DOI: https://doi.org/10.38027/ICCAUA2024IN0334

A Comparative Study of Topic Modelling Approaches for User-generated Point of Interest Data

* [1] Ph.D. Candidate **Ravi Satyappa Dabbanavar**, [2] Assoc. Prof. Dr. **Arindam Biswas**
[1 & 2] *Sustainable Urban & Regional Analysis Lab (SURREAL), Department of Architecture and Planning, Indian Institute of Technology Roorkee, India*
E-mail [1]: rs_dabbanavar@ar.iitr.ac.in E-mail [2]: arindam.biswas@ar.iitr.ac.in

Abstract
This study aims to enhance urban planning and management by harnessing the power of machine learning (ML) and big data. We focus on Urban Functional Zones (UFZs), the fundamental units for human socio-economic activities. Our methodology involves compiling Point of Interest (POI) data from various sources for comprehensive analysis. We employ various topic modeling approaches such as Latent Dirichlet Allocation (LDA), Latent Semantic Index (LSI), Hierarchical Dirichlet Process (HDP), and Top2Vec. Our principal results reveal significant differences in the performance and coherence of these models on short text documents. Consequently, our major conclusion is identifying the better-performing topic model for classifying UFZs from POI data. We also explore four text preprocessing steps to optimize the performance of the topic models. This study contributes to the field by providing a nuanced understanding of UFZs, paving the way for future data-driven urban planning and management.
Keywords: Big Data; Machine Learning; Point of Interest; Topic Model; Urban Functional Zone.

1. Introduction

The pace of urbanization is on the rise, and the population of these areas is growing tremendously. Over 50% of the global population lives in urban areas, and that share is expected to increase further in the coming decades (Profiroiu et al., 2020). This is facilitated by several factors: opportunity for better economic prospects, improving health and educational standards, and living conditions (Syaodih, 2019). However, rapid urbanization is also a massive challenge for local bodies in streamlining cities. Urban areas face various problems, such as congestion, pollution, limited availability of infrastructure, and social inequality Amen, 2021; Amen et al., 2023; Amen & Nia, 2020; Aziz Amen, 2022. City management is concerned with these challenges, focusing on sustainability (Lyu et al., 2018).

Land use planning, thus, is an integral part of city management, specifically to current and future city growth and income generation. Land use planning is the land allocation process for residential, commercial, industrial, or recreational functions. This would ensure that land resources are used sustainably and optimally to cater to the growing population's requirements and conserve the environment (Deng et al., 2018). Urban mapping and analysis, therefore, is a comprehensive and complex process because urban growth is dynamic, and the requirements of the urban populations are varied. Hence, urban planning must be updated with changes in patterns of demography and technological improvements to reduce the complications in this process.

In the modern world, there are continuous improvements with the use of technology, which has an increasing population. The technology-driven interaction generates an extensive amount of data, which is very hard to assimilate and process using the traditional approach, and such data is termed 'Big Data' (Paes et al., 2023; Zhao & Zhang, 2020). More specifically, using any technological platform provides a record and reflection of human activities through user-generated big data. This data has been very lucrative in determining and analyzing the urban landscapes (Hall, 2020; Lapointe et al., 2020).

Urban Functional Zones are the basic unit of land use and human socio-economic activities within a city. This plays a vital role in urban planning and management. Correctly classifying and understanding UFZs can provide crucial insights into effective urban management (Liu et al., 2021; Xu et al., 2019). The voluminous significant data users generate and its complexity challenge traditional data processing techniques. This is where advanced machine learning techniques and big data analytics come to the rescue. Topic modeling approaches such as LDA, LSI, HDP, and Top2Vec can effectively analyze and classify UFZs from user-generated POI data (Farid et al., 2021; Schindler et al., 2018). These models are developed to bring out the latent patterns in the data and pull some meaningful information from the text data sets.

Recent studies have proved that ML topic modeling algorithms perform better than traditional models. However, their test was supported by large text data sets. However, the POI data carries small snippets of text data, and it is unsuitable to consider any topic modeling algorithms unthinkingly. Therefore, we meant to evaluate and compare the performance of different topic modeling approaches in classifying UFZs out of POI data from traditional and ML topic modeling algorithms. By this, we look for the most effective model for this purpose. In addition, we further investigate different ways to pre-process text, i.e., tokenization and normalization, to enhance the performance of these topic models (Chen, 2021; Gunko et al., 2021). This research contribution gives a rich understanding of UFZ mapping, paving the way for future data-driven urban planning and management.

Conclusion Our findings will ease the work of urban planners and policymakers by helping them make informed decisions with comprehensive analyses of user-generated big data (Hou et al., 2019). The paper is structured in six

comprehensive sections, beginning with an introduction to the background and identification of the problem in this work in section 1. In section 2, we compare data preparation and different topic models. Section 3 elaborates on the methodology and steps taken in preprocessing text data. Section 4 elaborates on the performance of the topic model and the extraction of Meaning Functional Zones labels for the topics. Section 5 concludes the paper with a summary of the research and contributions to urban studies and planning that have been made. This composition ensures that coherence is maintained as one reads through the paper in which the complexities and innovative dynamics at the junction of UGBD and urban functional zoning are systematically unpacked.

2. Data Preparation and Materials

This study has used user-generated POI data from Mumbai City. The search for sources of POI data was driven by the principle of being open, allowing us to tap into the rich repositories of open data sources. It encompasses the worldwide popular Open Street Map, Foursquare's location data platform, and the ArcGIS Developer's extensive geospatial capabilities. 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). Finally, 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. Then, the POIs from various sources were gathered and prepared for the text preprocessing step.

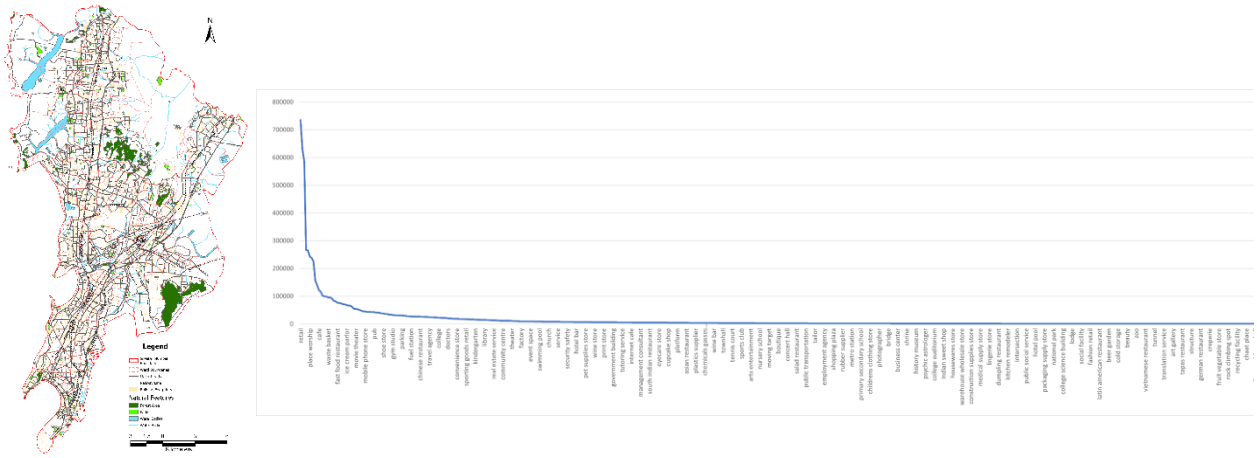

Figure 1. Different categories of places of interest are found in the Greater Mumbai city area.

2.1. Latent Dirichlet Allocation (LDA)

LDA is an unsupervised generative probabilistic model used to uncover the hidden topic structure in a collection of documents. LDA treats every document as a mixture of different topics, and each is again a mixture of words. One can go about it by first picking a distribution over topics for each document and then picking a word from the corresponding topic's distribution over words. One tries to infer the hidden topic structure, which involves per-document topic distributions, per-word topic assignments, and per-topic word distributions. Inference can be done with methods like Gibbs sampling or variational inference. Among many uses, LDA is widely used in text classification, information retrieval, social media analysis, and so on to detect different kinds of patterns and trends in large text datasets. The model can cope with topics on social networks like tweets, topics for group discussion, and other very similar topics (Muzafar et al. A Kundroo & Agarwal, 2020).

2.2. Latent Semantic Index (LSI)

LSI, also known as Latent Semantic Analysis (LSA), is another unsupervised topic model that relies on a technique using linear algebra to extract patterns in the relationship between terms and documents. LSI builds a term-document matrix, with each entry showing the occurrence of a term in a document. It further deconstructs this matrix using singular value decomposition into three matrices: a term matrix, a singular value matrix, and a document matrix. In this process, the dimensionality goes down by maintaining only the top k singular values to keep the most important relationships. It preserves latent semantic structure within the context of terms and documents globally and hence deals with the issues of synonymy and polysemy. Despite its computational intensity, LSI is effective in retrieving and classifying documents and is hence widely used in various applications such as information retrieval, spam filtering, and clustering bibliographic databases (Lan et al., 2021).

2.3. Hierarchical Dirichlet Process (HDP)

HDP is another extension to the LDA model and permits an infinite number of topics, hence more flexibility and scalability over complex datasets. It is a hierarchical model in which a Dirichlet Process on the top level specifies a global distribution over an infinite set of topics. Another DP specifies a distribution over the topics that belong to a

document at the document level, conditioned on the global distribution. This model is constructed so that when more data is presented, it can introduce new topics into the model without setting the number of topics a priori. HDP uses Gibbs sampling or variational inference to infer the topic distribution. The flexible model is amenable to large-scale datasets, video classification, and open-set recognition tasks that employ scalable topic modeling (Terenin et al., 2020).

2.4. Top2Vec

Top2Vec Top2Vec is an unsupervised approach to recent topic modeling that leverages joint document and word semantic embeddings in obtaining topic vectors. Unlike traditional models, Top2Vec does not require the number of topics to be set a priori or the number of steps for stop-words, stemming, or lemmatization. It uses pre-trained models to transform documents and words into a high-dimensional vector space, for instance, Doc2Vec or Word2Vec. Clustering of these vectors is then performed to obtain the topic vectors, where each topic is represented by the centroid of that vector in the vector space. Additionally, the number of topics is set automatically by the density of the vector space to provide a more natural and interpretable way to represent the topics. This model has shown strong performance in extracting coherent and informative topics from large text corpora and, as such, is well-suited to applications in social media analysis and customer service chat modeling (Angelov, 2020).

3. Methodology

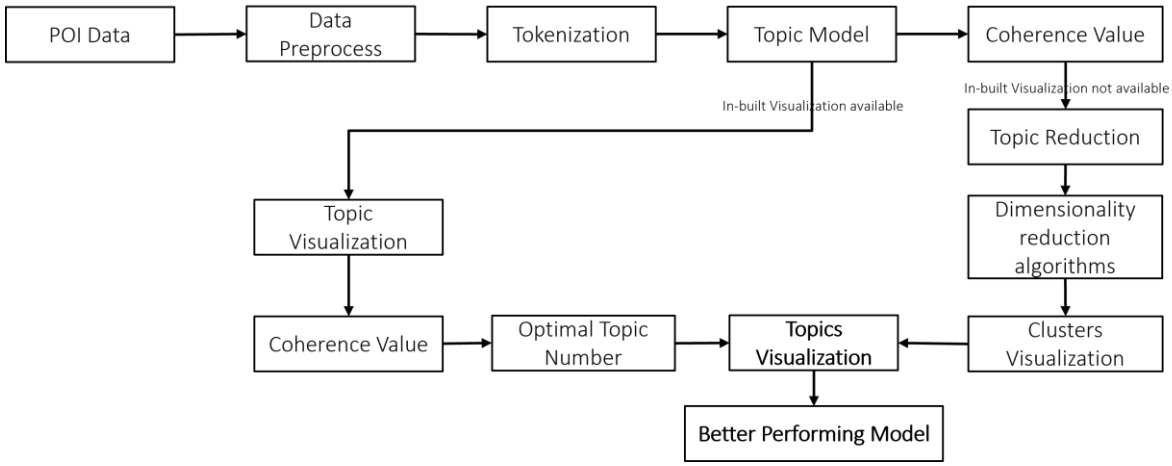

Figure 2. Structure of the Study (Developed by Authors).

3.1. Text Preprocessing

After collecting POI data, the extracted text data underwent four pre-processing stages for cleaning. We selected a technique to process the text data properly in each case, depending on the situation. Below explains the steps:

1. Transformation: This stage encompasses several subprocesses we applied to standardize the text data. The first task was converting all text in the documents of each POI to lowercase to make the text data uniform and hence become case-insensitive, thereby simplifying the rest of the processing steps. The second operation was to remove accents from characters to avoid differences that will arise between different representations of the same word, like converting "café" into "cafe." Also, we filter URLs within the text because these URLs often do not add meaningful information to the topic modeling process.
2. Filtering: This step cleans the text data by removing elements that may interfere during analysis. Words that are usually used and do not add any meaning to the content of the text are termed stopwords, such as "and," "the," and "and is," along with others. Numbers are also customarily removed from the text data since they are generally of little semantic value unless there is a specific requirement for the same. Next is regular expression filtering, done by removing unwanted characters or sequences in the text. This involves special characters, punctuation marks, and other non-alphabetic and non-numeric elements.
3. Tokenization: In this step, tokenization refers to breaking down text data into smaller units. Word tokenization splits the text into words or tokens, which is a crucial step in generating input for the topic models that analyze the frequency and co-occurrence of the tokens. That sentence tokenization must be done based on requirements; it is splitting the text into sentences. This becomes more relevant in the context of the models considering sentence structure and context within more extensive texts.
4. N-gram Range: The last pre-processing step is N-gram. The N-gram range was configured to create a broader context in the text data collected. Instructions were given to consider any N-gram range restricted to single words or unigrams and sequences of bigrams and trigrams. Such an approach helps catchphrases and expressions whose meanings are much more specific than the presentation of the individual words. For instance, the phrase "machine learning" as a bigram is much more informative than the words taken

separately: "machine" and "learning." This summary demonstrates how considering different N-gram ranges within the POI data allows one to understand the context and its nuances further.

These preprocessing steps have indeed made the text data from POIs clean, standardized, and appropriately segmented to become ready for further effective topic modeling. Such preparations are essential for correct and meaningful results in further analyses with different topic modeling approaches, such as LDA, LSI, HDP, and Top2Vec.

3.2. Topic Modeling and Evaluation

Once the preprocessed data was ready, the next phase was topic modeling. The process relied on coherence values for every model evaluated to determine the best number of topics. The following is the methodologies diagram, which shows workflow—data preprocessing, model evaluation, and visualization. We have taken pre-processed data on POIs and have applied LDA, LSI, HDP, and Top2Vec techniques for the topic modeling exercise. We tested for several topics from 3 to 50 for each model within this wide range to ensure we picked the potential topic structure. For every number of topics, an accompanying coherence value is calculated. The coherence values of topics are used to gauge how interpretable and meaningful the topics are by quantifying the degree of similarity between high-scoring words in a topic. More interpretable topics are often associated with a higher number of coherence values.

We have picked the number of topics that returned the highest coherence value for each topic modeling approach. It is critical because the correct number of topics ensures the model will catch the data's most relevant and distinct themes. Afterwards, the best number of topics chosen is used for further analysis and visualization. Usually, topic models come with built-in visualization functions that allow for direct observation and interpretation of the topics. Such direct visualizations make it possible to understand how the topics are distributed and related to each other and to get insight into the latent themes and patterns of the documents under study.

We have directly visualized the topics for models with built-in visualization functions to see their structure and coherence. Not all have this feature. We used t-SNE (t-distributed Stochastic Neighbor Embedding), an unsupervised non-linear dimensionality reduction algorithm for models without built-in visualization capabilities. A popular unsupervised non-linear dimensionality reduction technique, t-SNE, is used for projecting high-dimensional data into a lower-dimensional space to enable easier visualization and comprehension of the structure and clustering of the topics. It is crucial to provide value for clustering and find out how topics get grouped and their relationships. Visualizing topics allows us to see how distinct or overlapping topics are, providing a further understanding of thematic structure in POI data.

The secondary analysis included additional analysis of the topic clusters identified by each modeling approach. The word cloud visualizes the distribution of the most frequent words in a specific topic. In a word cloud, the size of each word indicates its frequency of use or importance. We started making word clouds by first computing the frequency of words appearing in documents assigned to each topic. Word clouds were then generated for each topic produced. It is a very intuitive and super-fast way to understand the main themes. The key terms were identified to identify the main topics (or themes) for each cluster, as in the review of the main topics. We then looked at the keywords in the word clouds that would indicate the main topics and subjects for the cluster. For example, if a word cloud for a topic has many instances of the words "restaurant," "menu," and "dining," we may infer that the topic is related to food and dining establishments. Comparison of word clouds for the topics across different topic models (LDA, LSI, HDP, Top2Vec) helps us ensure the consistency and distinguishability of the extracted topics between models and, therefore, to decide the most coherent and interpretable topics.

Such preprocessing and modeling steps ensure that derived information from the text data of POIs is appropriately prepared and analyzed, which brings credible and meaningful results. The approach is thus holistic and allows the identification and visualization of latent themes, which is very much needed to make valuable insights for urban planning and management available. Empowering urban planners and policymakers with an in-depth understanding of topics and their inter-relationships leads to the development of evidence-based decisions, improving the efficiency and effectiveness of urban management strategies.

However, The study focuses on point of interest (POI) data, which consists mainly of brief text snippets. Therefore, the results will only be relevant to small text data.

Analyzing small text snippets in the POI data enables the in-depth study of this detailed example. Thus, the capacity of the outcomes to be employed on like data types. This is a crucial point to note when interpreting results from the study or applying its methods in other research areas and practical domains.

4. Results

Based on the methodology, coherence scores were calculated for the number of topics, from 3 to 50, on every single model provided for each topic modeling approach: LDA, LSI, HDP, and Top2Vec. The analysis was comprehensive—across different ways of tokenization and n-gram ranges—to make sure we received the results robustly. Among the selected topic models, only LDA had topic visualization capability, and for other topic models, we utilized t-sne visualization.

4.1. LDA Topic Model Performance.

In this case, the LDA model with basic unigram or word tokenization produced the highest coherence measure recorded at 0.75 for the twentieth topic. Besides the coherence measures obtained, the model had less perplexity (-27.28). It is a statistical measure that helps understand how good a sample prediction is in terms of probability model. In particular, LDA helps one understand how well the model describes the documents present in the dataset. A lower perplexity value denotes that the model is the right one.

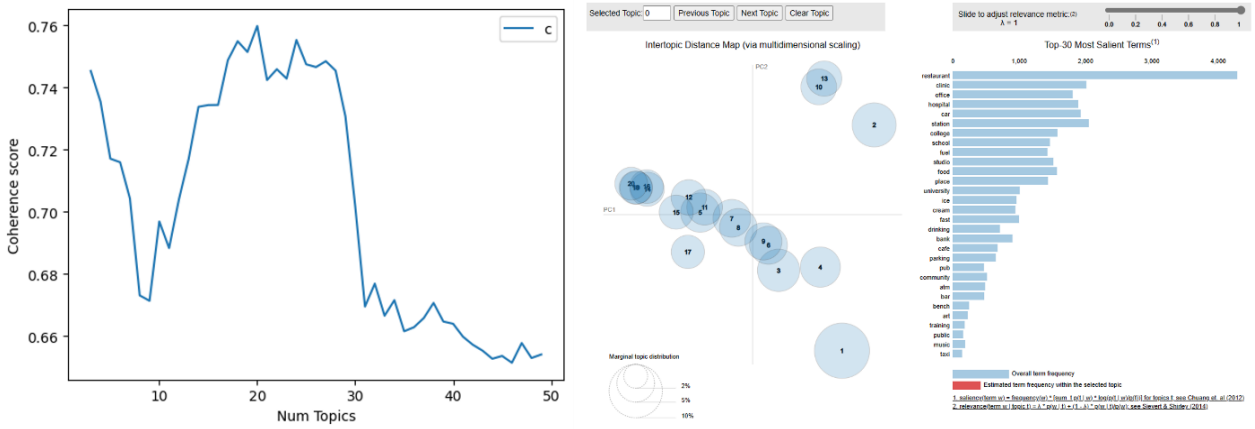

Figure 3. LDA Topic Model with word tokenization or uni-gram.

Further, data with bi-gram filtering showed the highest coherence value of 0.75 with a perplexity value of -27.28 for topic 20.

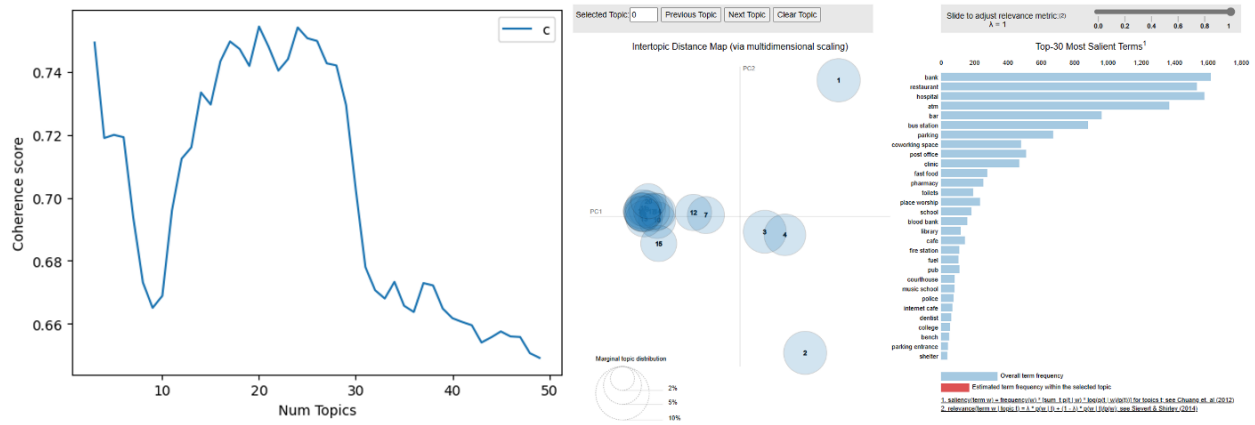

Figure 4. LDA Topic Model with bi-gram.

Meanwhile, data with bi-gram filtering showed the highest coherence value of 0.82 with a perplexity value of 6.24 for topic number 6.

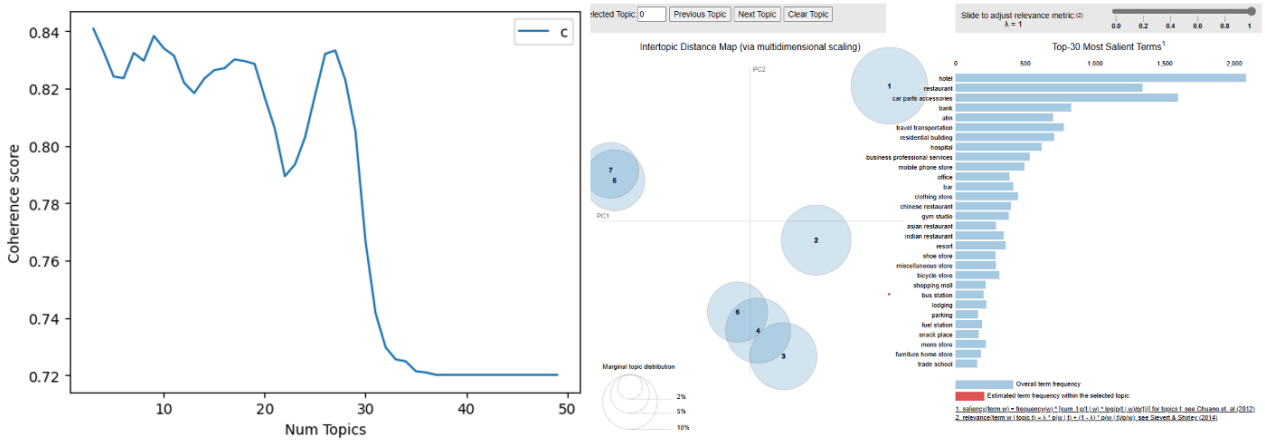

Figure 5. LDA Topic Model with sentence tokenization.

The overall performance of the LDA model with sentence tokenization was superior, as evidenced by its highest coherence value and lowest perplexity value. This suggests it is a better model for this task, providing more accurate

and meaningful results. We have represented the topics and documents present in each topic through the word cloud in Figure 6.

Topic 0 focuses on banking and worship services, highlighted by terms such as "bank," "place of worship," "fuel station," "gym," and "pharmacy." This indicates areas dedicated to financial services, religious activities, and essential amenities like fuel and health. Topic 1 centers on automotive and financial services, with key terms like "financial service," "automotive repair," "fast food," "cafe," and "bar." This suggests zones for financial transactions, car services, and quick dining options. Topic 2 is characterized by health and residential services, featuring terms such as "hospital," "residential building," "university," and "library," representing healthcare, housing, and educational institutions. Topic 3 emphasizes dining and healthcare services, with prominent terms like "restaurant," "clinic," "healthcare," and "medical center," indicating areas focused on food and health services. Topic 4 is related to retail and professional services, highlighted by terms such as "retail," "business professional services," "restaurant," and "transportation," suggesting zones dedicated to commerce, professional activities, and dining. Topic 5 centers on beauty, health, and automotive services, with key terms like "clinic," "car parts," "beauty," and "health," indicating areas for medical care, vehicle maintenance, and personal grooming.

Figure 6. Word cloud for all six topics from LDA sentence tokanization.

4.2. LSI Topic Model Performance.

In this case, the LSI model with basic unigram or word tokenization produced the highest coherence measure recorded at 0.75 for the twentieth topic. However, the LSI does not support perplexity value calculation. Therefore, the LSI model's coherence value was the only performance evaluation matrix. After getting a suitable topic number, we manually used t-sne to visualize and cluster similar topics. Below Figure 7 Shows the coherence value of all 50 and visualization topics and clusters for the 26th topic.

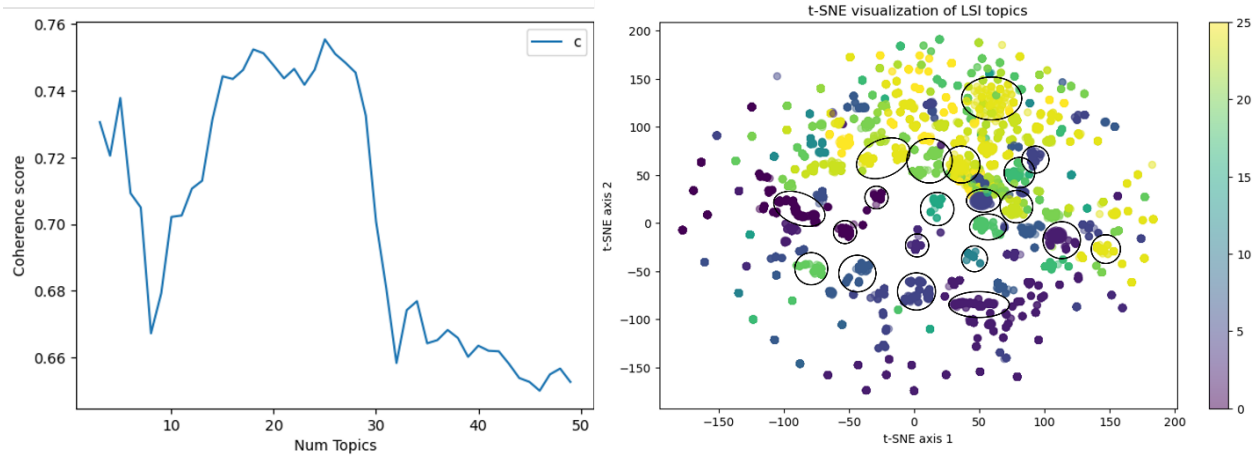

Figure 7. LSI Topic Model with word tokenization or uni-gram.

Similarly, for LSI with a bi-gram range, the model produced the highest coherence value of 0.755 for topic 24. Figure 7 shows the same with visualization. For the LSi model with the sentence tokenization model, the highest coherence value was 0.84 for the reduced topic number to 9. Figure 9 shows the visualization for topics and clustering for nine topics of LSI sentence tokenization.

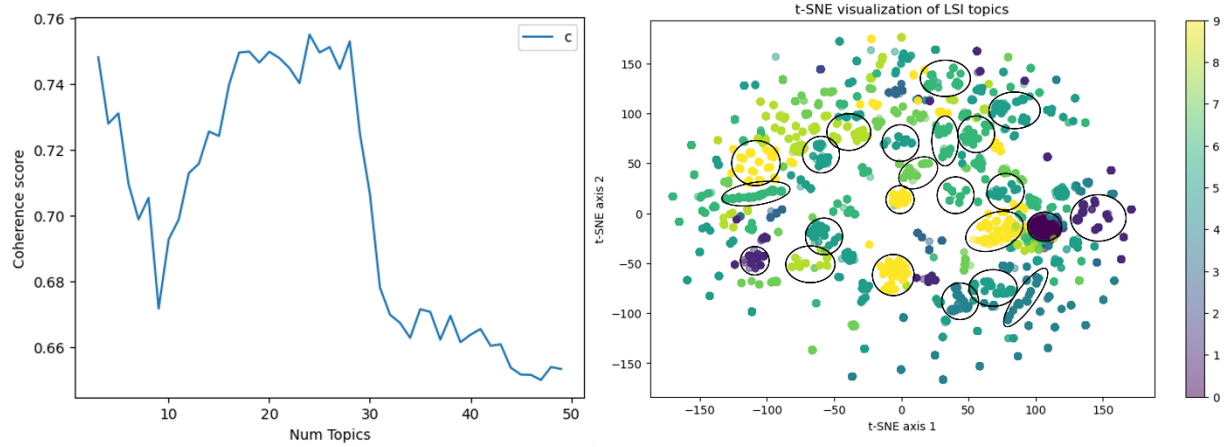

Figure 8. LSI Topic Model with bi-gram range.

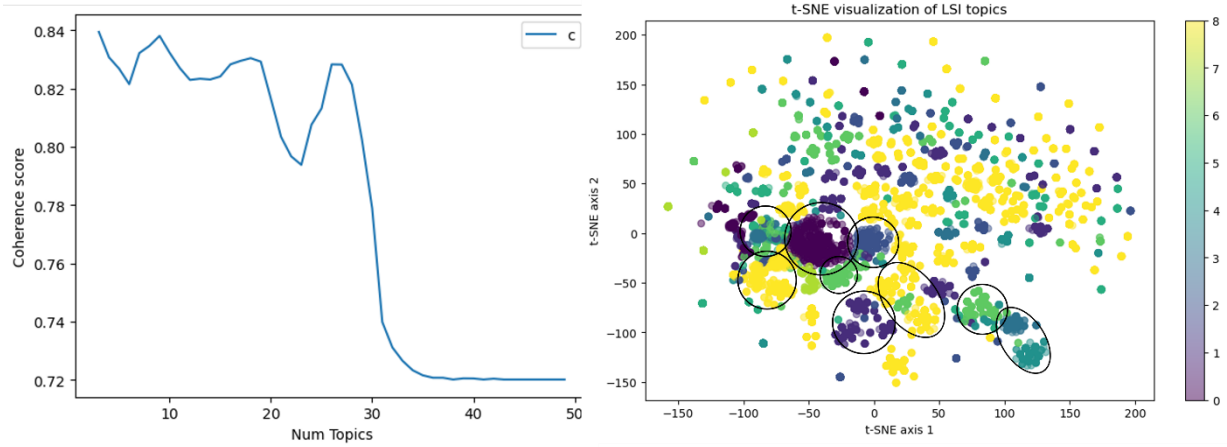

Figure 9. LSI Topic Model with sentence tokenization.

Figure 10. Word cloud for all nine topics from LSI sentence tokanization.

The highest coherence score of the LSI model with sentence tokenization indicates that it performed better overall. This implies that it is a more suitable model for this specific task, yielding more precise and significant outcomes. In Figure 10, we have created a word cloud representing the subjects and documents associated with each topic.

Topic 1 focuses on retail and professional services, highlighted by terms such as "retail," "business professional services," "financial service," and "restaurant," indicating areas dedicated to commerce, professional activities, and dining. Topic 2 centers on educational and hospitality services, with key terms like "high school," "clinic," "hotel," "engineer," and "kindergarten," suggesting zones for education, healthcare, and accommodation. Topic 3 is characterized by high school and convenience stores, featuring terms such as "high school," "food court," "donut shop," and "convenience store," representing educational institutions and small retail outlets. Topic 4 focuses on recreational and food services, with terms like "ice cream," "BBQ joint," "playground," and "park," highlighting leisure and dining facilities. Topic 5 emphasizes a mix of convenience and educational services, with prominent terms such as "convenience store," "high school," "restaurant," and "parking entrance," indicating areas that blend retail, education, and accessibility. Topic 6 is related to performing arts and grocery services, highlighted by terms such as "performing arts," "grocery store," "theatre," and "venue," suggesting zones dedicated to entertainment and food shopping. Topic 7 focuses on repair services and financial access, featuring terms like "computer repair," "ATM," "service," and "clinic," highlighting technology repair and healthcare services. Topic 8 emphasizes hospitality and healthcare, with critical terms like "hotel," "clinic," "ATM," and "restaurant," representing accommodation, medical services, and dining. Topic 9 centers on personal care and food services, with terms such as "hair salon," "cupcake shop," "clinic," and "construction", indicating zones for beauty, dining, and building services.

4.3. HDP Topic model performance

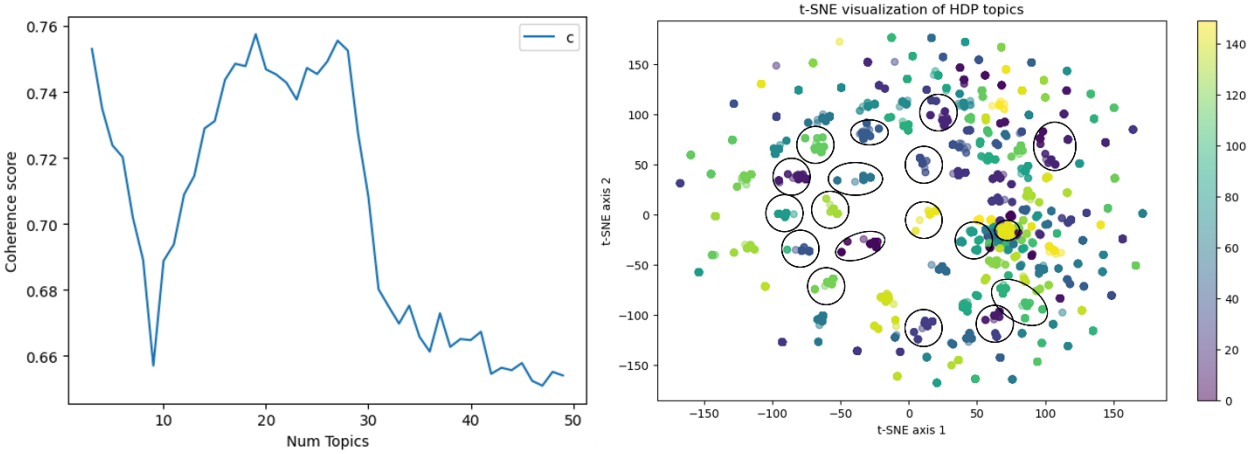

Figure 11. HDP Topic Model with word tokenization or uni-gram.

In this case, the HDP model with basic unigram or word tokenization produced the highest coherence measure recorded at 0.75 for the eighth topic. Similar to LSI, HDP also does not support perplexity value calculation. Therefore, the coherence value was also the only performance value matrix for the HDP model. After getting a suitable topic number with the highest coherence, we used t-sne to visualize the topics and manually clustered similar topics. Figure 11 Shows the coherence value of all 50 and visualization topics and clusters for the 18th topic.

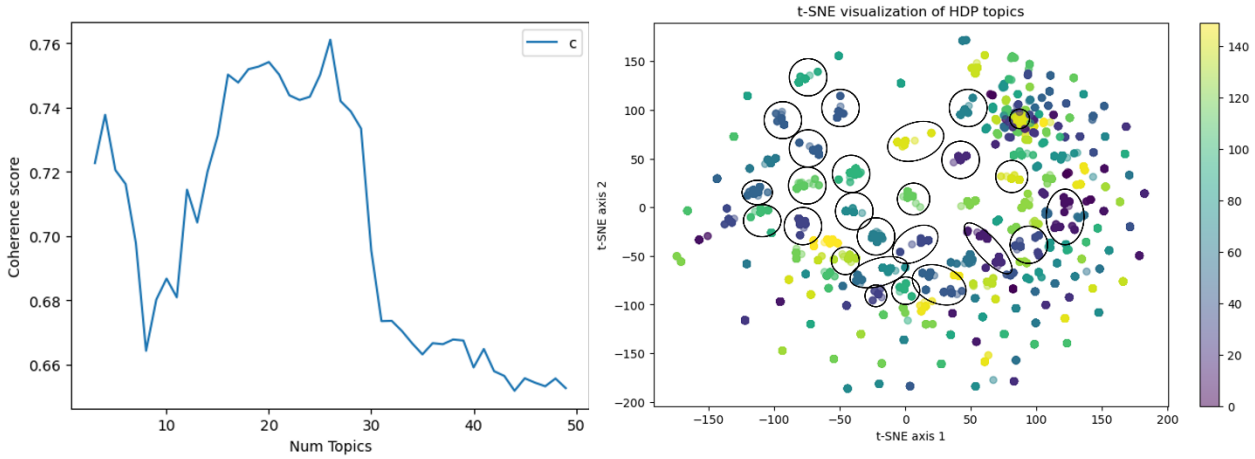

Figure 12. HDP Topic Model with bi-gram range.

The HDP bi-gram range model showed a value of 0.76 for topic number 26. Figure 11 shows the t-sne visualization for the same. The clusters in the model had fewer documents on all the topics and had many outliers. However, the sentence tokenization model showed a coherence value of 0.84 for topic number 9. Figure 13 Shows that for the HDP sentence tokenization, through visualization, we could observe that all the clusters of topics had a close relation with each other and had more outliers than previous HDP-filtered topic models.

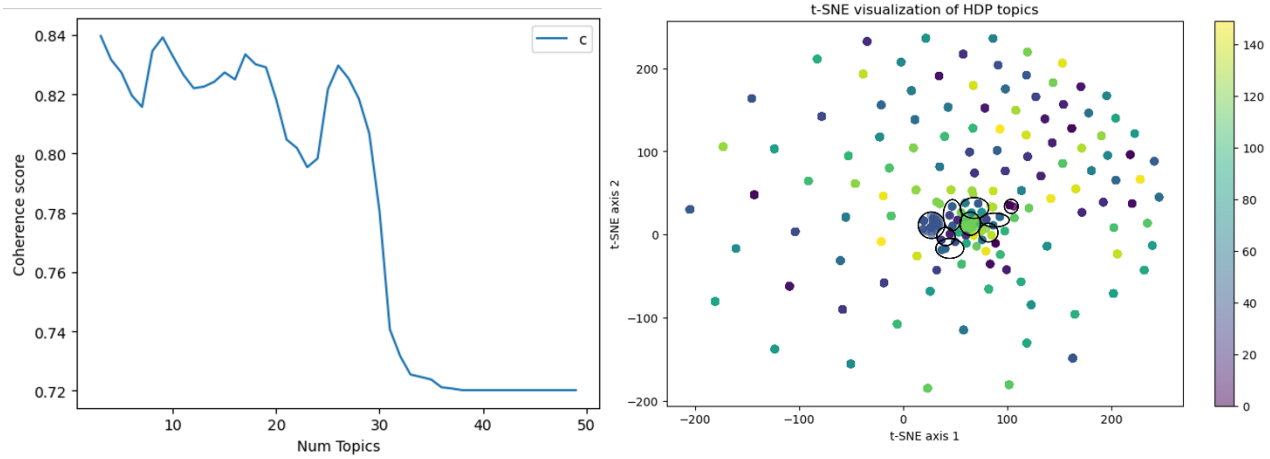

Figure 13. HDP Topic Model with sentence tokenization.

Figure 14. Word cloud for all nine topics from HDP sentence tokanization.

Like the above traditional models, the HDP model with sentence tokenization achieved the highest coherence score and performed better. This implies that it is a more suitable model for this specific task, yielding more precise and significant outcomes. In Figure 14, we created a word cloud that represents the subjects and documents associated with each topic for the HDP sentence tokenization topic model.

Topic 1 focuses on educational and community services, highlighted by terms such as "private school," "community center," "college," "wine bar," and "church," indicating areas dedicated to education, community activities, and social gatherings. Topic 2 centers on commercial and educational buildings, with key terms like "warehouse," "engineering building," "casino," "gift store," and "school," suggesting zones for commerce and higher education. Topic 3 is characterized by retail and event spaces, featuring terms such as "liquor store," "jeweller," "event space," "photography studio," and "park," representing shopping and entertainment venues. Topic 4 focuses on public and educational services, with terms like "garden center," "science museum," "stable," "campaign office," and "college," highlighting educational institutions and public amenities. Topic 5 is centered on retail and educational spaces, with prominent terms such as "college," "public art," "luggage store," and "water treatment service," indicating areas that blend retail and public services. Topic 6 emphasizes cultural and administrative services, with terms like "cultural center," "human resources agency," "garden center," "career counsellor," and "distillery," suggesting zones dedicated to cultural activities and administrative functions. Topic 7 is related to educational and public facilities, highlighted by terms such as "college," "bench," "public art," and "luggage store," indicating educational institutions and public art installations. Topic 8 focuses on parks and recreational areas, featuring terms like "Latin American restaurant," "boat," "ferry," "publisher," and "fire station," highlighting leisure and emergency services. Topic 9 centers on transportation and publishing services, with key terms like "boat," "ferry," "publisher," "beer," and "travel agency" representing transportation hubs and publishing services.

4.4. Top2Vec Topic model performance

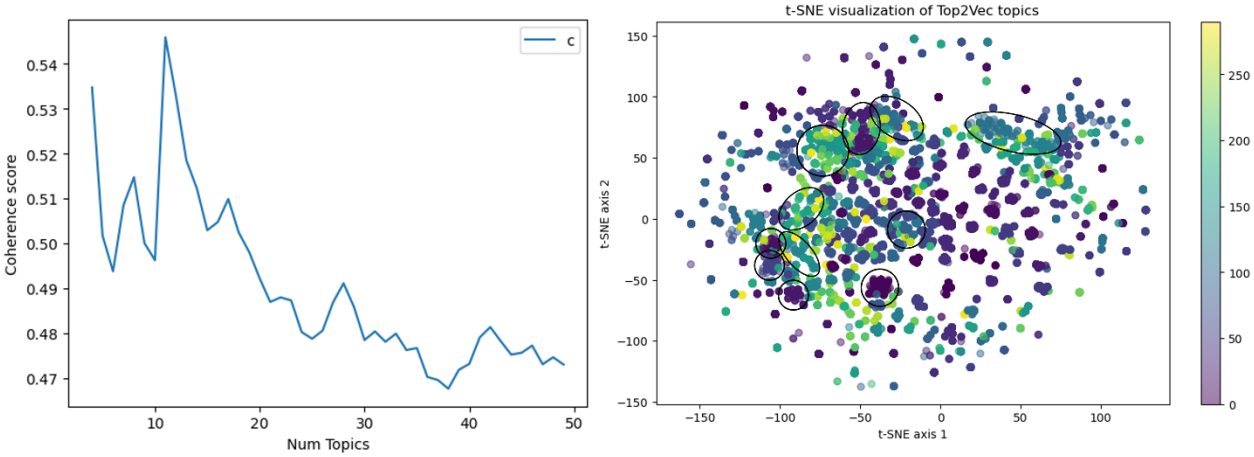

Figure 15. Top2Vec Topic Model with word tokenization or uni-gram.

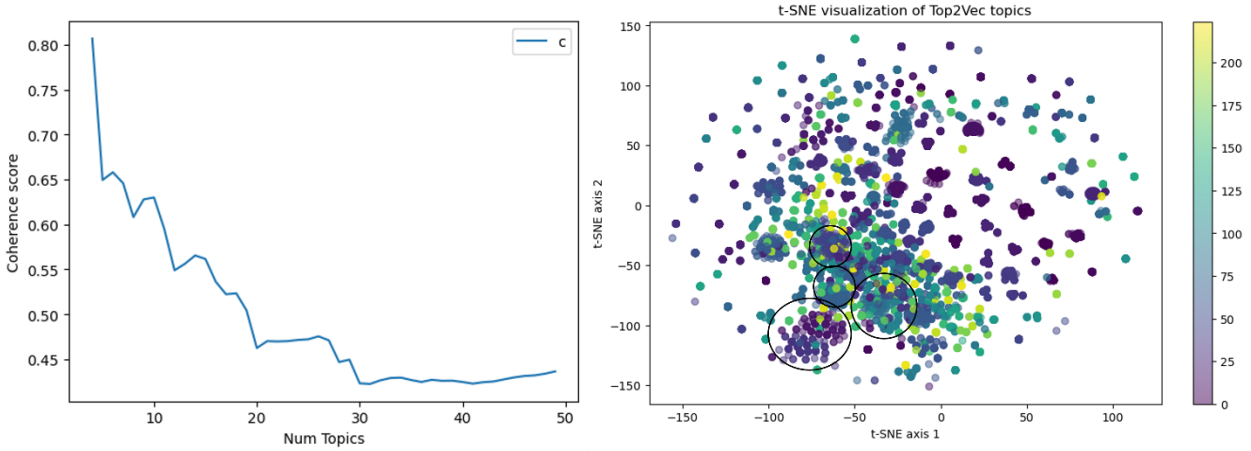

Figure 16. Top2Vec Topic Model with bi-gram range.

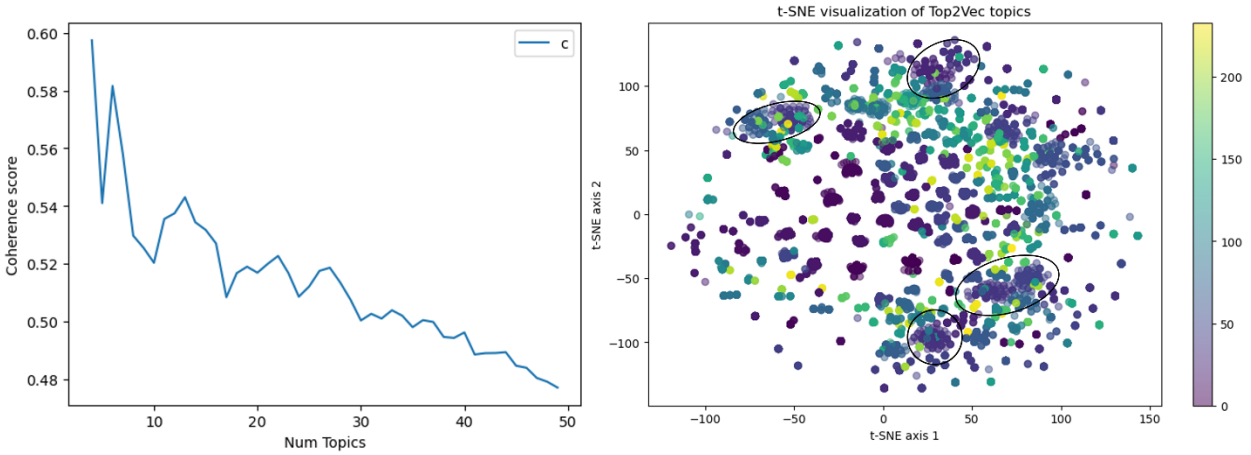

Figure 17. Top2Vec Topic Model with sentence tokenization.

In the case of the Top2Vec model, unlike previous models, Top2Vec is an ML topic modeling algorithm. Top2Vec topic model does not have the option to provide the topic model number. Therefore, initially, the model provided us with 119 topics, the maximum number of topics it could provide. Further, with iteration to calculate the coherence value with basic unigram or word tokenization, the model produced the highest coherence measure recorded at 0.55 for the eleventh topic. Like traditional topic models, Top2Vec does not support perplexity value calculation. Further, We have utilized the topic reduction function in the Top2Vec model and reduced the topics to 11. Figure 15 shows the t-sne visualization for the Top2Vec model with uni-gram or word tokenization. This model also provided us with many outliers.

Following the same methodology, we have calculated the coherence value for the bi-gram tokenization Top2Vec model. This provided us with a value of 0.81 for the Four topics. Applying the topic reduction function to the model and visualizing it with t-sne, this model in Figure 16 Showed more outliers than word tokenization.

The top2Vec model with sentence tokenization through iteration for coherence values showed the highest coherence value of 0.60 for the four topics. However, this model in the Figure 17 Showed fewer outliers compared to top2vec word and bi-gram range models.

Figure 18 Shows the frequency and significance of documents in all four topics through word cloud for the Top2Vec sentence tokenization. Topic 1 focuses on retail and stores, highlighted by terms such as "retail," "store," "dealership," "trade," "supplier," "shopping," and "mall," indicating commercial retail spaces like shopping malls, grocery stores, and pharmacies. Topic 2 centers on business and professional services, with key terms like "services," "business," "professional," "agency," and "department," suggesting areas dedicated to professional agencies, office spaces, and restaurants. Topic 3 is characterized by banking and financial services, featuring terms such as "bank," "banking," "finance," "ATM," and "office," representing financial institutions and related services. Topic 4 focuses on food and dining establishments, with terms like "restaurant," "cafe," "food," "diner," "bakery," and "pizzeria," highlighting diverse dining options. These word clouds provide a quick and intuitive understanding of the primary functions and activities within different urban zones, aiding urban planners and policymakers in effective planning and management.

Figure 18. Word cloud for all four topics from Top2Vec sentence tokanization.

5. Discussions

Sentence tokenization consistently outperforms word and N-gram tokenization for every model applied to the LDA, LSI, and HDP models. This, in turn, will reiterate the importance of methods of tokenization selection to improve the performance of topic models. Sentence tokenization mainly retains context within a sentence, giving a rich representation of data, which in turn leads to more coherent topics that are understandable. The LDA and HDP performed better on the coherence score and showed perplexity results, which generally found a better model through these metrics. This conclusion aligns well with the generally accepted understanding in the field, which states that the LDA and the HDP are adapted to capturing thematic structures underlying text data.

Table 1. Coherence and Perplexity values for all the topic models with different tokenization and N-gram ranges.

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	10	-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

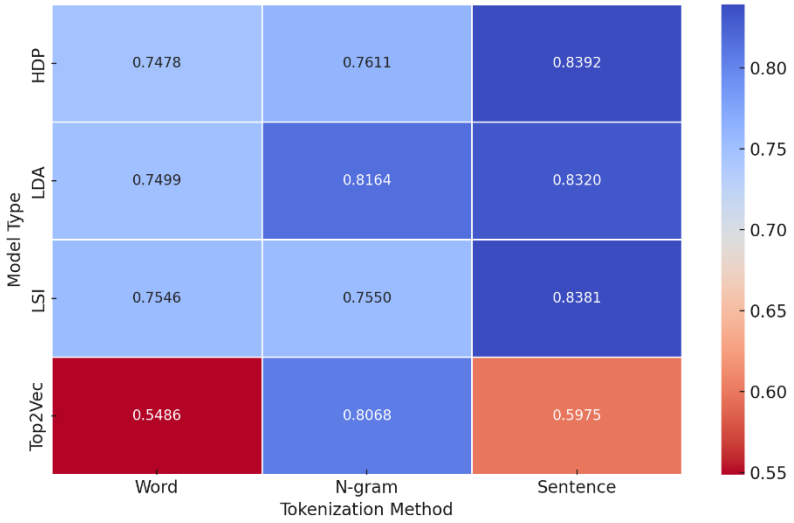

Figure 19. Heatmap of coherence scores by model type and tokenization method.

The comparison pinpointed high variability in the productivity of different topic modeling techniques and tokenization methods. On the other hand, the LDA and HDP models generally outperformed other models regarding the coherence of topics and ease of interpretation. The Top2Vec model had mediocre to poor coherence. The only exception was the model Top2Vec-trained with bi-gram tokenization, which showed competitive coherence under this tokenization. Such results only confirm the idea that there is no universal approach to apply to topic models, traditional models perform better with small text data, and the choice of model and pre-processing steps must be reasonable and provided in strict correspondence with the peculiarities of the dataset and the objectives of the analysis.

However, applying traditional unsupervised topic models usually leads to many outliers. Those will compromise the invested strength and trustworthiness of the information. Despite these challenges, topic modeling has proved a powerful and effective urban planning and policy tool. It guides insights from data concerning cities' spatial arrangement and contributes to identifying urban functional zones. It enables the heterogeneous science of urban landscapes by analyzing user-generated point-of-interest data. In this sense, the evidence could help urban planners, policymakers, or those who can embrace those decisions toward better and more efficient urban management strategies.

6. Conclusion

This paper aimed to compare the various topic modeling methods on user-generated Point of Interest data and the efficiency of the most efficient model in classifying Urban Functional Zones. The study, therefore, tried to simulate the performance of various topic modeling methodologies of LDA, LSI, HDP, and Top2Vec, where each of these models had different scores when tokenized using different forms: word (uni-gram), bigram, and sentence tokenization applications. This experiment showed much better performance than the other tokenization approaches. Specifically, the sentence tokenization of the LDA, LSI, and HDP models, having an excellent coherence score, keeps the context of the entire sentence, enabling the models to better catch the underlying thematic structure of the data and produce coherent and interpretable topics. On the other hand, the word and bigram tokenization, much closer to sentence tokenization in performance success, often led to lower coherence scores of the models and higher fragmentation of the topic representation. LDA and HDP emerged as the best-performing models in this study. Both scored best within the highest coherence scores when the sentence tokenization approach was conducted.

What can be said about this research is that all models are highly dependent on the choice of tokenization and the set of models itself. The LDA and HDP models showed high performance under different constituent choices. Although the LSI model is very efficient in data handling, it cannot compute the perplexity value needed to make a study on performance. Moreover, the tendency of the traditional models to throw outliers, even though they are at a very high level of performance in unsupervised mode, may also damage the credibility of the result.

One practical implication of the study is using topic modeling for city planning and administration. Research of user-generated POI data helps to gather information on a city's structure and its functional parts' organization. As the data shows, the established use of this method will help the urban planners outline commercial, public, entertaining, and residential parts of the city, rendering the city development and planning procedures more information-based. The potential of the topic model to analyze great amounts of information and trace the results based on the identified patterns further validates its use for urban studies. Additionally, the established combination of advanced ML procedures ascertains the ability of these methodologies to act as the driving force for modern urban studies. The

competency of the models, in particular, LDA and HDP, shows the importance of urban data complexities and fact-based information delivery. As such, it shows how data undergoing proper preprocessing, such as tokenization and normalization, among others, sets forth a possibility of attaining a better performance of topic models. Better results were reported with sentence tokenization, highlighting contextual information as crucial to ensuring interpretability and coherence. These critical insights can help guide future research and practical application on how text preprocessing techniques may be further fine-tuned with the specifics of analytical needs.

This paper is a major contribution to urban planning and big data analytics as a comprehensive comparison of different topic modeling approaches. Conversely, the superiority of sentence tokenization shown in the paper is further evidence of the robustness of the LDA and HDP models. It sets a clear path for future studies targeting studies that use user-generated data in urban analysis. The implications that this study opens in the field are endless, enabling more sophisticated and effective data-driven approaches to urban planning and management. Further research can continue to delineate how the advanced topic modeling techniques can be embedded into other machine learning algorithms to improve the accuracy and reliability of urban data analytics. Moreover, the scope of data sources should be broadened to include real-time data streams and supervised models to provide a more dynamic and responsive framework for urban planning. The ongoing big data analytics and machine learning revolution promises further innovations and enhancements within the field, ultimately leading to smarter and more sustainable urban systems.

This paper accentuates the vital role of machine learning and big data in shaping the future of urban planning. With its help, advanced analytics would permit urban planners and policymakers to understand those small nuances of urban life and be allowed to build dynamic, efficient, and sustainable cities. The outputs derived here have academic and practical implications in tools and methodologies that might apply to any regular concern about urban issues. Our cities' constant growth and dynamism invite the infusion of data-driven approaches that are better suited for facing the multifaceted issues confronting our urban areas that they will continue to face.

Acknowledgements

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interests

The Author(s) declare(s) that there is no conflict of interest.

References

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *ArXiv*, *abs/2008.09470*. https://api.semanticscholar.org/CorpusID:221246303

Amen, M. A. (2021). The assessment of cities physical complexity through urban energy consumption. Civil Engineering and Architecture, 9(7). https://doi.org/10.13189/cea.2021.090735

Amen, M. A., Afara, A., & Nia, H. A. (2023). Exploring the Link between Street Layout Centrality and Walkability for Sustainable Tourism in Historical Urban Areas. Urban Science, 7(2), 67. https://doi.org/10.3390/urbansci7020067

Amen, M. A., & Nia, H. A. (2020). The effect of centrality values in urban gentrification development: A case study of erbil city. Civil Engineering and Architecture, 8(5), 916–928. https://doi.org/10.13189/cea.2020.080519

Aziz Amen, M. (2022). The effects of buildings' physical characteristics on urban network centrality. Ain Shams Engineering Journal, 13(6), 101765. https://doi.org/10.1016/j.asej.2022.101765

Chen, P. (2021). Research on the Mode Transformation and Innovation of Urban Planning Management. *Journal of World Architecture*, *5*(5). http://ojs.bbwpublisher.com/index.php/JWA

Deng, Y., Fu, B., & Sun, C. (2018). Effects of urban planning in guiding urban growth: Evidence from Shenzhen, China. *Cities*, *83*, 118–128. https://doi.org/https://doi.org/10.1016/j.cities.2018.06.014

Farid, A. M., Alshareef, M., Badhesha, P. S., Boccaletti, C., Cacho, N. A. A., Carlier, C.-I., Corriveau, A., Khayal, I., Liner, B., Martins, J. S. B., Rahimi, F., Rossett, R., Schoonenberg, W. C. H., Stillwell, A., & Wang, Y. (2021). Smart City Drivers and Challenges in Urban-Mobility, Health-Care, and Interdependent Infrastructure Systems. *IEEE Potentials*, *40*(1), 11–16. https://doi.org/10.1109/MPOT.2020.3011399

Gunko, M., Batunova, E., & Medvedev, A. (2021). Rethinking urban form in a shrinking Arctic city. *Espace Populations Sociétés*. https://doi.org/10.4000/EPS.10630

Hall, P. (2020). Urban population. *Africa's Urbanisation Dynamics 2020*. https://doi.org/10.32388/j85ynp

Hou, B., Nazroo, J., Banks, J., & Marshall, A. (2019). Are cities good for health? A study of the impacts of planned urbanization in China. *International Journal of Epidemiology*, *48*(4), 1083–1090. https://doi.org/10.1093/ije/dyz031

Lan, D., Tian, A., Wang, Y., & Li, Y. (2021). An overview of the principle, algorithm improvement and application based on the theory of latent semantic indexing. *Academic Journal of Computing & Information Science*, *4*(5). https://doi.org/10.25236/ajcis.2021.040510

Lapointe, M., Gurney, G. G., & Cumming, G. S. (2020). Urbanization alters ecosystem service preferences in a Small Island Developing State. *Ecosystem Services*, *43*, 101109. https://doi.org/https://doi.org/10.1016/j.ecoser.2020.101109

Liu, S., Liao, Q., Liang, Y., Li, Z., & Huang, C. (2021). Spatio–Temporal Heterogeneity of Urban Expansion and Population Growth in China. *International Journal of Environmental Research and Public Health*, *18*(24). https://doi.org/10.3390/ijerph182413031

Lyu, R., Zhang, J., Xu, M., & Li, J. (2018). Impacts of urbanization on ecosystem services and their temporal relations: A case study in Northern Ningxia, China. *Land Use Policy*, *77*, 163–173. https://doi.org/https://doi.org/10.1016/j.landusepol.2018.05.022

Muzafar Rasool Bhat Majid A Kundroo, T. A. T., & Agarwal, B. (2020). Deep LDA : A new way to topic model. *Journal of Information and Optimization Sciences*, *41*(3), 823–834. https://doi.org/10.1080/02522667.2019.1616911

Paes, V. de C., Pessoa, C. H. M., Pagliusi, R. P., Barbosa, C. E., Argôlo, M., de Lima, Y. O., Salazar, H., Lyra, A., & de Souza, J. M. (2023). Analyzing the Challenges for Future Smart and Sustainable Cities. *Sustainability*, *15*(10). https://doi.org/10.3390/su15107996

Profiroiu, C. M., Bodislav, D. A., Burlacu, S., & Rădulescu, C. V. (2020). Challenges of sustainable urban development in the context of population growth. *European Journal of Sustainable Development*, *9*(3), 51–57. https://doi.org/10.14207/ejsd.2020.v9n3p51

Schindler, S., Mitlin, D., & Marvin, S. (2018). National urban policy making and its potential for sustainable urbanism. *Current Opinion in Environmental Sustainability*, *34*, 48–53. https://doi.org/https://doi.org/10.1016/j.cosust.2018.11.006

Syaodih, E. (2019). *The Challenges of Urban Management in Indonesia*.

Terenin, A., Magnusson, M., & Jonsson, L. (2020). Sparse Parallel Training of Hierarchical Dirichlet Process Topic Models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2925–2934). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.234

Xu, Q., Zheng, X., & Zheng, M. (2019). Do urban planning policies meet sustainable urbanization goals? A scenario-based study in Beijing, China. *Science of The Total Environment*, *670*, 498–507. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.03.128

Zhao, Z., & Zhang, Y. (2020). Impact of Smart City Planning and Construction on Economic and Social Benefits Based on Big Data Analysis. *Complexity*, *2020*, 8879132. https://doi.org/10.1155/2020/8879132