EGU General Assembly 2025

27th April – 02nd May 2025

Integrating User-Generated POI Data and Satellite Imagery for Enhanced Urban Land Use Classification: A Topic Modeling Approach

Ravi Satyappa Dabbanavar and Arindam Biswas

Supplementary Material

Sustainable URban & REgional Analysis Lab

Dr. Arindam Biswas
Associate Professor at the Department of Architecture and Planning, Indian Institute of Technology Roorkee.

Mr. Ravi Satyappa Dabbanavar
Doctoral scholar at the DAP, IIT Roorkee.
Bachelors in Planning,
Topic: Approach for essential urban land-use mapping by integrating Remote Sensing and User-generated Big Data

+6

Research scholars

- Urban Inequality
- Polarization
- Knowledge Economy
- Regional growth
- Big Data Analysis
- AI in urban analytics
- Urban & Regional governance

CONTENTS

Definitions

Point of Interest (POI):

Refers to any specific point location that someone may find useful or interesting, such as a landmark, restaurant, or any place of significance, often used in mapping and GPS navigation systems for identifying locations *(ESRI, 2023)*.

Topic Model:

Topic models are statistical models used to discover hidden topics or themes within a collection of documents. These models analyze patterns of word co-occurrence in the text and group words that frequently appear together into topics *(Blei et al., 2003)*.

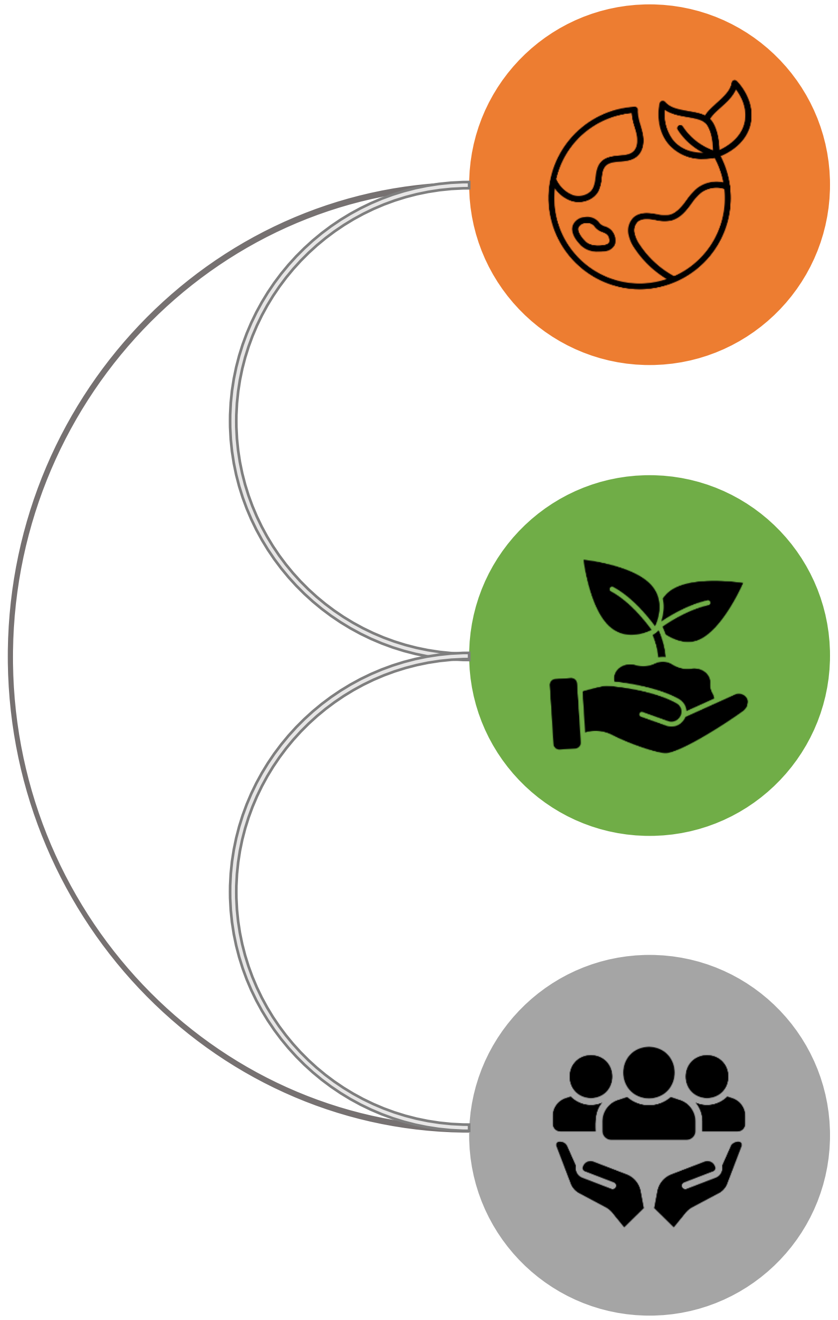
Coherence value:

Used to evaluate the quality of the topics generated by a Topic model. It assesses how interpretable the topics are by examining the semantic similarity between the words in each topic *(Stevens et al., 2012)*.

Global Challenges: The Importance of Land-Use Mapping

50% of the population now lives in cities, projected to reach 70% by 2050.

Rapid urban growth creates a need for precise land-use mapping



Mitigating Environmental Impact:

Helps in identifying urban green spaces, preserving ecosystems, and planning for climate resilience.

Supporting Sustainable Development Goals (SDGs):

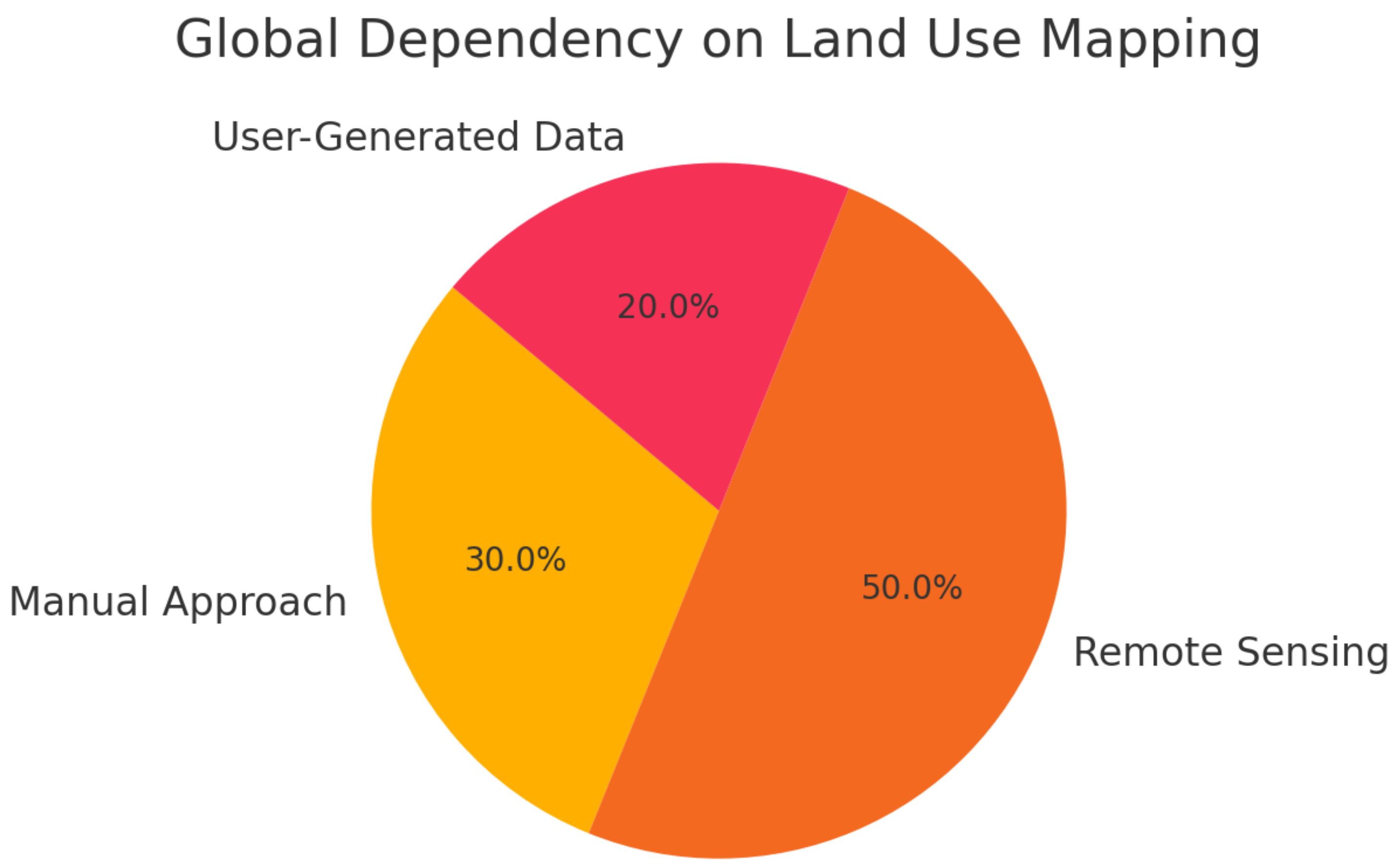
Ensures equitable access to resources and promotes balanced urban-rural development.

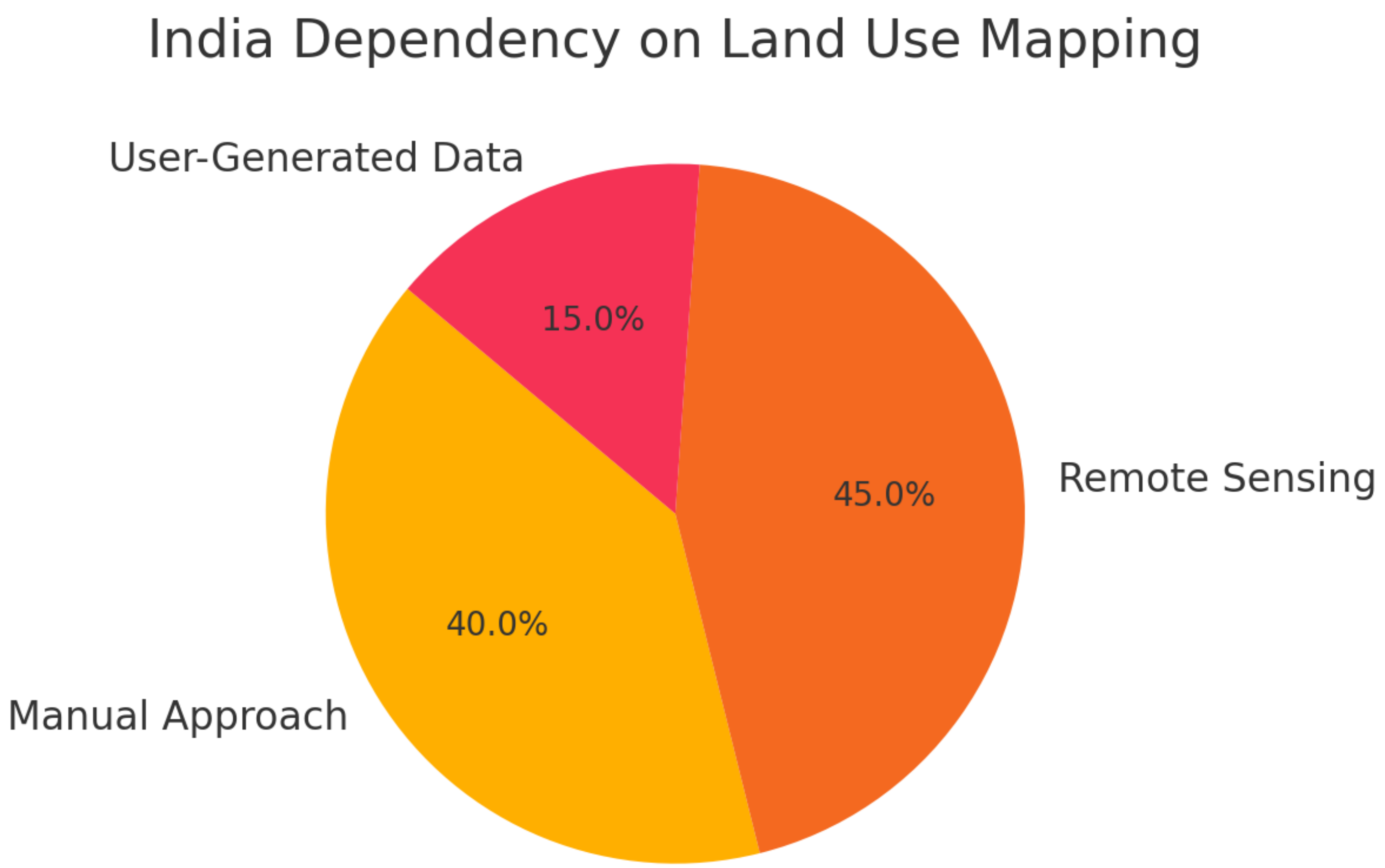
Optimizing Public Services:

Facilitates the efficient allocation of utilities such as transportation, water supply, and waste management.

Source: Yu & Fang, 2023, Gong, P., Li, X., & Zhang, W. 2019, United Nations. 2023

Different Land-use Mapping Approaches





- **Manual Approach:** Relies on human observation and on-ground surveys to map land use.
- **Remote Sensing:** Utilizes satellite imagery or aerial photography to capture and analyze data about land use patterns on a large scale, enabling quick and accurate mapping.
- **User-Generated Data:** Involves contributions from individuals or communities through platforms like crowd-sourcing real-time and localized land use information.

Source: UNEP Annual Report 2023, See, L., et al. (2016), Wulder, M. A., & Coops, N. C. (2014).

Challenges Across Different LU Mapping Approaches

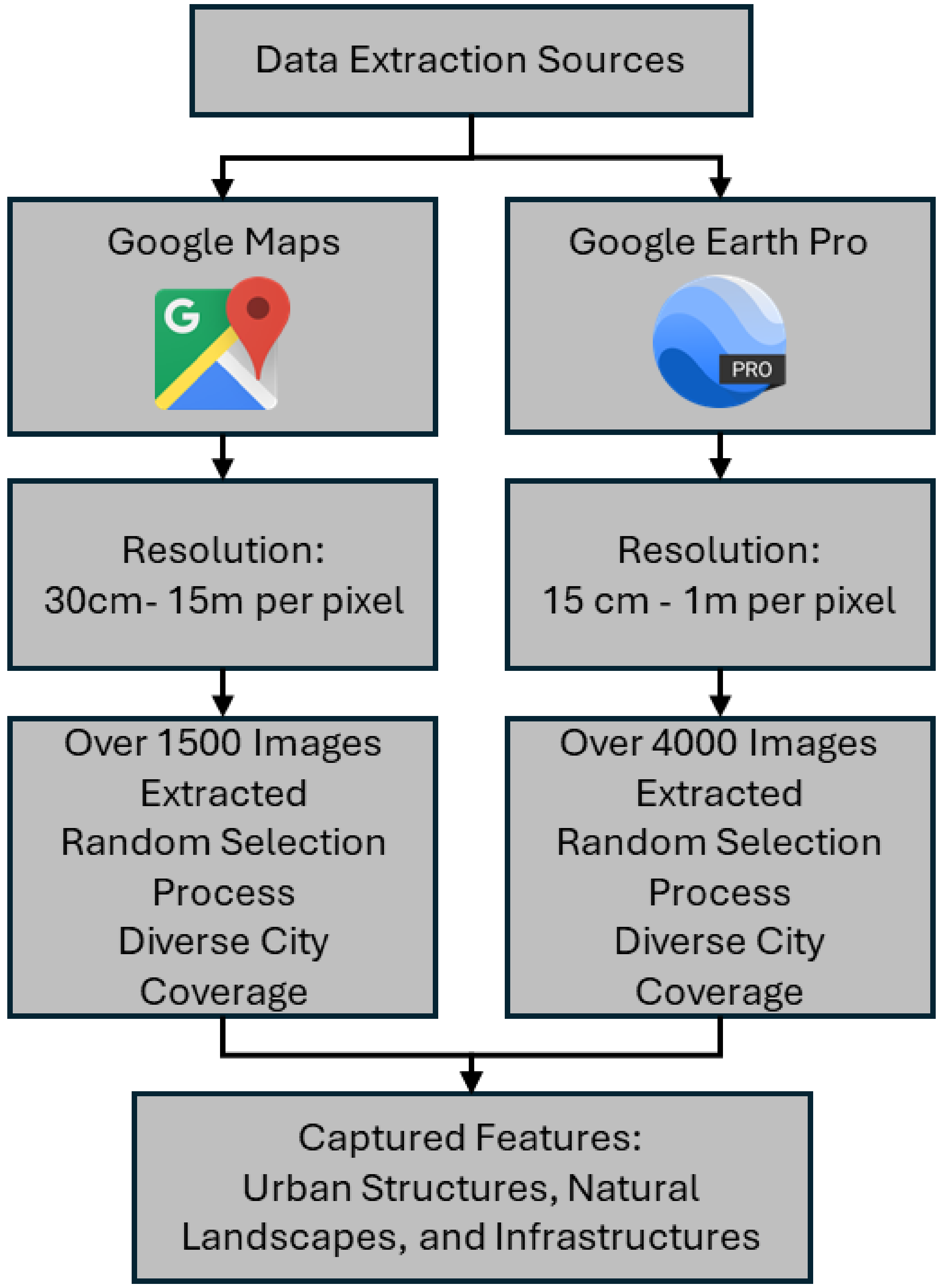
Approach	Challenges Identified	References
Manual Approach	• **Labor-intensive and time-consuming:** Requires physical effort for data collection and constant updates. • **Scalability issues:** Inefficient for managing growing urban regions. • **Subject to human error:** Inconsistent measurements and subjective interpretations.	Harley (1987), Kain & Baigent (1992), Waldhoff & Bareth (2009)
Remote Sensing-based Approach	• **Limited real-time analysis:** Data collection is not fast enough to respond to dynamic urban changes. • **Struggles with complex urban features:** Difficulty in distinguishing mixed land-use areas. • **High cost of sensors and processing:** Expensive equipment and technical expertise required.	Govindu et al. (2019), Gong et al. (2013), Liu et al. (2018)
User-generated Data-based Approach	• **Volume**: Massive datasets are difficult to store, process, and analyze efficiently. • **Variety**: Integrating structured, semi-structured, and unstructured data from diverse sources is challenging. • **Complexity**: Advanced tools are needed to analyze spatiotemporal patterns and relationships. • **Real-time Data Handling**: Processing live data streams is resource-intensive and may cause delays. • **Fragmented Data**: Data spread across multiple systems is hard to consolidate and interpret effectively.	Gandomi & Haider (2015), Assur & Rowshankish (2024), Gil (2022)

Aim

To develop an efficient, scalable, and integrated framework for urban land-use classification by combining remote sensing techniques and user-generated data to address urban growth challenges and governance needs.

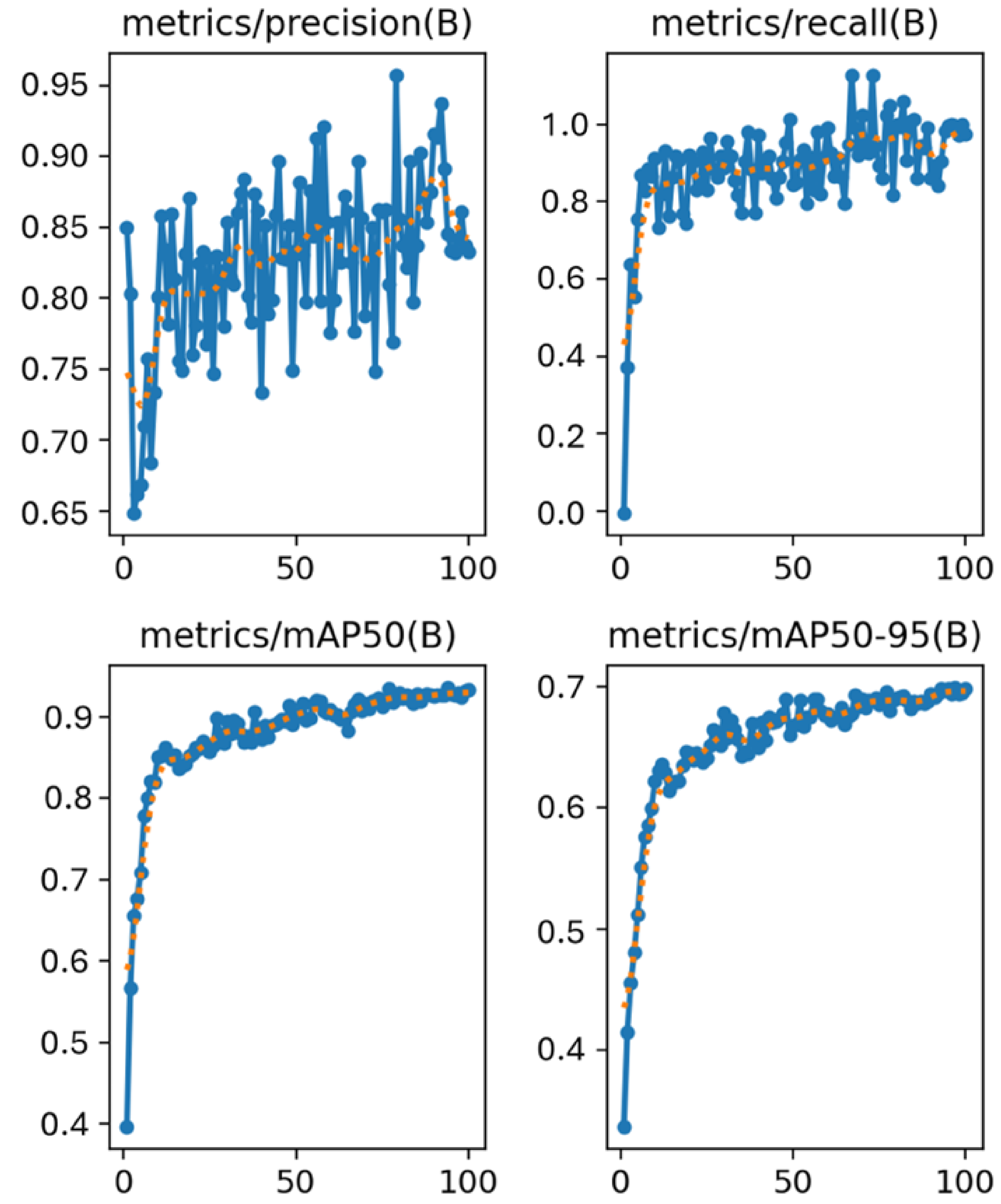
Remote sensing data to more readable unified format

Object detection/segmentation models comparison results



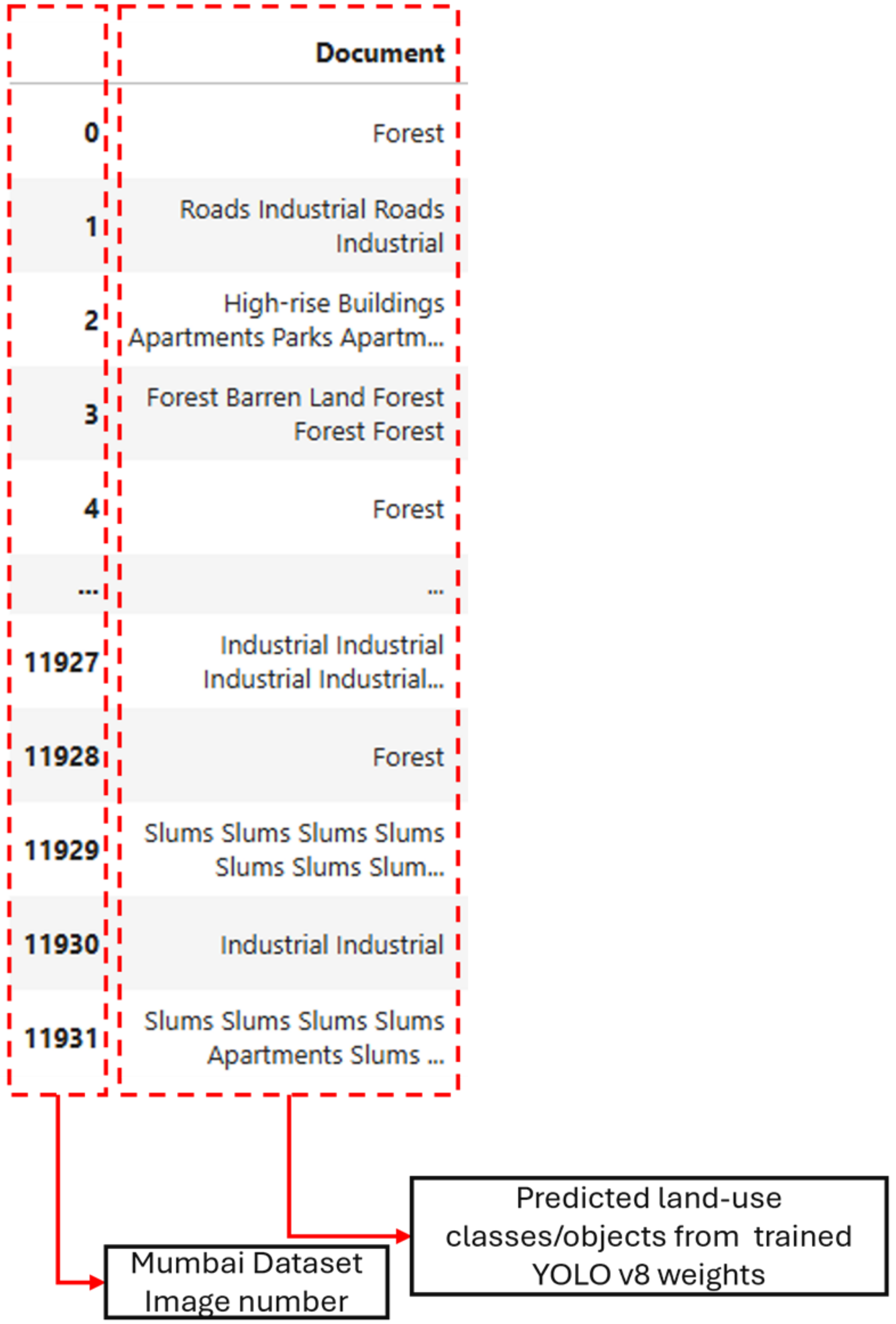
Model	Type	Average Precision (AP)	Mean Average Precision mAP@0.5	Average Inference Time (seconds/image)
		Precision the model can attain at various levels of recall	Average value of AP over all classes at an IoU threshold value of 0.5	Time taken per image on average by the model for processing and fulfilling the demands it must deliver against the image
		$AP = \sum n\,(R_n - R_{n-1})P_n$	$mAP@0.5 = \frac{1}{N}\sum_{i=1}^{N} AP_i$	$Average\ Inference\ Time\ (seconds/image) = \frac{Total\ Inference\ Time\ for\ All\ Images\ (Seconds)}{Number\ of\ Images\ Processed}$
Mask R-CNN with ResNet-50	Segmentation	0.41	0.43	0.058
Mask R-CNN with ResNet-101	Segmentation	0.53	0.56	0.064
Faster R-CNN with ResNet-50	Object Detection	0.49	0.50	0.049
Faster R-CNN with ResNet-101	Object Detection	0.55	0.57	0.054
YOLO v5	Object Detection	0.75	0.78	0.02
YOLO v8	Object Detection	0.89	0.848	9.40

Object detection/segmentation models comparison results



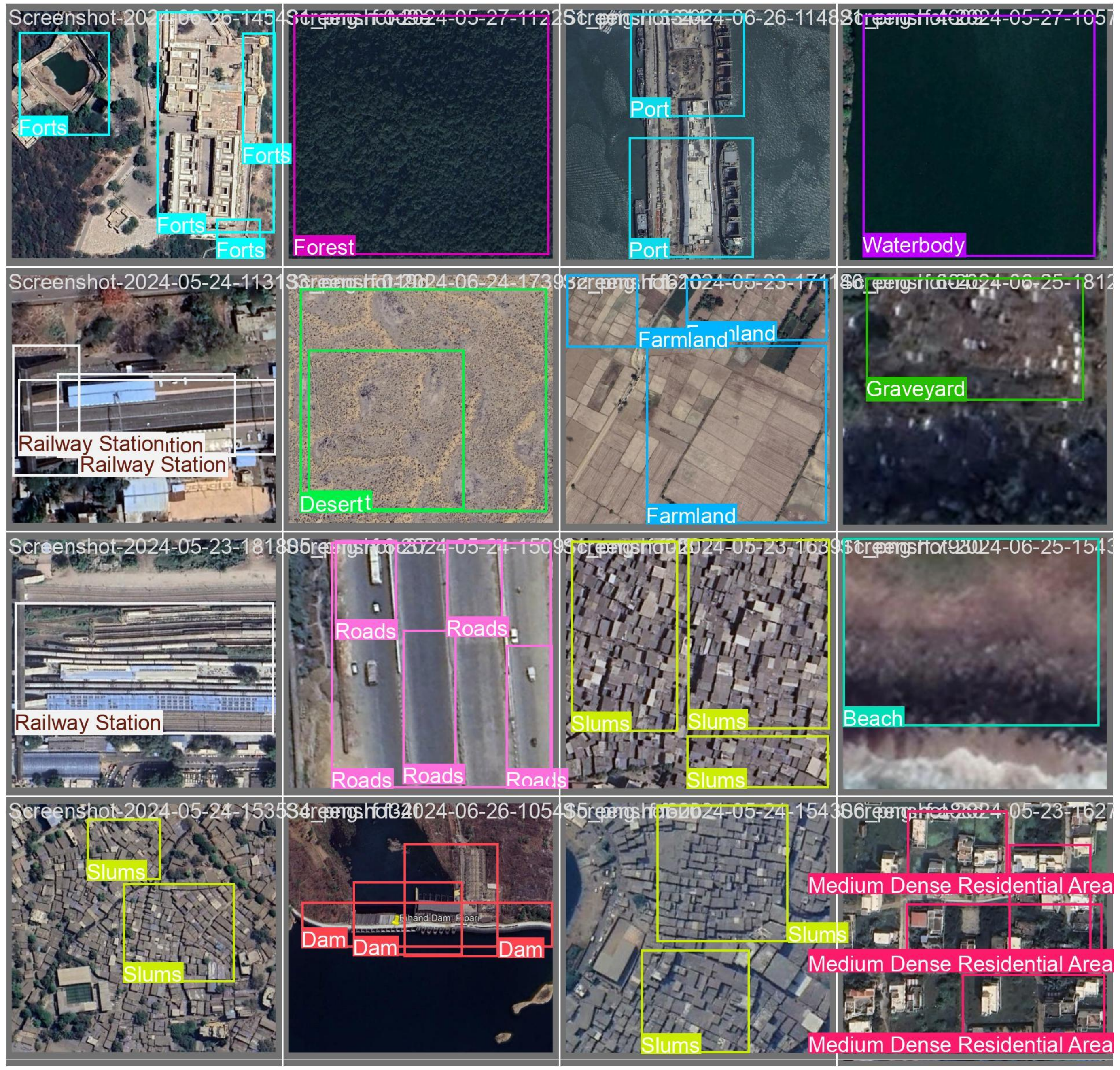
Performance Metrics (Precision, Recall, mAP):

- **Precision and recall** metrics steadily improve, with recall reaching the high precision, indicating that the model is very effective at finding most objects.
- **mAP@0.5** and **mAP@0.5-0.95** metrics show steady growth, with mAP@0.5-0.95 nearing 0.7, reflecting solid detection performance across different IoU thresholds.

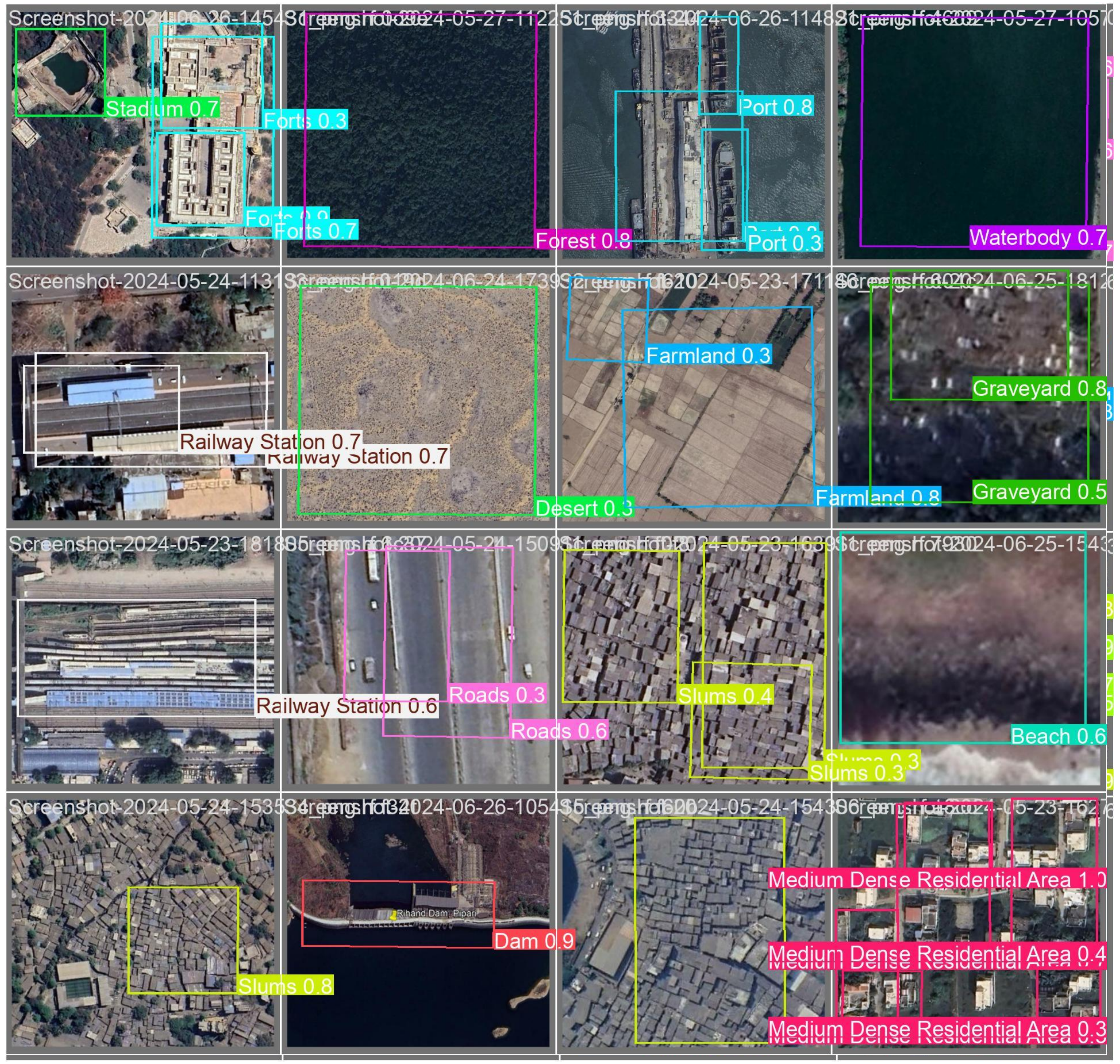


	Document
0	Forest
1	Roads Industrial Roads Industrial
2	High-rise Buildings Apartments Parks Apartm...
3	Forest Barren Land Forest Forest Forest
4	Forest
...	...
11927	Industrial Industrial Industrial Industrial...
11928	Forest
11929	Slums Slums Slums Slums Slums Slums Slum...
11930	Industrial Industrial
11931	Slums Slums Slums Slums Apartments Slums ...

Object detection/segmentation models comparison results



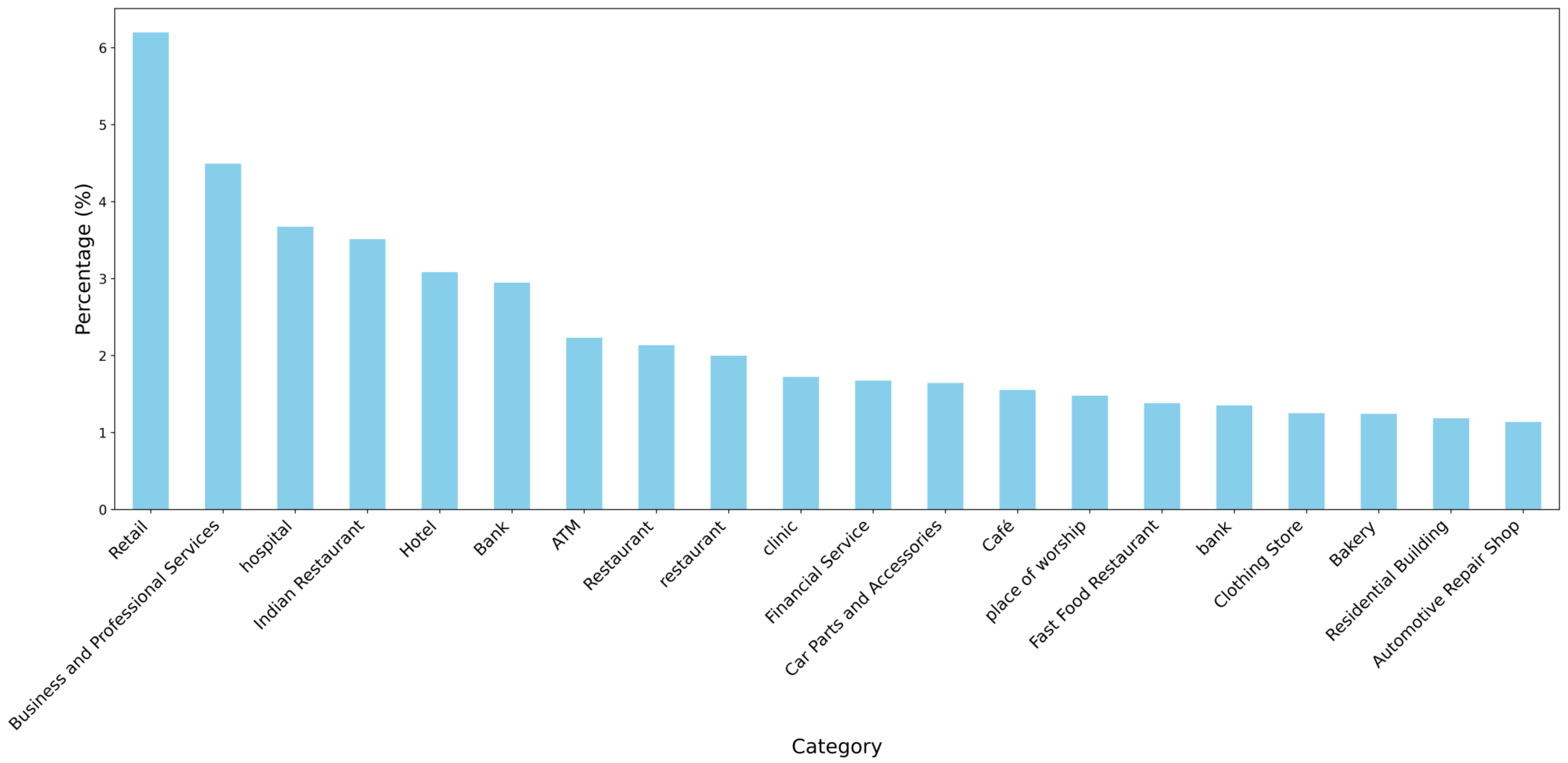
Trained Dataset



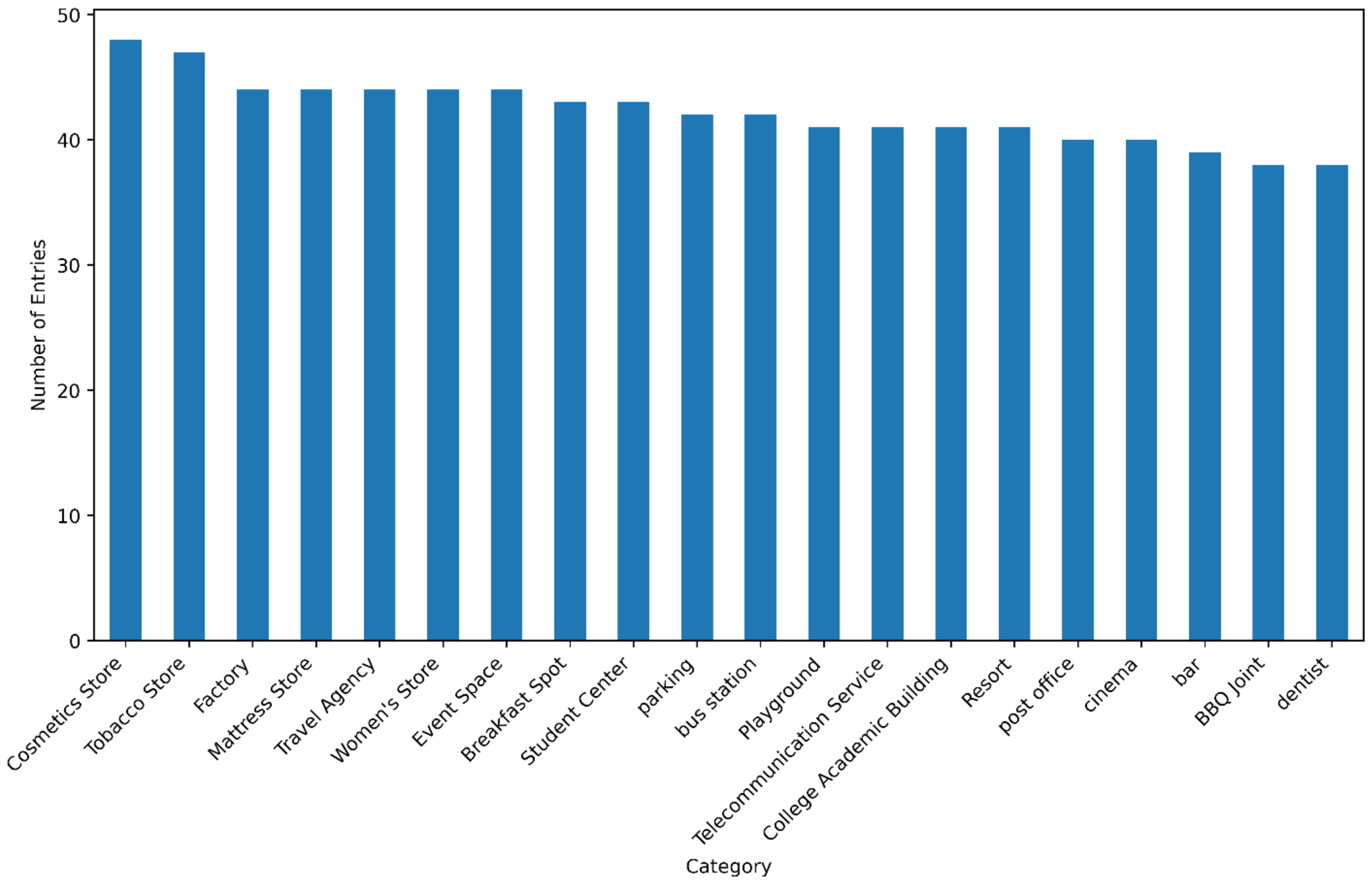
Predicted Dataset

User-generated data collection and to convert to single format

Data Preparation and Materials



Top 20 Proportional Representation of Categories

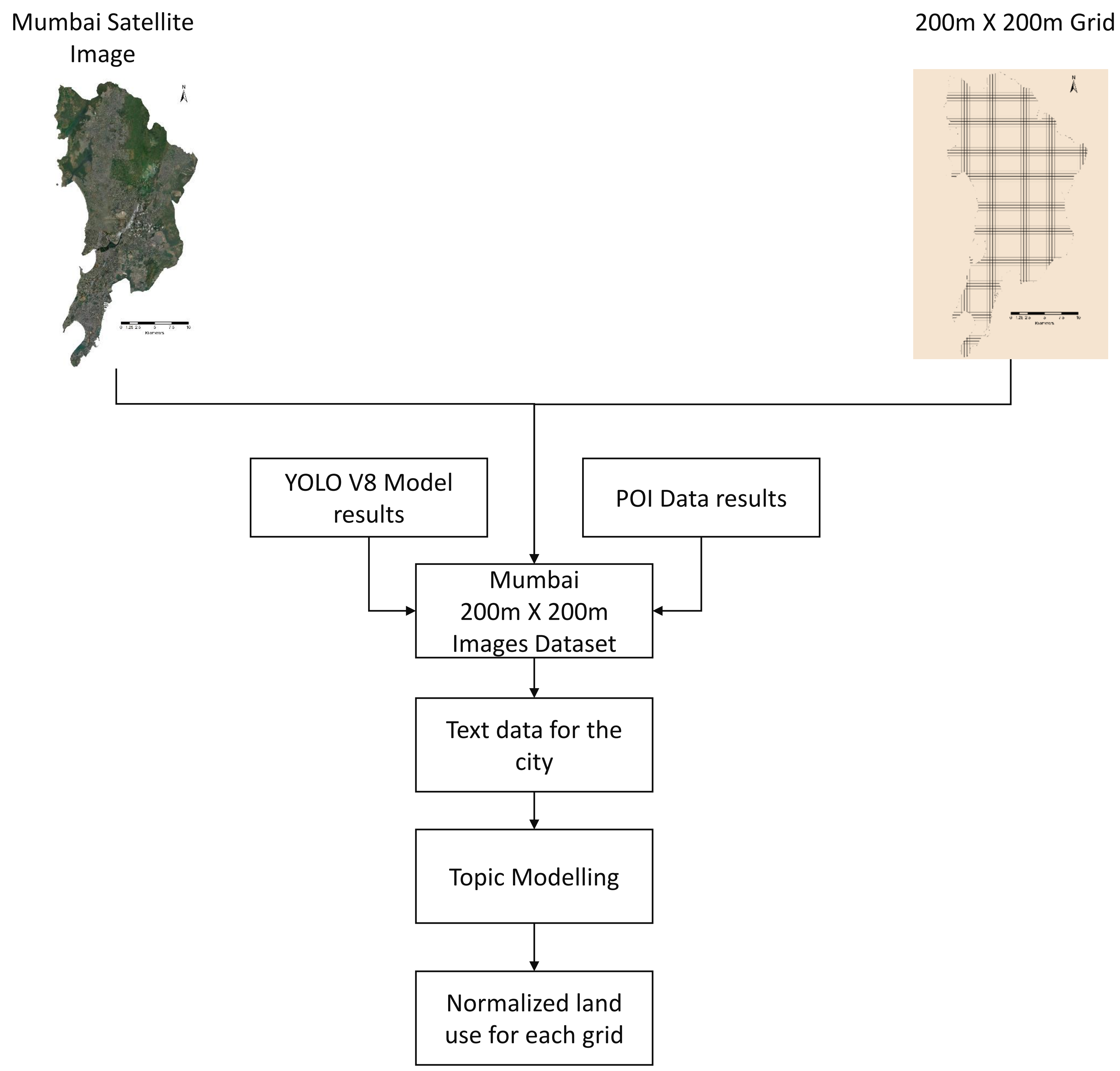


Top 20 Underrepresented POI Categories – Discussed alongside the paragraph detailing category imbalance.

- A total of 48,641 points were extracted from OSM's complex mesh (OSM accessed on 18th Dec 2023).
- Foursquare yielded 38,239 points from a massive database of user-generated places (accessed on 1st Dec 2023).
- Using the nearby search, the advanced mapping feature and API of ArcGIS Developer contributed to our data pool with 10,642 points (ArcGIS developers accessed on 13th Dec 2023).
- In total, we acquired **97,522** POI data. Further, after various cleaning steps the we were able to use **85,328** POI data.

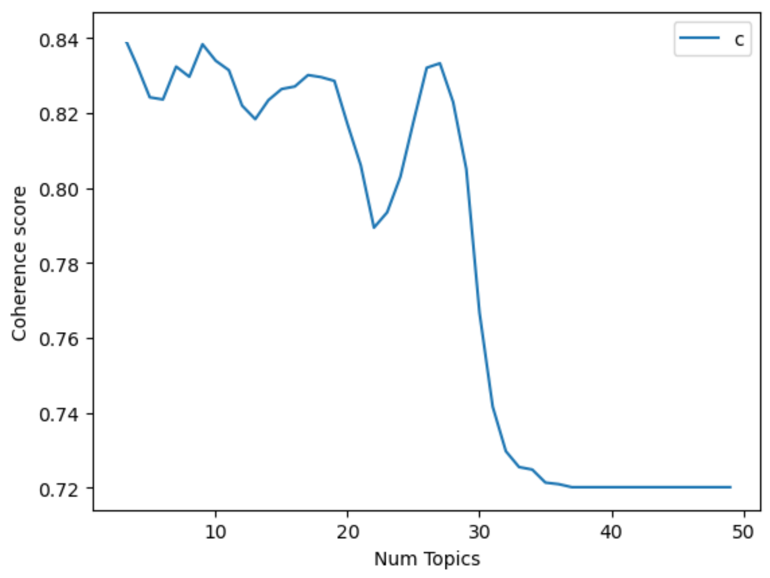
Source: UNEP Annual Report 2023, See, L., et al. (2016), Wulder, M. A., & Coops, N. C. (2014).

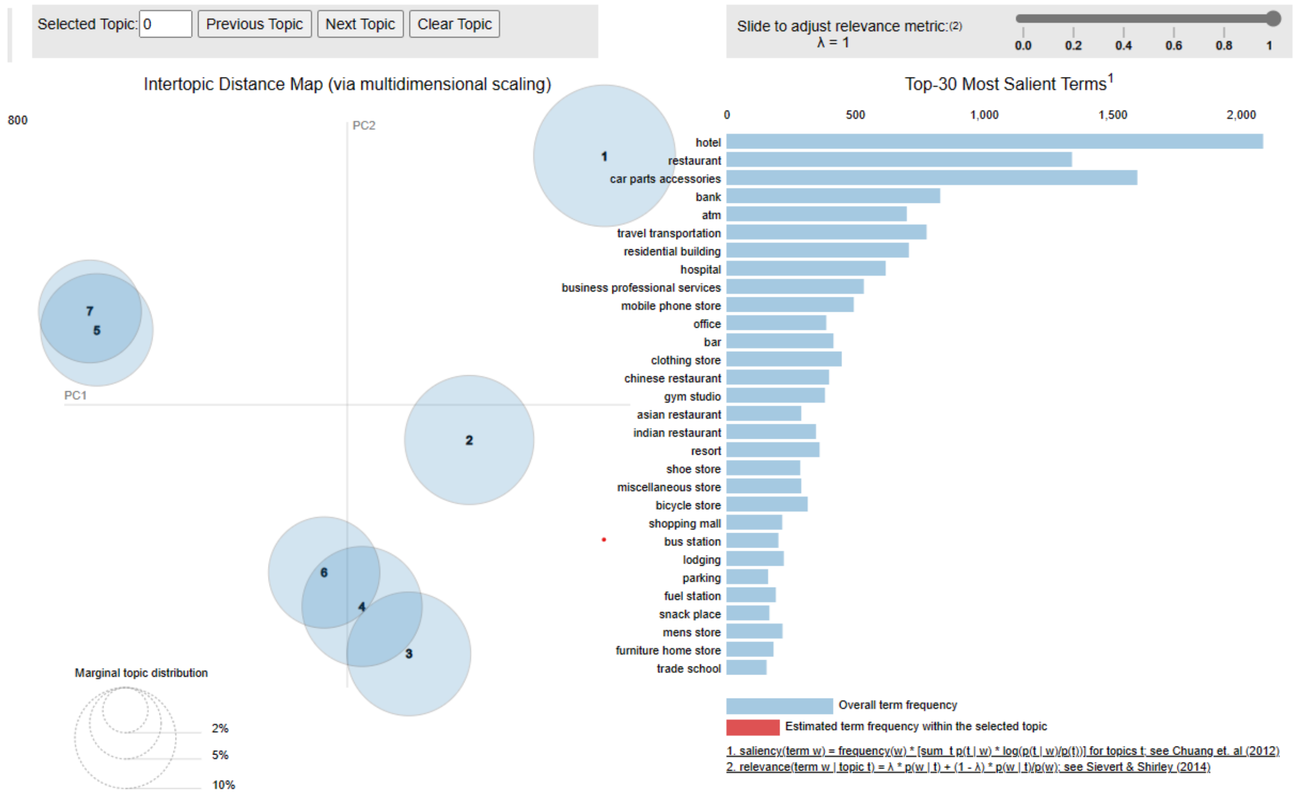
Collecting Text data



LDA Topic Model

Sentence Tokenization

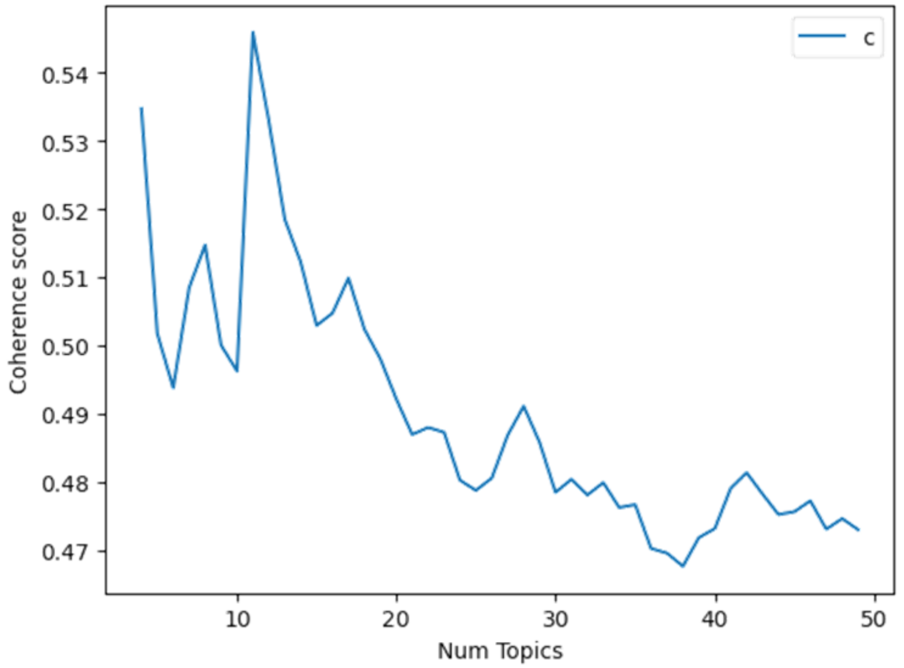
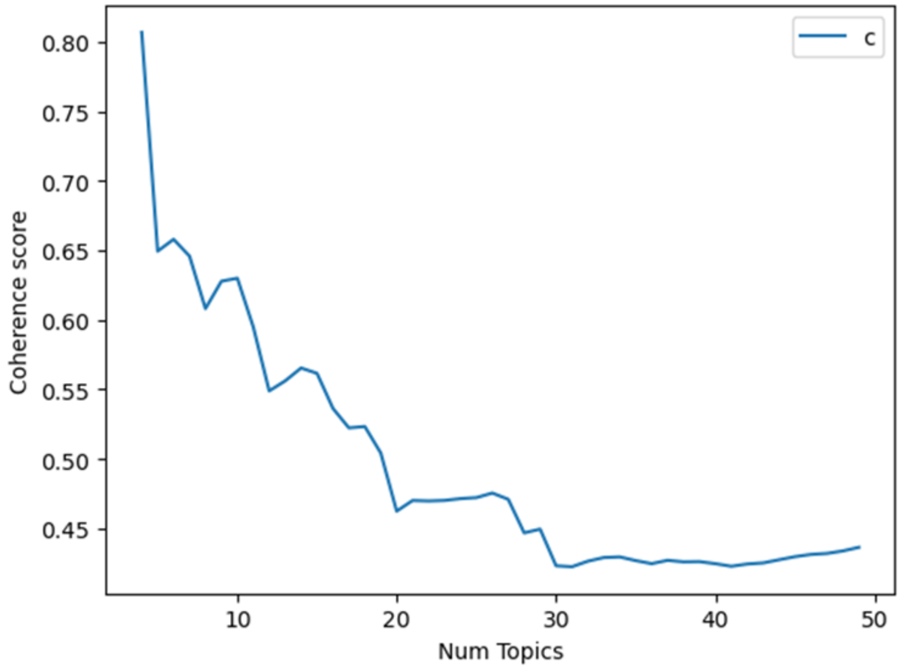
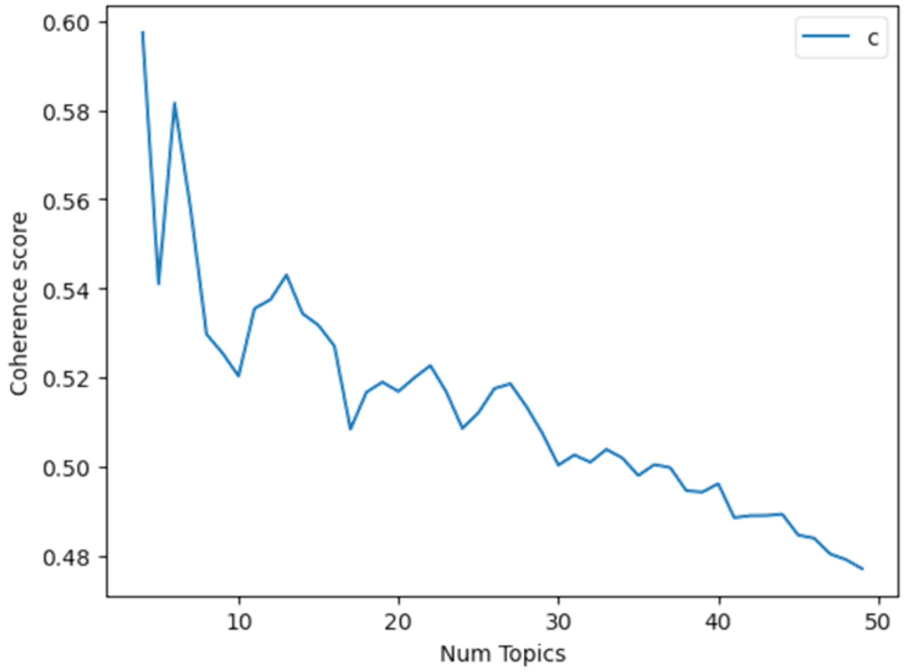
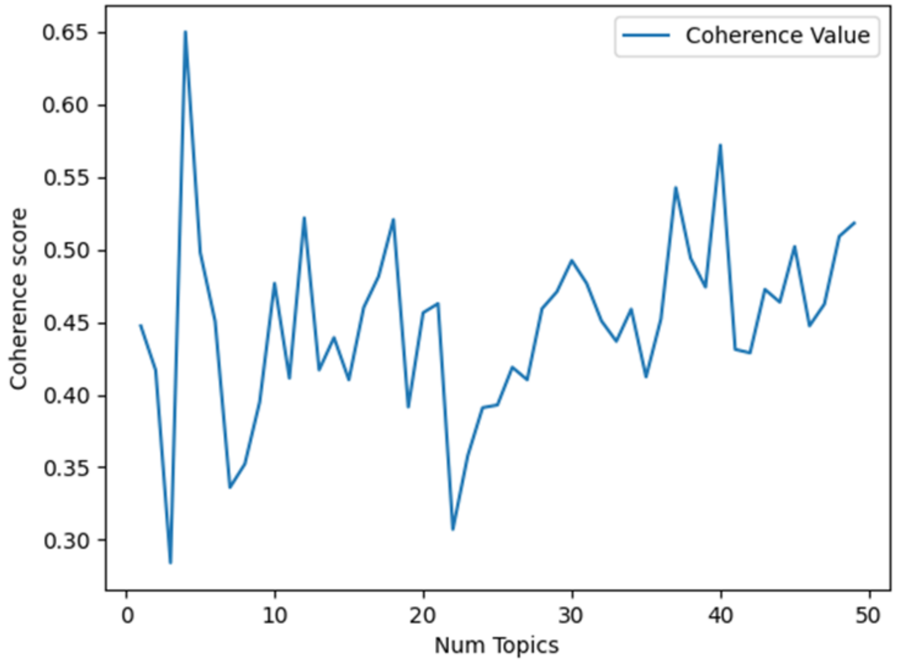
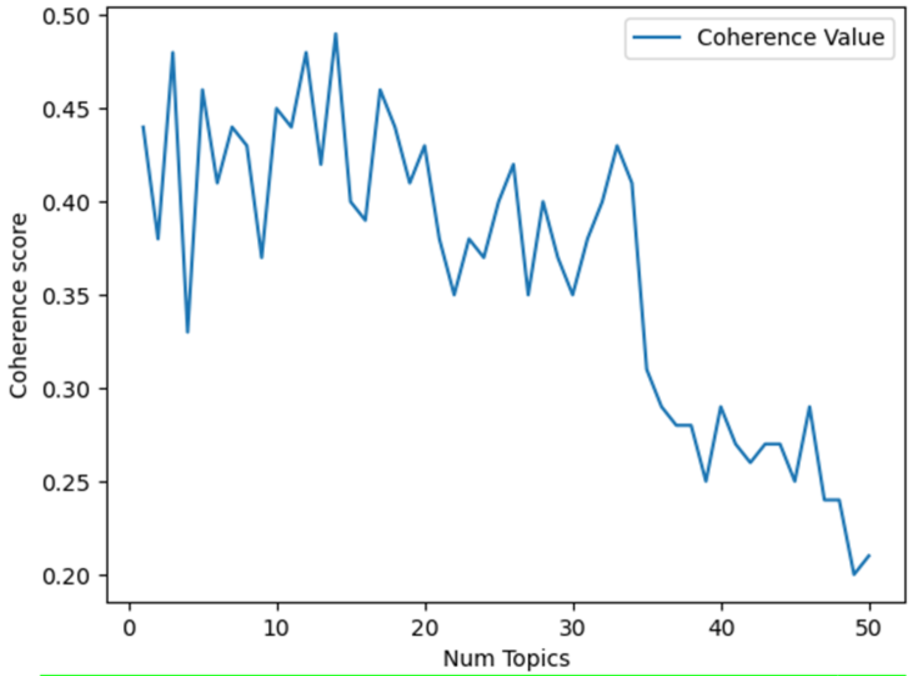
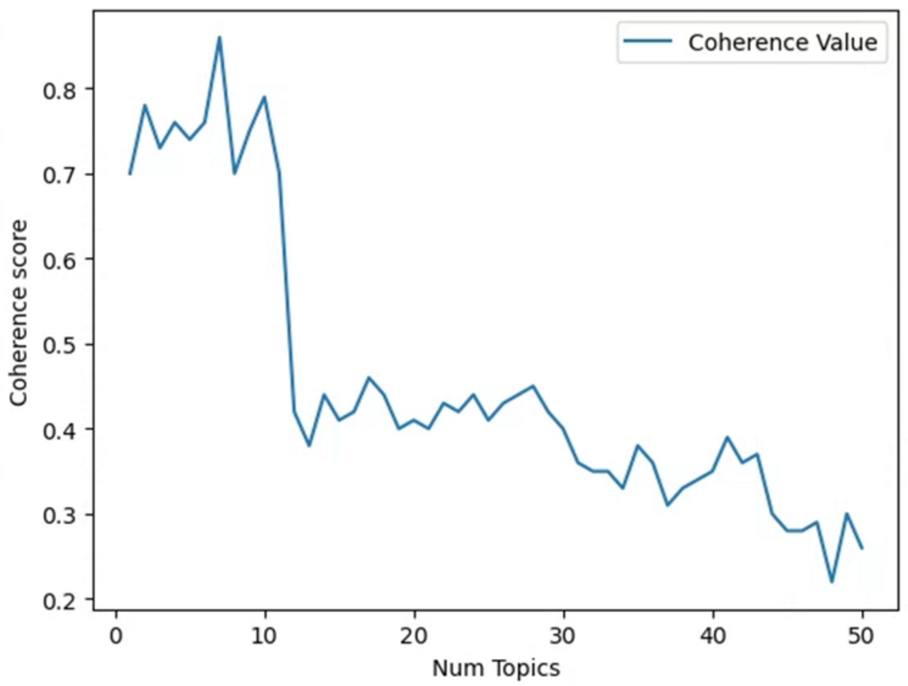




Topic 9-
Perplexity: -6.243973416757705
Coherence Score: 0.826662738335143

Source: Jason Chuang, 2012, Sievert & Shirley (2014)

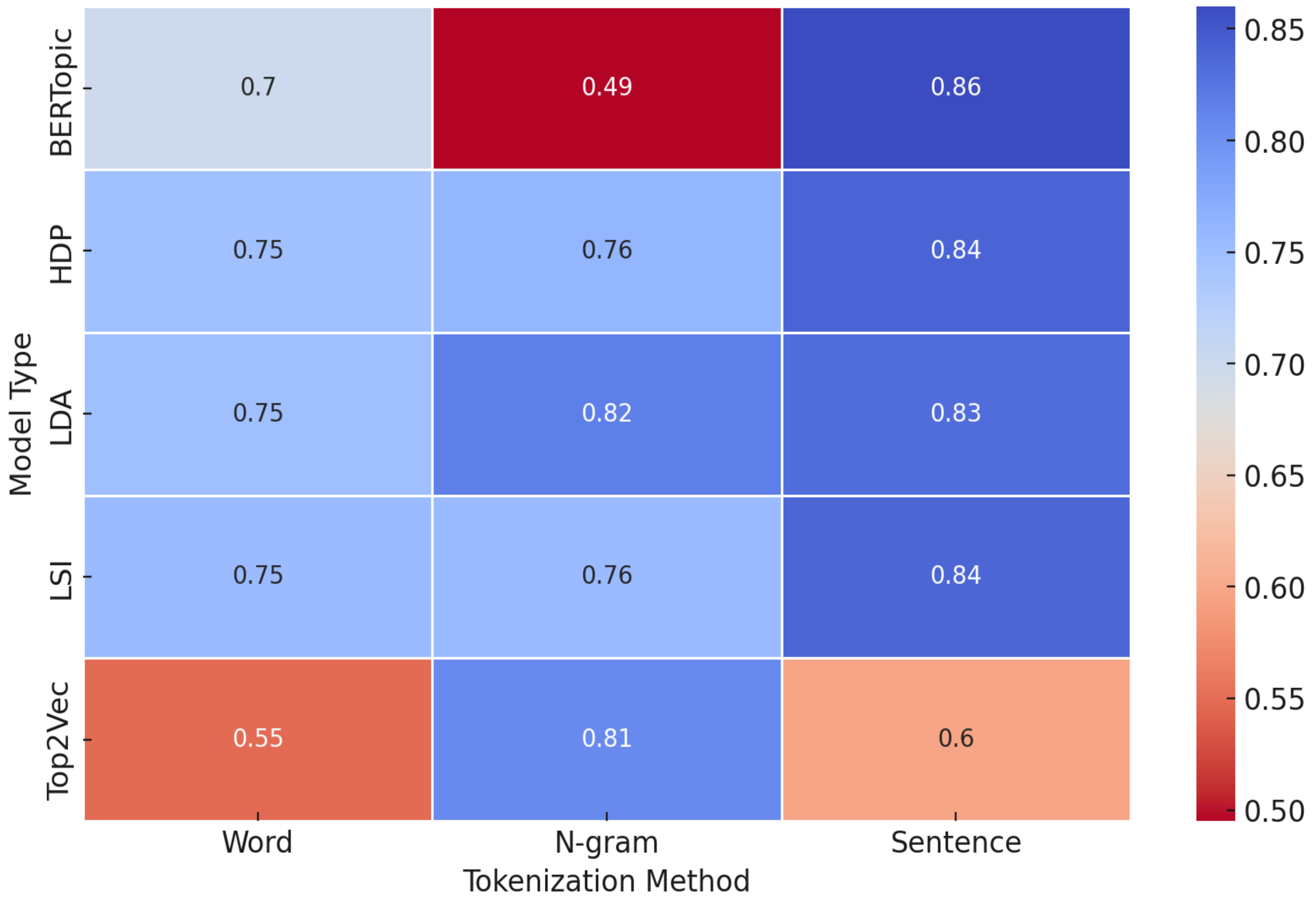
Topic Models comparison Results

Topic Model	Tokenization		
	Word	Bi-gram	Sentence
Top2Vec Topic Model		 Bi-gram Tokenization, Topic 4, Coherence Score: 0.8068	
BERTopic Model		 Sentence Tokenization, Topic 7, Coherence Score: 0.86	

Topic Models comparison Results

Model Type	Tokenization Method	Coherence Score	Number of Topics	Perplexity (Best Score)
LDA	Word	0.7499	20	-27.2774
LDA	N-gram	0.8164	20	-31.0310
LDA	Sentence	0.8320	9	-13.9060
LSI	Word	0.7546	26	N/A
LSI	N-gram	0.7550	24	N/A
LSI	Sentence	0.8381	9	N/A
HDP	Word	0.7478	18	N/A
HDP	N-gram	0.7611	26	N/A
HDP	Sentence	0.8392	9	N/A
Top2Vec	Word	0.5486	11	N/A
Top2Vec	N-gram	0.8068	4	N/A
Top2Vec	Sentence	0.5975	4	N/A
BERTopic	Word	0.65	4	NA
BERTopic	N-gram	0.495	13	N/A
BERTopic	Sentence	0.86	7	N/A

Heatmap of Coherence Scores by Model Type and Tokenization Method



SuRREAL
Sustainable URban & REgional Analysis Lab

BERTopic Topic Model: Sentence Tokenization

Topic 0

Topic 1

Topic 2

Topic 3

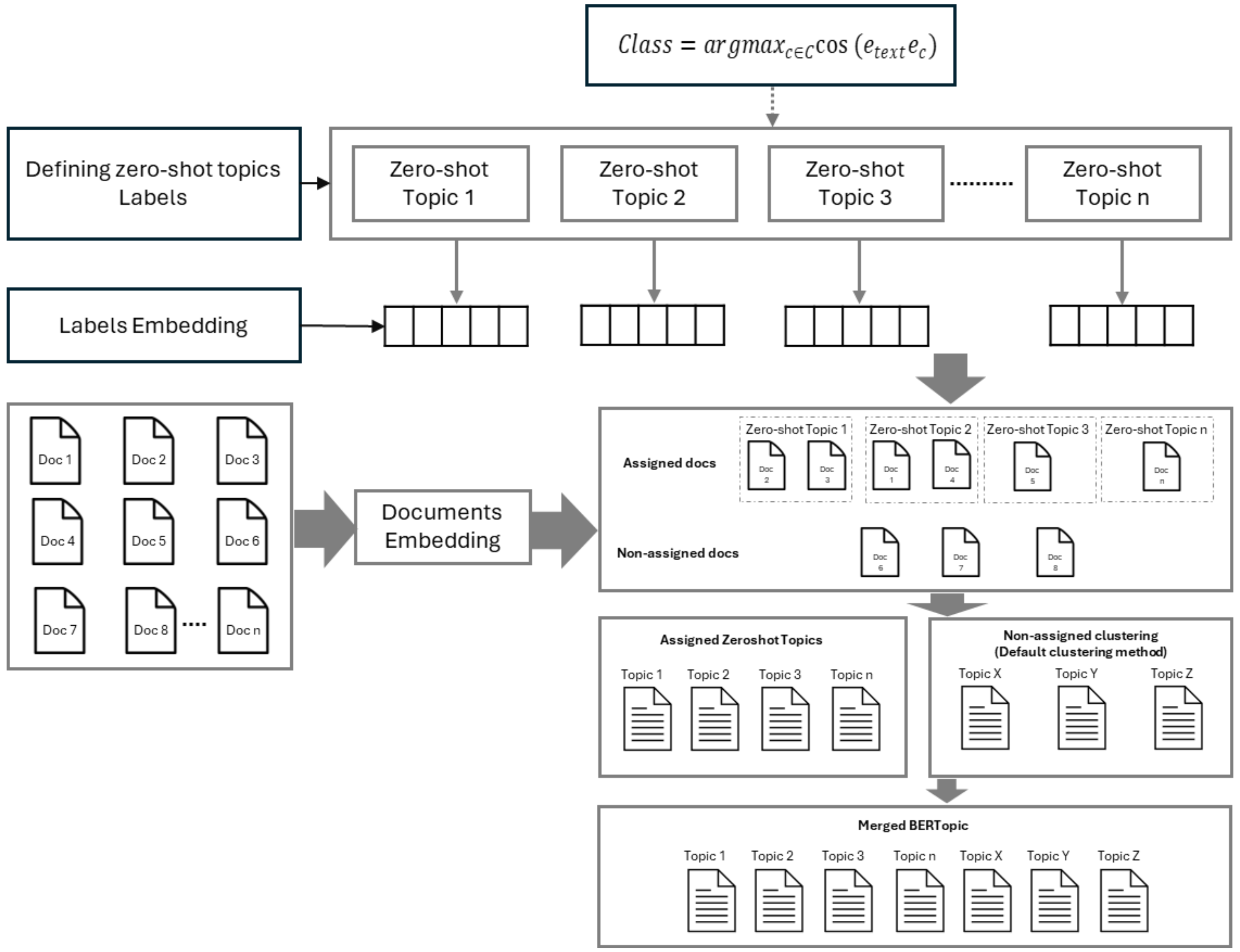
Topic 4

Topic 5

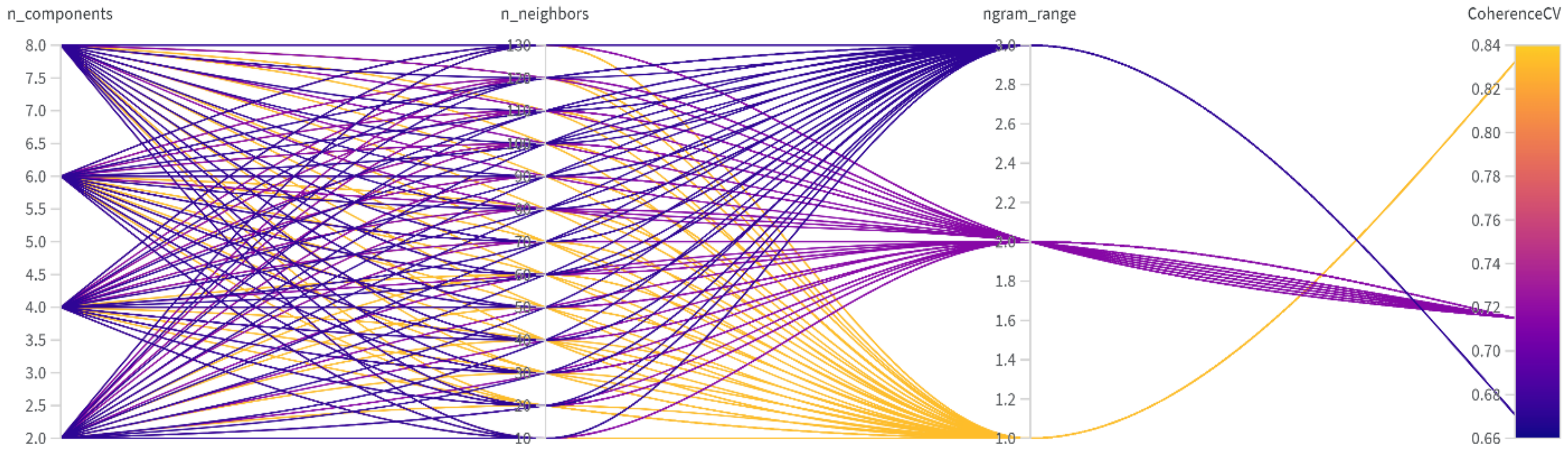
Topic 6

Source: Grootendorst, M., 2022, Jason Chuang, 2012, Sievert & Shirley (2014)

Zero-shot BERTopic Topic Model



- 110 sub classes of Atal Mission for Rejuvenation and Urban Transformation (AMRUT) as labels (Ministry of Urban Development, Govt. of India).
- We experimented with a hyperparameter of clustering and tokenization (Number of components, neighbors, and n-gram range.)
  - For components, we considered 2 to 8 with steps of 2.
  - For neighbors, we considered 10 to 130 with steps of 10.
  - For the n-gram range, we considered words, bi-gram, and tri-gram.
- The highest coherence value observed was **0.8415.**
- We observed that the hyperparameters, values of components, and neighbors were not affecting the model.
- Only the N-gram showed importance in the model, with a negative correlation of **-0.967**.

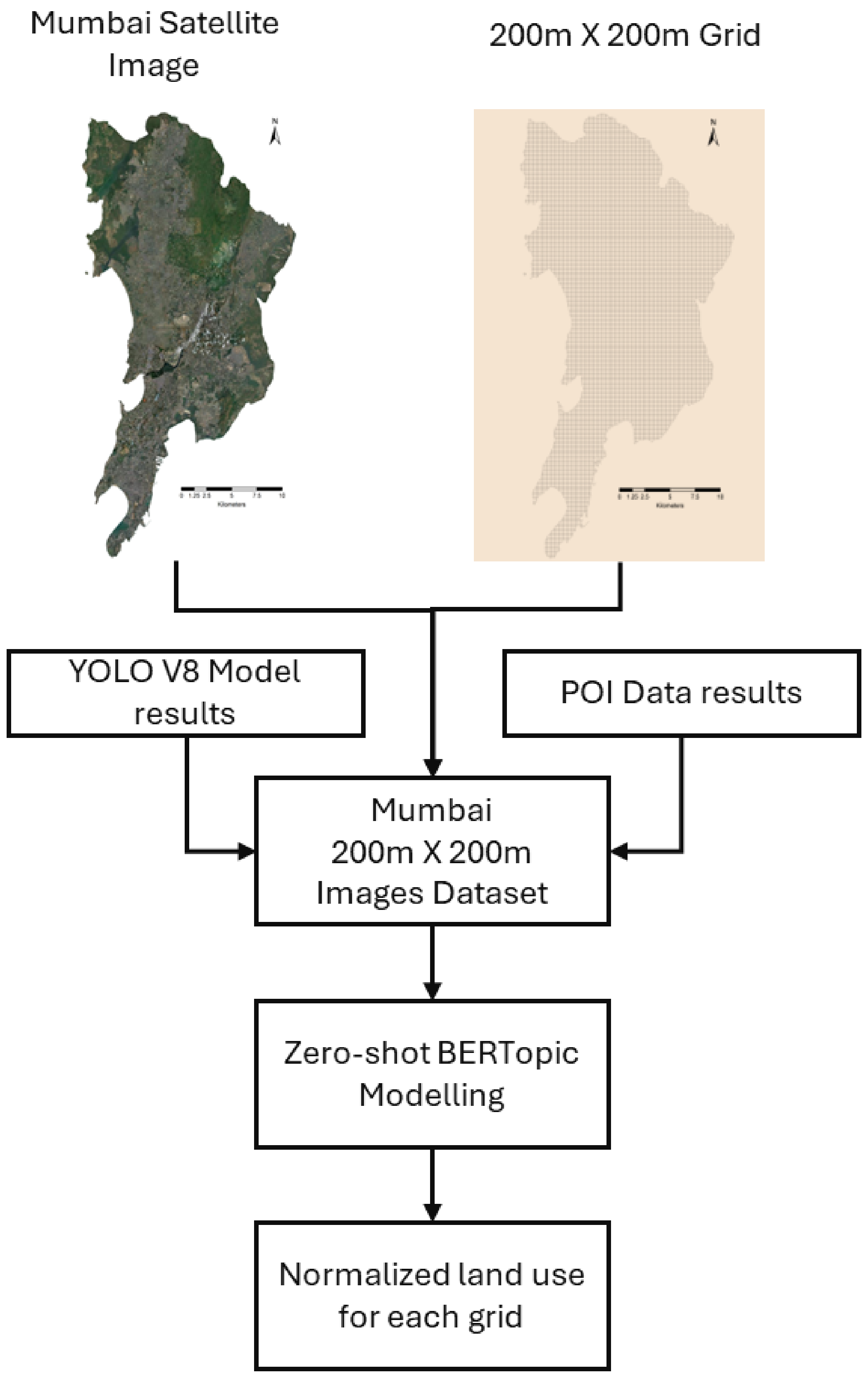


	Topic	Count	Name	Representation	Representative_Docs
0	0	5639	Restaurant	[restaurant, gastropub, diner, burger, food, c...	[Fast Food Restaurant, Fast Food Restaurant, F...
1	1	3685	Retail	[retail, retailer, supermarket, grocery, drugs...	[Retail, Retail, Retail]
2	2	2011	Banks	[bank, banking, loans, finance, atm, , , , ,]	[Bank, bank, Bank]
3	3	1413	Hospital	[hospital, medical, medicine, doctors, laborat...	[hospital, hospital, hospital]
4	4	1274	General Business	[professional, services, business, consultant,...	[Business and Professional Services, Business ...
...	...	...	...	...	...
105	105	1	Electric Power Plant	[utility, company, , , , , , ,]	[Utility Company]
106	106	1	Apartment	[condo, apartment, or, , , , , ,]	[Apartment or Condo]
107	107	1	Nursing Home	[nursing, , , , , , , ,]	[Nursing Home]
108	108	1	Golf Course	[golf, course, , , , , , ,]	[Golf Course]
109	109	1	Race Course	[track, , , , , , , ,]	[Track]

110 Topics assigned in the Zero-shot model (With 75% similarity)
Coherence Value: 0.92

Source: Grootendorst, M., 2022, Jason Chuang, 2012, Sievert & Shirley (2014)

Zero-shot BERTopic Topic Model



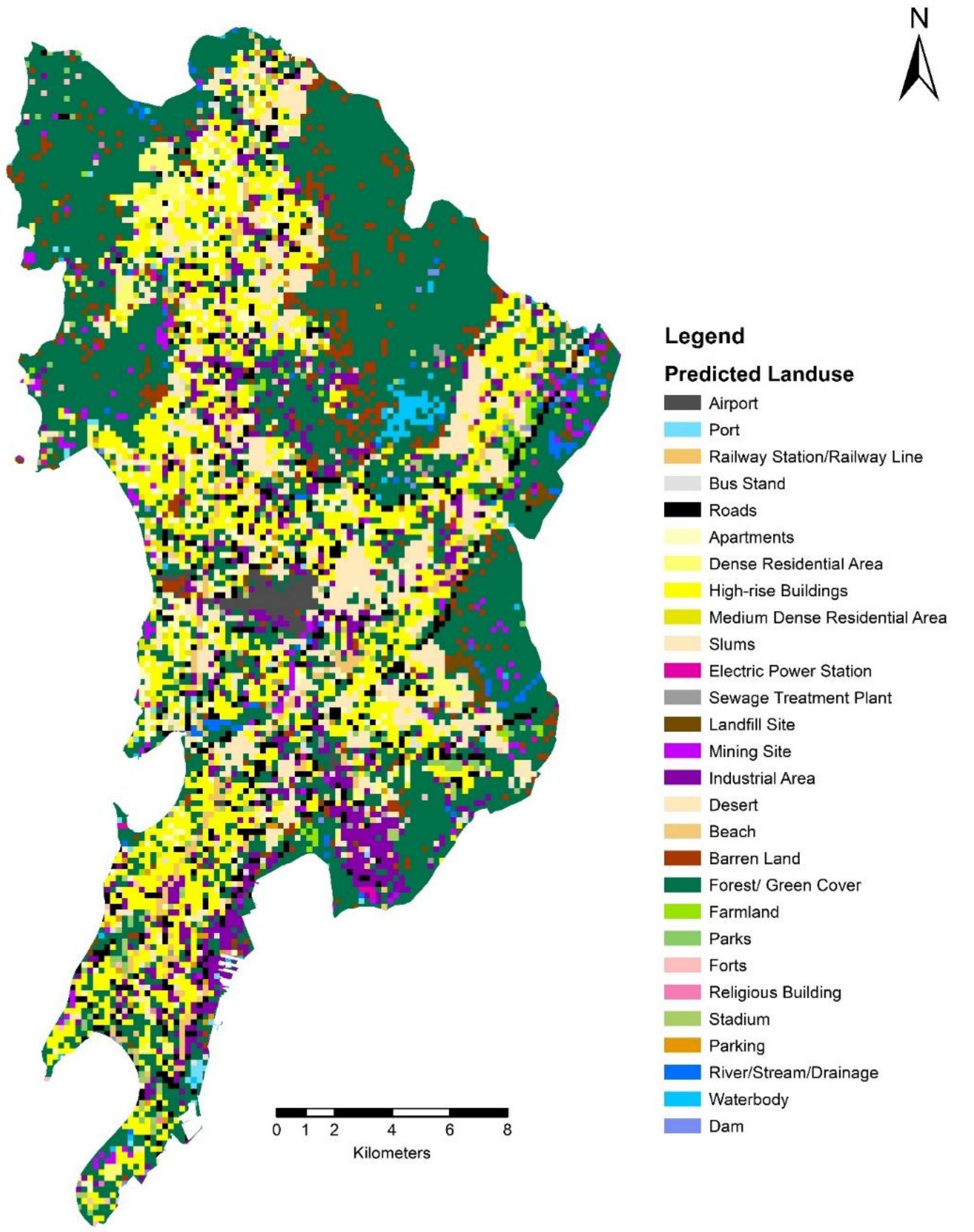
	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Representative_document
0	Forest	0	Forest	[forest, land, parks, mining, river, , , , ,]	[Forest, Forest, Forest]	forest - land - parks - mining - river - - -...	True
1	Roads Industrial Roads Industrial	4	Roads	[roads, industrial, mining, railway, buildings...	[Roads Roads Roads Roads Industrial, Roads...	roads - industrial - mining - railway - buildi...	False
2	High-rise Buildings Apartments Parks Apartm...	1	High-rise Buildings	[buildings, apartments, building, high, roads,...	[High-rise Buildings High-rise Buildings Apa...	buildings - apartments - building - high - roa...	False
3	Forest Barren Land Forest Forest Forest	0	Forest	[forest, land, parks, mining, river, , , , ,]	[Forest, Forest, Forest]	forest - land - parks - mining - river - - -...	False
4	Forest	0	Forest	[forest, land, parks, mining, river, , , , ,]	[Forest, Forest, Forest]	forest - land - parks - mining - river - - -...	True
...	...	...	...	...	...	...	...
11927	Industrial Industrial Industrial Industrial...	7	Railway Station	[industrial, railway, station, airport, buildi...	[Railway Station Industrial Railway Station ...	industrial - railway - station - airport - bui...	False
11928	Forest	0	Forest	[forest, land, parks, mining, river, , , , ,]	[Forest, Forest, Forest]	forest - land - parks - mining - river - - -...	True
11929	Slums Slums Slums Slums Slums Slums Slum...	2	Slums	[slums, buildings, roads, land, , , , , ,]	[Slums Slums Slums Slums High-rise Buildin...	slums - buildings - roads - land - - - - -...	False
11930	Industrial Industrial	3	Industrial	[industrial, building, buildings, , , , , , ,]	[Industrial Industrial Industrial, Industria...	industrial - building - buildings - - - - ...	False
11931	Slums Slums Slums Slums Apartments Slums ...	2	Slums	[slums, buildings, roads, land, , , , , ,]	[Slums Slums Slums Slums High-rise Buildin...	slums - buildings - roads - land - - - - -...	False

Mumbai Dataset Image number

Predicted Land-use objects and POI data

Assigned land-use topics from Zero-shot BERTopic

Results



Limitations:

- Data quality inconsistency in user-generated contributions.
- Limited real-time updates from current data sources.

Future Directions:

- Considering more user-generated data sources.
- Expand the framework to accommodate dynamic urban changes.
- Explore additional AI models for better performance in multi-class classification tasks.

References

1. **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Link
2. **Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D.** (2012). Exploring topic coherence over many models and many topics. *EMNLP-CoNLL*, Link
3. **Sievert, C., & Shirley, K.** (2014). LDAvis: A method for visualizing and interpreting topics. *Workshop on Interactive Language Learning, Visualization, and Interfaces.*
4. **Gandomi, A., & Haider, M.** (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
5. **Assur, M., & Rowshankish, H.** (2024). Real-time Big Data Challenges in Urban Analytics. *Urban Computing and Sustainability*.
6. **Gil, J.** (2022). Crowdsourcing and the role of citizen-generated spatial data in urban studies. *Cities*, 129, 103825.
7. **Harley, J. B.** (1987). The Map and the Development of the History of Cartography. *Imago Mundi*, 38(1), 35-45.
8. **Kain, R. J. P., & Baigent, E.** (1992). *The Cadastral Map in the Service of the State: A History of Property Mapping*.
9. **Waldhoff, G., & Bareth, G.** (2009). Land use/land cover mapping using high-resolution remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*.
10. **Govindu, V., et al.** (2019). Remote Sensing for Urban Land-Use Classification. *Remote Sensing Applications: Society and Environment*.
11. **Gong, P., Li, X., & Zhang, W.** (2019). 40-year retrospective of remote sensing in China. *ISPRS Journal of Photogrammetry and Remote Sensing*.
12. **Liu, Y., et al.** (2018). Classification of Urban Land Use Using Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*.
13. **See, L., et al.** (2016). Crowdsourcing and citizen science: A review of uses in the context of Earth observation. *Remote Sensing*, 8(6), 509.
14. **Wulder, M. A., & Coops, N. C.** (2014). Satellites: Make Earth observations open access. *Nature*, 513, 30–31.
15. **Grootendorst, M.** (2022). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. *arXiv preprint.*
16. **Chuang, J.** (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*
17. **UNEP Annual Report** (2023). *United Nations Environment Programme*.
18. **United Nations** (2023). *World Urbanization Prospects*.
19. **Yu, Y., & Fang, H.** (2023). The impact of urban growth on land use change. *Urban Studies Journal*.

THANK YOU