent embedding and clustering analyses, supporting 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.

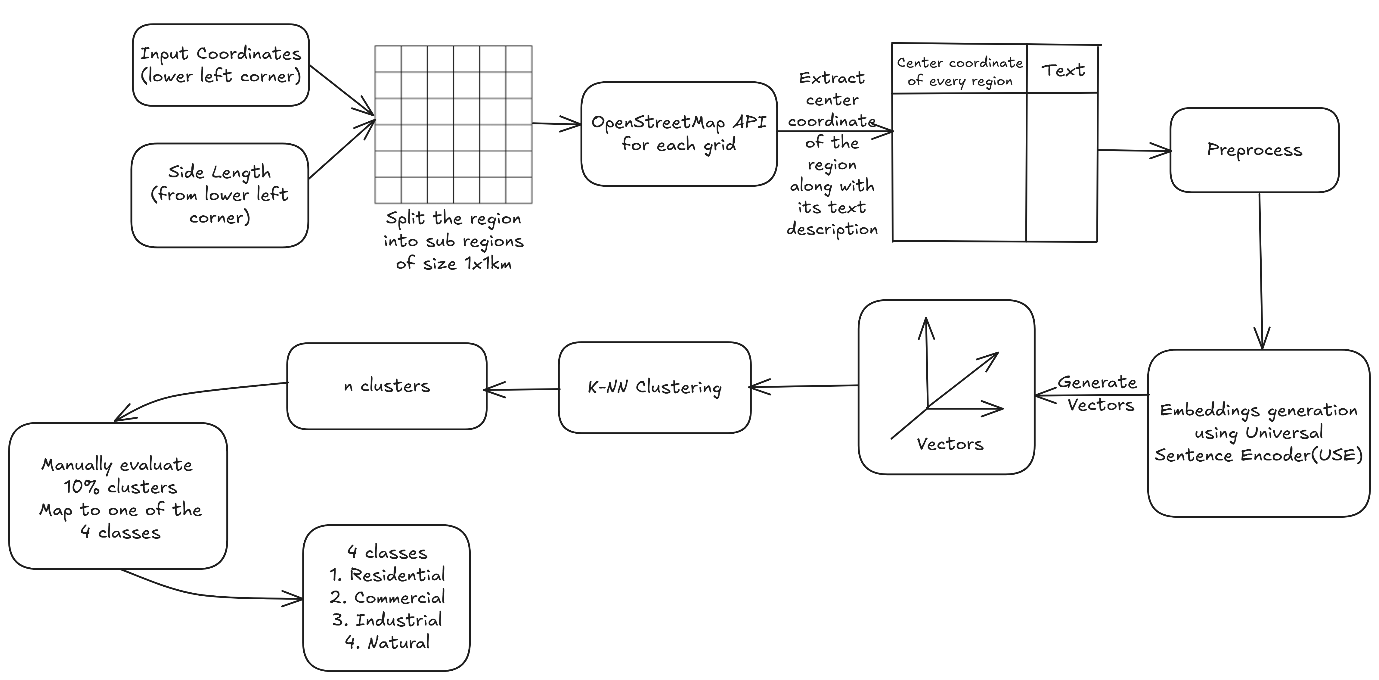


Fig. 1 – Flowchart highlighting the framework for functional map region classification. – add title

**2. Implementation Framework**

Following our established methodology, as shown in the Fig. 1, we partitioned the study area into 1km × 1km tiles, generating a dataset of 2,500 distinct spatial units. This resolution was selected to maintain sufficient granularity for meaningful functional analysis, capture local variations in land use patterns, enable efficient computational processing, and facilitate practical validation of results. During exploratory data analysis, we identified and addressed several key considerations specific to the MMR context. Tiles containing no text data underwent additional verification, particularly in coastal areas and regions with large natural features, helping distinguish between data gaps and legitimate natural areas. The MMR case study provided an ideal opportunity to test our framework's ability to handle complex urban environments, with varied development patterns, mixed land uses, and distinct natural boundaries offering appropriate challenges for validating our automated classification methodology.

**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.

**5.1. 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.

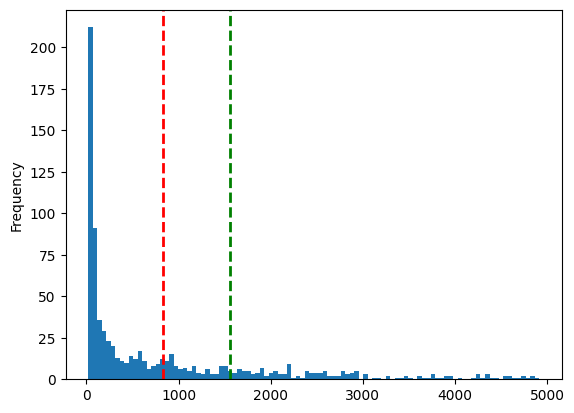


Fig. 1 – Binned frequency distribution of number of text strings vs. length of string, before pre-processing.

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.

**5.2. Void Analysis and Verification**

Approximately 15% of tiles contained no text data. Upon investigation, these regions primarily comprised water bodies (including the Arabian Sea and Thane/Vasai creek), protected mangrove areas, and undeveloped land parcels.

A frequency analysis of key terms across tiles revealed characteristic vocabulary for different functional zones. Commercial zones showed high frequencies of retail, office, and service-related terms, while residential areas were characterized by references to apartment complexes, housing societies, and community facilities. Industrial zones consistently featured manufacturing, warehouse, and logistics terminology, and natural areas were identified through references to parks, forests, and water bodies.

**5.3. 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.

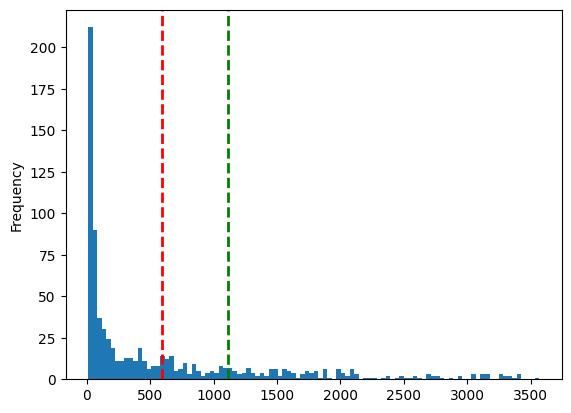


Fig. 2 – Binned frequency distribution of number of text strings vs. length of string, after pre-processing.

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 transforms words, sentences, or documents into numerical vectors that capture semantic meaning, allowing machines to understand and compare text mathematically. While traditional methods like TF-IDF create sparse vectors based on word frequency without understanding context, transformer-based embeddings generate dense, context-aware representations using deep learning. Notable models include Google's Universal Sentence Encoder (USE) and Sentence-Transformers (all-MiniLM-L6-v2). For generating text embeddings in this study, three methods were selected: TF-IDF, USE, and sentence transformer, enabling mathematical comparison where similar texts have similar vector representations.

**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.

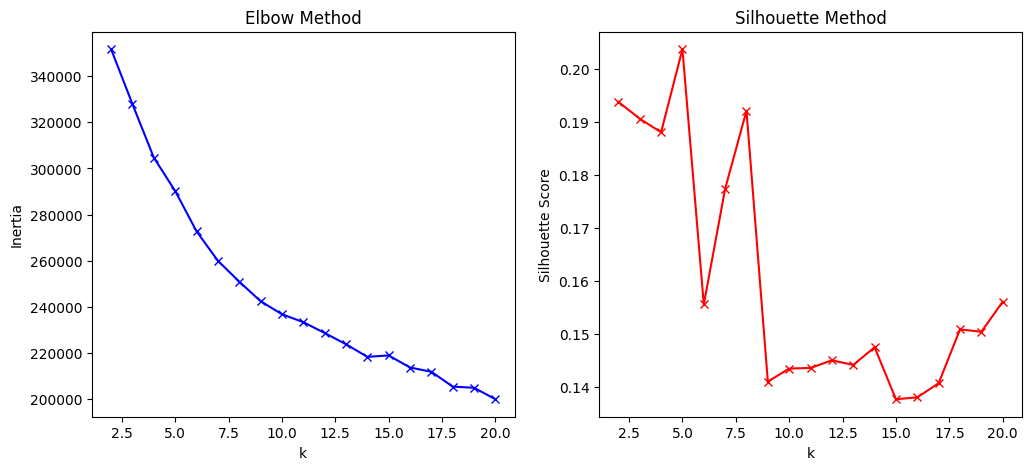


Fig. 3 – LEFT – Elbow method to identify the optimal value of K (K = 9); RIGHT - Silhouette method to identify the optimal value of K (K = 9), for Sentence Transformer based embedding.

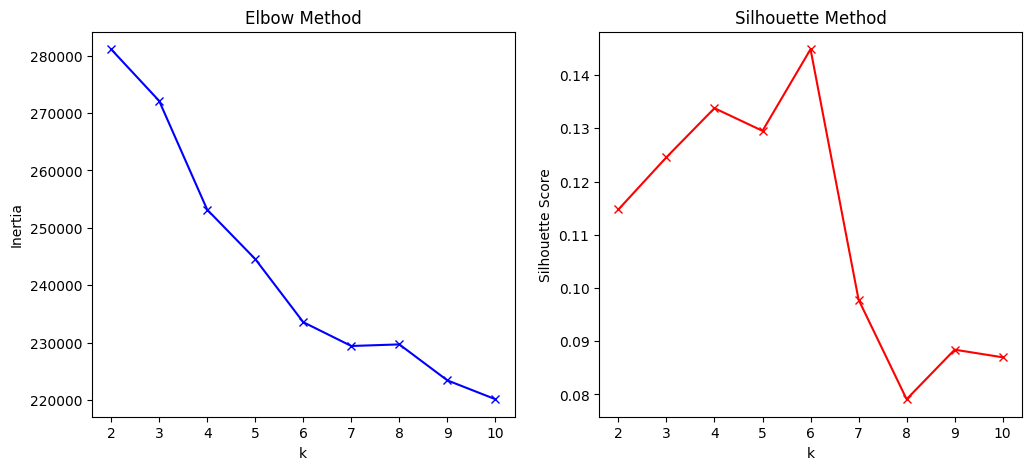


Fig. 4 – LEFT – Elbow method to identify the optimal value of K (K = 6); RIGHT - Silhouette method to identify the optimal value of K (K = 6), for Universal Sentence Encoder (USE) based embedding.

**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 assigns points to nearest centroid and updates them,

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

**Results**

**1. Evaluation Metrics**

To evaluate our automated functional map generation framework, we implemented a comprehensive set of metrics assessing both clustering quality and classification accuracy. These complementary metrics provide a thorough assessment of the model's practical utility for urban planning applications. We utilized confusion matrices to visualize performance by displaying true positives, true negatives, false positives, and false negatives across all functional zones. While accuracy measured the overall proportion of correctly classified tiles, we recognized its limitations with class imbalance in urban landscapes. Therefore, we incorporated precision and recall score to assess the model's ability to identify all instances of particular functional zones. The F1-score balanced precision and recall considerations, providing a unified metric that accounts for both false positives and negatives. Following equations define the mentioned evaluation metrics:

Table 1 presents a comprehensive comparison of performance metrics across the three distinct text embedding methodologies.

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 |
| Sentence Transformer | 94.672 | 93.501 | 94.428 | 93.976 |
| Universal Sentence Encoder | 97.056 | 97.068 | 97.052 | 97.049 |

Fig. 5 illustrates the normalized confusion matrix for the classification pipeline that utilizes the Universal Sentence Encoder (USE) embedding method. This visual representation provides a detailed analysis of classification performance across all functional zone categories.

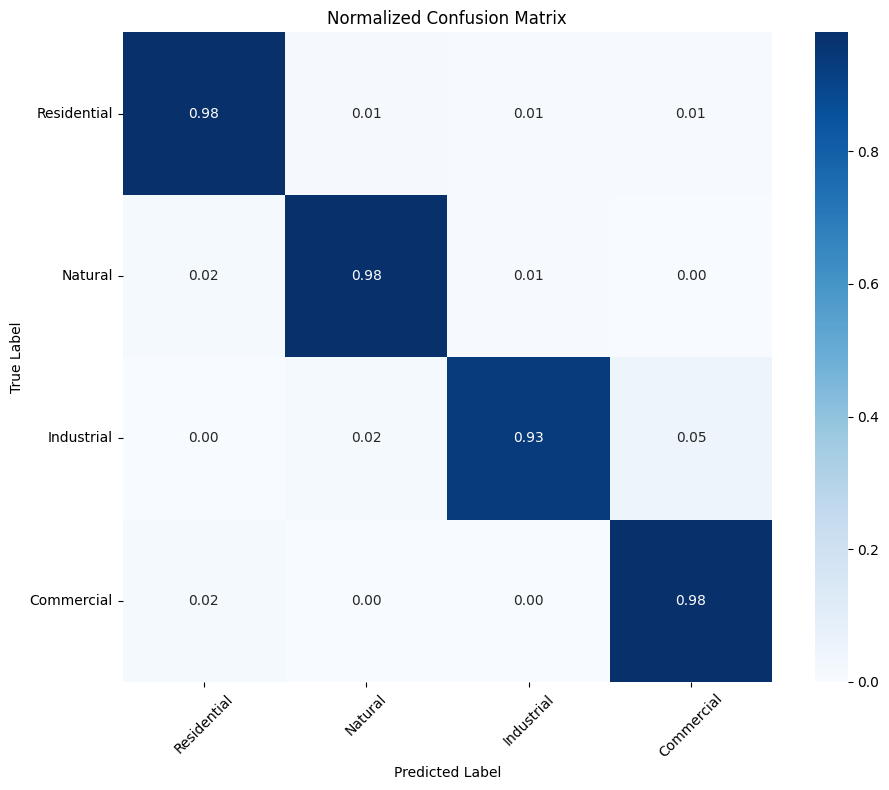


Fig. 5 – Normalized confusion matrix for USE based classification pipeline.

**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.

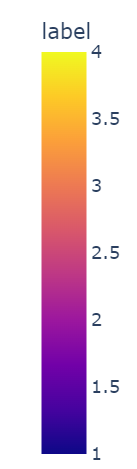
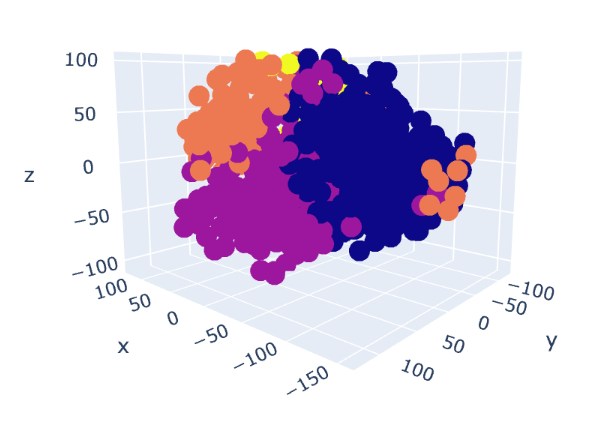


Fig. 6 – USE based embeddings projected in 3D using t-SNE

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.

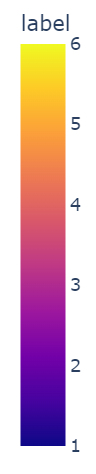
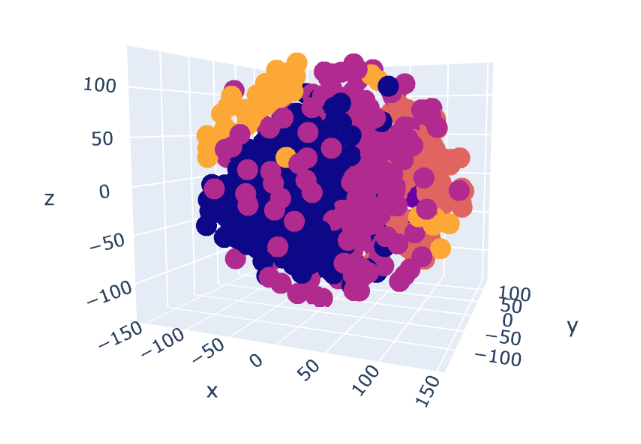


Fig. 7 – TF-IDF based embeddings projected in 3D using t-SNE

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.

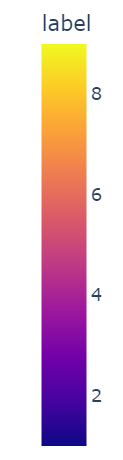
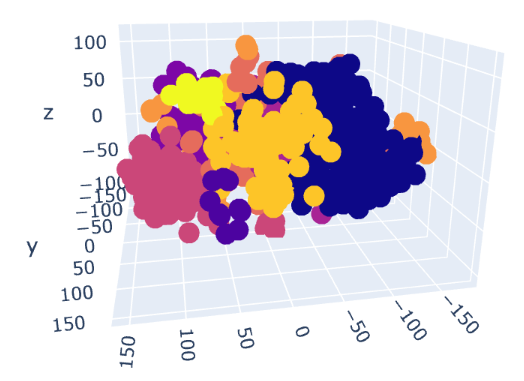


Fig. 8 – Sentence Transformer based embeddings projected in 3D using t-SNE

**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 presents a novel approach for generating functional urban maps through text-based clustering of OpenStreetMap data. By applying advanced NLP techniques and sophisticated text embeddings, our framework efficiently classifies urban spaces into distinct functional zones. Implementation in Mumbai demonstrates its capability to process complex urban environments accurately. Key contributions include a scalable methodology that captures semantic relationships between urban features, providing a viable alternative to manual mapping. Future work will incorporate temporal analysis to track and predict functional zone changes, enabling more dynamic urban planning and development forecasting.

**References**