

C02047: Doctor of Philosophy
CRICOS Code: 058666A
33875 PhD Thesis: Computer Systems
January 2025

A Study on
Data Analysis of Spatio-temporal Modeling
in Location-Based Social Networks

Naimat Ullah Khan

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Angela Huo

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Naimat Ullah Khan* declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Shanghai University.

This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Naimat Ullah Khan]

DATE: 02nd January, 2025

PLACE: Sydney, Australia

ABSTRACT

In today's digital age, Location-Based Social Networks (LBSNs) have become a crucial source of data for understanding user behavior and preferences. Mining data from these platforms has emerged as a significant research focus due to its potential to study patterns in user activities and preferences. The proliferation of LBSN-based applications generates vast datasets that yield valuable practical insights in areas such as public transport analysis, route optimization, disaster management, and location recommendations. The interactive features of these platforms allow users to share interests, activities, and multimedia content, producing rich datasets. These services capture and store user information alongside real-time data, enriched with metadata, textual content, multimedia, and geo-locations, facilitating comprehensive research into various aspects of user behavior patterns.

The dissertation first introduces a three-step analytical framework that combines statistical, temporal, and spatial modeling to analyze LBSN data. Previous related analysis approaches typically focus on single application domains, such as venue popularity, this dissertation introduces a comprehensive framework incorporating the three-step modeling across multiple domains, applied through case studies with analysis of user behavior in Shanghai using from Sina Weibo (Weibo), a famous LBSN in China. To enhance the accuracy of activity pattern study, Kernel Density Estimation (KDE) is implemented for anomaly detection and employs point pattern analysis to examine venue proximity effects. This approach reveals relationships between different venue types and enables detailed investigation of specific user categories, such as patterns in individuals associated with educational institutions and restaurants, providing deeper insights into activity patterns.

To address the limitations of traditional LBSN analysis, researchers either acquire domain-specific data or manually filter through millions of records. This dissertation presents Machine Learning (ML) approaches and develops Deep Learning for Location classification (Deep-Loc) models for venue classification and prediction. The proposed ML models achieve significant accuracy of 85-93% in tourism venue prediction, while the Deep-Loc methods demonstrate 99% accuracy in venue classification. These results help in find common reasons behind the specific behavior (previously addressed in literature as “maybe” or based-on assumptions). Furthermore, the study includes a hybrid group recommendation model that integrates collaborative filtering with context-aware features. This model, validated using Gowalla data, outperforms existing methods across different metrics, establishing a robust framework for LBSN data analysis applications.

The research methodology encompasses detailed documentation of the implementation stages, platform specifications, design, and is supported by comprehensive pattern analysis and empirical validation using case studies with real-world datasets.

DEDICATION

To my beloved parents, family and teachers.

ACKNOWLEDGMENTS

Truth, respect, and focus have been the most significant contributors to my success, especially during the challenging journey of my Ph.D., which began online at the start of the COVID-19 pandemic and included a recent year long leave of absence due to health issues. I would like to express my deepest gratitude to my supervisors Dr. Angela Huo provided exceptional guidance throughout my doctoral studies. Despite her demanding schedule, she consistently prioritized our fortnightly meetings and offered timely, insightful advice that profoundly supported my academic progress. I extend my highest respect and heartfelt thanks to Dr. Angela Huo and my co-supervisor, Professor Guandong Xu. I am grateful to the University of Technology Sydney UTS for providing an opportunity to pursue my Ph.D. and for creating a supportive, productive research environment.

My sincere appreciation also goes to Professor Wan Wanggen from Shanghai University and my friends Liu Ze Xin, Muzahid, and Wang Zhou for their invaluable research support and companionship during this journey. I am profoundly thankful to my family: my parents, siblings (Habib Ullah Khan, Nasib Ullah Khan, Abdullah Khan, Farid Ullah Khan, Akram Ullah Khan, and Muhammad Rafiq), and my wife, Rubina Riaz. Their unwavering financial, moral, and emotional support, along with the constant prayers and encouragement from my sisters and friends (including Adnan, Asif, Latif, and Adil), provided me the strength and confidence to overcome numerous challenges throughout my academic pursuit.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. **Khan, N.U.**, Rubina R., Jutao H., Xianzhi W., and Huan H., (2025) Enhanced Group Recommendation System: Enhanced Group Recommendation System: A Hybrid Context-Aware Approach with Collaborative Filtering for Location-Based Social Networks. The International Journal of Systematic Innovation. Accepted.
2. **Khan, N.U.**, Wan, W., Riaz, R., Jiang, S., and Wang, X., (2023) Prediction and Classification of User Activities Using Machine Learning Models from Location-based Social Network Data. Applied Sciences. 13(6): p. 3517.
3. **Khan, N.U.**, W. Wan, and S. Yu, (2020) Location-based Social Network's Data Analysis and Spatio-Temporal Modeling for The Mega City of Shanghai, China. ISPRS International Journal of Geo-Information. 9(2): p. 76.
4. **Khan, N.U.**, Wan, W., Yu, S., Muzahid, A., Khan, S., and Hou, L., (2020) A Study of User Activity Patterns and the Effect of Venue Types on City Dynamics Using Location-based Social Network Data. ISPRS International Journal of Geo-Information. 9(12): p. 733.
5. Ali Haidery, S., Ullah, H., **Khan, N.U.**, Fatima, K., Rizvi, S.S., and Kwon, S.J., (2020) Role of Big Data in The Development of Smart City by Analyzing the Density of Residents in Shanghai. Electronics 9(5): p 837.
6. Ullah, H., W. Wan, S. A. Haidery, **Khan, N.U.**, Z. Ebrahimpour and A. Muzahid (2020). Spatiotemporal Patterns of Visitors in Urban Green Parks by Mining Social Media Big Data Based Upon WHO Reports. IEEE Access 8: p 39197-39211.
7. Ullah, H., W. Wan, S. Ali Haidery, **Khan, N.U.**, Z. Ebrahimpour and T. Luo (2019). Analyzing The Spatiotemporal Patterns in Green Spaces for Urban Studies Using Location-based Social Media Data. ISPRS International Journal of Geo-Information 8(11): p 506.

OTHERS :

1. Rubina R., Guangjie H.,*, Kamran S., **Khan, N.U.**, Hongbo Z., and Lei W., (2025). A Novel Approach for Pattern Classification within Imbalanced Datasets in an industrial Internet of Things Environments. Springer Nature: Scientific Report Journal. Accepted.
2. Rubina R., Guangjie H.,*, Kamran S., **Khan, N.U.**, and Hongbo Z., (2025). A Novel Ensemble Wasserstein Generative Adversarial Network for Effective Anomaly Detection in Industrial Internet of Oxford: Oxford: The Journal of Computational Design and Engineering. Accepted.
3. Rubina R., Guangjie H.,*, Kamran S., **Khan, N.U.**, and Lei W., (2025). A Robust Framework for Anomaly Detection in Industrial Internet of Things. IEEE Transaction: Sensors Journal. Under Review.
4. **Khan, N.U.**, and W. Wan (2018). A Review of Human Pose Estimation from Single Image. 2018 International Conference on Audio, Language, and Image Processing (ICALIP), IEEE.
5. Khan, S., L. Han, H. Lu, K. K. Butt, G. Bachira and **Khan, N.U.**,. (2019). A New Hybrid Image Encryption Algorithm Based On 2D-CA, FSM-DNA Rule Generator, and FSBI. IEEE Access 7: p 81333-81350.
6. Muzahid, A., W. Wan, F. Sohel, **Khan, N.U.**, O. D. C. Villagómez and H. Ullah (2020). 3D Object Classification Using a Volumetric Deep Neural Network: An Efficient Octree Guided Auxiliary Learning Approach. IEEE Access 8: p 23802-23816.

TABLE OF CONTENTS

List of Publications	vii
List of Figures	xii
List of Tables	xv
1 Introduction	1
1.1 Research Domain	1
1.1.1 Background	1
1.1.2 Context of Study	2
1.1.3 Analytical Challenges and Proposed Framework	3
1.2 Research Gaps and Objectives	5
1.2.1 Research Gaps	5
1.2.2 Research Questions	6
1.2.3 Hypothesis and Proposed Methods	7
1.3 Research Contribution	7
1.4 Thesis Organization	8
2 Background Study	10
2.1 Spatial-Temporal data analysis	11
2.1.1 Trajectory Data	12
2.1.2 Temporal Data	13
2.2 Location-based Social Networks	15
2.2.1 Weibo (LBSN) Check-in Analysis	17
2.2.2 LBSN and Machine Learning	19
2.2.3 Venue Class Prediction and Analysis from LBSN	21
2.3 Hybrid Recommendation System	27
2.4 Summary	35

3	Analysis of Spatio-temporal Modeling of Location-based Social Network Data	37
3.1	Introduction	37
3.2	Dataset and Methodology	39
3.2.1	Dataset	39
3.2.2	Methodology	39
3.3	Analysis and Discussions	43
3.3.1	Temporal Patterns	43
3.3.2	Statistical Analysis	46
3.3.3	Comparison of Point Density and Kernel Density Estimation . . .	46
3.3.4	Anomaly Detection with Venue Clustering	49
3.3.5	Density Estimation	51
3.4	Summary	55
4	Investigation of the Effect of Venue Types on City Dynamics	56
4.1	Introduction	57
4.2	Analysis Framework	60
4.2.1	Venue Classification	61
4.2.2	Clustering-based Point Pattern Analysis	65
4.2.3	Nearest Neighbor Distance	68
4.2.4	Spatial Analysis with Venue Types	72
4.3	Summary	76
5	Venue Classification and Prediction	77
5.1	Introduction	78
5.2	Analysis Methods	79
5.2.1	Data Source	81
5.2.2	Feature Selection	82
5.2.3	Machine Learning Models	82
5.2.4	Model Evaluation	87
5.3	Results and Discussion	87
5.3.1	Statistical Modeling	87
5.3.2	Classification Into Multiple Venue Types	89
5.3.3	Predicting Tourism Class	93
5.3.4	Contribution and Comparison	99
5.4	Summary	100

6	Analysis of LBSN Group Recommendation System	102
6.1	Introduction	103
6.1.1	Challenges in LBSN Recommendations	103
6.1.2	Hybrid Recommendation Systems	104
6.2	Hybrid Recommendation System for LBSN	107
6.2.1	Overview of Dataset	108
6.2.2	Data Preprocessing	109
6.2.3	User Influence Modeling	110
6.2.4	Collaborative Filtering	111
6.2.5	Group Recommendation	114
6.2.6	User Behavior Analysis	114
6.2.7	Clustering Based on User Preferences	114
6.2.8	Temporal Behavior Analysis	115
6.2.9	Evaluation	115
6.3	Results	116
6.3.1	User Influence Modeling	116
6.3.2	Temporal Distribution	117
6.3.3	Category-Wise Interaction Analysis	118
6.3.4	Temporal Behavior Distribution	119
6.3.5	Clustering Based on Preferences	120
6.3.6	Hybrid Recommendation System	122
6.4	Ablation Study and Error Analysis	124
6.5	Summary	125
7	Conclusion and Extensions	127
7.1	Concluding Remarks	127
7.2	Future Research	129
	Bibliography	131

LIST OF FIGURES

FIGURE	Page
1.1 Scope of Research	5
1.2 Chapters' Contributions and Connections	9
2.1 Evolution of LBSN Data Research	11
2.2 LBSN Data Generation	16
2.3 General Research Framework	31
2.4 Study Area	32
2.5 Characteristics of Data Source	33
2.6 Data Acquisition Process Using Weibo API	34
3.1 Chapter's Research Framework	40
3.2 Temporal Analysis	40
3.3 Kernel Density Estimation and Point Density	41
3.4 Check-in Frequencies for 24 Hours	44
3.5 Check-in Frequencies for Days of The Week	45
3.6 Check-in Frequencies for 180 Days	45
3.7 Point Density Vs Kernel Density	48
3.8 Comparison Between Density Methods: (a) Point Density (PD) and (b) Kernel Density Estimation (KDE).	48
3.9 Venue Clustering Steps	49
3.10 Venues Clusters	50
3.11 Clustering with Outlier Detection	51
3.12 Density Estimation for Check-ins in Shanghai	52
3.13 Density of Check-in Data	53
3.14 Weekly Density a) First Week of April, b) Second Week of April, c) Last Week of January, d) First Week of February	54
4.1 Analysis Framework	60

4.2	Venue Filtering Process	62
4.3	Check-in Venue Distribution	63
4.4	Venue Categorization Statistics	64
4.5	Category-wise Temporal Analysis, a). Time by Category, b). Weekday by Category, c). Date by Category,	65
4.6	Clustering Method	66
4.7	Spatial Distribution of Residential Category by Clusters	67
4.8	Spatial Distribution of Entertainment Category by Clusters	67
4.9	Bimodal distribution of Distance from Residential to Entertainment Venues .	68
4.10	Cumulative Distribution of Distance from Residential to Entertainment Venues	69
4.11	Distribution of Distances from Entertainment Venues to Nearest Residential Venues	70
4.12	Nearest Neighbor Distance Distribution	71
4.13	Class-Based Density Estimation	72
4.14	Location of Venues in Different Categories in Shanghai	73
4.15	Category-wise Density a) Educational, b) Entertainment, c) Food, and d) General Location	74
4.16	Category-wise Density, a) Hotel, b) Professional, c) Residential, d) Shop- ping&Services, e) Sport, f). Travel	75
5.1	Pictorial Representation of Classification	80
5.2	Machine Learning Framework	80
5.3	Generalized Linear Model for Venue Classification	83
5.4	Logistic Regression for Classification	84
5.5	Boosted Trees Architecture	85
5.6	Deep-Loc Model for LBSN Venue Classification	86
5.7	Venue Classification into 10 Classes Using Machine Learning	90
5.8	Confusion Matrix for Generalized Linear Model	91
5.9	Confusion Matrix for Logistic Regression Model	91
5.10	Confusion Matrix for Gradient Boosted Trees	92
5.11	Confusion Matrix for Deep-Loc Model	93
5.12	Machine Learning Methodology for Venue Class Prediction	94
5.13	ROC of The Proposed Machine Learning Models	95
5.14	Graphical Representation of AUC	96
5.15	Accuracy of The Candidate Models	96
5.16	Performance Matrices. a) Precision, b) Recall, c) F-score and d) Sensitivity . .	97

LIST OF FIGURES

5.17	Lift Charts of Proposed Models, a) Deep-Loc, b) GLM, c) LRM, d) GBT	98
6.1	Proposed Framework	106
6.2	Dataflow Diagram of the proposed Method	108
6.3	User Influence Modeling	117
6.4	Temporal Distribution of User Interaction Across Categories	118
6.5	Category-Wise Interaction Analysis	119
6.6	Temporal Behavior Distribution Across Categories	120
6.7	Clustering Based on User Preference	121
6.8	Performance Comparison of Recommendation Methods	123

LIST OF TABLES

TABLE	Page
1.1 Important Terminologies	2
2.1 COVID-19 Era Research Topics and Questions	26
2.2 The Sample of Attributes in Initial Dataset	34
3.1 Sample Variables	39
3.2 Regression Summary	46
3.3 Final Multiple Linear-Regression	47
3.4 ANOVA	47
4.1 Research Contributions	58
4.2 Attributes of Categories	62
4.3 Check-in Venue Categories	63
5.1 Multiple Linear-Regression	88
5.2 Correlations Matrix	89
5.3 Comparison with Related Work	99
6.1 Overview of the Dataset	109
6.2 A comprehensive comparison of the proposed method with top performing methods	124
6.3 Ablation Study Results of the Hybrid Model	125

INTRODUCTION

This chapter is an overview of the research work undertaken within the scope of this Ph.D. in 'Communication and Information Systems' research dissertation. The choice of a significant research problem domain is defended by an extensive literature review to support the goal. Previous studies that motivated the devised research are presented in detail in the next chapter. The hypothesis and key objectives are then introduced, pertinent to Ph.D. level research and investigation, followed by future research in the same domain. This chapter provides the thesis organization for the following chapters in this dissertation as well.

1.1 Research Domain

1.1.1 Background

Most high-tech applications implemented today take the form of online communication a set of heterogeneous, discrete, interdependent systems interfaced with individual diverse mobile devices, laptops, or desktops. These LBSNs generate data enriched with metadata including multimedia content, textual information, and geo-locations, along with demographic elements such as age and gender [1]. The complexity and richness of LBSN datasets make their qualitative and quantitative modeling particularly valuable for researchers. This research has significant implications across multiple domains, especially in the development of intelligent applications and Smart City.

1.1.2 Context of Study

The diverse attributes of user data, including gender, demographic information, and temporal patterns, enable researchers to explore behavioral patterns across both individual users and larger populations. Current online technologies generate Big Data at unprecedented scales, containing complex patterns and relationships within their structure. This data complements traditional information-gathering methods such as surveys and questionnaires, providing opportunities to investigate novel patterns in user activities, interests, and preferences. These insights, when properly analyzed and presented, support practical applications across multiple domains such as venue popularity analysis for restaurants and tourist attractions, recommendation system development, smart city planning, and urban infrastructure optimization and many more [2].

Therefore, modeling the rich contextual data from LBSNs provides insights not only into users' overall behavior but also into specific categories like transportation, education, and tourism. In modern cybernetic environments, data mining reveals complex interactions and behavioral patterns that can be analyzed across multiple temporal reference points. LBSN databases now extend beyond simple numeric and alphabetic values, offering opportunities for patterns, trend analysis, and behavioral traits and variations. When properly modeled, these patterns can rapidly identify significant events, anomalous behaviors, and previously unobserved trends. These analytical insights contribute to improved venue management, event planning, and transportation optimization, ultimately supporting the development of intelligent solutions for modern urban challenges. Some of the important terminologies are given in Table 1.1 along with their definitions.

Table 1.1: Important Terminologies

Term	Definition
Location-based Social Network	A PC/mobile-based online social networking which utilizes and stores location information.
Check-in	A user confirms his/her location on the LBSN while engaging in an activity at a specific location or automatically shares location information when sending a message on LBSN.
Density	The degree of compactness of things.
Venue	The Point of Interest (POI) or location from where the users' check-in.
Spatio-temporal	The analysis with respect to space and time.
Kernel Density Estimation	A multivariate method that uses random sample data for estimating the density.

This thesis explores the following, which are proven to be merits of the proposed research methodology.

- For LBSN data significance, this research provides evidence of more efficient data sources in various fields.
- The classification of data into relevant groups makes it more efficient for researchers by eliminating the filtering of thousands of venues individually.
- Our proposed classification model can provide a baseline for further research in this domain.
- Venue prediction methods can be used in recommendation systems and research in to specialized domains.
- For data analysis, we present an effective three-step (Spatial, temporal and class based) analysis method to explore various valuable patterns within the data.
- City planners and activity related personnel can benefit from the knowledge about the activities and preferences of residents and tourists within the city.
- Integrates location, time, and personality for user-specific, context-aware suggestions.
- Leverages external data (e.g., social networks) to recommend POIs for new users and places.
- Combines CF and contextual features for better recommendations.
- Utilize individual preferences for cohesive group suggestions.

1.1.3 Analytical Challenges and Proposed Framework

The LBSN data analytics in this thesis addresses multiple analytical challenges and research problems. Firstly, the modeling of multi-user and multi-class data aims to identify common population characteristics, providing two key benefits: uncovering previously undetected patterns and validating datasets through confirmation of established human behavioral norms [3]. While existing research literature supports spatial and temporal analysis for pattern extraction, such studies often either present only a broad overview of human behavioral patterns or reduce multi-class data analysis to single-class observations [4]. This research begins with the overall activity trends and extends venue based analytics for finding the effect of different venue types on overall patterns. Secondly, designing a computational model that verifies the significance of each variable considered within the available data used for analysis before preceding to the implementation to show the efficiency of using LBSN data for such analysis.

Thirdly, the development of machine learning model that classifies users into categories based on their activities, eliminating the need to repeatedly filter through vast amounts of heterogeneous LBSN data. This approach enables the extraction of multi-user data and analysis of relationships between different venue types across various datasets. The classification system addresses a key limitation in current research by enabling targeted analysis of specific domains: for instance, allowing educational research to focus exclusively on data from educational venues, or transportation studies to analyze data specifically from transport locations. This proposed framework provides a template for domain-specific feature extraction, representing a substantial improvement over existing LBSN analytical approaches. Finally, using the proposed approaches as components for designing an enhanced group recommendation model for LBSNs taking into account the contextual information, temporal and spatial features along with commonly used collaborative filtering (CF) techniques.

This work proposes enhancement and efficiency in pattern analysis using LBSNs data, not limited to the overall abstract patterns in population but with proper specified classes critical to the progressive way of life in modern research. The enhanced approach enables more precise and targeted analysis, supporting evidence-based smart city planning and development. This research primarily focuses on developing analytical models, classification systems, recommendation system and observation of methodology in the case study of LBSN analysis. The importance of this research domain reflects the multidimensional nature of LBSN datasets, which capture diverse user interactions across multiple platforms. The study presents a systematic analysis of LBSNs datasets, examining behavioral trends across spatial and temporal dimensions. The findings reveal both common and some interesting distinctive patterns, demonstrating the significance of this analytical approach for understanding user behavior in LBSN contexts.

This research project, "A Study on Data Analysis of Spatio-temporal Modeling in Location-Based Social Networks" presents a comprehensive analysis of LBSN-generated Big Data analytics. The dissertation explores the conceptual, theoretical, and practical frameworks for extracting actionable insights from LBSN data with real-world implications and applications. The feasibility and significance of the project, with relationship between this study and the wider field of LBSN studies is mapped in Figure 1.1, highlighted as the research domain.

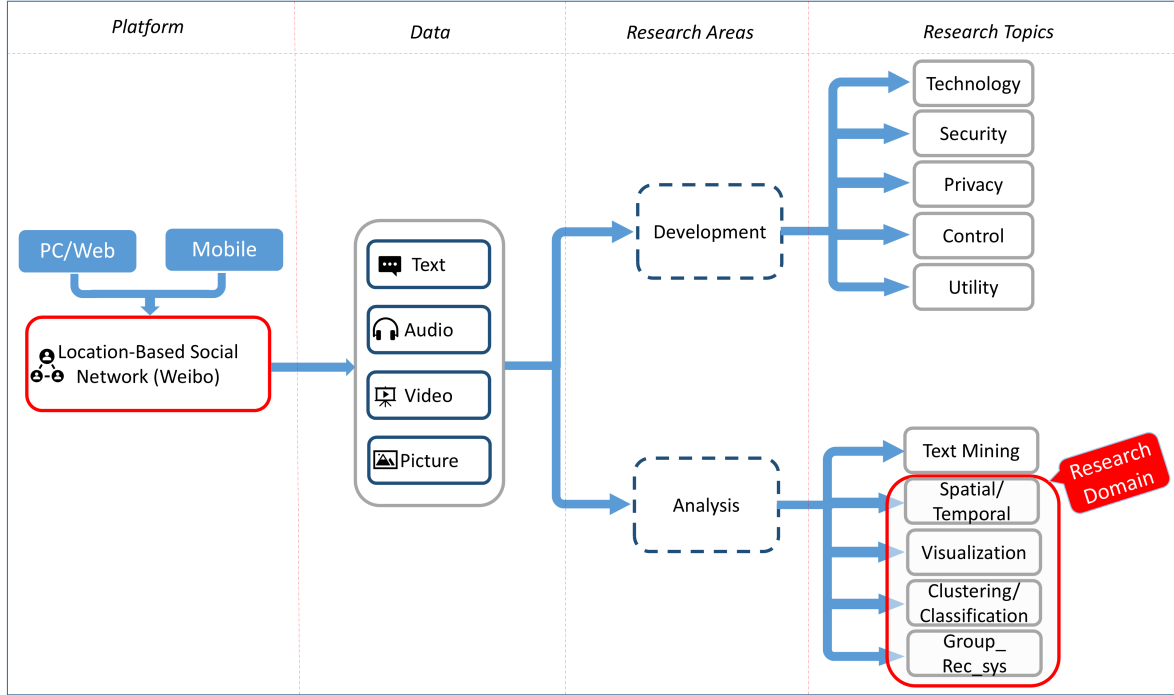


Figure 1.1: Scope of Research

1.2 Research Gaps and Objectives

The comprehensive literature review identifies critical gaps in current methodologies, clarifies research objectives, and outlines potential advantages of the proposed approach. These findings, detailed in the following subsections, serve as benchmarks for developing the research framework.

1.2.1 Research Gaps

The analysis, classification, and modeling of Big Data extracted from LBSNs addresses key research requirements identified through a systematic literature review. The following sections detail these critical aspects.

- Need for validation of LBSN data patterns against established human behavioral trends.
- Role of enormously generated LBSN data and its efficiency instead of manually collected data.
- Analyzing and Develop analytical models to uncover latent patterns within LBSN Big Data.

- limited studies on integration of anomaly detection models with density-based methods for optimization and accuracy.
- Need for robust venue prediction methods for domain-specific venue information extraction.
- Limited understanding of user preferences through spatiotemporal analysis various venues types.
- The need to filter thousands or sometimes millions of records to find subject specific data.
- Finding users preferences with the help of spatiotemporal analysis of historical check-ins.
- limitations in robust integration of multidimensional contextual factors having limited support for group recommendations in LBSN environments.

1.2.2 Research Questions

Based on the identified gaps in LBSN Big Data analysis and pattern recognition, the following research questions have been formulated to guide this investigation. These questions are structured to address specific aspects of data validation, pattern analysis, classification, and contextual integration within LBSN environments.

- How can LBSN data patterns be effectively validated against established behavioral trends?
- What is the comparative efficiency of LBSN-generated data versus manually collected data?
- How can the reliability of LBSN based big data patterns be systematically verified?
- What is the optimal integration method for anomaly detection and density-based approaches?
- Which methods can be used to efficiently extract domain-specific information from large-scale LBSN datasets?
- What methods best facilitate accurate classification of venues into different classes?
- How can filtering mechanisms be optimized for large-scale data processing in recommendation systems?
- How can multiple contextual factors be effectively integrated into LBSN analysis?
- What methods best support group recommendations while maintaining individual preference consideration?

1.2.3 Hypothesis and Proposed Methods

Analysis of research gaps leads to the hypothesis: "The modeling of spatial and temporal data from Location-based Social Network Big Data across classified categories can enhance analytical capabilities and implementation effectiveness across diverse research domains including recommendation systems."

Therefore, this research presents a comprehensive three-step framework incorporating classification, statistical analysis, and density estimation to extract meaningful patterns and anomalies from LBSN data. The methodology integrates Multiple Linear Regression (MLR) for data verification, Machine Learning models for activity-based venue classification, and the development of an LBSN group recommendation system. Additionally, the research validates the effectiveness of Kernel Density Estimation (KDE) for geo-tagged data analysis through comparative assessment of famous density methods, providing a robust approach to pattern identification with anomaly detection in LBSN datasets.

- Implementation of Multiple Linear Regression for data verification.
- Representation of spatio-temporal analysis methods based on real data.
- Anomaly detection framework development using KDE with DBSCAN.
- Development of comparative analysis frameworks namely Point Density and KDE.
- Creation of efficient classification mechanisms with the help of ML.
- Development of activity-based venue classification models i.e. Deep-loc.
- Development of user preference modeling systems with Collaborative Filtering.
- Implementation of context-aware suggestion system for LBSN.
- Creation of group recommendation frameworks.

1.3 Research Contribution

The contribution of this research as a part of the Ph.D. degree is to develop models for the effective utilization of data for research and analysis and modeling of data patterns defined in the research using LBSN. The patterns are then interpreted and explained in accordance with examined trends (actual events based on the case studies of Shanghai), which are identified from the data under consideration during this research. The abstracted items are:

- Development and validation of LBSN data modeling frameworks for user behavior analysis, incorporating spatio-temporal analysis to extract multi-dimensional user preferences.

- The research presents statistical models demonstrating the significance of these datasets for behavioral pattern analysis.
- Comparative analysis of methodological approaches, establishing the feasibility of the proposed framework and proposing classification models beneficial for LBSN applications, particularly in venue recommendation systems.
- Implementation of ML models for venue classification and prediction, enhancing the accuracy of location-based analytics.
- Creation of a hybrid recommendation framework for LBSNs that incorporates multi-dimensional contextual factors, advancing the capabilities of location-based recommendation systems.

Besides the theoretical knowledge and research skills, this Ph.D. journey has yielded the following practical implications:

- Exhibit a clear understanding of up-to-date critical thinking in big data analytics and state-of-the-art methodologies in analysis and processing actual data.
- Developed specialized expertise in Big Data analysis techniques, Machine Learning applications for 3D image processing, Human Pose Estimation and recommendation system for LBSNs.
- Acquired proficiency in modern development platforms and tools essential for advanced research implementation.
- Created innovative data description, analysis methods for real-world applications.
- Internalized and applied ethical principles in the conduct of independent research.

1.4 Thesis Organization

After the introduction in this Chapter 1, this dissertation presents a comprehensive literature review in Chapter 2, followed by a detailed study design outlining research areas, data sources, and methodologies. Chapter 3 explores the implementation and comparative analysis of spatio-temporal methods for LBSN data in Shanghai, focusing on KDE with anomaly detection. Chapter 4 examines classification approaches and analyzes the impact of venue types on city dynamics. Chapter 5 proposed Machine Learning models for venue classification and specific venue class prediction. Chapter 6 details of the hybrid group recommendation system for LBSNs, combining the CF and Singular Value Decomposition (SVD), which incorporates context-aware modeling and group recommendation capabilities. The thesis concludes in Chapter 7 with a summary of the research and future directions.

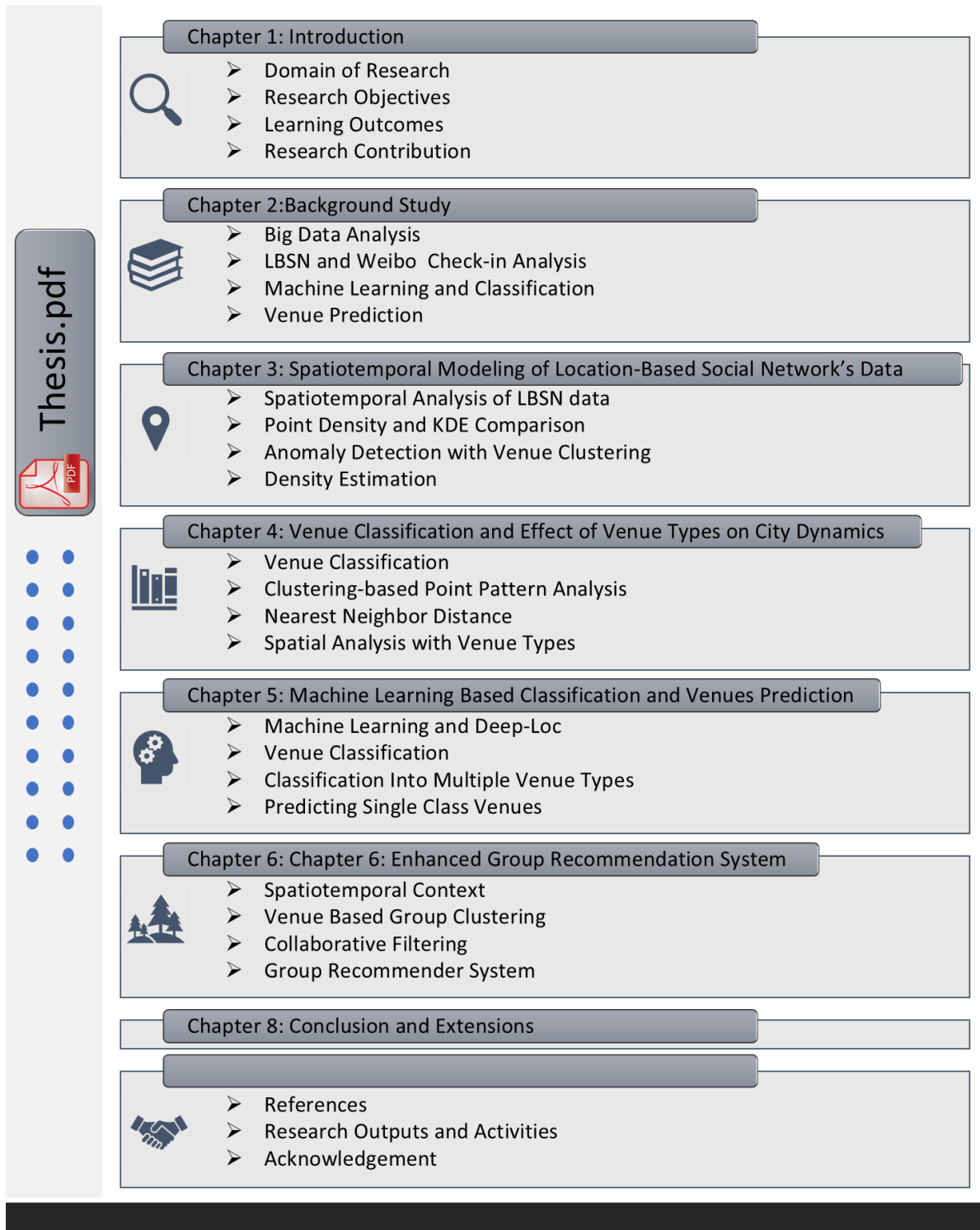


Figure 1.2: Chapters' Contributions and Connections

BACKGROUND STUDY

Data analytics has been one of the most popular and focused topics among researchers in recent decades. The top enterprises in the world, such as Amazon, Google, and Facebook, emphasize on using big data analysis for decision making can prove to be significantly beneficial and is believed to be of value to most analysts, managers and executives [5]. Smart decision making based on analysis is considered as the first step of a data-driven society which provides a vital role in the development and effective advancement of organizations. Several benefits of big data have been described by McAfee [6][16], considering the role of big data in smart decisions and predictions based on actual data rather than intuition which proved to be more beneficial for the organizations in competition against rivals with the examples of digital enterprises like Google already taking advantage of using the big data analytics. They found that these organizations show 5% more productivity and 6% more profit than others by performing better due to matching the needs of their customers based on evident data. LaValle et al. [7] presented that the best performing organizations use data not just for decisions about finance and operations but for other fields, including marketing, customer experience, customer care and many more. According to De Mauro et al. [8], 'Big Data' has gained value by creating partitions of population in different groups for customization, identifying preferences, finding trends, enhanced business models, services, and products. Therefore, we can conclude that 'Big Data' is not just a fancy word, even not simply a reference to volume but a valuable asset that helps enterprises for better understanding their environment, employees and customer's needs, behavior, activities, and companions.

2.1 Spatial-Temporal data analysis

The rapid advancement in technology has resulted in the exponential growth of 'Data' at a remarkable speed. This made the handling of such huge piles of data much more difficult, mainly due to its growing volume in contrast with the development of computing resources. The term 'Big Data' has always been considered a part of Computer Science since early computing. The basic definition of 'Big Data' states that it's the data that cannot be processed using traditional tools. The formal definitions of 'Big Data' still include the term "Structured data" however most researchers are considering and utilizing the huge amounts of unstructured data as a source of extracting some of the most interesting and valuable information in different research domains.

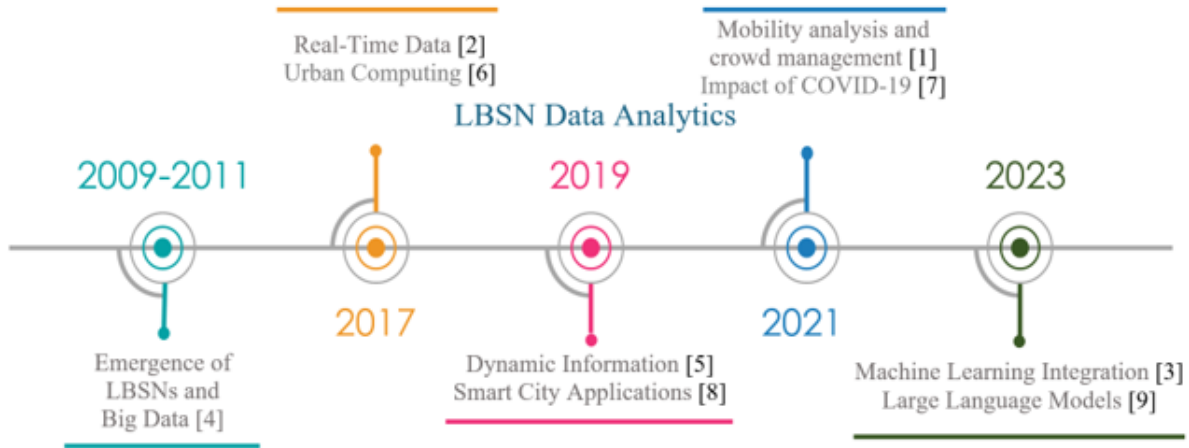


Figure 2.1: Evolution of LBSN Data Research

In the last few decades, the interest of researchers in big data has increased exponentially and research into big data, compared to other fields of computer science, has gained tremendous attention, as shown in Figure 2.1¹. The term itself and articles like "Big data is opening doors, but maybe too many" [18] and "Big data: the greater good or invasion of privacy?" [19] suggest a perception of volume; however, there are more features to be considered regarding big data, such as complexity, structure, behavior, tools, techniques, and technologies used to process and analyze it. Dumbill discussed three different dimensions of big data: "volume", "velocity" and "variety" of contents [20]. Providing the insight that big data analysis is not that simple, Mayer-Schonberger and Cukier highlighted the three main challenges of big data: populations instead of samples, messy instead of clean data, and correlation instead of causality [21]. The authors first explained that big data is always the collection of as much data as possible

¹ref: [9–17]

and suggested that collecting, storing and analyzing massive data will be allowed in real time for data scientists with the advancements in technologies. They called the sampling an outdated deterrent and the method of using population for big data analysis to get patterns claiming that the “random sampling of the bouts might have failed to reveal it.” However, it seems quite beneficial to use all the data for analysis. However, even with today’s technological power, it is difficult to capture all the data of all times. Another dimension of big data discussed by the authors is "Messy" meaning measurement error that is an obstacle for data scientists and the main concern of any data science. The authors argued that the use of data analytics would allow analysis of more aspects related to the problem hence leading to better results, but we must relax the standard for accepting measurement errors. However, they failed to explain the difference between precision and validity, as analysis of huge amounts of data may be precise but not valid because of the use of systematically biased data.

The authors finally talked about the “Correlation” predicting that “big-data correlations will routinely be used to disprove our causal intuitions, showing that often there is little if any statistical connection between the effect and its supposed cause” in the future. The authors believed that causality is based on correlation; we can find the cause of an action or an event in a pattern by finding the correlation between the data related to the action/event. This may not entirely be true as bias can easily alter the actual causal effect. Additionally, Miller and Goodchild defined big data as data that cannot be analyzed using traditional tools. The authors wrote that due to the potential impact of data sciences, scientists are paying a great deal of attention to the methods used for data-driven analysis. In 2013, Ovadia and Librarian emphasized the importance of big data for social scientists and librarians and suggested that it is much too important to be ignored, as most social science research is based on huge amounts of data and enormous datasets [16].

2.1.1 Trajectory Data

As a central focus for many study fields, including time and space geography, urban functionalities, human mobilities and more, big data analysis is a vast research field that was initially studied using statistical data from surveys, interviews, travel diaries, questionnaires and other manual collections of datasets [22]. Statistical data collection may not be an efficient way to determine patterns in said fields and related studies. Therefore the data from mobile devices, smart cards, GPS navigators, and location-based and online APPs containing users’ activities with geo-locations are widely used and have

been found to be more efficient for such studies [23]. The rapid advancement of mobile technologies and excessive use of mobile devices, it is easy to track users' locations from their devices and activities. For example, Gonzalez et al. [24] introduced a dataset that contained data from 100,000 users over six months, arguing that despite the variety in individual human travel records, following reproducible and simple patterns. These similarities in travel trends can be critical such as in urban planning, agent-based modeling, and epidemic/pandemic prevention to emergency response. Although the data only contained the nearby location of the mobile phone towers from where the phone call originated, it still proved to be very helpful in estimating the approximate locations of users with a certain margin of time and was subsequently used in the prediction of human movement [25] where the authors used Entropy because it is the most basic probabilistic measure that captures predictability patterns in time series data. The goal was to find an answer to a very fundamental question, that is, What is the extent to which we can predict human behavior, using mobile communication to foresee the mobility and whereabouts of individuals. The authors concluded that by using suitable data-mining algorithms can be used to calculate the actual mobility predictions.

Various properties of Geographical Information Systems (GIS) functionalities with the potential role in urban mining research were reviewed through a discussion on how GIS data can be utilized to analyze, visualize, report and mine the temporal or spatial or features for urban land use, its cover charges, and future possibilities [26]. The modern digitized world allows researchers to conduct quantitative analysis of activity patterns and related factors, like living areas, social contacts, and personal references. Chang et al. used police crash data, loaded it with X and Y coordinates and applied the suggested methods to study the spatial distribution of road crashes in Shanghai [27]. Huo et al. categorized user activity into three different sections, namely location prediction, trajectory mining, and location recommendations [28]. These authors also emphasized its role in the understanding of user activity patterns and how it can be beneficial in many areas like traffic control, disaster relief, marketing, and city planning.

2.1.2 Temporal Data

The research on human mobility and activity analysis initially took place as early as 1885, referring to the publication by E.G. Ravenstein [29]. The research was based on census data for analyzing patterns in human migration motivated by the negative and conflicting statement by Dr. William Farr, who stated that there appears to be no definitive law or similarity pattern in human migration [30]. The author based his claims

on millions of records of census data about the migration to analyze patterns in the British, Irish and Scottish all over the United Kingdoms. The conclusions included three main statements that are: the majority of migration are over short space, longer are mostly to urban regions, and human migration is increasing with the development in the economy. This research opened new doors in understanding human mobility and activity studies using empirical data by highlighting patterns in human movements, such as the effect of distance and economy on travel preferences.

The research direction was followed by many researchers in order to understand, model and analyze patterns in human activity and mobility along with their traits in a large number of behavioral studies [31]. The datasets used in these studies were collected manually and therefore provided a static view of patterns in human mobility. This kind of research can be helpful with insights about a city or county of users or even some spatial information about the mobility of individuals, however, the census data can never provide the exact time of each and every visit to the smaller granularity. Therefore, the data acquired through census are considered to suffer from limited granularity with respect to temporal and spatial features to describe human activities. Despite that, the census data can provide useful information about the overall patterns in human mobility on a large scale; however, it fails to capture a large portion of important short-term or daily based movement and activities within small distances, especially in the case of urban environments and studies based on developed urban cities where people have several activities on a regular basis which is important to understand human behavioral studies, urbanization, and smart cities. Some of the main challenges produced by modern urbanization include planning and management of transportation and other administrative services based on the use and movement of people within the city [32].

The most common method for studies related to the human population has been representative surveys which made the acquisition of information possible about the origin, transportation, and destination of each trip within a city [33]. These survey-based studies of the mid-1970s laid the foundation of modern urban mobility theories with a huge impact on similar studies today. With the deployment of Second Generation (2G) technologies in the early 1990s changed the source of communication and this kind of study immensely. This made the tracking of the movement over a small period of time even up to great granularity, like within seconds, with relatively high precision in terms of geography, population and time. Among the first large scale study of human mobility using Call Record Detail (CRD) was published in the year 2008 [24]. Even if the CRD records made it possible to predict the location of given users with 93% accuracy [34],

however, this accuracy is based on a threshold of a few hundred meters approximation. Therefore, it was a big step from the manual survey but failed to capture the exact movement, as stated earlier, in urban environments and short distances. This led to the need for data sources with the exact locations (latitude/longitude) at any specific time, which is provided by the modern LBSNs, along with other valuable records with much greater information potential.

2.2 Location-based Social Networks

Online social networks are considered as some of the most significant sources for data analysis because of their widespread and ever-growing use in almost every part of the globe. To show the importance of understanding human mobility patterns, Cohn et al. analyzed the behavior of about 10,000 different LBSN users considering the types of visited places and the observation in time, proving that combining temporal similarities of user behavior is more beneficial in various applications [35]. LBSNs enable users to post their interests, activities, and locations, (Figure 2.2 [36]) creating data that researchers can analyze across multiple fields.

In some famous early article, Alrumayyan et al. studied peoples' patterns related to various venue categories in the capital city of Saudi Arabia, Riyadh, with the support of Foursquare data [37]. They used sentiment analysis and Latent Dirichlet Allocation (LDA) model for finding the popularity of venues with positive comments and low ratings. The study was more focused on the 'Food' category because people are more interested in sharing their experiences and leaving comments while visiting food venues. Similarly, A study by Lin et al. in San Francisco New York, and Hong Kong, from check-ins of about 19,000 Swarm (an APP of Foursquare) users, discussed the user preferences and associations between different venue categories in different cities, at different times of the day [38]. They concluded the common human behavior among all three cities by correlating the data on venues with respect to time. LBSN data is used in several critical domains such as Graham et al. [39] studied the importance of LBSNs in assisting local governments by conducting a survey of more than 300 government officials from the United States. They discussed the contribution of LBSNs for managing a crisis, resulting in positive relationships with the ability of users to control the crisis situation. Other similar studies highlighting the ability and use of LBSN tools in crises include articles on the wildfires in California [40], suggesting that people from society will use any available source for information according to their needs and actions while facing disaster,

presenting the case study of Hurricane Sandy and evaluating the initial emergency response based on Twitter along with government response in the face of disaster, and the earthquake in Haiti which evaluated that how the social networks can be additional source of data for disaster management [41].

Various studies and analytical methods focusing on human activities derived from mobile data have been explored in a number of article [42]. Myers et al. studied human mobility patterns from datasets of two LBSN's refer not to share their location while checking-in from various places due to multiple reasons, raising a concern that the LBSN data may not be suitable for human mobility studies [43]. However, it was followed by various studies about activity patterns from LBSN data [44], proving the significance of using check-in data from LBSN. Ezequiel et al. presented personalized geo-social recommendations based on LBSNs by using two different datasets (Foursquare and Gowalla) and observed similar patterns in both datasets by studying the privacy preserved personalized geographical influence of locations [45]. Wang et al. conducted a broader study by using data collected from Foursquare and Yelp to identify factors that help to uncover the reasons for venue popularity [46].

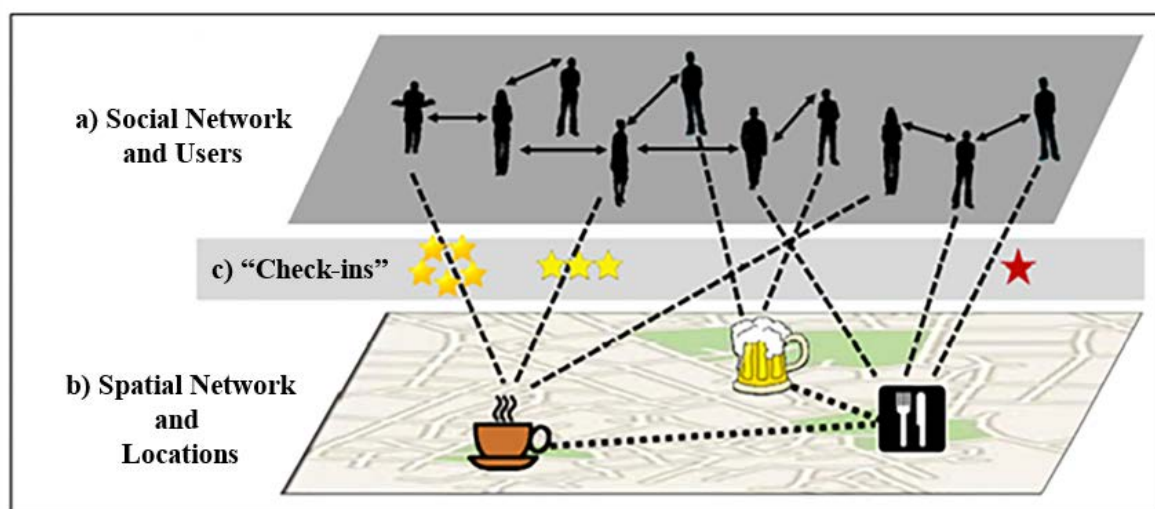


Figure 2.2: LBSN Data Generation

The popularity of a venue was calculated based on the frequency of check-in from various users at the specific venue. 2.2 demonstrates the concept of Check-ins where a) represents the users, b) locations and c) Check-ins. It was concluded that there are three main reasons influencing the popularity of a venue; 1) venue profile information, as venues with complete profile information are undoubtedly more popular; 2) venue age, as people tend to visit known and famous places and 3) venue category, as venues

under the 'Food' category were found to have the highest number of check-ins. Nowadays, social network applications are commonly used, and the generated data implies a significant and certain trace of useful information for the analysis of urban dynamics. This demonstrates the opportunities; how the massive sources help to acquire a better understanding of social dynamics and can be used in urban interferences [47].

2.2.1 Weibo (LBSN) Check-in Analysis

Lots of research has been done uncovering different features in and from LBSN data in the last few years. Most of the previous researchers studied information from LBSNs like Twitter and Foursquare to investigate a variety of patterns, like user activity and mobility, urban planning, and venue categorization. Weibo is a famous LBSN in China and has proven to be an efficient source of data for this type of analysis. Long et al. used human mobility and activity patterns based on Weibo along with transit smart-card data to analyze the growth of urban boundaries for the city of Beijing [48]. It was accomplished by developing three measures: by quantifying the effectiveness of urban growth boundaries with user activities, the correlation between planned population and human mobility, and urban activities. Rizwan et al. used Weibo data from early 2016 to observe check-in behavior and gender differences [49]. The authors identified different parts of the city preferred by either gender at different times. The key contribution of the article was identifying the density of people all over the city using KDE. Our paper [50] covered a similar task with the additional classification of the venues to identify the major influence of user preferences within the city using Weibo dataset.

Hu et al. used Weibo check-ins for detecting urban hotspots and commercial venues [51]. The authors detected the commercial areas by clustering the hotspots using geo-distribution metrics and calculated the rough boundaries of the commercial hotspot by calculating the center of the hotspot along with the distribution of an ellipse around it, with is further processed by a proposed algorithm to find the actual boundaries of the commercial hotspot. The method was further tested with time sequenced data of Wuhan city from Weibo for extraction and detection of the commercial hotspots. Fang et al. [52] presented a framework for analyzing the impact of disaster through LBSN data and use of Weibo messages for analysis of response and recovery in a case study of Wuhan, in flood and rainstorm disaster in 2016 . The temporal analysis was done using word frequency in the messages along with the major impacts of that disaster, and created hotspots using the spatial data of major flood sites, showing the effectiveness of LBSN analysis for disaster studies and recovery in the mega city. Another article by Zhao et

al. [53] considered the significance of LBSNs, specifically Weibo, in disaster situations and proposed a similar framework for physical and emotional damage assessment in a network environment in real time. The model captures detailed messages related to disaster and assess them in real time to estimate the damage losses by using natural language processing and semantic analysis for identifying the scale and duration of the damage during a disaster by making use of a large amount of LBSN data.

Another famous method that can be used for LBSN data analysis is called The Density-Based Spatial Clustering of Applications with Noise (DBSCAN). This is a notable anomaly detection providing a robust approach inspired by Garcia-Rubio et al., who used DBSCAN with k-means clustering for anomaly detection [54]. It is a sophisticated and widely used clustering method recognized for the ability in detecting clusters of different sizes and shapes. Unlike other conventional algorithms including k-means, DBSCAN does not impose the pre-specified number of clusters, presenting a more adaptive and flexible method for clustering various datasets. While DBSCAN offers flexibility in discovering arbitrarily shaped clusters and automatically identifies noise points, it requires careful parameter tuning (epsilon and minimum points), which can be challenging in high-dimensional LBSN datasets. Conversely, k-means is computationally efficient and easy to implement but assumes spherical clusters and needs the number of clusters predefined, limiting its adaptability to complex urban mobility data. Both methods have distinct advantages and limitations, suggesting hybrid or adaptive clustering strategies may better capture diverse spatial patterns inherent in LBSN data.

The DBSCAN method is an important and famous approach in cluster analysis. Many studies in LBSN studies used DBSCAN such as Yu et al. [55] proposed a recommendation model named NGPR which implements DBSCAN and Node2Vec technique along with integration of the preferences, geographical distances and POI's popularity for the suggestions. In another study Karagoz et al. used DBSCAN algorithm for finding the starting point for their graph based recommendation system by predicting the trust score of LBSN users for recommending venues based on the users' previous check-ins and their social networks [56]. Similarly, Xia et al. studied clustered check-ins by implementing DBSCAN on combination of study from two different datasets i.e., combining mobile phone data with LBSNs data to understand the activity of the users representing the population at an urban level with big data fusion in Shanghai [3].

In the famous article by Karayazi et al., they developed a novel approach by combining multi-source data i.e., heritage data, supporting data such as attraction, museum, shopping etc. and LBSN data to explore attractiveness of tourist locations and their

attraction characteristics. The authors also conducted cluster study with DBSCAN to detect POIs [57]. Hu et al., [58] identified the mobility patterns of tourist by combination of three methods. Firstly, collecting Twitters check-in data posted by tourists. Then, they applied the DBSCAN for retrieving clusters of the tourist data and finally, implemented an analysis method including indegree, outdegree, and association degree to retrieve popular venues. Cantero et al. proposed the methodology to identify patterns both indoor and outdoor users' activities using LBSN data through clustering with respect to time using DBSCAN for analyzing their daily activities with minimum infrastructure [59].

In a more recent study, Li et al. used linear regression on a dataset acquired from Weibo to show the positive relationship between the officially reported cases of COVID-19 and Weibo posts in the Hubei province of China which were also predictive of the cases COVID-19 percentage increase. People started talking about "unknown virus pneumonia" at the beginning of the outbreak, and when the virus was officially announced, the terms "Wuhan coronavirus" began to emerge and rapidly evolved to "novel coronavirus" or "COVID-19". The study demonstrates the role of Weibo during the progress of the pandemic and the trends among people on social media [60].

2.2.2 LBSN and Machine Learning

A major sources of data used in different kinds of analysis is LBSN with the help of ML methods requiring huge datasets. It is a valuable research field considered as the center of a variety of research domains like geography analysis, human behavior, activities, preferences etc. This kind of research initially used manual data collection approaches like surveys, interviews, questionnaires, and other statistical methods [22, 61]. However, with progressing time, the manual collection of data is not deemed appropriate due to the requirement of big data in the true sense for finding significant patterns within the data. The data collection method evolved into the use of global positioning systems coordinates, location-based online services and smart cards with the developments in mobile technology [20, 21].

As communication devices became more and more portable, the data collected about the user activities through these devices became easier and more accessible to researchers. One of the earlier research studies conducted by Gonzalez et al. [24] used data from about one hundred thousand users. It was the early-stage introduction of portable devices, and the technology to pinpoint the exact location of users was not mature enough; however, it provided a reasonable approximation of the users with the nearest base tower while making a call. The importance of using location-based data for

user behavior and activity patterns is discussed by numerous researchers, including the article [62]. The user activity analysis then shifted to use of LBSNs as a major source of data because portable mobile devices became more readily available almost everywhere in the world [35]. The facility of posting activities and preferences with locations provided by the online services not only interests users to share their life with friends but also works as a tool for generating large amounts of data which is used to find patterns in the general user behavior by exploring the similarities and differences in these activities as discussed in the article [63]. Many research articles are published comprising the study of users' behavior, including user mobility, geo-social recommendations, recommendation systems based on studies of two different cities in the United Kingdom [64] etc.

One of the major research fields in the study of LBSN is exploring patterns in the data with respect to venues [65]. In an earlier study, an enormous dataset was used by Li et al. [66] containing about 2.4 million sites in 14 different countries from Foursquare to find the popularity features of venues. The authors concluded three core features: venue's profile, age and nature of the activities as the most influential factors in the popularity of venues. Similarly, Bawazeer et al. [37] conducted a study about 'Food' venues and the common users' behavior in the capital of Saudi Arabia, Riyadh and suggested that people share their experiences more frequently from food venues as compared to other venues. A study based on Foursquare containing nearly 19,000 users from New York, San Francisco and Hong Kong was conducted by Xie et al. [38] for the purpose of finding out preferences among various types of venues during different times of day.

Most of the above-mentioned literature is based on data from mobile phones or some of the most famous LBSNs used almost all over the world, such as Twitter, Foursquare etc. The study of patterns within the LBSN data is based mostly on these internationally recognized platforms, which represent most of the common users' behavior and trends [67]. However, these renowned applications are not commonly used in China. One of the most frequently used LBSN in China is called Sina Weibo (or Weibo), which is utilized by the majority of people and, therefore, famous amongst researchers for LBSN data analysis. Some examples of studies based on Weibo include finding the attraction feature of famous tourism venues in Shenzhen [68], the research on human mobility and activity patterns for analysis of urban borders [69], sentiment analysis of users' opinions from within the contextual data for finding tourism attraction features and many more. In our work, we used similar Weibo check-in data for analysis of user behavior with respect to time and venues, the contribution of different types of venues in city dynamics with the user preferences and the comparative analysis of the behavior of residents and tourists

within the Shanghai city [50, 70, 71].

The automated data collection and analysis can provide more efficient ways for the exploration of big data. Some of the significant applications of ML in exploring different aspects of web based data include disease diagnosis using IoT, web mining, data extraction on webpages, mobile localization in 3D wireless sensor networks, channel propagation, socio-technological analysis of cybercrime, cyber security detection and many more [72]. However, Previous research studies in the LBSN data analysis domain are based on huge data collected online automatically in order to get more insights, but the classification of data into multiple activities or venues has been done manually by searching through thousands or millions of records of users' data which takes a lot of time and effort by the researcher. Therefore, in this study, we propose different implementations of various ML methods for venue classification into multiple classes and predicting a desired class using the same data from Weibo for Shanghai applied in our previous research. Although machine learning enables automated venue classification, several challenges remain. Social media data is often noisy and imbalanced across venue categories, which can bias models toward dominant classes. Additionally, labeled training data is costly and scarce, limiting supervised learning effectiveness. Model generalizability across different cities or LBSNs also poses difficulties due to varying user behavior patterns. Future directions include semi-supervised learning, data augmentation, and hybrid approaches to improve robustness and scalability in venue classification.

2.2.3 Venue Class Prediction and Analysis from LBSN

In today's era, urban tourism takes place in a variety of venues throughout a city, like theme parks, historic places, and museums, and also extends to shopping malls, local neighborhoods, markets etc. [73]. Modern cities are multi-functional in nature, and various users, including tourists and residents make use of different facilities like transportation, accommodation, and restaurants; that are not exclusive for tourists. The major problem of the past in data collection and lack of knowledge can be solved by the availability of LBSN, which has proved to be a valuable resource for knowledge discovery in many fields. Vassakis et al. discussed the significance of using data from LBSN for extracting tourists' behavior and presented a case study of Crete in Greece from posts, nationality, reviews, place ranking, photos and engagements [74]. Dietz et al. combined data from multiple sources, including Foursquare data about 23,418 trips from 77 countries, and calculated various metrics to ensure the validation and significance of the considered data for extracting patterns in tourists' behavior. The focus of this

research was on identifying the countries that are visited in a single trip, the seasonal visits to different countries and the duration of travel in each country [4]. Two different groups classified as foreign and domestic tourists were compared by Maeda et al. to identify the preferences of both groups based on Foursquare and Twitter data in Japan [75]. The authors defined the tourist destination as an area that attracts tourists, for example, a theme park, a historical place, resultant or hotel. By comparing the data of local and foreign tourists in both temporal and spatial features, the authors concluded that foreigners prefer that nightlife spots should be located near the tourist hotspots.

In most cases, tourists and residents are not isolated and rather mostly share same venues and facilities in a city, as observed by analysis and comparison of the spatio-temporal patterns of both groups within the city [76]. Vu et al. [77] presented a model for extracting geographical data from geotagged images from LBSN and used the geo-data to examine tourist travel patterns in different tourist destinations in Hong Kong. They conducted a more specific analysis by identifying several key areas of interest for tourists, mostly situated in the areas near the downtown. Tourists' behavior was examined by Talpur and Zhang through sequential pattern mining for mining trends and establishing patterns for future predictions in tourism activities using Foursquare data in Singapore. The authors focused on finding the similarities between the activity of different tourists in order to identify the regularities in the patterns for better planning and management of these more consistent activities at various tourist attractions [78].

The efficiency of check-ins from Foursquare for analyzing the tourists' and residents' behavior, instead of surveys and other traditional methods, was presented by Silva et al. [79] with the application of such analysis for the most famous cities from four continents, namely, Landon, Tokyo, New York and Rio de Janeiro. The results of the analysis identified locations that are more correlated with tourists and some related to residents along with the most frequently visited locations by both the tourists and residents within the cities. A study of users in eight European cities conducted by Garcia-Palomares et al. [80] proved the high concentration of tourists' activities at tourist hotspots as compared to the residents' activities by comparing LBSN data from Foursquare containing images and GPS information. They also found the patterns within the eight cities and concluded that some of the cities with world famous tourist destinations, like Paris and London, have more dispersed tourist as compared to the other considered cities. Paldino et al. [81] and Kotus et al.[82] also stated that tourists are mostly active in central areas of the city, whereas residents are active in socializing places like squares, parks, and sports facilities by comparing both the tourists' and residents' data. Paldino et al. used different

datasets validating the findings from a source other than the geo-information provided by the LBSN. The dataset contained about 100 million publicly shared images collected over the period of 10 years, and all the images were labeled as tourists, residents or unknown to America and Europe. While the article of Kotus et al. considered Poland for a similar study, with the goal of finding the patterns in the development of tourism related structures and the tourists' behavior in comparison with the residents.

These studies analyzed tourist behavior based on data from various LBSNs in different parts of the world. At the same time, Weibo has been considered by many researchers for such kind of analysis for Chinese tourists. For example, Gu et al. [66] explored different tourism spots attraction features using check-in data from Weibo in a mega-city China called Shenzhen. They concluded that theme parks from early 90s were observed to be more popular compared to the latest established parks, geographic accessibility effects the number of check-ins significantly from the different hotspot and that the results from data analysis provide more useful insights for tourism management and marketing strategies. In urban tourism, the activities of tourists and residents are not only unevenly distributed in space but also in time [83].

The study of Chinese tourist activities in Lijiang, China, by Li et al. [84] based on a famous LBSN platform, including photos of 356 Chinese tourists of about 8 years and used geo-information to perform statistical analysis in order to determine various patterns of tourist mobility and tourist hotspots all over the city. A similar study was conducted for tourists by [84] Tang & Li in the Chinese city of Xi'an to examine the patterns in tourist flow behavior, and function of traveled areas using users data [85]. The study revealed that tourism aggregation in some specific areas of the city and hierarchical features in these patterns classified the travel nodes and routes into different categories based on their spatial distributions. A more recent study from Liu & Shi about Chinese tourism in Hangzhou city by using check-ins from Weibo analyzed the influence of intercity high-speed trains on tourism the community in the region. The authors concluded that travel accessibility and tourism are highly correlated with each other by using the temporal and quantitative analysis of the data along with other factors like weather, and proximity of nearby cities [86]. Another study by Wang et al. used Weibo data to mine the tourism crowd in Shanghai [87]. This study initially analyzed Weibo data to find the popularity of tourism venues, followed by sentiment analysis of tourist opinions from Weibo contextual information.

The application of Weibo data for tourists' travel patterns through LBSN data analysis with a case study of Chinese tourists in Australia, five popular analysis methods

are discussed by Chen et al., including Geographic Information Systems, Gravity Model, Markov Chains Model, Sequential Rule Mining and Social Network Analysis [88]. The history of visited tourist destinations and the sequence of tours are very important and crucial for tourism management and planning, especially in helping to predict the route choices and travel demand of various tourist spots. Hasnat & Hasan used Twitter data to analyze the tourist choice from the longitudinal travel data of the tourists for the study area of Florida [89]. They used the Conditional Random Field method on the sequence of locations visited by the tourist to predict the next expected destination. The authors further explained the possibility of expansion of this method by adding more content-based characteristics to get more flexible and accurate predictions. A valuable opportunity for prediction through big data analysis is provided by the extensive use of LBSN by the tourism community. One of the most common methods used by researchers for the prediction of tourists' next location based on their preferences is the Markov model. For example, Liu & Wang proposed a recommendation system by applying Markov Model to the data acquired from Weibo for predicting the next Point of Interest (POI) based on patterns in historical travel data of the tourists and their preferences while considering their current locations [72]. While LBSN data offers valuable insights, integrating mobility data has the potential of complementing LBSN data for exploring the activities residents in urban environment. Cagliero et al. proposed a multi-dimensional approach that characterized urban activities in locations, time, and data source type that allows for a comprehensive analysis of urban activities. By mining common activity trends in both LBSN and mobility data, a deeper understanding of urban activities can be achieved [90]. To gain a more extensive understanding of patterns in travel and location attractiveness, Pezenka et al. provided the linkage of geo-tagged images with reactions from users. Geo-tagged images offer insights into user travel behavior, while likes and comments act as proxies for user awareness of common locations, providing a more holistic view of travel patterns [1]. These articles provide evidence of using LBSN data for activity analysis. However, our research extends these studies by analysis into the contribution of venues in the activity patterns with the help of ML models.

The application of LBSN data in tourism research has emerged as a valuable tool for studying long-term assessments of tourism. LBSN data provides a rich source of long-term data used for studying tourism-related activities and tourist flows. According to an article by Senefonte et al., geo-tagged data from platforms like Twitter and Flickr can be leveraged to extract spatial signatures, enabling study of the spatial distributions of city tourism and assessing tourism development over time [91]. Similarly, a study

by Senefonte et al. states that comparing LBSN data with official tourism data has demonstrated a high correlation, particularly in representing the global tourism network. LBSN data, particularly the platforms like Foursquare, can effectively capture the actual tourism flows, making LBSNs a valuable sources for understanding global tourism [92].

In specific urban contexts by Wang et al. and Terzi, LBSN data has been utilized to quantify and categorize cultural attractions from tourist preferences and travel trajectories. By analyzing LBSN data, such as Instagram geotags, researchers have identified clusters of cultural attractions of tourism and presented typologies for historical and contemporary based features. These categorizations enhance our understanding about tourist preferences and present valuable insights for destination management [76]. The relevance of social data, including LBSN data, has been explored by Bhatt and Pickering in countries with low income, which rely on tourism, like Nepal. Analysis of metadata related to pictures on platforms such as Flickr, researchers can study spat-temporal patterns in tourism, identify specific tourism attributes, and evaluate tourist sentiments and satisfaction. This research discussed about the understanding about how social media data can effectively be utilized in developing countries for tourism research [93]. These studies use tourism related data by going through the heterogeneous data provided by the LBSNs and don't provide powerful and efficient machine learning. Therefore, we proposed machine learning models for prediction and classification of LBSN data.

The past decade has seen a surge in research on urban tourism using various innovative analytical approaches and data sources. An recent article demonstrated the utility of geo-tagged social media data in exploring tourism flows, with one study presenting an approach involving the weighted inclusion of comments and likes [1]. Another study used geo-tagged social media records for analyzing the spatial behavior and distribution of tourists in Tokyo's major tourism sites [94]. These research findings highlight the value of social media which is rich resource for information about tourism patterns and behaviors. Significantly, studies have used LBSN data to provide an overview of tourist activity and mobility in urban environments. This approach allows researchers to monitor changes in city tourism geography over time. Interestingly, the application of LBSN data in a sustained manner has been found to inform urban planning and design in tourism destinations [91]. Research has also ventured into the interpretation of tourism intensification using innovative concepts. For instance, one study used fractals and fractal dimension to explore tourism boosting patterns in Lisbon and Oporto, revealing significant associations between the dimensions and tourism intensification across urban spaces [95].

Table 2.1: COVID-19 Era Research Topics and Questions

Consumer Behavior	<ol style="list-style-type: none"> 1. Identifying how consumer preferences are changing in relation to the reliability, cleanliness, and safety of a location or service provider. 2. Improving our knowledge of consumer and employee psychology to comprehend how they perceive risk and behave when faced with it. 3. Comparing consumer attitudes and behavior between hotels and Airbnb before and after COVID-19. 4. Using evolutionary psychology to comprehend and explain how short- and long-term effects of the pandemic on tourism include disease dangers, social isolation, and economic hardship. 5. Examining the need for travel as the most important need in Maslow's hierarchy. 6. Investigating how travelers manage uncertainty in their trip plans. 7. Measuring and monitoring shifts in the perception of the destination (some places are riskier than others due to poorer hospitals, stricter sanitary regulations, etc.). 8. Recognizing the pandemic's contribution to the rise in stereotypes, xenophobia, and ethnocentrism. 9. Analyzing visitor motivations and travel habits, as well as the growing significance of rural tourism. 10. Highlighting the significance of tourists' impressions of crowded places. 11. Examining how the epidemic has affected global attitudes of tourism. 12. Analyzing locals' perceptions of the influx of tourists to less inhabited rural areas. 13. Determining that domestic or non-traveling tourism is a result of a financial crisis, a conscious decision to lessen tourism's harmful effects (on the environment and in the case of a pandemic), or both.
Performance Modeling	<ol style="list-style-type: none"> 1. Analyzing the effect of COVID-19 on revenue, employment, demand, company transformation, career change, pay, and benefits. 2. Analyzing the impact of COVID-19 on the socioeconomic circumstances of a destination. 3. Evaluating destinations' regional disparities in their COVID-19 resilience. 4. Evaluating COVID-19's geographical effects at multiple spatial scales, including the national, state/provincial, municipal, and site levels. 5. Analyzing how (geo)politics affect destination management and policy. 6. Examining how tourism demand in the future may be shaped by the rise of adaptive workspaces (such as digital nomads).
Forecasting	<ol style="list-style-type: none"> 1. Predicting demand for tourism in the face of serious public health emergencies like COVID-19. 2. Investigating more economic and judgmental approaches to scenario forecasting. 3. During COVID-19, there will be a greater emphasis on tourism forecasting. It is crucial to know when visitors intend to return so that hotels and other service providers may be better organized and use the appropriate methods. 4. Analyzing booking patterns using big data from social media, travel websites, and search engines.
Destinations	<ol style="list-style-type: none"> 1. Investigating whether touristic establishments, like hotels, may be redesigned as multipurpose buildings that can be converted into healthcare facilities. 2. Increasing tourism stakeholders' and destinations' resilience. 3. Increasing familiarity, loyalty, and trust between travel locations and their guests as a way to combat rising travel risk perceptions.
IT	<ol style="list-style-type: none"> 1. Evaluating the effects of technological advancement on business success and employment. 2. Evaluating the effects of technological advancement on consumer satisfaction and safety. 3. Examining the potential use of blockchain technology and cryptocurrencies for global travel. 4. Examining how e-tourism and digital media might provide opportunities and solutions in a world afflicted by a pandemic. 5. Investigating how machine learning and artificial intelligence could improve the actual travel experience. 6. Improving online interactions, such as through augmented reality excursions and online social interactions.
Sustainability	<ol style="list-style-type: none"> 1. Analyzing the shift in emphasis towards the welfare of customers, residents, and employees. 2. Researching sustainable and safe tourism development. 3. Better balancing the needs of residents, visitors, and vacationers. 4. Promotion of environmentally and socially responsible tourism products that is intelligent, not ideological. 5. Investigating the expanding significance of health tourism. 6. Downplaying the importance of tourism research that primarily intends to promote mass tourism with negative environmental effects.

In the wake of COVID-19, the focus has shifted towards exploring the future of tourism. A study presented for developing an reference for post-COVID-19 tourism analysis, soliciting expert opinions on potential future topics, existing research areas of relevance, and recommendations for data collection [96]. Moreover, research has examined the impact of increasing tourism flows on cities. Another study focused on the transformation of the Vila de Gracia neighborhood in the Barcelona, exploring residents' attitudes toward changes in tourism oriented businesses, housing market prices, socio-demographic changes, and the use of public space and nightlife activities [97]. However, due to COVID-19 restrictions and suspension, most of the tourism venues were closed and to the best of our knowledge, not many articles as compared to normal days were published during the pandemic in LBSN and tourism domain. According to Assaf et al. [96], the possible and suitable research topics during the pandemic vary in nature and suitable directions are listed providing insights to the significant research in tourism analysis effected by the COVID-19 pandemic [97, 98]. The list of topics and questions attracting the interest of researcher [99, 100] for this era is given in the Table 2.1.

In short, the reviewed articles highlight the increasing importance of LBSNs as a source of data for activity analysis and extracting useful patterns into various research domains such as tourism. LBSN data offers long-term insights into tourism, complements other data sources like mobility data, and provides valuable information on travel patterns, location attractiveness, and tourist preferences. However, the most common drawback in most of these articles is extracting data of specific category for research into different research domains like restaurants, parks etc. manually. Therefore, this thesis presents Machine Learning approaches for LBSN data classification and prediction, along with density estimation with anomaly detection, and a unified framework for spatio-temporal analysis and tourism. A brief description experimental design of the proposed methodology and devised solutions for class based LBSN analysis in the domain of communication and information systems is provided in the following section.

2.3 Hybrid Recommendation System

The research on efficient recommendation engines is becoming more and more necessary as LBSNs emerge as a primary platform for real-world user interactions. LBSNs present unique recommendation challenges by integrating contextual factors such as user location, check-in timing, and personal preferences, extending beyond traditional user-item interaction models. The CF approaches widely used in most recommender systems [101],

excel at generating similar based recommendations tailored with user behaviors. However, the contextual aspects inherent in LBSNs, when the user preferences dynamically shift with location and time, presents significant limitations for traditional CF techniques. Traditional CF methods face significant challenges in LBSNs due to the sparse user-POI interaction matrices, which limit recommendation accuracy. Incorporating contextual factors such as time, location, and social trust improves personalization but increases model complexity and computational cost. Hybrid recommendation systems attempt to balance these issues but often struggle with scalability in large, dynamic LBSNs. Moreover, multi-criteria preferences and evolving user behavior demand adaptive models that remain active research areas.

The complexity of LBSNs has motivated research into integrating contextual information to enhance recommendation accuracy and relevance. While CF techniques have been successfully applied across domains like media streaming and e-commerce [102], their direct application to location-based networks remains challenging. Matrix factorization methods, including SVD, have improved CF effectiveness by reducing user-item interaction matrix dimensionality and uncovering hidden factors. Nevertheless, the data sparsity problem in LBSNs, where users visit few POIs, creates substantial gaps in interaction matrices [103]. These constraints underscore the necessity of developing innovative recommendation methodologies that can effectively capture real-world contextual features and manage sparse data environments.

To address limitations in conventional CF for large-scale social networks, context-aware recommendation systems have been proposed by researchers like E. Ezin et al. [104]. These systems integrate external contextual factors such as user time and location into recommendation mechanisms. LBSN users typically prefer locations spatially closer to their current position, making geo-influence a critical system component. As discussed by Yuan et al. [105], users visit different places at varying times, establishing temporal factors as essential in LBSN recommendations. Some studies have expanded contextual considerations beyond spatial and temporal parameters. For example, individual personality differences significantly impact location preferences extroverted individuals might seek active environments, while introverted users prefer calmer settings [106]. For instance, people with different personality traits like extroverted individuals typically gravitate toward vibrant environments, while introverted people seek serene settings. Researchers like Hossein [107] and Arab [108] discusses integrating personality traits can enhance recommendation, potentially improving overall user satisfaction in LBSNs.

LBSN research shows social connections as a valuable information source for im-

proving recommendation quality. Social network analysis, as discussed by Wang et al. [109], can be integrated into LBSNs to leverage user relationships, particularly in scenarios with limited user-item interaction data. The authors suggested that users' preferences are significantly influenced by their social circles, friends, coworkers, and contacts can inform and shape location choices. Social trust-based models, introduced by researchers like Kanfade et al. [110] and Bhaumik et al. [111], extend traditional Collaborative Filtering by incorporating trust scores between users, enabling more nuanced preference generalization. Wang et al. [112] proposed a circle-based recommendation system that generates suggestions by grouping friends into distinct circles, focusing on preferences within these social networks. Recent developments in recommender systems increasingly recognize that user preferences in LBSNs are shaped by multiple factors beyond traditional user-item interactions. Recently, Decision-making in these systems now incorporates multi-criteria considerations such as cost, ambiance, accessibility, and user reviews. Multi-criteria rating systems (MCRS) [113] enable more complex, detailed recommendations compared to traditional collaborative filtering techniques. Unlike the conventional CF approach by Nian et al. [114], multi-criteria CF [115] extracts multiple dimensions of user preferences to capture the trade-offs users consider when selecting locations. In LBSNs, this approach allows users to prioritize criteria like distance, reflecting the nature of location-based decision-making.

Since traditional CF methods do not consider handle contextual features, leading to the development of hybrid recommendation approaches that combine multiple strategies to enhance overall performance [116]. These hybrid systems integrate advantages from CF, context-aware modules, and content-based strategies to mitigate individual method limitations. CF and context-aware approaches, including SVD, are frequently used to capture complex interactions between users and items. This approach is especially vital in LBSNs where user-item interactions are typically sparse and latent factors can mask user-POI relationships [117]. The increasing trend of collaborative destination planning has shifted focus from individual to group recommendation systems [118]. Group recommendation systems face the complex challenge of satisfying multiple users' preferences, often complicated by conflicting individual desires. Various grouping methods have emerged to address these challenges, including social preference analysis and voting processes. These aggregation techniques aim to compute group-level recommendations while carefully balancing individual preferences [119]. By incorporating contextual and social aspects, group recommendation algorithms become more sophisticated in suggesting relevant choices for a group along various venue categories.

Therefore, the limitations of traditional CF in handling sparse data and contextual factors like time and location in LBSNs, advances in hybrid systems integrating CF and social trust modeling have improved personalization but still face challenges in combining contextual and social dimensions, especially for group recommendations. In line with these findings, propose a hybrid recommendation framework that not only leverages CF and SVD for handling data sparsity but also integrates contextual factors such as time and location.

This section provides the framework and brief description of the study design for this research, from the study area, and data collection to processes and outputs. Each of the methods presented in this chapter is explained in detail in the following chapters, along with their implementation and results. The dataset is collected from Weibo, one of the most popular micro blogs in China, using a Python based Application Programming Interface (API). The extracted data is preprocessed, and the proposed methodologies are implemented, yielding to our findings that are presented here in this thesis. In the initial stages, the dataset is explored using traditional statistical methods to show its efficiency for further analysis. We repeatedly used multiple linear regression and other famous methods such as ANOVA, adjusted R square and others (explained where used), appropriate at each stage of our research for showing the efficiency and finding suitable variables for different types of analysis presented in the current thesis.

The statical modeling is also used to find the frequency of user check-ins, highlighting multiple aspects of trends within the dataset. Some of these aspects include the tendency of users with respect to time, including daily, weekly and frequencies in longer time periods, i.e., six months or the whole year. We further expanded our study by finding check-in patterns with respect to space with the help of density estimation. For this research, we studied two famous methods called point density and kernel density estimation comparatively and found that kernel density provides more insights about the check-in patterns and therefore used it in the following studies. Another research problem that we addressed was finding the effect of various types of venues on city dynamics with the help of spatio-temporal analysis of check-in from multiple venue types. We initially classified the data manually (a very tiresome and time-consuming process), which led to our next research problem, i.e., how to use machine learning to classify the data into multiple studies for such research and predicting single venue types for specialized studies such as Tourism, thus leading to the last chapter of the thesis in which we utilized our proposed models in a real-world use case. The overall framework for this research is shown in Figure 2.3.

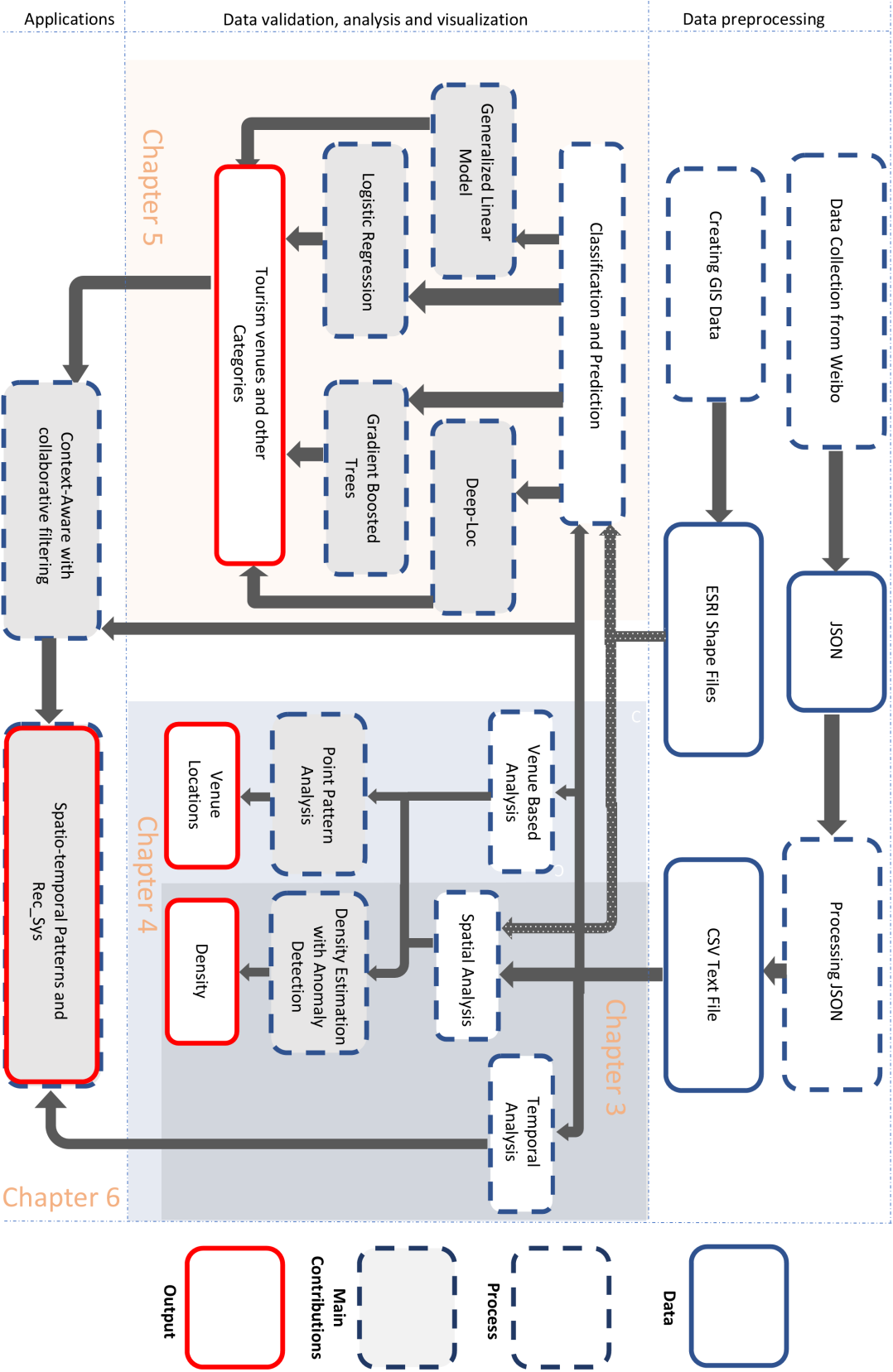


Figure 2.3: General Research Framework

Study Area This study of spatio-temporal modeling was conducted on Weibo data taken from Shanghai, China, which is situated on the Yangtze River between $30^{\circ}40' - 31^{\circ}53'N$ and $120^{\circ}52' - 122^{\circ}12'E$, with a total area of 8359 square kilometers, as shown in Figure 2.4. In 2016, Shanghai was divided into 16 districts and one county, namely Baoshan, Changning, Fengxian, Hongkou, Huangpu, Jiading, Jingan, Jinshan, Minhang, Pudong New Area, Putuo, Qingpu, Songjiang, Yangpu, Xuhui, and Chongming (which was not included as it is rarely visited by people) [120].

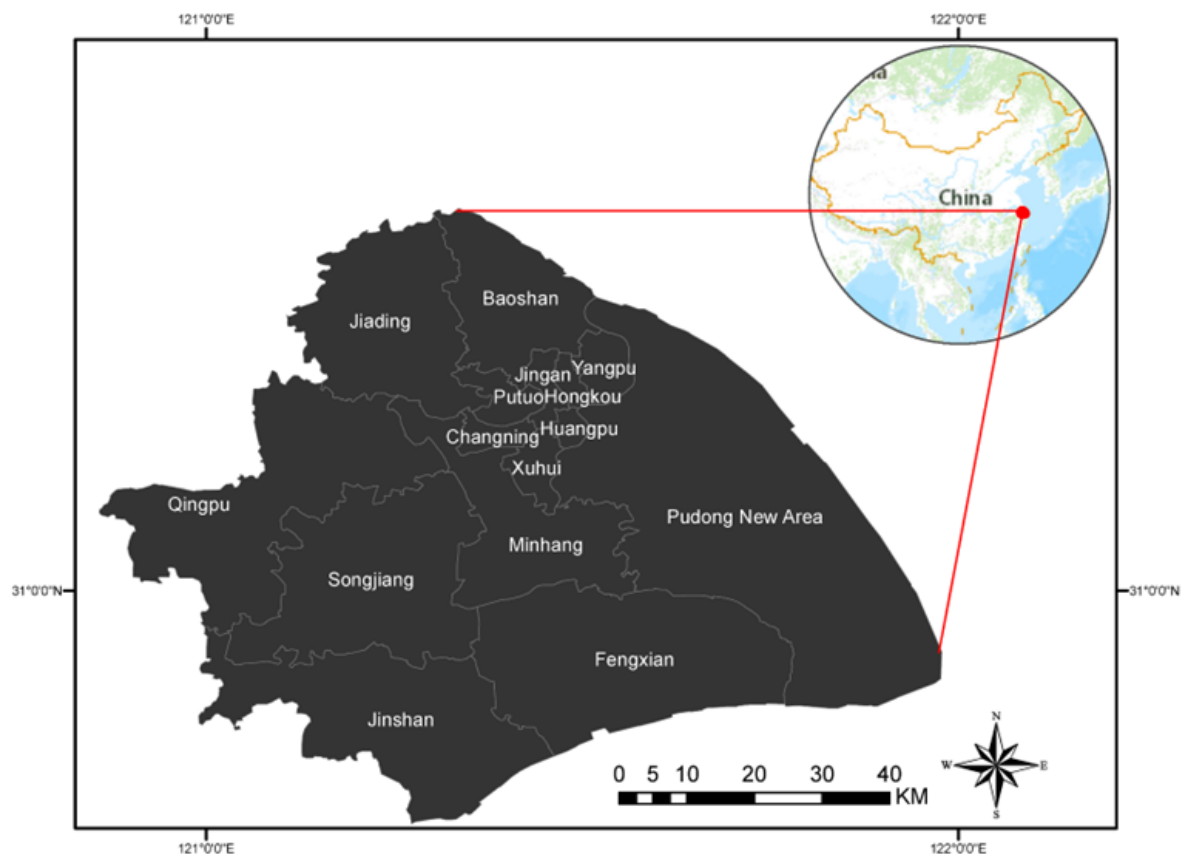


Figure 2.4: Study Area

As the economic city of China, Shanghai connects China to the global economy. The total Gross Domestic Product (GDP) of Shanghai in 2016 was 2.7 trillion Chinese Yuan, with an average 7.4% increase over the past 5 years, and the per capita GDP reached up to 15,290 USD (103,100 Yuan). With an average population of 3,854 people per square kilometer in urban areas, Shanghai has become the first city in China and fifth in the world regarding its population, with around 0.66 million people moving in annually. Its population reached from 16.74 million to 23.02 million during the last decade from 2000

to 2010, increased by 37.53% with more than 24 million residents at the end of 2015. The main reason for this growth is the large number of migrants and tourists, which made up 39% of the total population of Shanghai in 2010. The recent master plan greatly emphasizes on providing more facilities regarding development and administration for the betterment of tourists and residents of Shanghai (Shanghai Master Plan (2016–2035) [121]. Shanghai is a world-famous tourist destination with many renowned attractions, such as Oriental Pearl, Lujiazui, Century Park, Yu Garden, Jing'an Temple, Nanjing East Road, and the Bund etc., mostly situated in the city center, while there are more than 800 parks in different parts of the city.

Data Source: The data used in the current study are from one of the most popular Chinese microblogs, Weibo. Facebook and Twitter are the most popular LBSNs in the world. In China, Weibo, a hybrid of Facebook and Twitter, is one of the most dominant LBSNs [63]. It has become a major platform, enabling users to share their activities, opinions, preferences, and locations along with audio, images, and videos through checking and writing posts, alongside communicating with their friends. Since Weibo was launched on 14 August 2009, the number of users, check-ins, and activities has increased rapidly. Weibo provides different types of geo-spatial resources; three of the main resources include user-profile locations, places mentioned in posts, and sharing real-time locations through check-ins. By the end of 2018, the total number of users increased to over 500 million, reaching 462 million monthly active and 200 million daily active users. Weibo launched an international version in March 2017 and claims to have users in more than 190 countries [122]. Figure 2.5 shows information about Weibo according to the famous information provider Wikipedia [123].


Sina Weibo (新浪微博)	Type of Business	Type of Site	Available in	Owner	URL	Launched	Current Status
	Public Company	Microblogging	Simplified Chinese	Weibo (Operated by Weibo Corporation)	weibo.com	14-Aug-09	Active
			Traditional Chinese				
			English				

Figure 2.5: Characteristics of Data Source

Data Acquisition: The primary inspiration for the use of LBSN is to share interests and activities and thereby build new and close social relationships, enabling researchers to discover patterns in users' activities and preferences from the big data generated by the LBSN. The data source for this research is Weibo, one of the biggest and most

popular micro blogs in China, which allows users to 'check-in' using their mobile devices. According to the Weibo press release, the number of monthly active users in China reached 462 million by December 2018, about one third of its entire population.

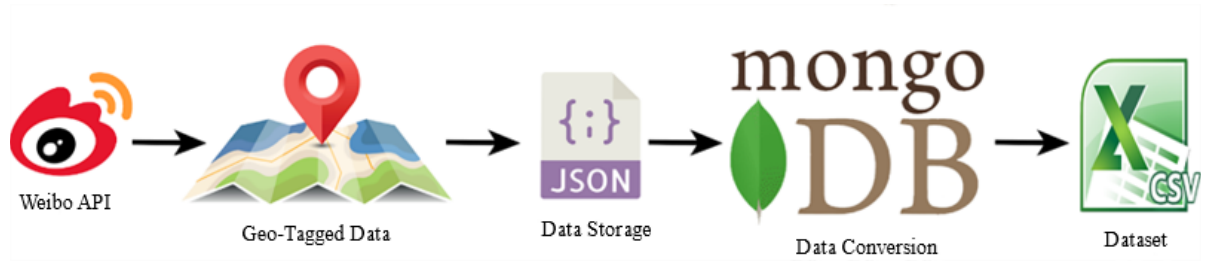


Figure 2.6: Data Acquisition Process Using Weibo API

Weibo provides a python based public Application Programming Interface (API) to search and download the geo-tagged data, which can be used for analysis of spatio-temporal patterns. We used the API to collect data in specific areas of China, specifically Shanghai, for 2017 (the year of my enrollment in Shanghai University). There were approximately 3.5 million check-ins of about 2 million users. The data was in the JavaScript Object Notation (JSON), which was preprocessed for analysis to Comma-Separated Values (CSV) with the help of MongoDB for the current study, as shown in Figure 2.6. Following Table 2.2 shows the sample of the dataset.

Table 2.2: The Sample of Attributes in Initial Dataset

User ID	Gender	Check-in Date	Check-in Time	Origin	Location-ID	Latitude	Longitude	Location Name
56xxx20	f	5/8/2017	2:23:12	Shanghai	B2xxx9A	31.29164	121.31098	Nanxiang Hospital
16xxx62	m	4/27/2017	12:17:47	Shanghai	B2xxx93	31.314793	121.45629	Lingnan Park
34xxx42	f	8/16/2016	15:44:45	Beijing	B2xxx9B	31.281199	121.507732	Siping Cinema
15xxx44	m	1/2/2017	15:22:32	Shanghai	B2xxx98	31.28861	121.537	Yanji Library

Spatial Density Estimation (Point Density vs KDE):

The density of total check-ins and the impact of various venue kinds on density estimation have been examined using geo-data on the map of Shanghai city. Spatial analysis is carried out utilizing the map characteristics and shape files of map features. streets, districts, metropolitan regions, and others. For the spatial analysis, we first compared the two commonly used methods called Point Density and KDE in our study. KDE provides more precision and smoothness. While Point Density offers a straightforward and computationally efficient way to estimate check-in concentrations, it suffers from abrupt boundary effects and sensitivity to the choice of radius, which can lead to misleading hotspots. In contrast, KDE applies smoothing via a kernel function, producing continuous

density surfaces that better capture underlying spatial patterns. However, KDE can be computationally expensive on large datasets and is sensitive to kernel bandwidth selection. Despite these limitations, KDE generally provides superior spatial insight for LBSN data with complex, overlapping user activity zones. Nevertheless, a hybrid approach or adaptive bandwidth KDE could be explored further to balance precision and computational efficiency, a research gap this thesis begins to address.

2.4 Summary

To sum up the discussion, the goal of this dissertation is to provide suitable models for the analysis, classification, and modeling of the data. And to analyze and extract the patterns of LBSN (Weibo) data for Shanghai, China, using various suitable temporal and spatial analysis techniques along with different check-in venues for the proposed real-world applications. Our research explores the use of LBSN data by examining the association between time, frequency of check-ins, user preferences and venue classes based on contribution and city's characteristics and a general framework for feature extraction of different venue classes. The extensive literature review helped us find the research gaps that include the need to propose a framework for data analysis as different research uses a variety of methods for the same kind of analysis. Also, most of the researchers use KDE for spatial analysis without comparison to other density estimation methods, such as point density, as presented in our research. Furthermore, to avoid filtering millions of check-ins to find the related venue type which are considered for various research, as there are hundreds or thousands of venues available within the data led to our proposed spatio-temporal analysis with anomaly detection, deep learning-based classification models and, finally, application in recommendation systems.

While numerous analytical techniques have been applied to LBSN data, each presents unique advantages and drawbacks. Statistical and spatial methods provide foundational insights but may lack scalability or granularity. Clustering algorithms like DBSCAN improve anomaly detection but require careful parameterization. Machine learning enables scalable classification but demands extensive labeled data and robust feature engineering. Hybrid recommendation systems address contextual challenges but introduce complexity and scalability trade-offs. This critical evaluation highlights the need for integrated frameworks that balance accuracy, efficiency, and interpretability, motivating the novel hybrid approaches developed in this thesis.

This chapter further contains a brief description of the study design of this thesis. We

started with the overall framework of this research depicting the sections and parts of analysis formulating the tourism analysis. The methodologies for the analysis explained in each following chapter are also described here, starting with an introduction of the study area, Shanghai, and some facts about this famous city in China. The details about the dataset and data acquisition process from Weibo, along with the attributes and records included in this study. The following chapters are based on the analysis and detailed studies of the proposed methodologies presented in the overall research framework including temporal and spatial methods, along with different methods used for our research. We provide the analysis methods used explored in this thesis, starting with temporal analysis, venue categories for the study of the contribution of venue types in city dynamics, defining ML models used for prediction and classification of venue categories, the MLR model used to access the feasibility of variables, and the spatial analysis mainly related to the KDE, proposed machine learning models, and finally the application of these contextual features with collaborative filter to develop an efficient group recommendation system for LBSNs.

ANALYSIS OF SPATIO-TEMPORAL MODELING OF LOCATION-BASED SOCIAL NETWORK DATA

The aim of the study in chapter is analyzing patterns using temporal and spatial analysis techniques within LBSN data from Shanghai, China. The research examines relationships between time and check-in frequency, considering both user behavior and urban characteristics. Statistical models and comparative spatial methods demonstrate the significance and effectiveness of the methodological approach. Results are presented through statistical visualizations, tabular data, and spatial heat-maps. The analysis is carried out with the help of SPSS for temporal analysis and Kernel Density Estimation (KDE) using ArcMap and OpenStreetMap for spatial analysis. The findings reveal diverse patterns in user check-ins across multiple dimensions: hourly, daily, and six-month periods, incorporating gender analysis, frequency distribution, and check-in density patterns throughout Shanghai.

3.1 Introduction

The analysis of spatio-temporal data has emerged as a crucial research focus, driven by the popularity of LBSNs. These platforms generate extensive user-generated content that provides valuable insights for practical applications. The extracted information supports diverse applications, including public transit flow analysis, location recommendation systems, population density mapping, route optimization, disaster management response

and many more [1]. Social networking platforms encourage users to share activities and interests within their social networks, generating rich datasets that enable researchers to analyze user behavior patterns and preferences more accurately. These platforms capture and store user information with real-time location data, producing comprehensive datasets enriched with multimedia content, textual information, geo-location data, and metadata. This multi-dimensional data enables detailed investigation into various aspects of human behavioral patterns.

Researcher are increasingly focused on analyzing and modeling human activities through spatio-temporal data analysis [37]. Check-in data from LBSNs has become most frequently used data source in research studies, despite the fact that they pose sampling problems, such as biases in gender, age, and social classes. A 'check-in' occurs when users either manually confirm their location during an activity or automatically share their location through messaging or posting contents online [124]. The exponential growth of location-based services across platforms like Facebook [125], Twitter [126], Foursquare [127], and Weibo[128] specifically in China, has enabled extensive research into user behavioral patterns. These studies examine relationships across various aspect of user activities, including gender, educational backgrounds, and age groups, revealing functional characteristics both within and across cities.

Weibo is not only famous among users but also among researchers, as the check-in data from Weibo can be utilized to carry out various kinds of studies in order to extract useful information based on the geo-location data it provides. For example, some recent studies have analyzed road crashes in Shanghai [129], investigated the growth of urban boundaries in Beijing [130], and analyzed tourism venue attraction features by using LBSN data from the period 2012–2014 [68]. Most of these studies have covered the check-ins of specific users or specific application fields, such as tourism, city boundaries, road crashes, spring festival rushes, gender etc. However, to the best of our knowledge, there is no comprehensive study based on the Weibo data of Shanghai combining both temporal and spatial analysis while considering the characteristics and nature of the check-ins from different venues. Therefore, we performed different kinds of analyses on check-in data from Weibo for six months (January 1st, 2017, to June 30th, 2017) in Shanghai, covering spatio-temporal analyses with distinct venue classifications and various aspects of the users' check-in data.

This Chapter makes Four key contributions. First, we present a temporal analysis from hourly to weekly patterns, and for the total study period (180 days), of check-in data acquired from the most famous Chinese LBSN, Weibo, and consider more than 220,000

check-ins gathered over six months. Second, we show the significance of our dataset using the proposed MLR model. Third, we show the effectiveness of using KDE instead of using Point Density by estimating density for our dataset. And finally, we developed and applied KDE with anomaly detection by combining the KDE with DBSCAN algorithm to uncover different patterns in check-ins including anomalies and outliers. The temporal analysis was carried out by using IBM SPSS 25, which is popular among researchers, to discover various patterns in the data with respect to time. We present both graphical charts and statistical results with detailed discussions for a clear understanding of the research findings. For the spatial analysis, we used ArcMap, the KDE technique [131], and geospatial data (Shape Files) from OpenStreetMap [132].

3.2 Dataset and Methodology

This section of the thesis includes the dataset and methodologies used in the spatio-temporal analysis. The details starting with the dataset, framework of the research, temporal patterns, statistical modeling, comparative analysis of density estimation methods, anomaly detection and spatial analysis using KDE as given below.

3.2.1 Dataset

This study mined check-in patterns and estimated the density for spatio-temporal data from Weibo in Shanghai from 1st January 2017 to 30th June 2017, including 222,525 check-ins from 166,898 users. The sample of the data attributes and variables used in this research is given in the Table 3.1.

Table 3.1: Sample Variables

User ID	Gender	Check-in Date	Check-in Time	Latitude	Longitude	Location Name
56xxx20	f	5/8/2017	2:23:12	31.29164	121.31098	Nanxiang Hospital
15xxx44	m	1/2/2017	15:22:32	31.28861	121.537	Yanji Library

3.2.2 Methodology

The overall framework of the methods and contribution of this chapter is given in Figure 3.1.

Temporal Patterns: The temporal check-in analysis further consists of three parts: a) daily patterns, b) weekly patterns, and c) check-in patterns for 180 days (the study period of our research). It can be implemented in Python, as shown in the following Figure 3.2.

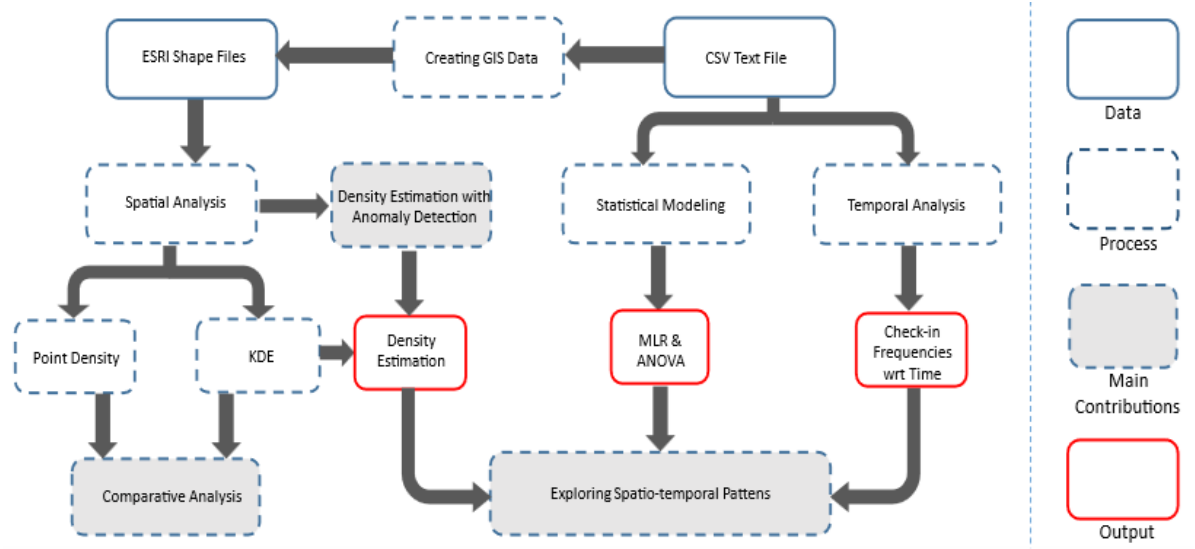


Figure 3.1: Chapter's Research Framework

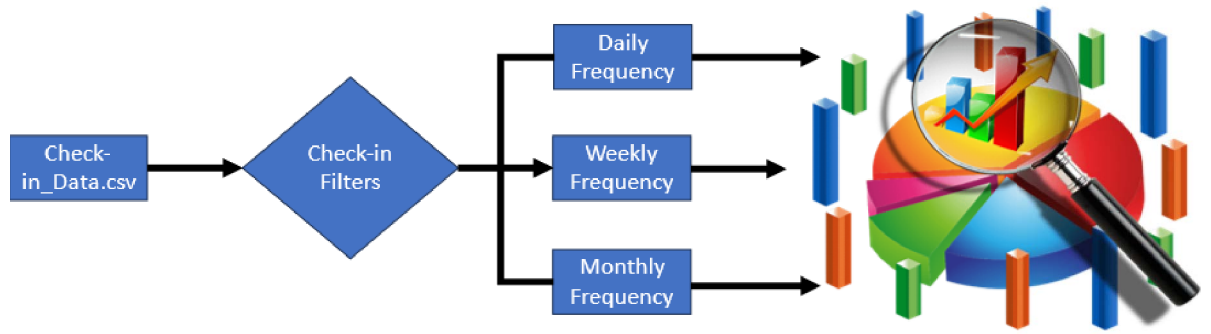


Figure 3.2: Temporal Analysis

Statistical Modeling: The statistical modeling in this research is based on multiple linear regression, through which we present the significance of our dataset and the preference of using KDE using the same dataset to study the 10 districts of Shanghai. as published in our article [2]. To find the significance of the dataset, we used Equation 3.1 for our model. The following Equations 3.2 represent the fitted value equations for our MLR model.

$$\begin{aligned}
 Y = & \beta_0 + \beta_1 \text{Baoshan} + \beta_2 \text{Changning} + \beta_3 \text{Hongkou} + \beta_4 \text{Huangpu} \\
 & + \beta_5 \text{Jingan} + \beta_6 \text{Minhang} + \beta_7 \text{PudongNewArea} + \beta_8 \text{Putuo} \\
 & + \beta_9 \text{Xuhui} + \beta_{10} \text{Yangpu} + \epsilon
 \end{aligned} \tag{3.1}$$

$$\begin{aligned}
 \hat{Y} = & b_0 + b_1 \text{Baoshan} + b_2 \text{Changning} + b_3 \text{Hongkou} + b_4 \text{Huangpu} \\
 & + b_5 \text{Jingan} + b_6 \text{Minhang} + b_7 \text{PudongNewArea} + b_8 \text{Putuo} \\
 & + b_9 \text{Xuhui} + b_{10} \text{Yangpu} + \epsilon
 \end{aligned} \tag{3.2}$$

Comparison of Point Density and Kernel Density Estimation: We demonstrate the efficiency of our spatial method by comparing the KDE with another commonly used method called Point Density. It can be implemented in Python, as shown in the following Figure 3.3.

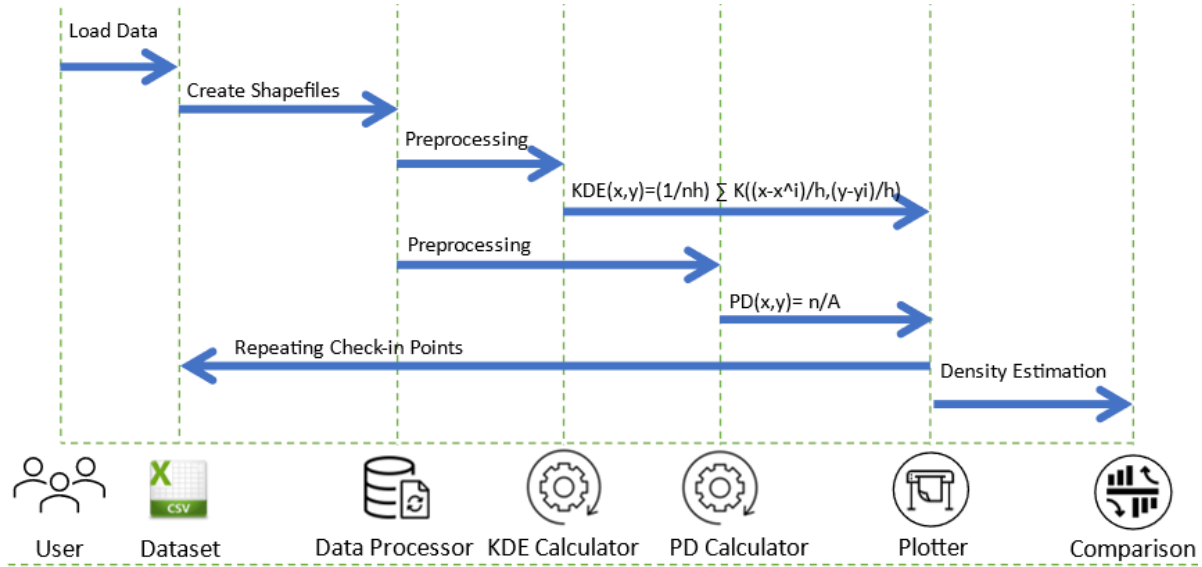


Figure 3.3: Kernel Density Estimation and Point Density

The point density method used in the current study calculates the frequency of the event intensity (density) within the neighborhood of a given point, accounting for a geo-spatial projection. The number of points per unit area at each location throughout an area of interest is referred to as the Point Density function. A “neighborhood” is defined for each point to calculate this density surface, usually by specifying a bandwidth (or search radius); the total number of points within the neighborhood is divided by the total area of the neighborhood. The point density function is expressed as Equation 3.3:

$$\lambda(a, b) = \frac{n}{|A|} \quad (3.3)$$

where λ is the point density at a location (a, b) , the number of events is represented by n , the area of the neighborhood is denoted by $|A|$, and $\lambda(a, b)$ is the unit of users per unit area. When neighborhoods overlap, the results are summed to indicate a higher density of users.

The KDE method is used for plotting more accurate and smooth density. The KDE can be defined as Equation 3.4, let D be the dataset used in the study and $D = \{d^1, \dots, d^n\}$, where $d^i = \langle x, y \rangle$ is the geo-location of each check-in from $1 < i < n$, at the time t .

$$f_{KD}(\mathbf{d} | \mathbf{D}, h) = \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{d}, \mathbf{d}^i) \quad (3.4)$$

where;

$$K_h(\mathbf{d}, \mathbf{d}^i) = \frac{1}{(2\pi h)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}^i)^\top \Sigma_h^{-1}(\mathbf{d} - \mathbf{d}^i)\right) \quad (3.5)$$

where h in Equation 3.5 is the bandwidth, which depends upon the resulting density estimation as shown in Equation 3.4, which generates a smooth density surface around \mathbf{d} on check-in [133].

Proposed Anomaly Detection Method: The proposed method for grouping the venue categories is based on DBSCAN. Considering the density of points, a dataset is grouped into clusters using this famous technique. In contrast to other clustering methods, such as k-means, it may find clusters of any shape and does not need the exact number of clusters to be specified beforehand. The neighborhood of a data point can be studied by determining its radius using the ε -Neighborhood or Epsilon (ε) distance threshold. The smallest number of points (*MinPts*) that must be situated in a point's ε -neighborhood in order for it to be classified as a core point. Hence, if a point's ε -neighborhood contains at least the *MinPts*, then it is the core. A border point is inside the ε -neighborhood but has less than *MinPts* within it. The condition for clustering is that two points p and q are in same clusters if there is a chain of points p_1, \dots, p_n such that $p_1 = p$, $p_n = q$ and $p_{i+1} \in N_\varepsilon(p_i)$ for $1 \leq i < n$.

To find clusters and outliers inside the high-density zones, we use a clustering technique based on DBSCAN. The approach will group data points into clusters according to density. These groups indicate locations that have a high check-in density over time, which are probably preferred locations or events. Unlike spherical assumption approaches like k-means, proposed method can locate clusters of any shape. It identifies outliers by successfully separating noise points from clusters and is mainly affected by ε and *MinPts*, which can be adjusted depending on specific circumstances.

Spatial Analysis: The spatial analysis is carried out with the help of map attributes collected from OSM to observe with geo-data on the map of Shanghai city. The shape files of these map attributes are used on ArcMap with a built-in Python interface to investigate the density of overall check-ins and the contribution of venue types in density estimation. OSM is a popular geo-information platform with multiple attributes, i.e., districts, streets, metro etc., used in geo-spatial modeling and research. For the spatial analysis, we first compared the two commonly used methods called Point Density and KDE in our study. The Point density is calculated as the number of points in a specified area called the radius. The calculated points and its neighbors are then divided by the total area to find the density of check-ins on the map. The results and comparison are

provided in Chapter 4 of this thesis. The proposed density estimation method is devised by combining the KDE with anomaly detection based on DBSCAN algorithm to find anomalies outlier.

3.3 Analysis and Discussions

With the advancements in online services, wireless communication, mobile devices, and location-sharing technologies, LBSNs such as Facebook, Foursquare, Twitter, and Weibo are attracting researchers' attention due to the huge amount of data generated by these LBSNs. The data can be used to extract very useful information for urban planning, crisis and disaster management, and other fields of study involving big data with a high spatio-temporal resolution. The current study had three different aspects of analysis: temporal, check-in venue classification, and spatial analysis of the Weibo data for Shanghai. This section includes results and discussion of these three aspects.

3.3.1 Temporal Patterns

The temporal check-in analysis is an important part of the pattern analysis as it reveals patterns with respect to time. There are different preferences of users when it comes to the utilization of LBSN in various parts of the day involving different crucial circumstances such as job, work, home and school routine and many more. The temporal patterns of user check-ins in our dataset are discussed as follows.

Daily Patterns: To investigate the check-in frequency pattern of the Weibo users, we observed the distribution of check-ins for 24 hours of the day, as shown in Figure 3.4. It can be observed that routine activities have a profound impact on the number and time of check-ins. For instance, the frequency of check-ins starts rising in the early morning, is considerable after 10 a.m. and is highest after 12 p.m., while the check-ins start declining after midnight. The peak of check-ins was 10 p.m. to 12 a.m., a typical time frame for the social activities of many people.

We can see in the Figure 3.4 that on the time scale from midnight to midnight (00 to 24 hours), the check-in frequency is more skewed towards the right, showing more check-ins in the afternoon, evening, and before midnight. The figure demonstrates a normal data distribution, shown by the kurtosis having a nearly normal value of 3. There are less check-ins after midnight and in the early morning because of the sleeping routine of Shanghai residents. As one of the most developed cities in China, the check-in frequency

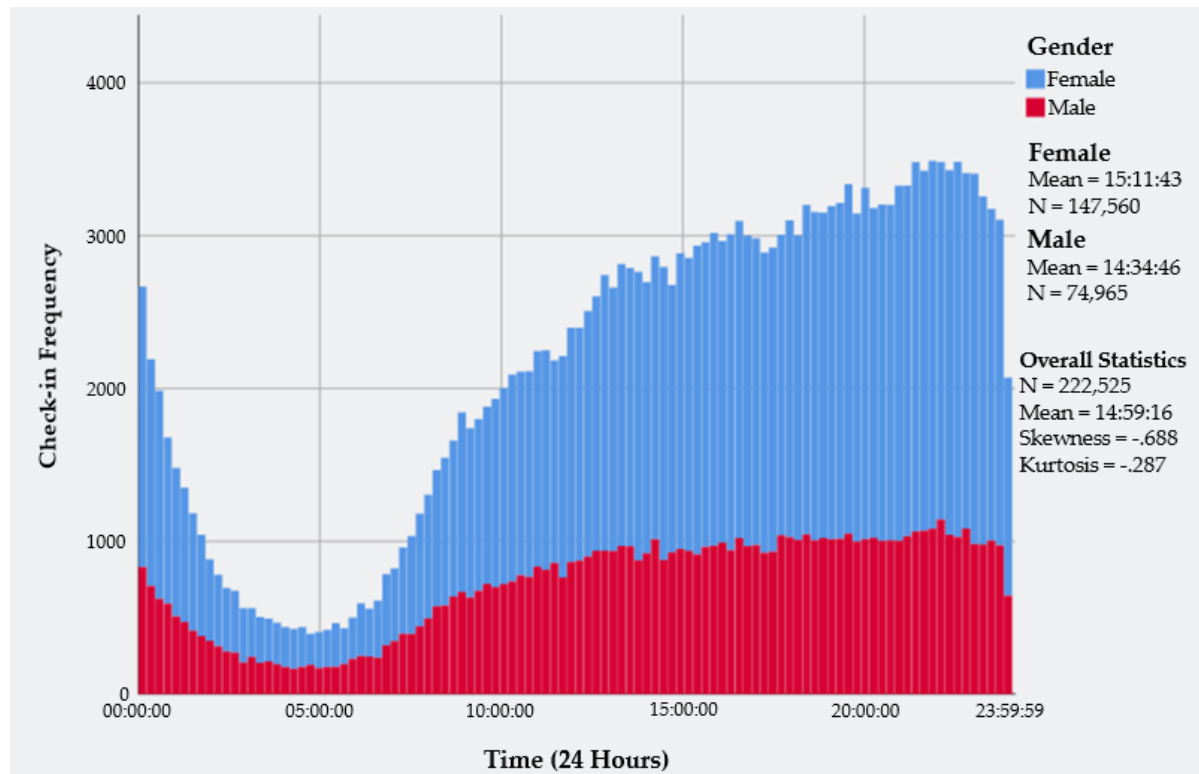


Figure 3.4: Check-in Frequencies for 24 Hours

of both male and female users is almost the same, but the number of check-ins differs because of the different numbers of males and females in our dataset. The frequency is normal until the afternoon because the people are mostly at work, and it increases as they finish work and as they meet their friends and families or visit places before eventually decreasing for the night period.

Weekly Patterns: This section analyzes the weekly rhythm of check-ins. Weekly patterns suggest that the user check-ins are predominant on the weekends when compared to the weekdays. The full view of the total number of check-ins for each day of the week is shown in Figure 3.5. It can be seen that most user check-ins took place on Saturday and Sunday, suggesting the behavior of people using LBSNs on the holidays. Users tend to increase their social activities after work on Friday, Saturday, and Sunday, and this increase sometimes lasts until night on Sunday; therefore, more activities occur from Friday night until Monday morning. This figure illustrates that the frequency of check-ins on Saturdays and Sundays is the highest, followed by Friday and Monday. Tuesday, Wednesday, and Thursday show the minimum number of social activities throughout the week.

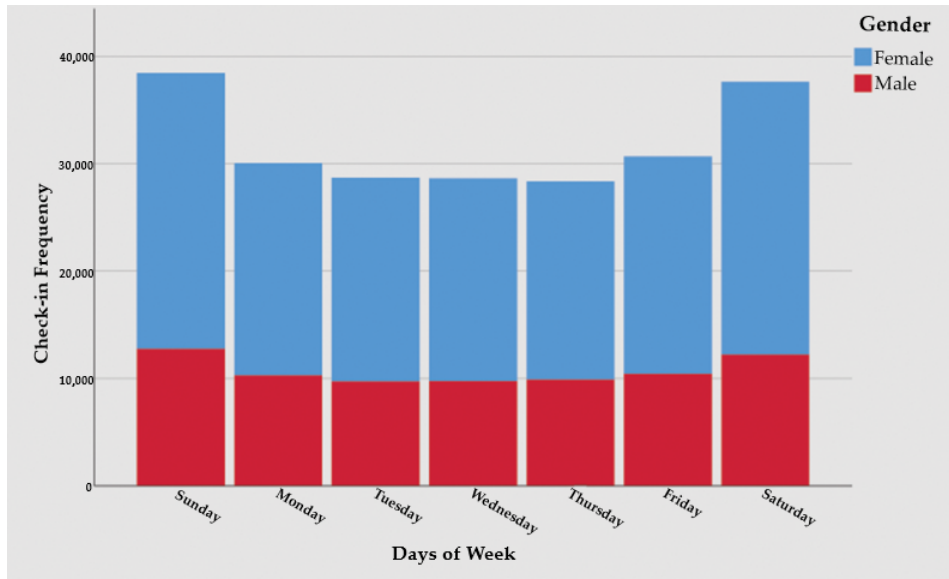


Figure 3.5: Check-in Frequencies for Days of The Week

Patterns by Date: This section represents the daily trends of the total number of Weibo users for 180 days (1st January 2017 to 30th June 2017) in Shanghai.

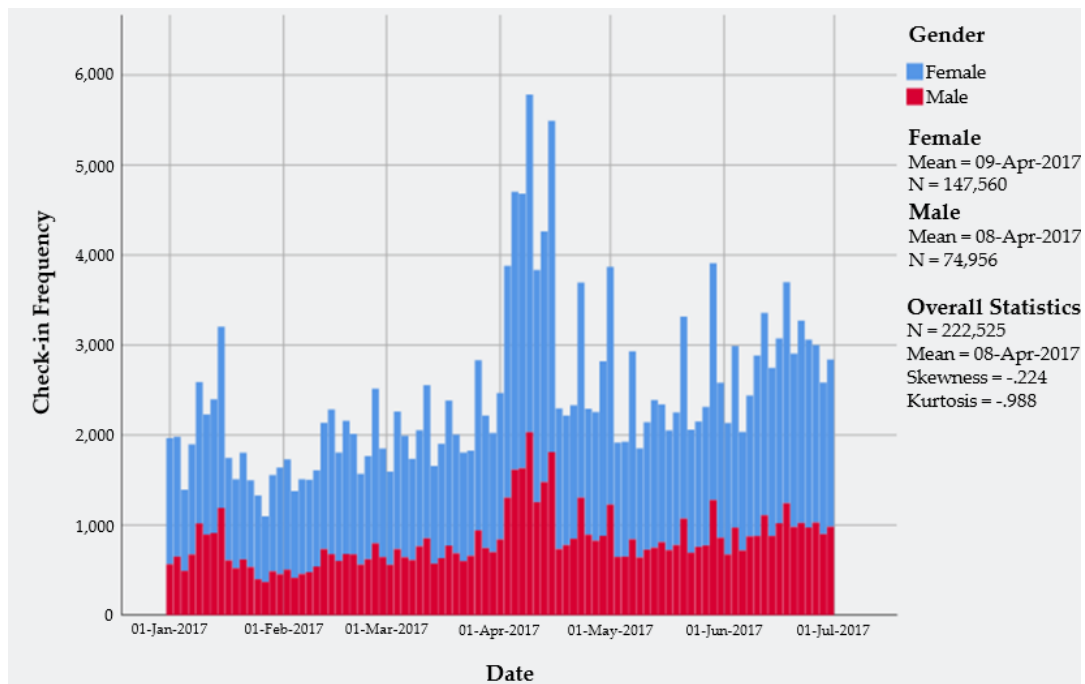


Figure 3.6: Check-in Frequencies for 180 Days

Figure 3.6 presents the variations of check-in frequencies for both males and females during the study period. Figure 3.6 demonstrates that the maximum number of check-ins occurred in the first two weeks of April 2017. Some of the main reasons for this include

Shanghai Fashion Week, the Shanghai Formula 1 Grand Prix, the Shanghai Ballet Company's 'Swan Lake' and Easter, all from the 'Entertainment' category, which had the highest impact and was also due to the number of venues related to this category in the dataset. The least number of check-ins occurred in the last two weeks of January 2017 and the first two weeks of February 2017 because of the periodic vast migration of people around the Chinese New Year or Chinese Spring Festival [134], wherein a massive number of Shanghai residents move back to their hometowns on vacation, accounting for 39% (in 2010) of the total population of Shanghai. The results also reveal that people tend to share their activities using LBSNs more on such occasions as visiting places and meeting with friends compared to being physically present at home or work.

3.3.2 Statistical Analysis

In this section, we present the significance of our dataset and the preference of using KDE using the same dataset to study the 10 districts of Shanghai. To find the significance of the dataset, The following Table 3.2 represent the fitted value equations for our MLR model. It can be observed that from the p-value of the model with regards to the inference of F score, the overall model seems significant [135].

Table 3.2: Regression Summary

Residual Standard Error	Degrees of Freedom	Multiple R-Squared	Adjusted R-Squared	F-scoere	p-Value
1.19	10299	0.3916	0.3951	397.3	$< 2.2 \times 10^{-6}$

As the model is developed with the highest Adjusted R-Squared, as shown in Table 3.2, some of the values do not show high significance.

The tables 3.4 represent the model's coefficients, showing that with each unit increment in the value, the number of user check-ins increases by about 1.6%, having very low p-value for Baoshan; similar increase can be seen for Huangpu, Putuo, Minhang, and Xuhui, approximately 1.5%, 0.8%, 1.5%, and 0.9%, respectively and low p-value indicating their significance.

3.3.3 Comparison of Point Density and Kernel Density Estimation

We demonstrate the efficiency of our spatial method by comparing the KDE with another commonly used method called Point Density.

Table 3.3: Final Multiple Linear-Regression

Summary Statistics					
	Min	1Q	Median	3Q	Max
	-23.7348	-0.6238	0.2834	0.8349	2.0153
Coefficients	Estimate	Std. Error	t Value	Pr(> t)	Significance
Intercept	4.8421088	0.0164305	294.703	$< 2 \times 10^{-16}$	***
Baoshan	0.0159932	0.0029441	5.432	5.69×10^{-8}	***
Changning	0.0079145	0.0024644	3.056	0.002245	**
Hongkou	0.0080145	0.0019644	3.211	0.001325	**
Huangpu	0.0147966	0.0019293	7.669	1.88×10^{-14}	***
Jingan	0.0016087	0.001976	0.814	0.415602	***
Minhang	0.014982	0.0028092	5.333	9.86×10^{-8}	***
PudongNewArea	0.0039008	0.0013275	3.717	0.003307	**
Putuo	0.0084851	0.0022825	3.717	0.000202	***
Xuhui	0.0091736	0.0019475	4.71	2.50×10^{-6}	***
Yangpu	0.0089936	0.0021494	4.184	2.89×10^{-5}	***

"Significance Codes: *** (p-value: [0, 0.001]), ** (p-value: [0.001, 0.01])."

Table 3.4: ANOVA

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)	
Baoshan	1	7238.2	7238.2	5109.7946	$< 2.2 \times 10^{-16}$	***
Changning	1	1251	1251	883.1285	$< 2.2 \times 10^{-16}$	***
Hongkou	1	307.7	307.7	217.2498	$< 2.2 \times 10^{-16}$	***
Huangpu	1	339.1	339.1	239.3547	$< 2.2 \times 10^{-16}$	***
Jingan	1	35.9	35.9	25.3324	4.91×10^{-7}	***
Minhang	1	77.7	77.7	54.8639	1.39×10^{-13}	***
PudongNewArea	1	34.2	34.2	24.1414	9.09×10^{-7}	***
Putuo	1	31.7	31.7	22.3638	2.29×10^{-6}	***
Xuhui	1	30.4	30.4	21.4392	3.70×10^{-6}	***
Yangpu	1	19.9	19.9	14.068	0.0001773	***

"Significance Codes: *** (p-value: [0, 0.001]), ** (p-value: [0.001, 0.01])."

The Figure 3.7 represents the density estimation using point density and DKE for the check-ins in the dataset. The above figure demonstrates the point density of user

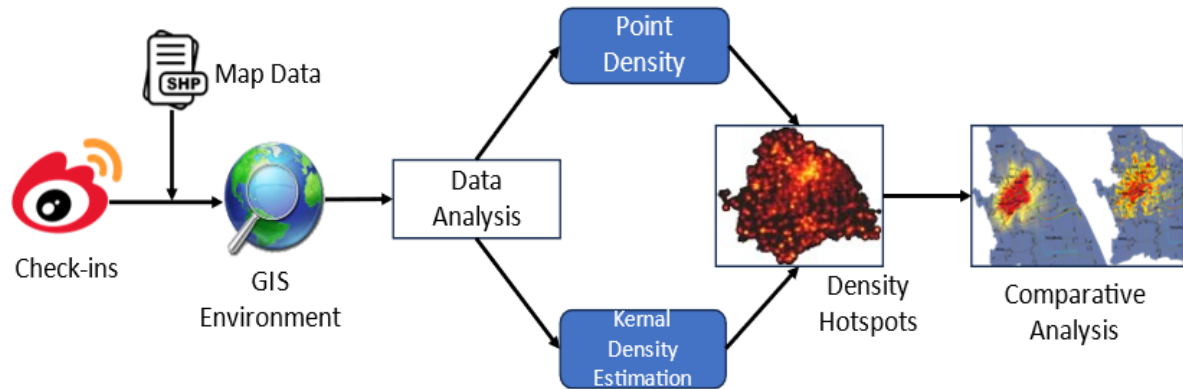


Figure 3.7: Point Density Vs Kernel Density

check-ins in the study area. It can be observed that although point density provides an overall view of the density showing the tendency of users in various areas within the city, however, it lacks the details about the areas where the check-ins are denser as compared to others. To address this issue, we used KDE to plot the density of check-ins more clearly, as shown in Figure 3.8.

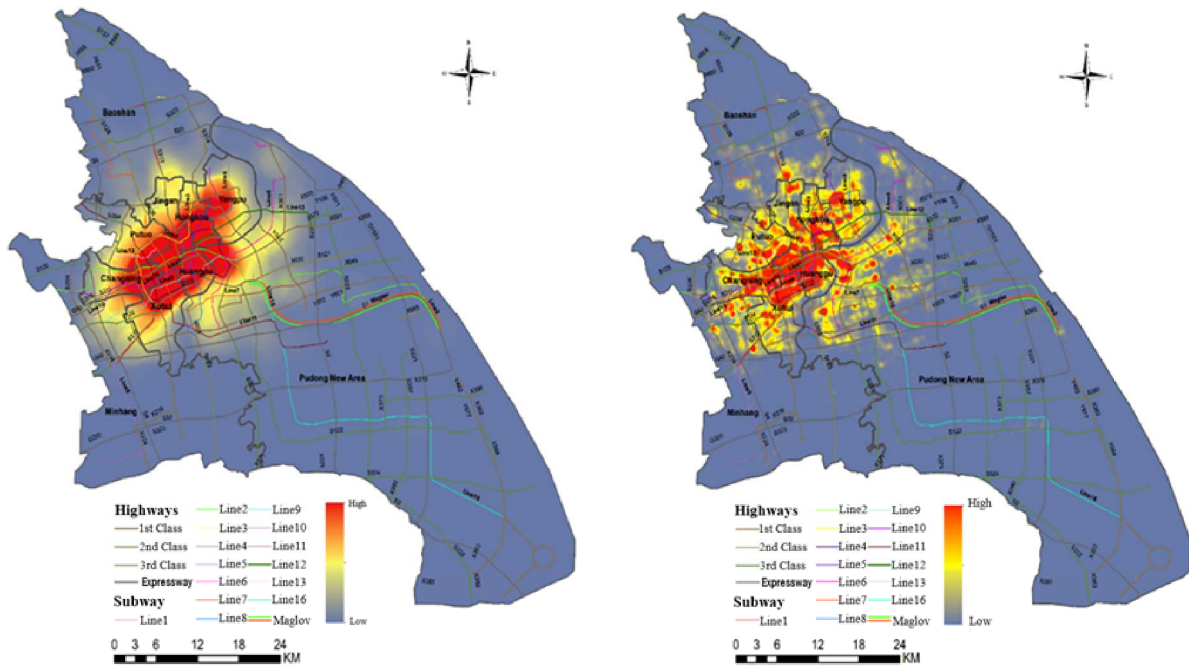


Figure 3.8: Comparison Between Density Methods: (a) Point Density (PD) and (b) Kernel Density Estimation (KDE).

Although the parameters for both the density estimation methods are kept default (The same) and both Figure 3.8.a and Figure 3.8.b shows higher density of check-ins in the center of town and lower density away from it. However, we can observe the details of the locations with fine granularity using KDE as compared to the Point Density. So,

we can identify more accurate density at specific areas rather than just the whole picture with abstract information.

3.3.4 Anomaly Detection with Venue Clustering

The proposed method for anomaly detection is based on DBSCAN for clustering the venue categories. It is a popular algorithm that groups a data into clusters by calculating the density of data points. It is explained as follows.

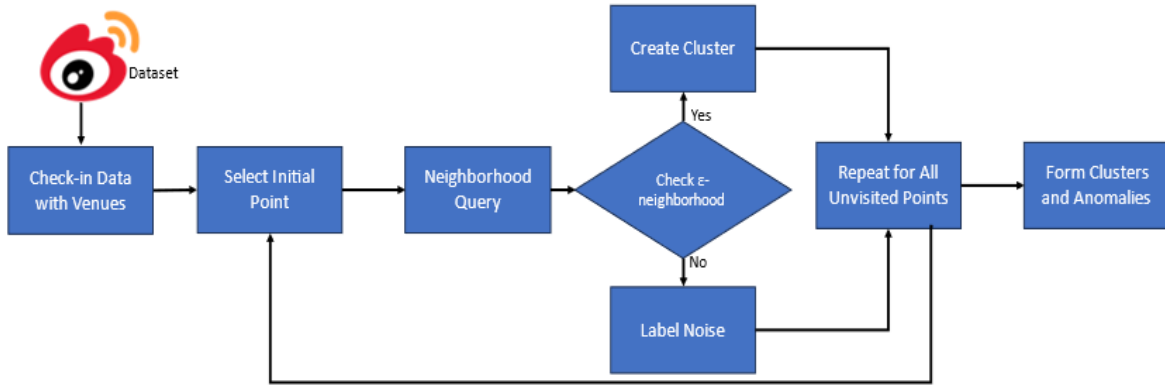


Figure 3.9: Venue Clustering Steps

ϵ -Neighborhood or Epsilon (ϵ) is a distance threshold which identifies the radius near a data point to study its neighborhood as given in the Equation 4.1.

$$N_\epsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \epsilon\} \quad (3.6)$$

where D is the dataset and $\text{dist}(p, q)$ is a distance metric.

The minimum number of data points ($MinPts$) involved within the (ϵ)-neighborhood of a point to be considered as the core point. So, a point is core if its ϵ -neighborhood has at least $MinPts$. The point p is a core if it follows the following condition as Equation 3.7.

$$|N_\epsilon(p)| \geq MinPts \quad (3.7)$$

A border point lies within the ϵ -neighborhood of a core but has less than $MinPts$ neighbors in its own ϵ -neighborhood. Any point that is neither a core point nor a border point is classified as noise. The condition for clustering is that two points p and q are in same clusters if there is a chain of points p_1, \dots, p_n such that $p_1 = p, p_n = q$ and $p_{i+1} \in N_\epsilon(p_i)$ for $1 \leq i < n$. First we estimate the density of check-ins at different locations by applying KDE to the check-in data for creating density as in Figure 3.10.

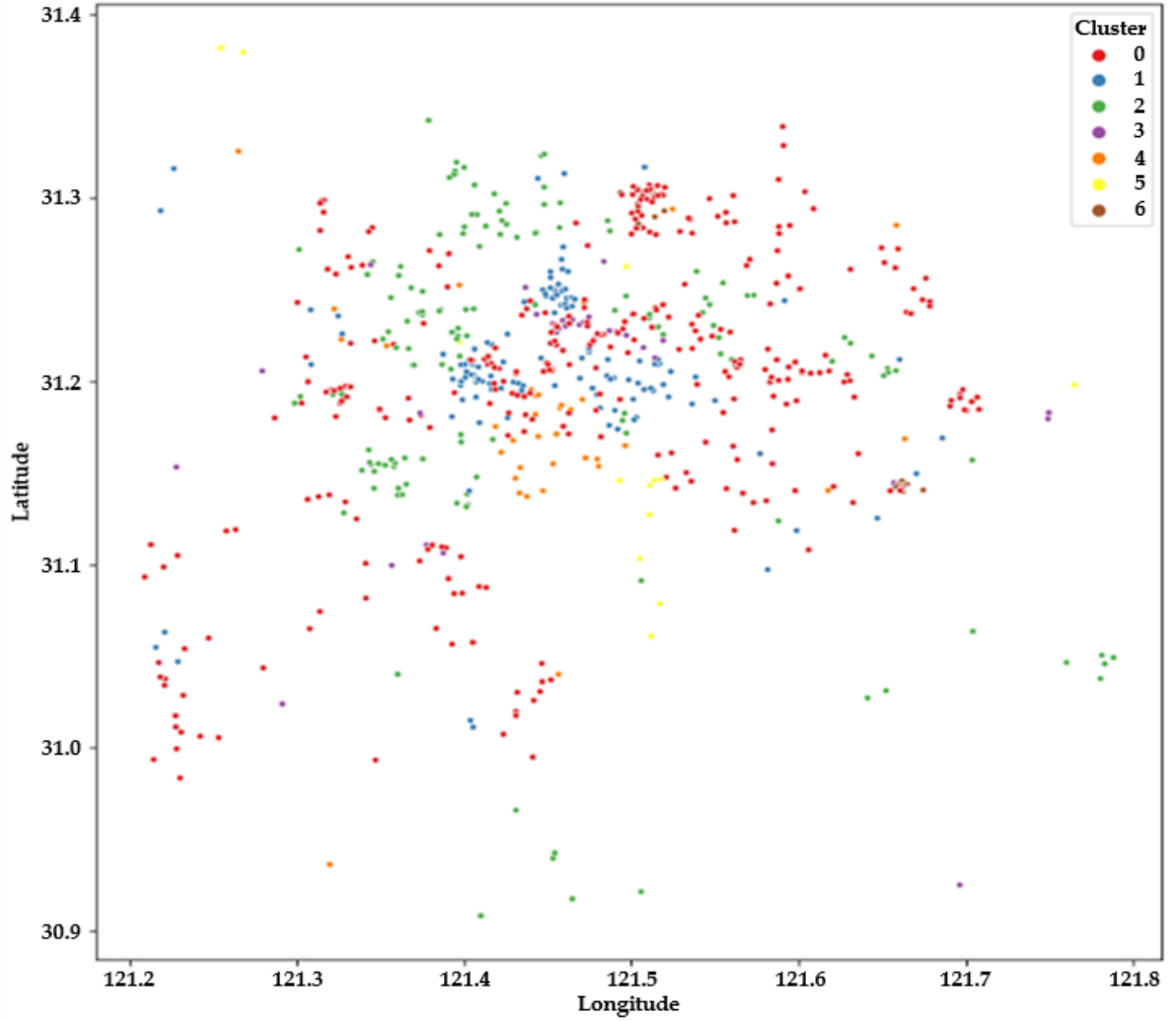


Figure 3.10: Venues Clusters

This map identifies the clusters or hot spots where check-ins are concentrated. We use KDE to identify areas where the density is above the threshold. These high-density areas will be the focus of the clustering algorithm. We apply clustering algorithm that uses DBSCAN to identify clusters and outliers within the areas. The algorithm will classify data points into clusters based on density as shown in Figure 3.10. These clusters represent areas with a consistently high density of check-ins, likely popular spots, or venues. The presented algorithm is capable of finding clusters of arbitrary shapes, unlike spherical assumption methods like k-means. It effectively separates noise points from clusters, identifying outliers and is primarily influenced by ϵ and $MinPts$, which can be tuned based on domain knowledge. Anomalies or outliers are locations that deviate from the general clustering pattern, representing unusual or rare check-in behaviors, and it will also identify outliers or anomalies as presented in Figure 3.11.

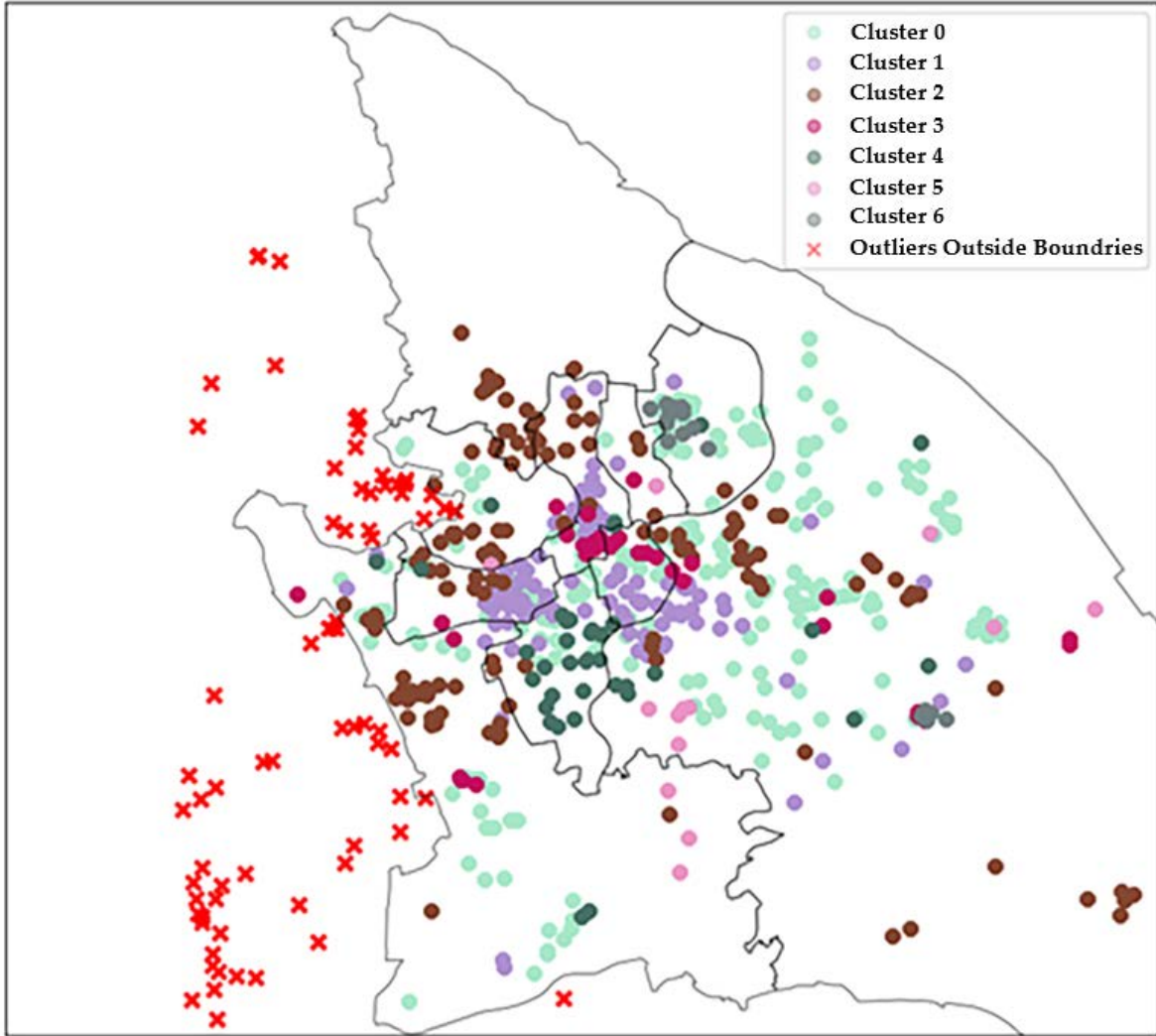


Figure 3.11: Clustering with Outlier Detection

Combining KDE with anomaly detection allows for a nuanced analysis of check-in data, identifying both the general trends (through KDE) and specific clusters and anomalies (through DBSCAN). This approach can help uncover areas of consistent popularity, as well as unusual or unexpected check-in behaviors, providing a richer understanding of the spatial patterns in your data. The parameters of both KDE and DBSCAN (like bandwidth for KDE and epsilon for DBSCAN) can further be fine-tuned to get the most meaningful and accurate results based on datasets and research objectives.

3.3.5 Density Estimation

The check-in patterns are further investigated by estimating the density of Check-ins all over Shanghai city. Therefore, we used KDE for finding the check-ins density using

ArcMap. The process of density estimation for the Weibo check-in data in Shanghai is given in Figure 3.12.

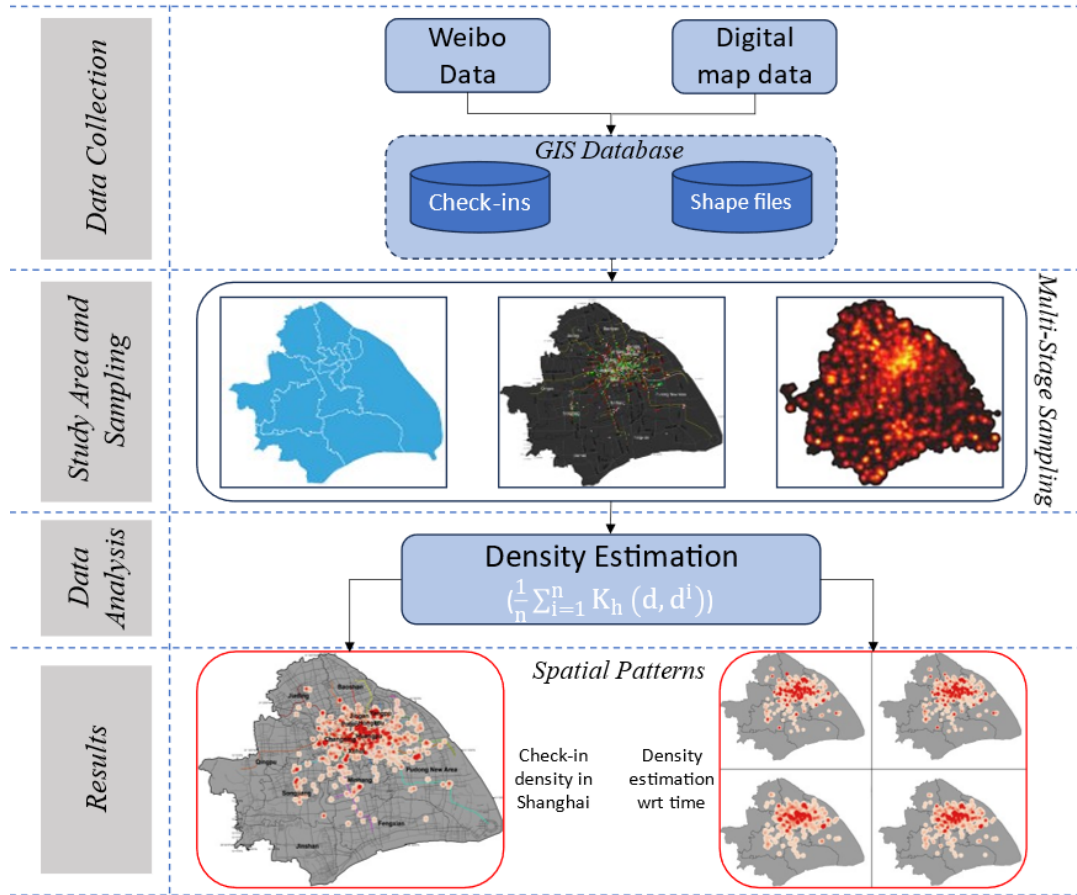


Figure 3.12: Density Estimation for Check-ins in Shanghai

The density estimation is calculated using KDE with the help of heatmaps to show different patterns of user activities and check-ins within Shanghai City. We calculated the density based on the check-ins by all users, providing us with more accurate results for further analysis, as presented in Figure 3.13. The following figure plots the density distribution of check-ins in various regions of Shanghai. Red represents the highest density and white represents the average, which eventually dissolves into the base color of the map according to the type of data. Check-ins that did not satisfy the minimum criteria in our dataset were not considered; thus, such data does not appear on the map.

This Figure 3.12 clearly indicates that check-ins in the city center are denser as compared to the regions away from the center of the city (as expected). The areas of Hongkou, Huangpu, and Jingan are the most dense areas as compared to the other districts. The spatial analysis was conducted by comparing weekly density for the first two weeks of April (having the maximum number of check-ins) with the last week of

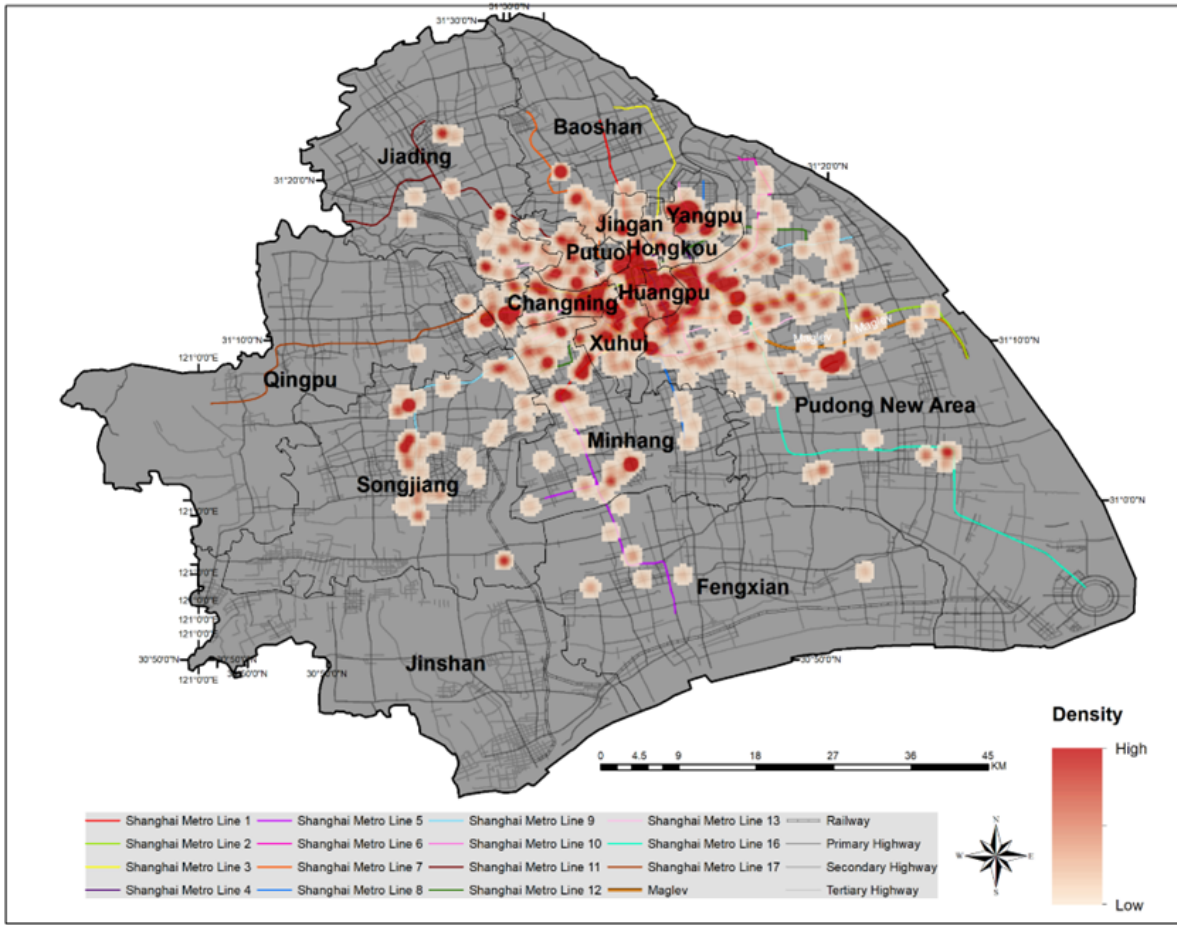


Figure 3.13: Density of Check-in Data

January and the first week of February (containing the minimum number of check-ins), as presented in Figure 3.14. The figure shows the density of check-ins in four weeks. Two of them (Figure 3.14a and Figure 3.14b) have the maximum number of check-ins (17,344 in the first week and 14,920 in the second week of April), and two (Figure 3.14c and Figure 3.14d) have the minimum number of check-ins (4699 in the last week of January and 5952 in the first week of February). It can be observed that, although the density varies in different areas all over the city, the downtown area remains the denser area even with a smaller number of user check-ins throughout the weeks of January and February; however, the overall check-in distribution covers a larger area during different periods of time.

It is important to consider that the downtown area is considered to be the commercial center of Shanghai; therefore, these areas have more facilities in almost every way, including transportation, food, shopping malls, government offices, and nightspots. However, as Shanghai is a considerably developed and modern city with lots of parks and

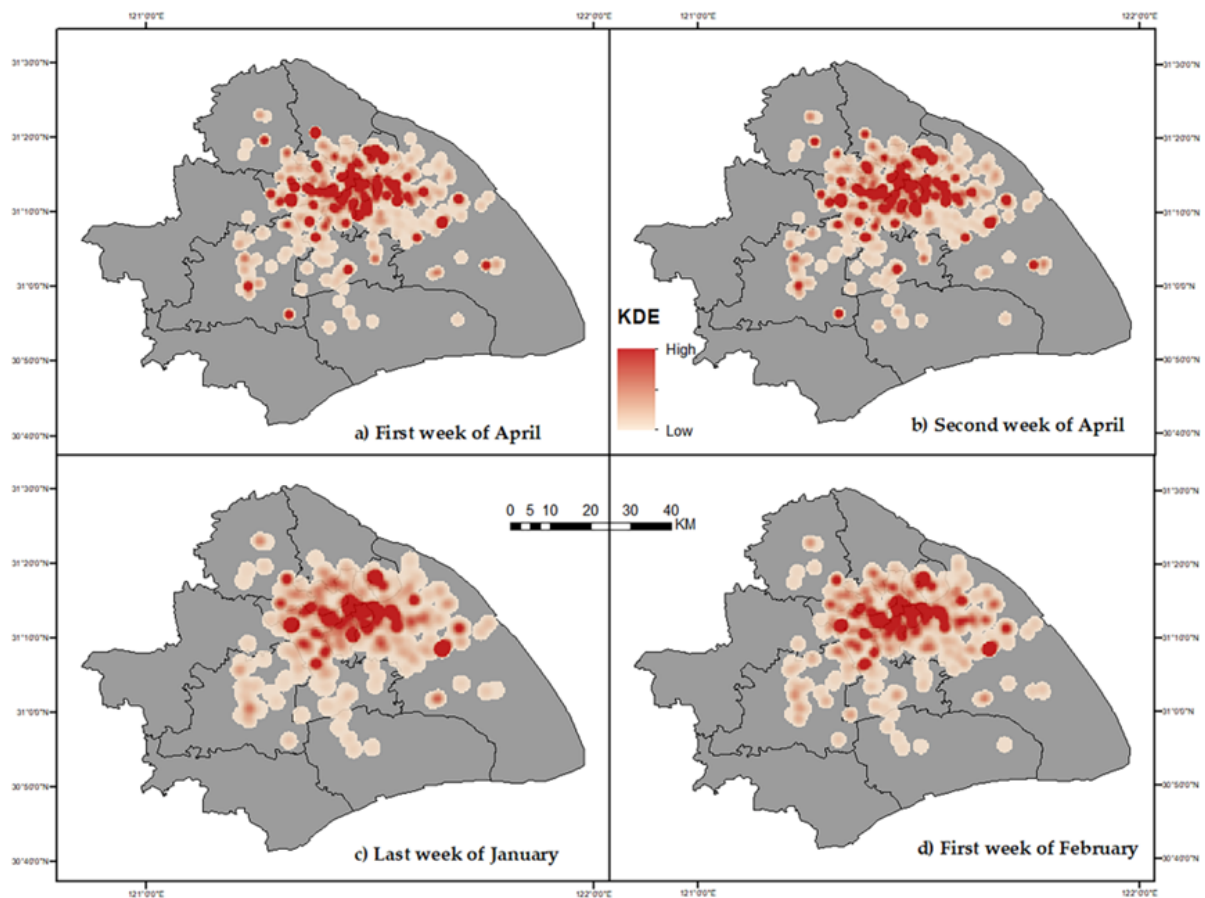


Figure 3.14: Weekly Density a) First Week of April, b) Second Week of April, c) Last Week of January, d) First Week of February

diverse Educational and Residential venues, the check-in clusters can be observed in different places throughout the city.

The analysis shows that data from Weibo are an efficient resource for analyzing the distribution of user activities and preferences in terms of spatio-temporal aspects. One of the benefits of using LBSN data for spatio-temporal analysis is that we can extract and visualize large-scale information for a megacity such as Shanghai in more detail. Some areas in downtown Shanghai are crowded, while other suburban areas have less visitors. This study intended to observe the behavioral traits of users by providing evidence that the dynamics of a megacity can be influenced by various facilities and the contribution of the nature of different venues.

We explored the spatio-temporal patterns in the check-ins to show the distribution of users in Shanghai. In this study, we performed an empirical analysis of check-ins using graphs, tables, and density maps based on LBSN data. The spatio-temporal patterns were studied from various perspectives, including hours, days, and venue categories.

From the chronological perspective, the results verified the frequency of check-ins rising from the middle of the day until late at night and the obvious increase in weekend activities as compared to weekdays. From the spatial point of view, the level of spatial intensity of users in the city center was higher in the downtown area, as this is the center of activity for most of the activities.

3.4 Summary

The study was carried out to look at four different aspects of analysis: a temporal analysis to reveal the patterns based on time, a model to show the significance of the data (MLR Model), a comparison of spatial analysis methods and spatial analysis using (KDE), by combining KDE with DBSCAN algorithm for anomaly detection, resulting in a clear observation of results through mapping. For the analysis presented in this chapter, we used check-in data from Weibo to analyze geo-spatial data to uncover various temporal and spatial patterns throughout the most famous places in Shanghai.

The findings demonstrated that LBSN data proves to be an efficient source for human behavior analysis. Specifically, people tend to use LBSNs more in the evening instead of the morning and workdays. We also observed that KDE is more accurate as compared to Point Density. Though many of the results are similar to what we expected, we obtained some interesting facts about the use of LBSNs. For example, the density extends to suburban areas more rapidly as compared to other areas within the city. The use of anomaly detection using DBSCAN with the KDE enables us to find the outlier within the data and also can be used to identify anomalies when applied to unseen data for highlighting the points that do not follow the observed common established patterns and not part of the existing clusters.

Data from LBSNs can play a strategic role in both the development and improvement of various aspects of mega cities' "smartness". The possibility to analyze the activities of urban agents has completely modified the representation of the relationship between activities and spaces. This can assist in urban planning by providing the tools to attain objectives of sustainability and make mega cities livable and more efficient. This study is further expanded in the next chapters.

INVESTIGATION OF THE EFFECT OF VENUE TYPES ON CITY DYNAMICS

The main purpose of research undertaken in this chapter is to study the classification of venues into various groups and the effect of these venue types on the density distribution of users and model check-in data from LBSN for the city of Shanghai, China by using combination of multiple temporal, spatial and visualization techniques after classifying users' check-ins into different venue categories. This chapter expands the research discussed in the previous chapter by exploring the relation between time, frequency, place, and category of check-in based on location characteristics and their contributions. The venue categorization is based on the nature of the physical locations within the city, and an effective model is presented for venue classification. The results of usage patterns from hours to days, venue categories and frequency distribution into these categories as well as the density of check-in and proximity between different venue types and contribution of each venue category in its diversity, are thoroughly demonstrated. Our findings uncover various benefits of studying the effects of different types of venues on the city in many aspects, which can be applied in studying individual venue types, predictions, recommendation systems and other interactive applications.

4.1 Introduction

The mining of LBSN data for useful patterns and insights has become an important and interesting topic among researchers in recent decades. Because of the huge number of applications based on LBSNs nowadays, an enormous amount of data is generated that is analyzed for extracting valuable information, particularly from a practical point-of-view, e.g., in areas including public transit flows, route planning, disaster management, etc. [136]. The online features of these applications encourage users to add and share their interests, activities, pictures, videos, etc., with their friends within the network, resulting in a massive amount of data that enables scholars to identify user activities and preferences more accurately through analysis. The online services provide and store user information along with their real-time locations, and the collected data is generally enriched with metadata, text, multimedia and geo-locations that can be applied to perform further research regarding several characteristics of human behavior.

Various studies have been conducted for analyzing and modeling human activities from geo-data. Most recent research to find the relationship and obtain patterns among users such as female and male, educated or less educated classes, age groups, etc., utilize data from internationally renowned LBSNs such as Facebook, Twitter, Foursquare, etc.[146]. Despite the exponential growth of Facebook, Twitter, etc. In countries around the world, most of these LBSNs are blocked or have limited use in China. Therefore, Chinese citizens tend to use national micro blogs, i.e., Weibo and hence check-in data from Weibo may be suitable for LBSNs data analysis here in China. These patterns include activity behavior, mobility, density, and also reproduce functional attributes in the city and between different cities. The word 'Check-in' represents a user who confirms her/her location using an LBSN application by performing an activity or sharing location with someone in a message. Weibo is famous not only among users but also in many scientists as they carry out numerous types of studies to extract valuable information from the geo-data provided by Weibo, e.g. some recent studies such as analysis of road accidents in Shanghai, analysis of the attraction features of the tourist hot spots through the data from Weibo, spatiotemporal analysis by gender for Beijing, etc. [147]. These studies are mostly based on check-in data analysis for particular users or specified application areas like tourism, road crashes, estimating urban boundaries, spring-festival rush, gender, etc. For gaining more useful insights, there is a need to relate these spatiotemporal patterns to the nature of venues from where the user check-in, and to the extent of our understanding, it has not been discussed previously by

CHAPTER 4. INVESTIGATION OF THE EFFECT OF VENUE TYPES ON CITY DYNAMICS

Table 4.1: Research Contributions

Attributes	Studies	Research Topic	Gaps	Our Contribution
Spatiotemporal Analysis	Veerandi et al., [137]	"Measuring Urban Deprivation from User Generated Content"	Daily or Weekly trends without finer-grained temporal analysis	Higher temporal resolution reveals nuanced behaviors, critical for urban dynamics and planning
	Rizwan et al., [138]	"Visualization, Spatiotemporal Patterns, and Directional Analysis of Urban Activities Using Geolocation Data Extracted from LBSN"		
	Li et al., [139]	"Construction and Adaptability Analysis of User's Preference Model Based on Check-in Data in LBSN"		
	Muhammad et al., [140]	"Spatiotemporal Analysis to Observe Gender Based Check-in Behavior by Using Social Media Big Data: A Case Study of Guangzhou, China"		
Venue Classification	Gao et al., [141]	"Extracting Urban Functional Regions from Points of Interest and Human Activities on LBSNs"	Single Type Venue / Activities not classified	Venue categorization allows for a finer-grained analysis of social dynamics and venue popularity
	Li et al., [66]	"Venue Classification and Exploring Venue Popularity in Foursquare"		
	Feng et al., [65]	"The Geographies of Expatriates' Cultural Venues in Globalizing Shanghai: A Geo-Information Approach"		
Nearest Neighbor Distance	Senefonte et al., [142]	"Regional Influences on Tourists Mobility through the Lens of Social Sensing"	User Preferences / Density Estimation	Enhanced spatial analysis capability, crucial for identifying clusters and understanding social gathering patterns
	Feng et al., [143]	"Deepmove: Predicting Human Mobility with Attentional Recurrent Networks"		
Density Estimation	Zhang et al., [144]	"Revealing Spatial Preferences Embedded in Online Activities: A Case Study of Chengdu, China"	KDE used without Anomaly Detection	Combining DBSCAN with KDE for spatial analysis and outlier detection offers a novel approach, enriching LBSN data analysis
	Rizwan et al., [120]	"Big Data Analysis to Observe Check-in Behavior Using Location-Based Social Media Data. Information"		
	Xia et al., [145]	"Spatiotemporal Residual Networks for Citywide Crowd Flows Prediction"		

other researchers. Therefore, we focused on three different aspects of the analysis of Weibo's check-in data for the period of one year, i.e., July 01, 2016 to June 30, 2017 from Shanghai City, for uncovering spatiotemporal patterns with venue classification, and density estimation. So, the current study presents two key aspects of the analysis as our contribution to the existing knowledge in this area of research. Which is the Classification of the dataset and study of 10 venue categories and applying the spatiotemporal analysis for modeling and density estimation of each category for a better understanding of typical check-in concentrations based on each venue category, demonstrating the role of venues in the diversity of users in Shanghai and presenting a model for future studies in these different venue types.

The studies presented in the previous research including these in Table 4.1 the integration of diverse LBSN data into the user activity studies framework provides unique perspectives into human mobility and social behavior., thus opening new ways of understanding user activities and preferences. Our study based on Weibo check-ins gives, therefore, an enriched view of urban life in Shanghai based on Weibo data with respect to the commonly used Foursquare or Twitter datasets, especially in China. As compared to previous studies, which are generally quite coarse-grained in temporal analysis, usually general daily or bias weekly timing, in the current chapter, we present a more comprehensive investigation with statistical analysis with hourly, daily, weekly, and long-term patterns to provide a more fine-grained description of city life preference. An improved time resolution of the data could shed light on even more detailed trends. The venue classification is improved with the proposed significant classes, which comprehend social dynamics well and help in estimating venue-related activities in a comprehensive manner. The spatial analysis is extended significantly due to the implementation of nearest neighbor distance to understand spatial clustering and our capability to recognize patterns in human activities and interaction more accurately. Moreover, the proposed method of high-resolution KDE-based density maps present a detailed explanation of the hot spots of spatial activities, which provided useful information for expanding studies about the effect on city dynamics and LBSN based urban studies. Such integration with KDE method for density estimation and anomaly detection was a useful kind of combination providing spatial analysis with the ability to examine more patterns and outliers at a higher level of detail within density estimation.

4.2 Analysis Framework

The overall analysis and contribution presented in this chapter are given as a framework for research, as shown in Figure 4.1. The methods used in this analysis are presented here starting with the method for venue classification, going through Spatial Clustering for showing the diversity of check-in as well as the density estimation based on the defined venue classes.

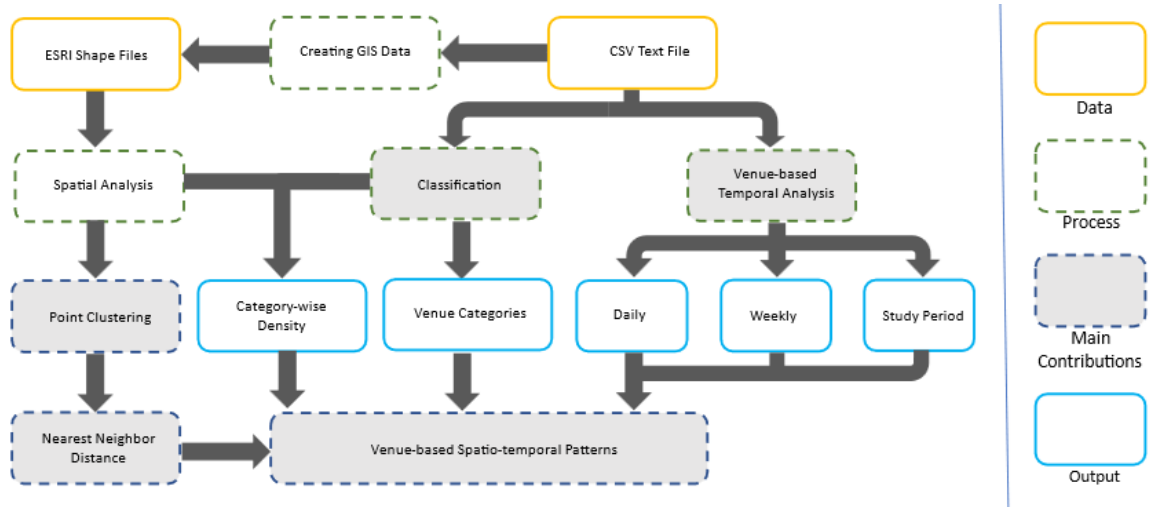


Figure 4.1: Analysis Framework

The accessibility and proximity of different services and facilities, from residential to entertainment venues, affects not only the quality of life but also the social and economic aspects of cities. Studying these spatial relations is critical for urban planning, development, and community services. This research extends our existing research on the effect of venue types by studying spatial correlations to investigate the relationship between various types of venues. The point pattern analysis is carried out by the proposed method that used point clustering with Nearest Neighborhood Distance for finding the spatial clusters. We start with a random unvisited point in the dataset, which is followed by neighborhood query, stating that for the current point, retrieve its ϵ -neighborhood and identify the points within this radius. We then move to the Cluster Formation in which there are two iterative steps. Firstly, if the ϵ -neighborhood of the current point contains at least $MinPts$, create a new cluster, and add all reachable points to this cluster. Secondly, if the ϵ -neighborhood contains fewer than $MinPts$, label the point as border points. These steps are repeated for the whole dataset, creating the desired point clusters. The next step in the point pattern analysis method is finding the nearest neighbor distance in which the distance between points from different categories is

calculated to analyze the spatial correlation between different venue types. For this purpose, the k-dimensional tree is utilized for effective nearest neighbor searching.

In spatial analysis with venue Types, we applied the KDE for spatial analysis with the help of point density estimation. The point is to calculate the density of the defined venue classes including ‘Educational’, ‘Entertainment’, ‘Food’, ‘General Location’, ‘Hotel’, ‘Professional’, ‘Residential’, ‘Shopping & Services’, ‘Sports’, and ‘Travel’ and compare them in order to find the effect of these venue types on the density and diversity of data points showing the preferences and trends in activities of users within the city based on the type of venues.

4.2.1 Venue Classification

There has been rapid development in mobile technology, wireless communications, online and location-based services in the past few decades. Therefore, services based on these elements i.e., LBSN such as Twitter, Weibo and Facebook, etc., are drawing more and more scholars to analyze the massive data collected by these services. The analysis proved to be very helpful to extract useful patterns about crucial jobs like crises and disaster management, urban planning, development of smart cities and other fields involving big data. Along with the previously discussed spatiotemporal analysis, the effect of venue classes or types are presented via statistical results and density estimation.

We filtered the dataset acquired from Weibo to find the most suitable variables for the current research. The filtered dataset used in this study includes 441,471 check-ins by 144,582 users from 20,171 venues. We initially classified the data manually based on their names and the nature of activities performed at each venue in order to analyze the effects of the venue types and the importance of the classification.

Our effort is to show the importance of venue classification. Although these venue categories were assigned by filtering through the data with the help of key words such as “School”, “Bank”, “Airport” etc., using MS Excel, an example of the step-by-step process of venue filtering according to the classes is given in the Figure 4.2.

Although the model performs well, it still needs enhancements to be able to implement. As we can see that the model successfully classifies the data into 10 groups; however, this is a work in progress, and as a collaborative research student at the University of Technology Sydney, we will try to explore more methods and evolution methods in future work to show its significance. Although the model works, it still needs more work to be implemented in real-time. Nevertheless, this proposed model and the following analysis open many doors for researchers to do more efficient research in this

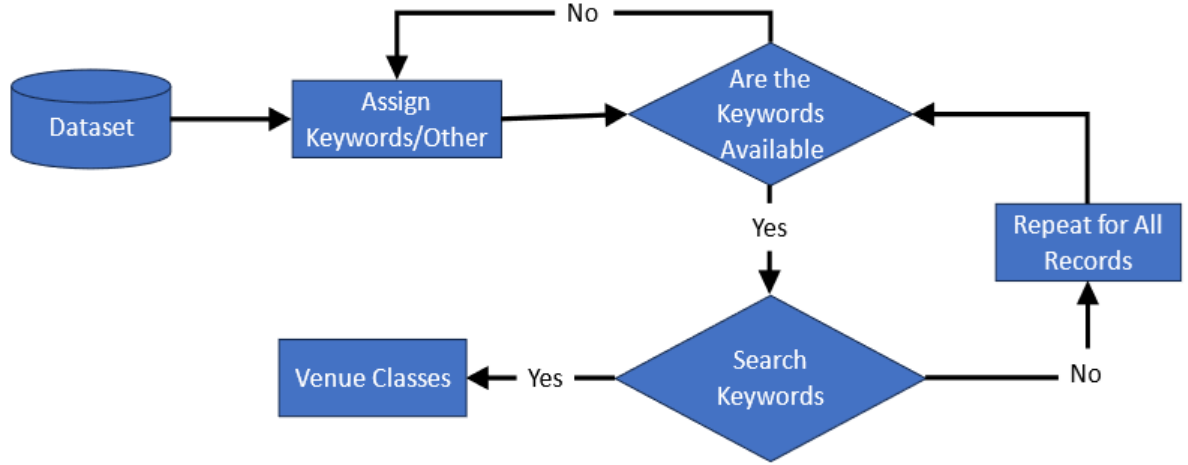


Figure 4.2: Venue Filtering Process

domain using ML instead of manual classification. After the classification, the detailed characteristics of the dataset are given in Table 4.2.

Table 4.2: Attributes of Categories

Category	Number of Locations	Percentage	Total Check-ins	Average	Gender		Total Number of Users
					Female	Male	
Educational	2535	12%	52645	20.76726	11948	6955	18902
Entertainment	2547	19%	79471	31.20181	16750	7964	24714
Food	2675	4%	22747	8.503551	3235	1746	4981
General_Location	1119	8%	23033	20.58356	5280	3084	8364
Hotel	1101	4%	16568	15.04814	3189	2523	5712
Professional	3030	7%	40527	13.37525	7760	5264	13024
Residential	2945	24%	98221	33.35178	19491	11919	31410
Shopping&Services	2263	12%	62155	27.4675	16494	7310	23804
Sports	594	4%	19590	32.9798	3309	3165	5474
Travel	1362	7%	26514	19.46696	4728	3469	8197

A major advantage of using LBSN data is the ability to find the location of the check-in activity, along with its purpose. Each check-in provides the latitude and longitude of the actual venue by the LBSN (e.g., Weibo). When searched for in the LBSN data, the latitude and longitude give a specific location on a geo-referenced map. This data about the location can be utilized to get other information about the visited venue. We use only the most general types of the hierarchy, containing 10 different venue types: ‘Educational’, ‘Entertainment’, ‘Food’, ‘General Location’, ‘Hotel’, ‘Professional’, ‘Residential’, ‘Shopping & Services’, ‘Sports’, and ‘Travel’. The categories and examples of the check-in locations are given in Table 4.3.

This presented venue filtering algorithm defines keyword lists for each of the 10 venue categories and uses a function `categorize_venue()` to assign each venue to the appropriate category based on whether its name contains any of the keywords in the

Table 4.3: Check-in Venue Categories

Category	Check-in Venue Example
Entertainment	Concert_Hall, Daning_Theatre Cinema_Shanghai_Paragon_Studios
Educational	Tongji_University_South_Campus Cao_Yang_Second_Middle_School
Food	Tang_Lian_Hot_Spring_Restaurant Starbucks_(Gubei_Store)
General Location	Wanda_Square Changle_Road
Hotel	Three_Star_Yunfeng_Hotel Five_Star_Shanghai_Fujian_Hotel
Professional	HSBC_Court_Buildings Ping_An_Bank_Headquarters
Residential	Residential_Area_Yonghe_Sancun Xinli_Greenland_Apartments
Shopping	Shopping_Mall_Baili_Life_Plaza Mall_Australia_Square
Sports	Playground_Happy_Valley Lushan_Golf_Club
Travel	Hongqiao_Airport Line_2_Xujing_East_Bus_Station

corresponding keyword list. The `apply()` function is then used to apply the categorization function to each venue name in the data. The distribution of check-in points in the categories is given in Figure 4.3.

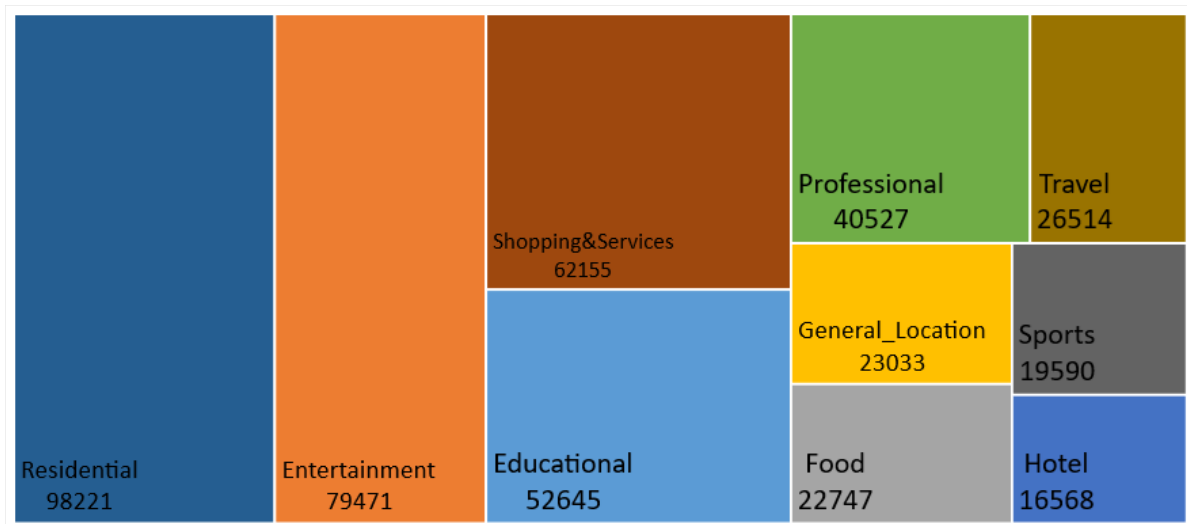


Figure 4.3: Check-in Venue Distribution

The total check-ins in the ‘Residential’ category are greater than all other categories, while ‘Hotel’, ‘Food’ and ‘Sports’ have minimum check-ins. It is interesting to observe

that although the number of locations in “Professional” is more than in every other category, check-ins and users in ‘Residential’ is more than the double of check-ins in the ‘Professional’ category. Similar patterns can be observed in the ‘Educational’, ‘Entertainment’, ‘Shopping&Services’ and ‘Food’ categories, like, the number of venues is almost the same, but the check-ins and users are significantly different in numbers. This reveals an interesting point, i.e., in the second case, it is an obvious fact that people tend to use LBSNs more frequently while having a good time at ‘Entertainment’ and ‘Shopping&Services’ venues, e.g., Concerts, Parks, Shopping Malls, etc. Similarly, they use LBSNs more at their homes as compared to being at work or professional places, for instance, hospitals, courts etc.

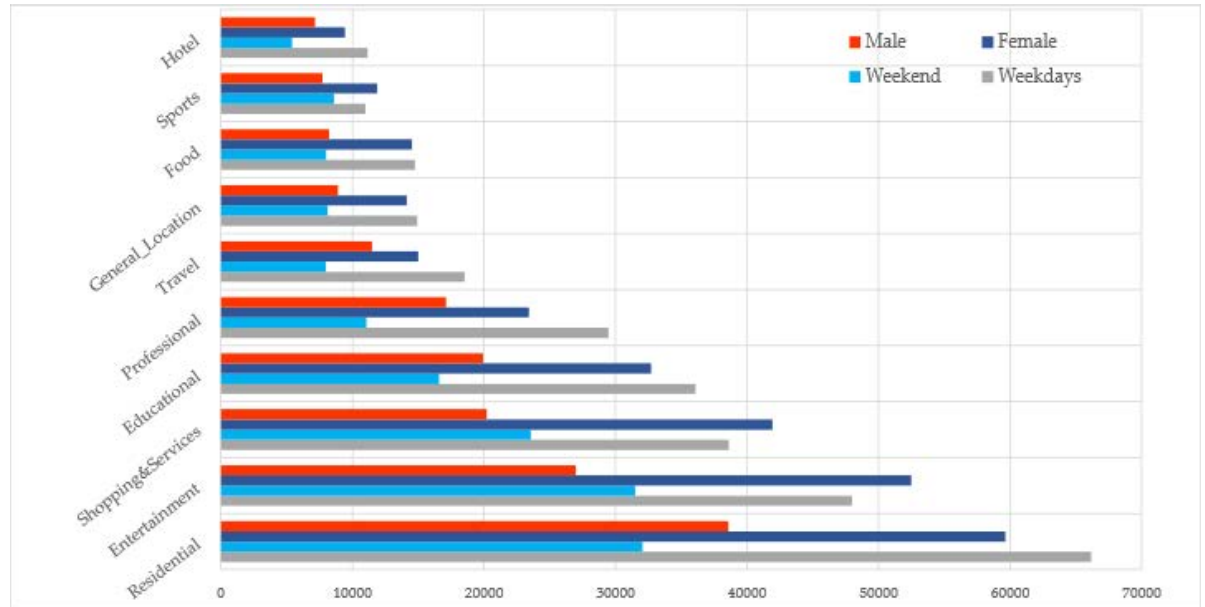


Figure 4.4: Venue Categorization Statistics

We provide temporal analysis to show the effects of different types of venues on the users’ behavior. The overall statistics of the venue categories are provided in Figure 4.4. To gain more insights into the contribution of these check-in venue categories, we expand our study of categorization with respect to time, week and date as shown in Figure 4.5. The category-wise temporal analysis shows some common behaviors like the decline of check-ins frequency in vocations and spikes in almost all categories after vocations in April, along with uncovering that winter vocations (Chinese New Year and Spring Festival) do not have that much effect on ‘Entertainment’ and ‘Shopping&Services’ categories as compared to winter vocations as in Figure 4.5.a. The Figure 4.5 .b indicates that the check-ins in the ‘Residential’ category start rising earlier in the morning and start declining later at night as compared to other categories (as expected). Additionally, other

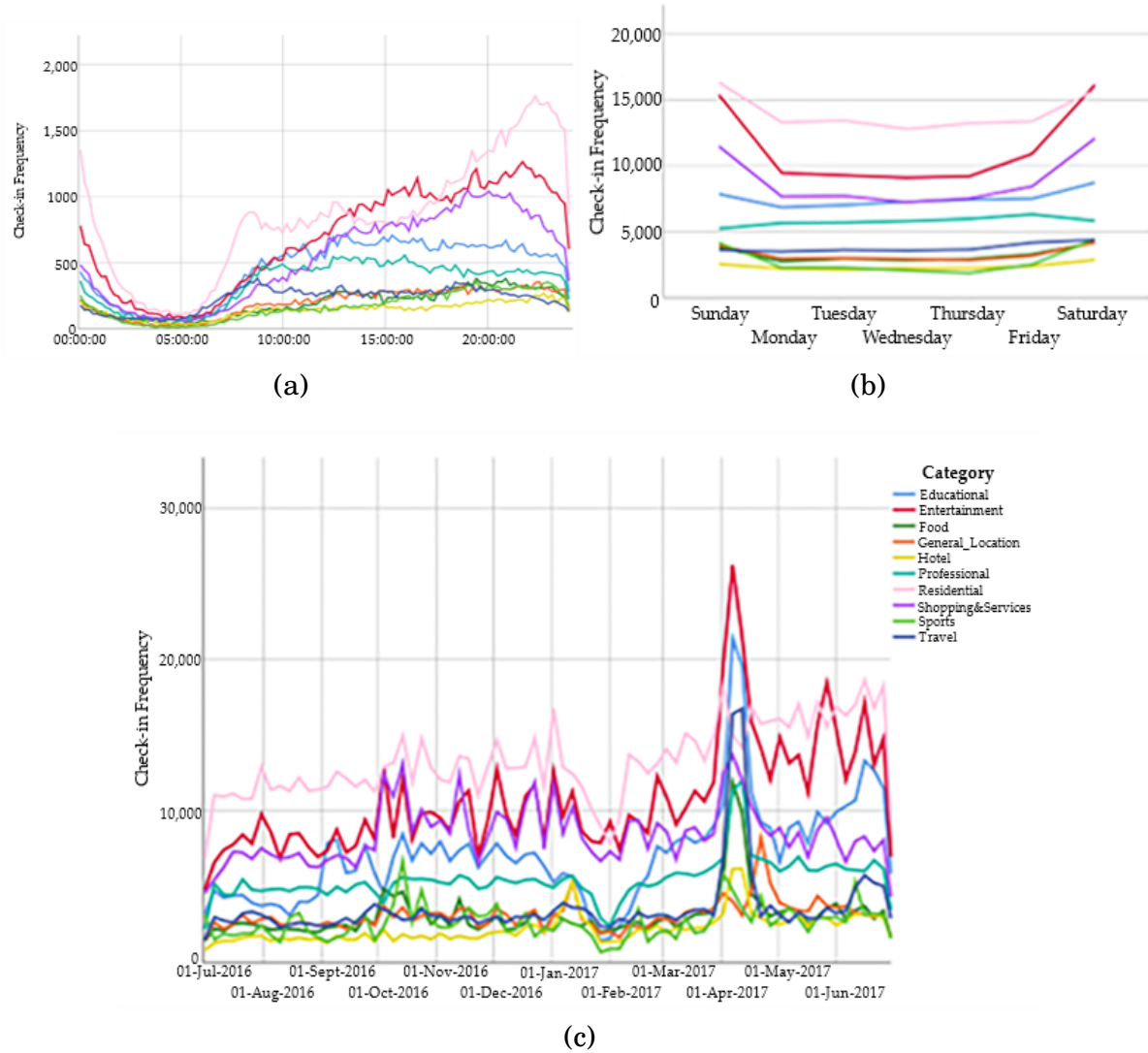


Figure 4.5: Category-wise Temporal Analysis, a). Time by Category, b). Weekday by Category, c). Date by Category,

categories start declining after working hours, but 'Entertainment', 'Shopping&Services' and 'Residential' check-ins are at the peak after working hours until midnight. The weekly analysis in Figure 4.5.c represents that weekend have a significant effect on the 'Professional' category as the check-ins are minimum on weekends and on 'Entertainment' and 'Shopping&Services' as they are at peak on weekends.

4.2.2 Clustering-based Point Pattern Analysis

In the evolving landscape of urban dynamics, the relationship between various types of venues plays an important role in modeling the experiences of users. The significance

of this study can be demonstrated as, we used the ‘residential’ category having the most number so check-in along with ‘entertainment’ (as a part of our tourism study). The accessibility and proximity of different services and facilities, from residential to entertainment venues, affects not only the quality of life but also the social and economic aspects of cities. Studying these spatial relations is critical for urban planning, development, and community services. This research extends our existing research on the effect of venue types by studying spatial correlations to investigate the relationship between various types of venues. The point pattern analysis is carried out by the proposed method that used point clustering with Nearest Neighborhood Distance for finding the spatial clusters. The details of the point clustering method are provided in this section, which is based on the following steps.

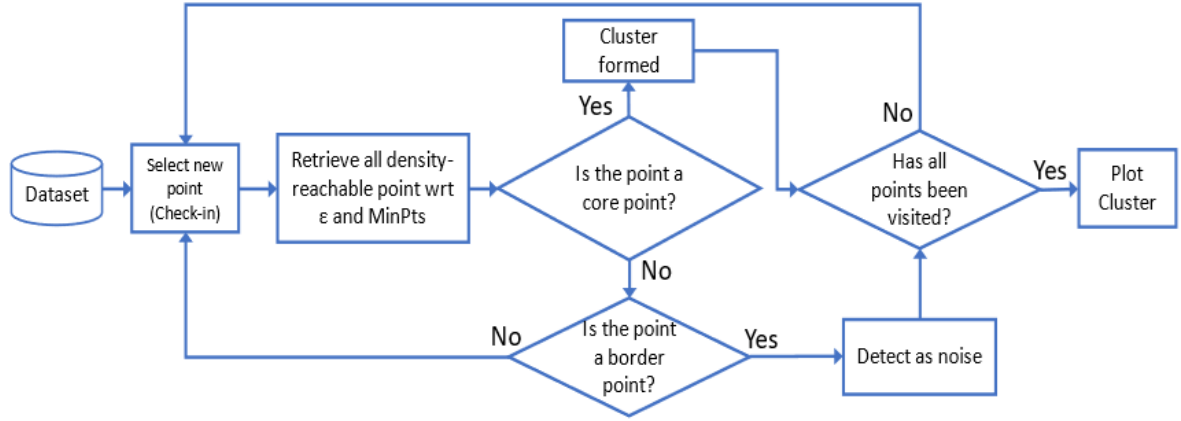


Figure 4.6: Clustering Method

We start with a random unvisited point in the dataset, which is followed by neighborhood query, stating that for the current point, retrieve its ϵ -neighborhood and identify the points within this radius:

$$N_{\epsilon}(p) = \{q \in D \mid \text{dist}(p, q) \leq \epsilon\} \quad (4.1)$$

where D is the dataset and $\text{dist}(p, q)$ is a distance metric. We then move to the Cluster Formation in which there are two iterative steps. Firstly, if the ϵ -neighborhood of the current point contains at least MinPts , create a new cluster, and add all reachable points to this cluster. Secondly, if the ϵ -neighborhood contains fewer than MinPts , label the point as border points. These steps are repeated for the whole dataset, creating the desired point clusters.

The next step in pattern analysis is finding the spatial clusters. The aim is to identify check-in clusters within each venue category based on geographic proximity. The

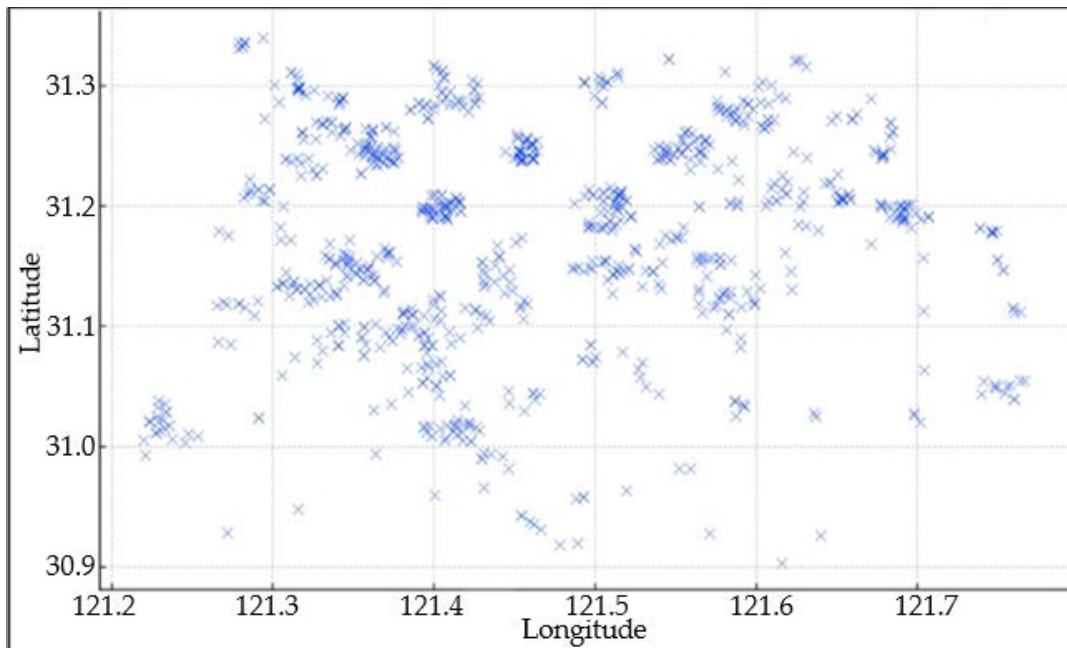


Figure 4.7: Spatial Distribution of Residential Category by Clusters

clustering algorithm is used for its capability to retrieve arbitrarily shaped clusters and manage outliers effectively.

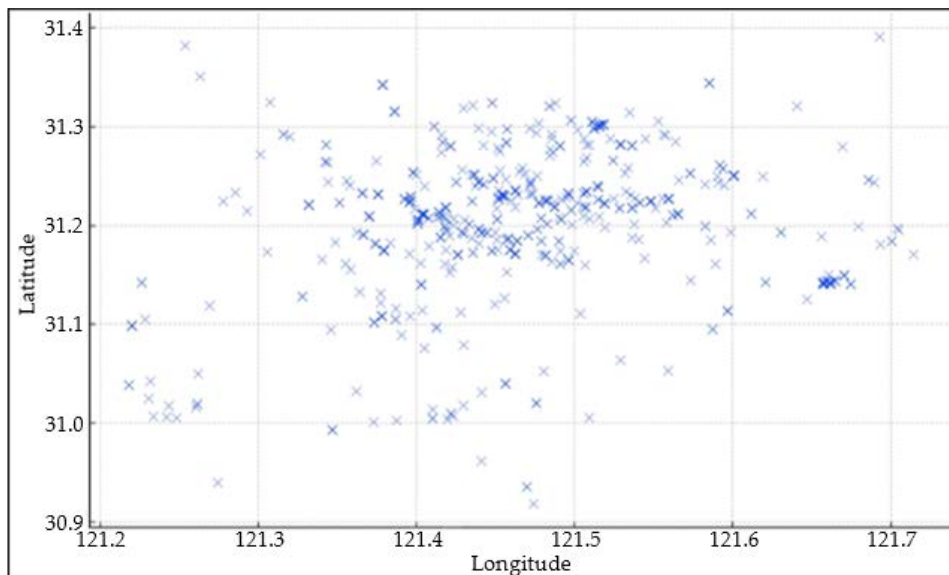


Figure 4.8: Spatial Distribution of Entertainment Category by Clusters

The above scatter plots show the spatial distribution of the 'Residential' in Figure 4.7 and 'Entertainment' category check-ins in Figure 4.8, reflecting the grouping of residential and entertainment-related check-ins in different areas of the city. These clusters can reveal areas with a higher level of residential check-ins. If these groupings

are close to each other, it suggests a higher level of spatial correlation among the residential or entertainment areas, respectively. These figures also reveal areas with a high concentration of entertainment venues, which can be compared to the residential clusters for analysis of spatial relationships between where people live and where they seek entertainment.

4.2.3 Nearest Neighbor Distance

The nearest neighbor distance between points from different categories is calculated to analyze the spatial correlation between different venue types. The distribution of nearest neighbor distances was visualized using a histogram to provide insights into the spatial correlation. Clustering of venues was inferred from a higher frequency of shorter distances, suggesting proximity between entertainment and residential venues.

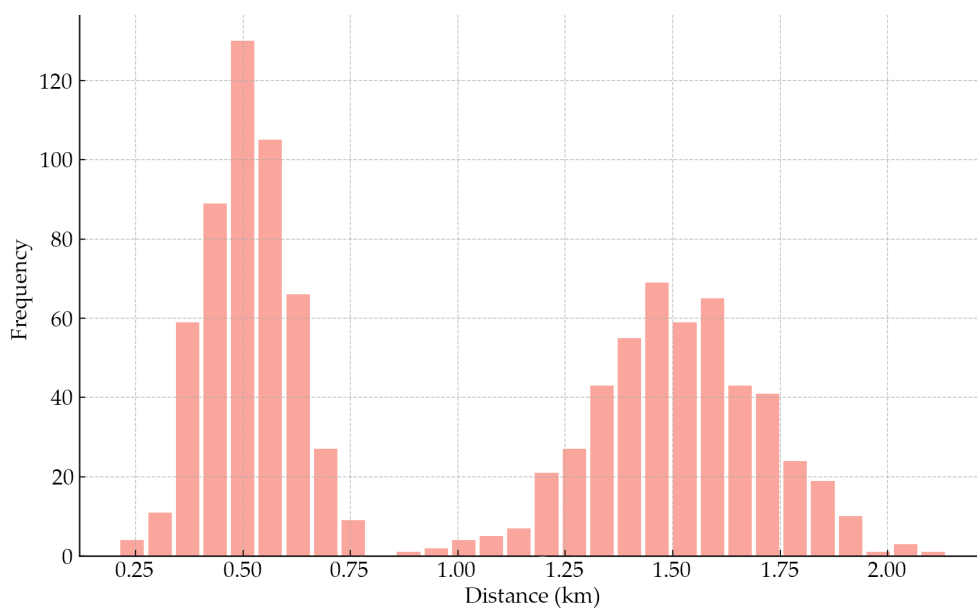


Figure 4.9: Bimodal distribution of Distance from Residential to Entertainment Venues

The following figures show a bimodal distribution of distances from entertainment venues to the nearest residential venues. Figure 4.9 clearly points out the bimodal distribution of the distances, which suggests that there are two common distances at which most entertainment venues are usually located relative to residential venues. That may mean that there were even two different urban settings or planning paradigms within the city: those of the city center, where often, due to higher density, short distances are being used; and those of suburban locations, where venues might be more diverse. Figure 4.10 shows a cumulative distribution of distances from entertainment venues to

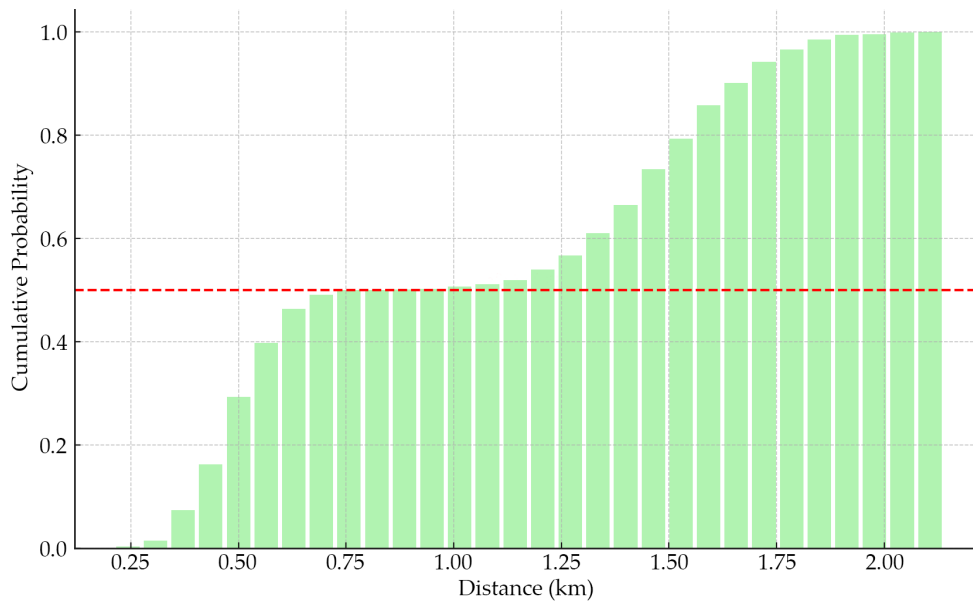


Figure 4.10: Cumulative Distribution of Distance from Residential to Entertainment Venues

the nearest residential venues. The bimodal distribution usually refers to the tendency of data distribution to assume two different distinct peaks; most entertainment venues are usually found at two different distances relative to residences. Such peaks would suggest two groups of venues for the provision of entertainment: one cluster close to residential areas, and the second one a bit far from the residential parts. This might bring some idea about the type of entertainment venues. For example, the smaller local ones, e.g., cafes or community centers, should be located nearer to the residencies, while the big ones, e.g., concert halls or sports areas.

The y-axis represents the frequency of a venue at a different distance. The higher the bar, the greater is the number of occurrences in which a venue is at that distance. The highest showing that the mode distance between the home venue and the nearest venue for entertainment is about half a kilometer. This would be indicative of some urban design where the districts have some of the entertainment areas meant for residential use, close and intimate to form an entertainment option and within, in some cases making it so that some would have larger districts for entertainment possibly located away by requiring transportation. This insight might be accounted for by urban planning policies or land use, or possibly by natural geographic features. It also clearly indicates on the graph the frequency extending towards the right, fewer entertainment venues found at further located places from resident places.

The next step in the point pattern analysis method is finding the nearest neighbor

distance in which the distance between points from different categories is calculated to analyze the spatial correlation between different venue types. For this purpose, the k-dimensional tree (k-d tree) is utilized for effective nearest neighbor searching which is defined as follows:

Given two sets of points S_r and S_e representing 'Residential' and 'Entertainment' venue locations, respectively, the nearest neighbor distance from an entertainment venue $e \in S_e$ to the closest residential venue is given by Equation 4.2.

$$d_{\min}(e, S_r) = \min_{r \in S_r} d(e, r) \quad (4.2)$$

where $d(e, r)$ is the Euclidean distance between points e and r , which can be computed on a plane as Equation 4.3.

$$d(e, r) = \sqrt{(x_e - x_r)^2 + (y_e - y_r)^2} \quad (4.3)$$

We used the point pattern analysis by using the 'Residential' and 'Entertainment' categories from our study to compute the nearest neighbor distances between these types of venues.

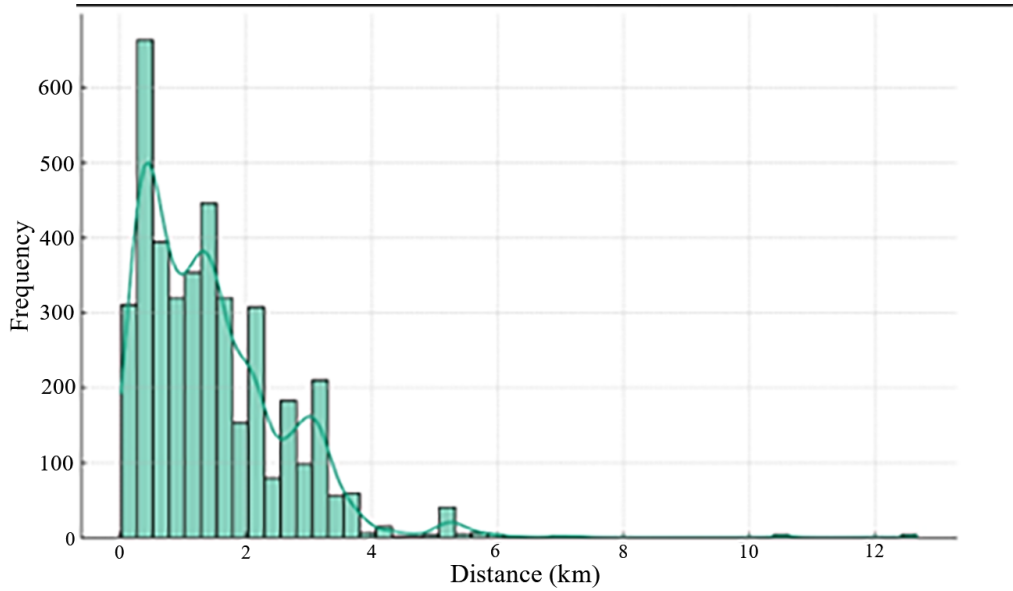


Figure 4.11: Distribution of Distances from Entertainment Venues to Nearest Residential Venues

The Figure 4.11 exhibits the distribution of distances from residential venues to the nearest entertainment venue, measured in kilometers. The peak at the lower end of the

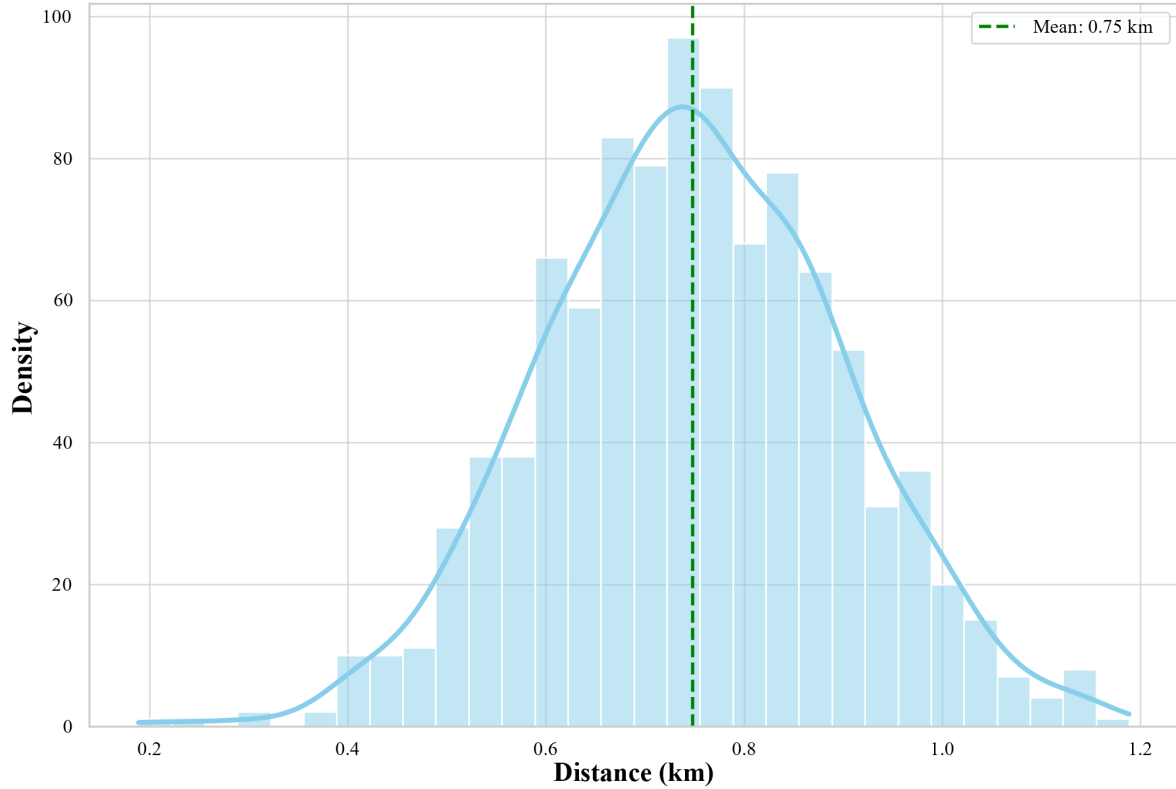


Figure 4.12: Nearest Neighbor Distance Distribution

distance range suggests that many residential venues are located quite close to entertainment areas, indicating a potential clustering effect which supports the notion that accessibility is a key factor in the placement of these venues in an urban environment. Figure 4.12 demonstrates the nearest-neighbor distances on an expanded scale between the facilities of entertainment and residential venues. The distance distribution shown in the figure represents the frequency with which the venues fall into a specified range of distances. The center of the distribution shows that most of the entertainment areas are around this distance from the residential venues. A smooth line clearly indicates a continuous probability density curve of the distances and enables one to identify the distribution shape as well as its central tendencies. The center shows the mean (average) nearest neighbor distance according to the check-ins in within the Weibo dataset, telling us the entertainment distance of venues to the closest residential venue. The venues of entertainment are located mostly nearer to the residential areas in distance and at a denser concentration around the mean distance, as represented in the above figure. The distribution seems to be slightly positively skewed; showing, most venues are clustered around lower distances, but some are also located around higher distances. Which can be very helpful such as the proximity pattern can advise companies about optimal locations

for their services, the city planners, local government, and community service providers can use this kind of insight for optimizing the allocation of land resources, confirming that entertainment facilities are easily accessible to residential areas, thereby enhancing the quality of urban life.

4.2.4 Spatial Analysis with Venue Types

In this section, we investigate spatial analysis by visualizing the locations of venue categories and the density of total check-ins by using the geo-location data from Weibo on a map of Shanghai. For this purpose, we used a map including features from Open-StreetMap because it contains the most recent updates of the map features. We can observe features such as city boundaries, districts and district boundaries, Shanghai Metro lines, and the road structure.

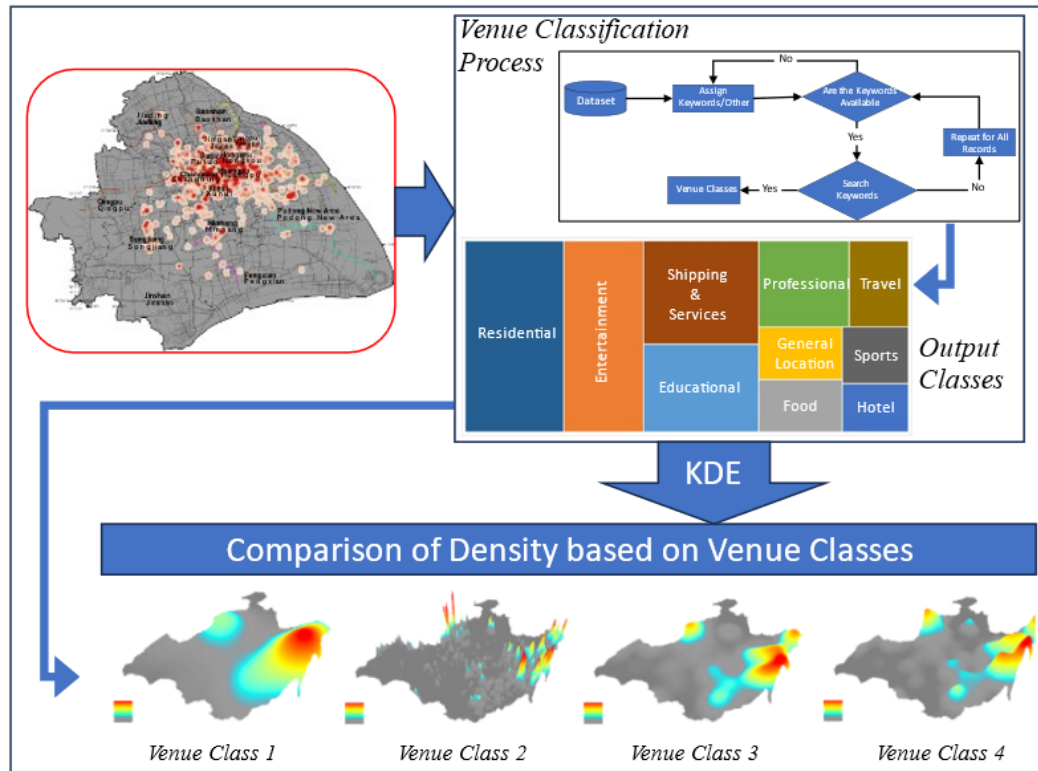


Figure 4.13: Class-Based Density Estimation

With the help of these features, it is easy to evaluate and recognize the different locations on the map. For spatial analysis, we first plotted the locations of all the famous venues in Shanghai, as shown in Figure 4.14.

It can be observed from the above figure that, as per the planning of one of the major

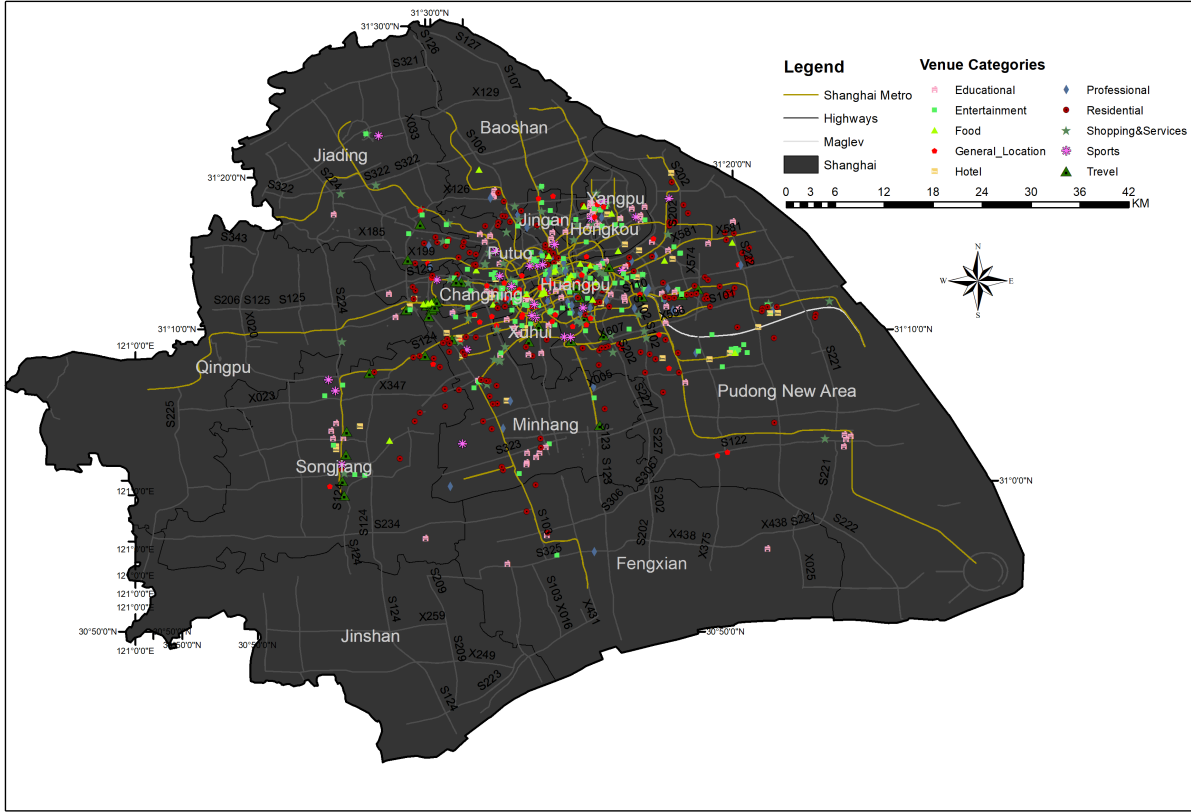


Figure 4.14: Location of Venues in Different Categories in Shanghai

cities of China and ease of access, most of these locations are situated either in the city center or near Shanghai Metro. The seven districts, namely Changning, Huangpu, Putuo, Hongkou, Xuhui, Jintan, and Yangpu, situated in Puxi (Huangpu West), are collectively called the downtown area or the city center of Shanghai [63]. The downtown has a higher concentration of famous places, as would be expected in any major city; however, the Educational and Residential venues are relatively dispersed within the city.

In order to show useful insights about the effect of activities in different venue categories on the diversity of check-in locations, we plot the density of each venue category in Figure 4.14. It can be seen in the following Figure 4.14.a,b,c&d that with approximately the same number of locations in ‘Educational’, ‘Entertainment’ and ‘Food’ categories, check-ins in ‘Educational’ and ‘Food’ are more concentrated as compared to ‘Entertainment’, showing that users visited variety of ‘Entertainment’ venues while they preferred specific ‘Food’ venues, and ‘Educational’ intuitions obviously have check-ins from specific locations. ‘General_Location’ are more diverse as compared to ‘Hotel’ because they are situated at specific places in the city.

As seen in below Figure 4.16.a, it is a common behavior that people barely use

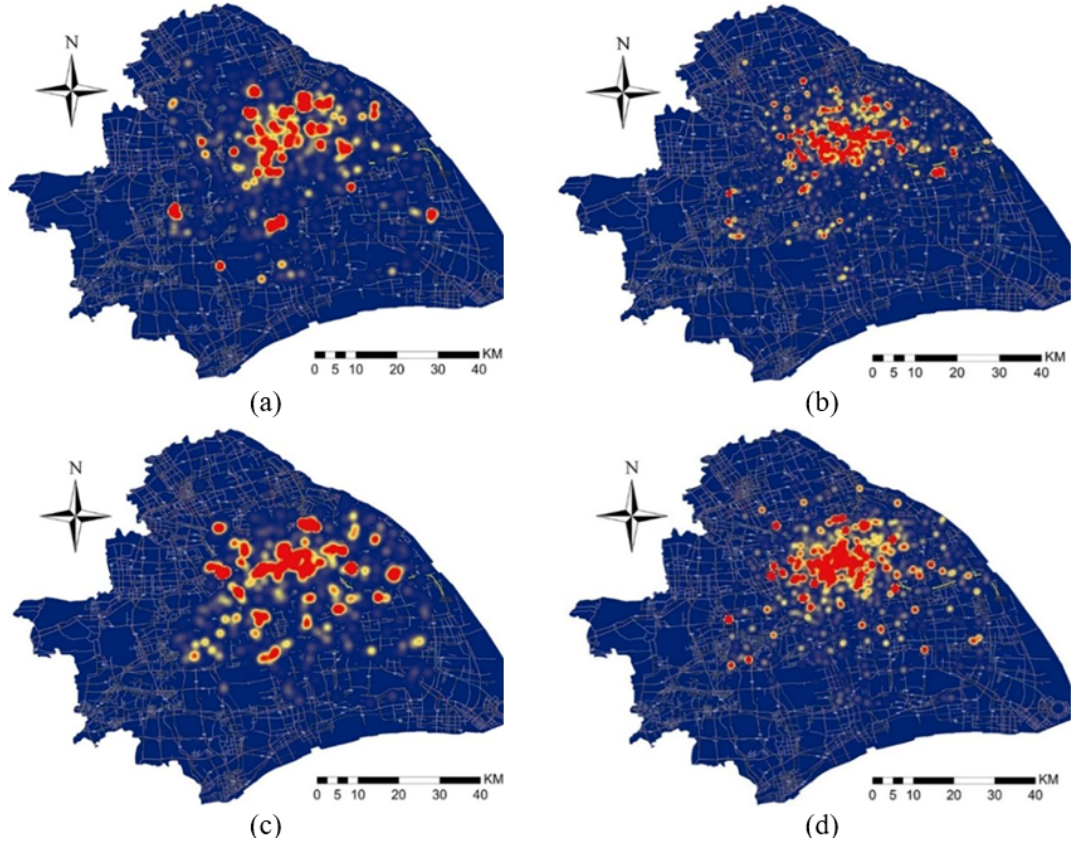


Figure 4.15: Category-wise Density a) Educational, b) Entertainment, c) Food, and d) General Location

LBSNs from ‘Professional’ locations, similar patterns can be observed in Figure 4.16.b, having the most average density unlike to all other categories. On the other hand, because of the huge residential apartments in the mega city of Shanghai, Figure 4.16.c demonstrates highly concentrated density of check-ins in the ‘Residential’ category. The ‘Shopping&Services’ in Figure 4.16.d (substantial check-ins) and ‘Sports’ in Figure 4.16.e (less check-ins) show diversity in the intensity of check-ins representing the interest of users to explore different shopping sites and various types of ‘Sports’ venues. The ‘Travel’ category in Figure 4.16.f, however, displays the density of check-ins alongside metro lines and metro/bus stops.

The contribution of each check-in venue category to the overall density of Weibo data in Shanghai can be elaborated as the locations in ‘Educational’, ‘Food’, ‘Hotel’, ‘Professional’ and ‘Travel’ categories mainly account for the concentration of density to these specific places while locations in ‘Entertainment’, ‘General_Location’, ‘Shopping&Services’ and ‘Sports’ adds diversity of check-ins in low density areas. However,

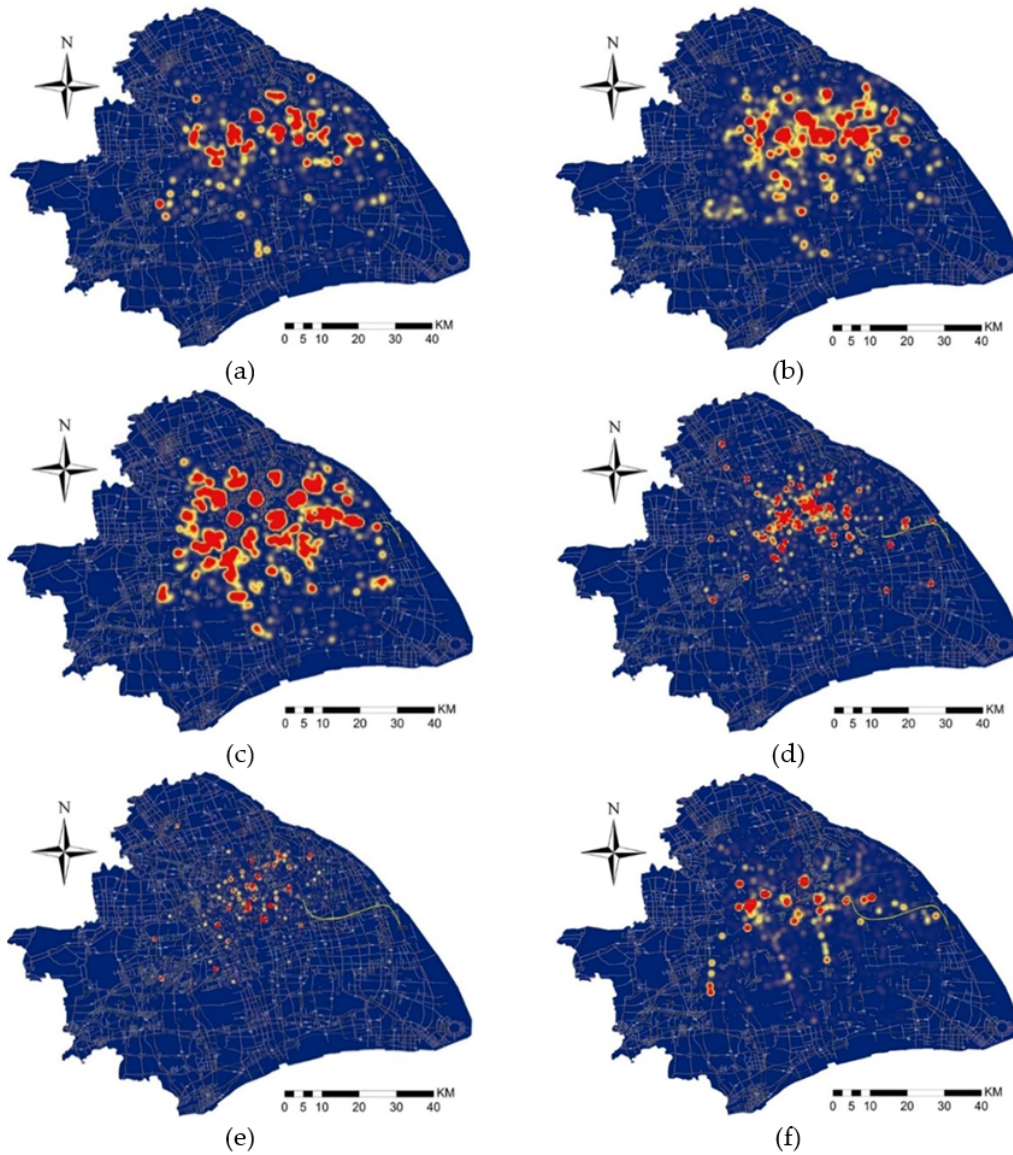


Figure 4.16: Category-wise Density, a) Hotel, b) Professional, c) Residential, d) Shopping&Services, e) Sport, f). Travel

the fact remains the same that city-center has denser density as compared to suburban areas.

4.3 Summary

In this chapter, we introduced the venue classification method for detailed analysis of spatiotemporal patterns, effect of venues and provided the opportunities for researchers to conduct various kinds of comparative studies between different venue classes or individual class. The data from Weibo for one year (July 2016 to June 2017) was used for spatiotemporal analysis to explore patterns in different activity categories in Shanghai. The research included analysis of check-in behavior based on time (daily, weekly, annual), check-in in different venue categories, and the location of various categories along with the contribution of each check-in venue category in the density. Temporal analysis revealed key behavior that was the significant effect of vocations and mass migration on the check-in frequency resulting in minimum number of check-ins during that period. The venue classification provided insights like maximum number of check-ins from residential, entertainment and shopping areas as compared to check-ins from others, and an additional fact that shopping and entertainment are not affected that much by summer vocations, Spring Festival or Chinese New Year. The point pattern analysis revealed the importance of conducting nearest neighbor distance to find the spatial proximity and accessibility between venue types as evident from the exiting close proximity between the residential and entertainment venues. The results of density estimation demonstrated by modeling density of venue categories like those containing professional and residential venues contribute to a more concentrated while the categories like shopping and entertainment extend the density to suburban areas.

VENUE CLASSIFICATION AND PREDICTION

The current research has aimed to investigate and develop ML approaches to be applied to classify LBSN data and predict user activities based on the nature of various locations (such as entertainment) and their preferences. The analysis of user activities and behavior from LBSN data is often based on venue types, which require the input of data into various categories. This has previously been done through a tedious and time-consuming manual method. Therefore, we proposed an efficient approach of using ML models to extract these venue categories. In this study, we used a Weibo dataset as the main source of research and analyzed ML methods for more efficient implementation. We first introduced four models based on well-known ML techniques, including the GLM, LRM, GBT and Deep-Loc model. We designed, tested, and evaluated these models. We then used assessment metrics, such as the ROC or AUC, Accuracy, Recall, Precision, F-score, Sensitivity and various evaluation matrices, to show how well these methods performed. We discovered that the proposed ML models are capable of accurately classifying the data, with our Deep-Loc model outperforming the other models, followed by GBT, GLM, and LRM, for multiclass distributions and single class predictions. We classified the data using our ML models into the 10 classes we used in our study and predicted tourist destinations among the data to demonstrate the effectiveness of using ML for LBSN data analysis, which is vital for the development of smart city environments in the current technological era. The comparison of our proposed model shows better performance as compared to recent state-of-the-art methods.

5.1 Introduction

The research on LBSN data has gained huge attention from scholars with the rapid growth of mobile technologies. The LBSN data has been used for analysis in variety of specialized fields such as study of people's behavior in festivals, shopping malls, food venues, tourism and many more [139]. This kind of data contains heterogeneous attributes about users, and researchers need to filter out the data relevant to specific venues to conduct more specialized studies. The dataset often includes thousands or sometimes, millions of records before the data filtering which is done manually which is a time consuming and troublesome issue for such kind of research. Therefore, some ML methodologies are required to classify the data based on some specific characteristics so that the multi-venue data can be classified without the need of manual work. With the interactive web-based interface of modern LBSNs, researchers have more opportunities to utilize the data about the majority of the population for various kinds of analysis. This data provides a sample of various aspects of human behavior and traits while interacting with the LBSN during variety of activities through check-ins from different venues. The study of these behaviors provides valuable insights about the general trends within the population for planning and development of events, festivals, parks, shopping malls, restaurants and, ultimately, development of a smart city. The LBSN data has also been used in more specialized studies like finding the popularity factors of restaurants, role of parks, tourism behavior and many more, which are proved to be tremendously valuable in these fields. However, for these specialized studies, it is crucial to consider the data relevant only to the venues of interest within the huge number of records and manually classify the specific data for each individual research. As one of the strengths of using LBSN data for human behavior is the availability of huge amount of data, it is often difficult and more time consuming to classify the data for finding suitable records [13].

The primary objective of this research is to find and develop ML methods that can be used to classify the LBSN data used in our previous research and predict the tourism venues for analyzing the activities of tourists and residents in Shanghai, while showing the efficiency of the proposed models for LBSN studies. This research question was formulated after finding the research gap from our rigorous literature review suggesting that many studies are conducted in the field of LBSN analysis with manual classification of data. The use of ML provides a more efficient way to conduct these studies while keeping the integrity and validity of the research intact so that the researchers and developers can focus on more beneficial analysis without worrying about going through

each record among the piles of ‘Big Data’ manually [15]. In the same context Wang et al.[46] pointed out the detriments of manual classification while discussing the imperfection, unreliability of classifications that are generated with the human eye. Therefore, the computerized, digitized, ML based classification of our data is proposed. to show the feasibility of the dataset used in this study, we initially applied statistical analysis using IBM SPSS 25 [148] followed by the proposed ML through Rapid Miner [149]. After consideration of the research gap, we addressed the following research question in this study.

- How can we use and develop ML models to categorize LBSN data into specialized fields? And which ML model best fits the LBSN data to predict a specific class of venues (tourism) for study of a particular research domain.
- Proposed Deep-Loc model specifically designed for efficient and effective venue classification instead of the traditionally manual categorization of LBSN data.

In the current research, we analyzed various machine-learning methodologies and proposed an efficient approach to the venue classification problem by using machine learning with the help of four models that show promising performance in the classification of data into multiple classes and predicting the designated class of users based on the information about activities performed at different venues from their check-in records. Once the models are trained and implemented, they can remove the overhead of manual classification in the field of venue based LBSN analysis.

5.2 Analysis Methods

In this section, we present and explain the overall methodology of our research and the steps involved during this study. Following Figure 5.1 illustrates the workflow of our classification methodology. The pictorial representation of the experimental setup for the location category-based activity predictions and classification. The first step is to collect the data and prepare it for analysis. This involves cleaning the data, removing any duplicates, and encoding categorical variables. The data used in this research can be acquired using Weibo API, as illustrated in Figure 2.6.

The data attributes used in this research are selected after verifying the variable significance using linear regression. The data is then divided into training and testing sets with K -fold cross validation which can be defined as: Let’s consider the dataset of N samples, denoted by $X = (x^1, x^2, \dots, x^N)$. The dataset is first split into k folds of

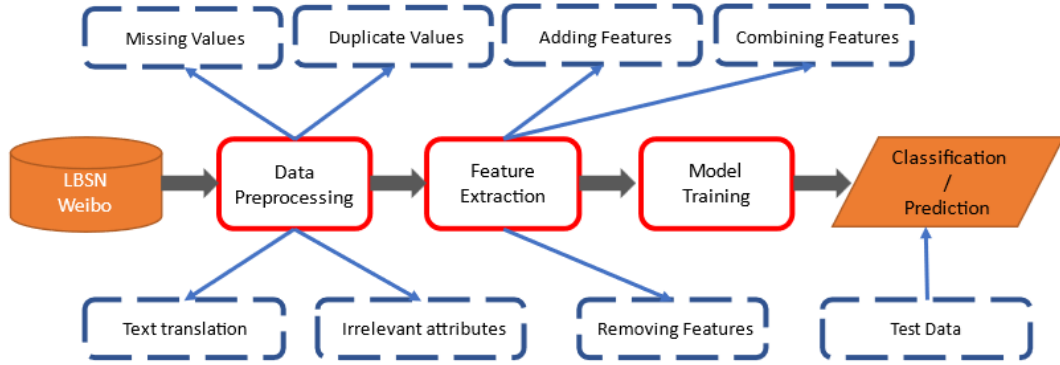


Figure 5.1: Pictorial Representation of Classification

equal size, denoted by X^1, X^2, \dots, X^k . For each fold $i = 1, 2, \dots, k$, train the ML models on the training data $X - X^i$ and evaluate their performance on the test data X^i . The training data $X - X^i$ consists of all the samples in X except those in X^i , and the test set X^i consists of the samples in X^i . The performance of the model on each test set X^i is measured using a performance metric, including accuracy, recall, precision, or F-score. The final performance metric for the model is obtained by averaging the performance metrics of each iteration, i.e., by averaging the performance metrics on each test set X^i . This provides an estimate of the model's performance on an independent dataset. The final performance metric gives us a robust estimate of their performance on independent data, as it is based on an average over iterations [150].

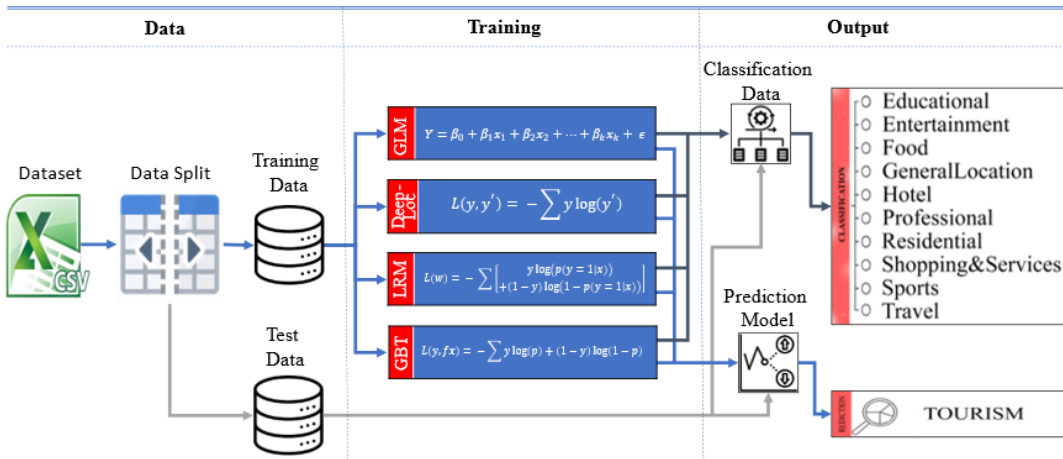


Figure 5.2: Machine Learning Framework

The model training phase consists of training several ML models are used to classify the location categories, including the GLM, GBT, LRM, and Deep-Loc model. Each model is trained on the training and evaluated on testing sets. The performance of each model is evaluated using metrics such as accuracy, recall, precision, F1-score, Receiver Operating

Characteristic (ROC) curve, Area Under the Curve (AUC), confusion matrices, and lift chart. And finally, the best-performing model can be highlighted based on its overall performance and ability to accurately classify and predict the location categories.

The detailed implementation of each model is provided in the next sections. The experiments were performed on a computer with an Intel Core i7 processor, 16GB of RAM, and NVIDIA GeForce GTX 1060 graphics card. The Deep-Loc model is structured with several hyper-parameters that play an important role in its training and performance. The training process is set to iterate over the dataset for 10 epochs, ensuring that the models have ample opportunities of learning intricate patterns from the data and the number of batches equals 7 with size of 1992. The model employs a learning rate of 0.005, a parameter that dictates the step size during weight updates in the optimization process. To prevent oscillations and ensure smooth convergence, a momentum of 0.9 is incorporated, with the added advantage of the Nesterov acceleration for faster convergence. The decay parameter is set at epsilon $1.0E-8$ and rho 0.99, gradually reducing the learning rate over time to refine the model's weight adjustments as it approaches the optimal solution.

The model's optimization strategy is governed by the Stochastic Gradient Descent (SGD) algorithm, a widely used method known for its efficiency and robustness. To quantify the difference between each predicted probability and the actual class label, the categorical cross entropy loss function is utilized, which is particularly suitable for multiple class classification tasks. For model evaluation, accuracy serves as the primary metric, providing a clear measure of the ratio of correctly predicted suggestions made by the model. To prevent over-fitting and ensure the models' generalization to unseen data, 10% of the training data is reserved for validation purposes. The software used included Rapid Miner 9.7, Python 3.8, and IBM SPSS. Further details of the dataset and methodology are presented in the following subsections.

5.2.1 Data Source

The data source used in this study is acquired from Weibo which was also presented in our previous research for venue classification. The dataset includes the following feature as extracted during the data acquisition and pre-processing:

- User ID (Unique for every user however available multiple times with subsequent check-ins).
- Gender of the user.

- Check-in Day (Day of the week including weekends/weekdays).
- Check-in Time.
- Check-in location name (like Shanghai University, Lingnan Park, etc.).
- Check-in Category (used for training during supervised learning).

The dataset used in this study is from Weibo to find the most suitable variables for the current research. The dataset used in this study includes 441,471 check-ins by 144,582 users from 20,171 venues. We initially classified the data manually based on their names and nature of activities performed at each venue in our previous study and used the same dataset for the supervised learning of our proposed classification and prediction models. While for the evaluation of these models, we utilized 10-fold cross validation method for dividing the dataset into the standard split of 80/20 for training and respective testing with stratification.

5.2.2 Feature Selection

The LBSN datasets include a number of features that have been used for research in a variety of domains. Although these features possess value in one way or another. However, it is imperative to choose the best possible features for research in any individual domain. There are numerous methods for the feature selection for example, forward elimination, backward elimination Bayesian, Akaike Information Criterion and so on. We used the famous Adjusted R^2 with backward elimination method to include p -values with a threshold of 0.05 [135]. Which can be defined as shown in Equation 5.1.

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \cdot \frac{(n - 1)}{(n - p - 1)} \quad (5.1)$$

where R^2 is the coefficient of determination, n is the sample size, and p is the number of predictors. The backward elimination method is a stepwise regression method that starts with all predictors and removes the variable with the largest p -value one by one until all predictors have a p -value less than a specified threshold (usually 0.05). The Adjusted R^2 is then re-calculated after each variable is removed to assess the impact on the goodness of fit.

5.2.3 Machine Learning Models

This analysis is based on four ML models that are provided below.

- *Generalized Linear Model (GLM)*: The GLM is a flexible and frequently used statistical model that can be implemented for both binary and multiple class classification problems. GLM is a generalization of the linear regression model that allows for non-linear relationship between the predictor and the response variables. For binary classification,

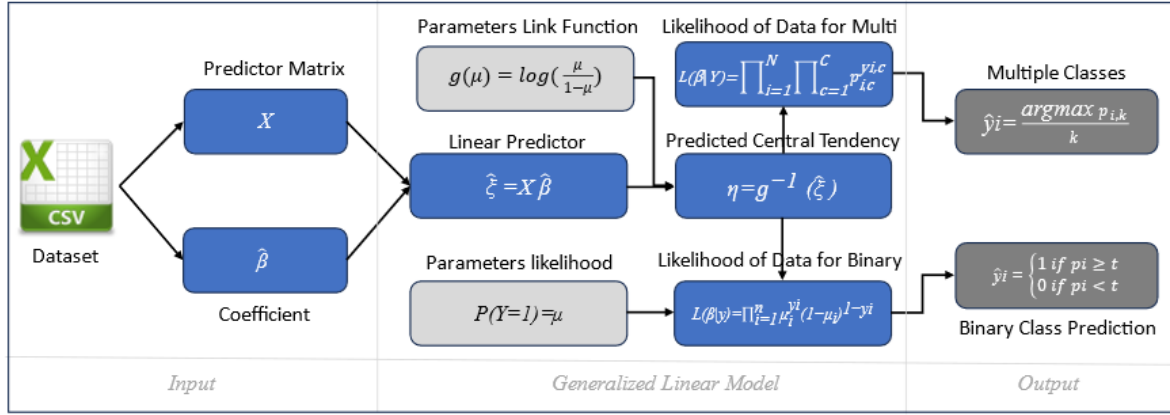


Figure 5.3: Generalized Linear Model for Venue Classification

GLM can be used to model the log odds of the binary outcome as a linear combination of the predictor variables, and then transform the log odds to probabilities using the logistic function. For multi-class classification, GLM can be extended to use a multinomial or ordinal response distribution, depending on the nature of the outcome variables. The mathematical form of GLM can be written as follows:

$$\eta = X\beta \quad (5.2)$$

where η is the linear predictor, X is the matrix of predictor variable, and β is the vector of regression coefficients.

$$g(\mu) = \eta \quad (5.3)$$

where g is the link function that relates the expected value of the response variable (μ) to the linear predictor.

$$f(y; \mu) = c(y) \exp((y\eta - b(\eta))/a) \quad (5.4)$$

where f is the probability density function of the response distribution, y is the outcome variable, μ is the expected value of the outcome variable, a , b , and c are constants that depend on the choice of the response distribution, and $g\eta$ is the linear predictor.

- **Logistic Regression Model (LRM):** LRM is simple and widely used statistical model that can be implemented for binary classification problems. LRM models the probabilities of the binary outcome as a function of the predictor variable, using a logistic function that maps the linear predictor to a probability value between 0 and 1. The mathematical

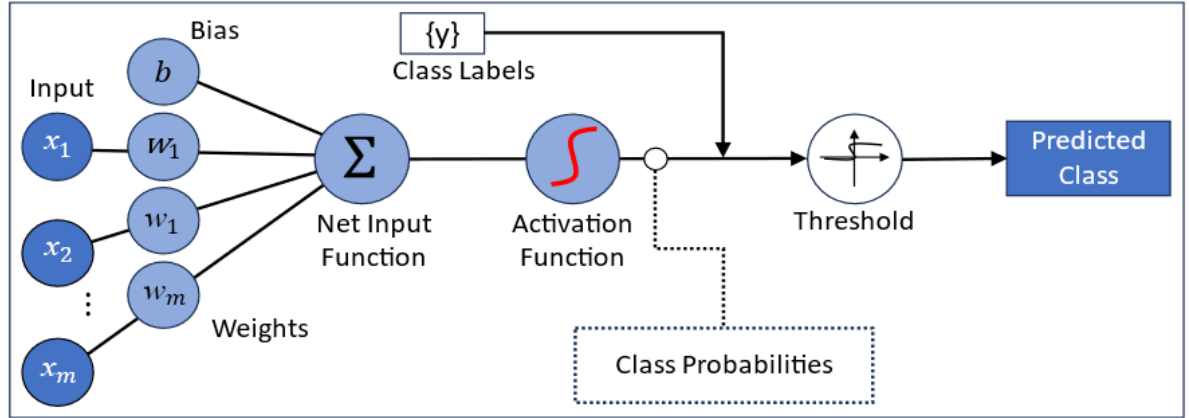


Figure 5.4: Logistic Regression for Classification

form of LRM can be written as follows:

$$p(y = 1|x) = \frac{1}{1 + \exp(-w^T x)} \quad (5.5)$$

where $p(y = 1|x)$ represents the probability of the binary outcomes ($y = 1$) given the predictor variables (x), w is the vector of regression coefficients, and T is the transpose operator.

$$p(y = 0|x) = 1 - p(y = 1|x) \quad (5.6)$$

where $p(y = 0|x)$ holds the probability of the negative outcomes ($y = 0$) given the predictor variables (x).

$$L(w) = - \sum [y \log(p(y = 1|x)) + (1 - y) \log(1 - p(y = 1|x))] \quad (5.7)$$

where L is the log-likelihood function used to measure the goodness of fit for the model.

- **Gradient Boosted Trees (GBT) Model:** GBT is a powerful and widely used ML algorithm that can be implemented for both binary and multiple class classification problem. GBT is an ensemble method which is the combination of multiple decision trees for improvement of the accuracy and robustness of the model. For binary classification, GBT models the probability of the binary outcome as a function of the predictor variables, using a

series of decision trees that are trained sequentially for minimizing the loss function. For multi-class classification, GBT can be extended to use a softmax output layer that assigns probabilities to each of the possible classes. The mathematical form of GBT can

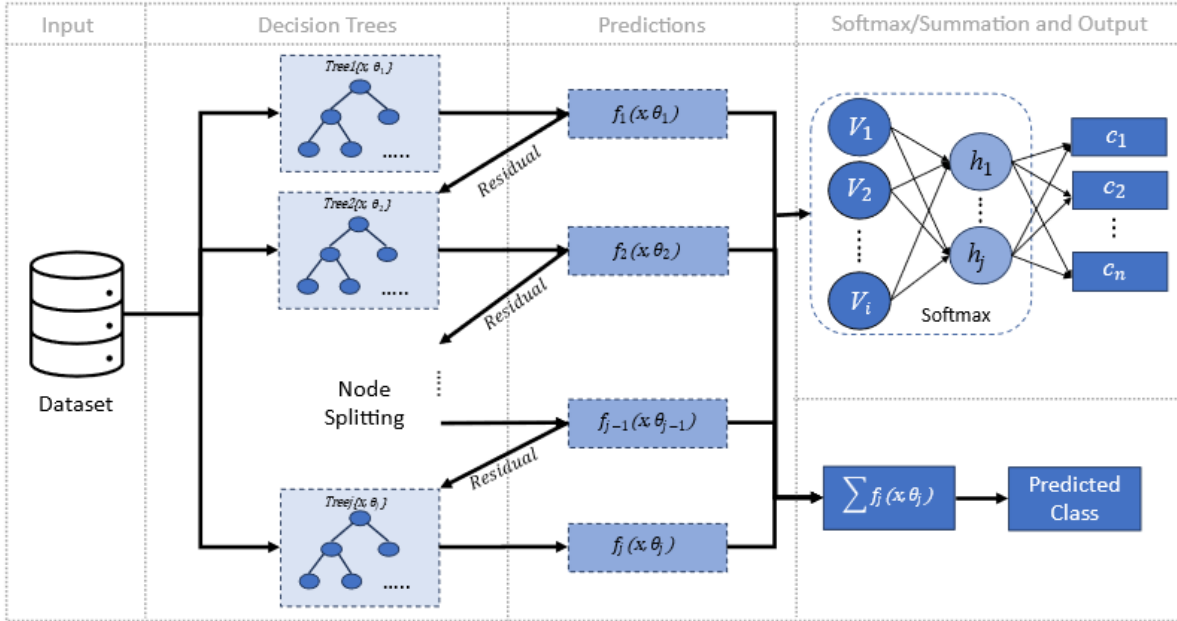


Figure 5.5: Boosted Trees Architecture

be represented as follows:

$$f(x) = \sum T(x; \theta_j) \quad (5.8)$$

where f is the prediction function, T is the decision tree function, θ_j is the set of parameters of the j -th decision tree, and the sum is over all the decision trees in the ensemble.

$$p(y = 0|x) = 1 - p(y = 1|x) \quad (5.9)$$

where $p(y = 0|x)$ is the probability of the negative outcome ($y = 0$) given the predictor variables (x).

$$L(y, f(x)) = - \sum [y \log(p) + (1 - y) \log(1 - p)] \quad (5.10)$$

where L is the loss function that calculates the difference between the predicted probabilities (p) and the true outcomes (y).

- *Deep-Loc*: The proposed Deep-Loc method is especially devised deep learning model for location classification. Deep learning is commonly used, powerful and flexible ML

method that can be used for both binary and multi-class classification problems. The architecture of the implementation of the ML for venue classification is given in the Figure 5.6.

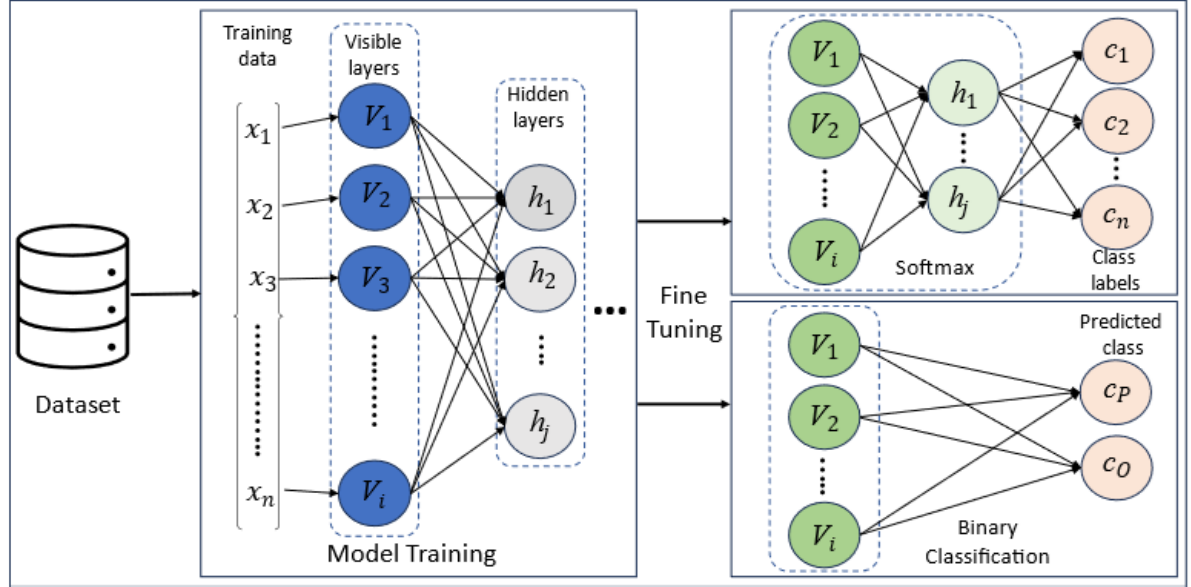


Figure 5.6: Deep-Loc Model for LBSN Venue Classification

The Deep-Loc model uses the check-in data from LBSN to classify the data into multiple classes or extract venue of specific type. The x represents the matrix set of the training data including various attributes for each check-in records and trains through the visible and hidden layers of the devised method. This output classes can either be the multi-class (such as Education for C_1 , Food for C_2 and more) or binary class (P for the predicted class and O for others) as required for the analysis. Deep-Loc model is based on artificial neural networks composed of multiple layers of interconnected nodes that can learn complex nonlinear relationships between the predictor and the response variables. For binary classification, Deep-Loc is used to model the probabilities of the binary outcomes as a nonlinear function of the predictor variables, using one or more hidden layers. For multiple class classification, Deep-Loc is extended to use a softmax output layer that assigns probabilities to each of the possible classes. The mathematical form of Deep-Loc can be written as follows:

$$z(l) = W(l)a(l-1) + b(l), \quad (5.11)$$

where $z(l)$ represents the linear transformation of the input data at layer l while the $a(l-1)$ denotes the activation of the previous layer, and $W(l)$ shows the weight matrix at l layer, $b(l)$ indicate the bias at l layer.

$$a(l) = f(z(l)) \quad (5.12)$$

where f is the activation function that applies a non-linear transformation to the linear transformation of the input data.

$$L(y, y') = - \sum y \log(y') \quad (5.13)$$

where L is the loss function that calculates the difference of the predicted probabilities (y') and the true outcomes (y).

5.2.4 Model Evaluation

The proposed models are implemented using the famous ML platform called RapidMiner. An important part of ML is the model evaluation to estimate the efficiency and effectiveness of the models used during the analysis. A portion of data is always used for training the methods and some portion of unseen data is kept justifying that the said method is good or bad, and the classification or predictions are correctly done. The evaluation techniques used in this study include accuracy and Confusion Matrices for Classification, and AUC or ROC, accuracy, precision, recall, F-score, and sensitivity for Tourism venue Prediction. Most of these famous evaluation methods are self-explanatory and some are described here. The AUC shows the association of true to false positive rates containing threshold each producing 2x2 contingency matrix. Precision is a measure of the true positive predations that are accurately classified, and the recall refers to the true positive prediction of all positive values in the data. The F-score captures both the recall and precision of a single value to show both these properties and the sensitivity is the true positive recognition ratio.

5.3 Results and Discussion

This section provides our results with detailed explanation with evaluation and comparison of the proposed models.

5.3.1 Statistical Modeling

To find the importance of variables used in this study, it was necessary to look at the predictors and their effect on the number of user check-ins statistically before imple-

mentation of the ML algorithms [2]. To implement this model, the following regression equation was used Equation 5.14.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon \quad (5.14)$$

The parameters and considered variables are shown in Equation 5.15:

$$\begin{aligned} Y = & \beta_0 + \beta_1 \text{User_ID} + \beta_2 \text{Gender} + \beta_3 \text{Time} + \beta_4 \text{Day} \\ & + \beta_5 \text{Location_Name} + \beta_6 \text{Educational} + \beta_7 \text{Entertainment} \\ & + \beta_8 \text{Food} + \beta_9 \text{General_Location} + \beta_{10} \text{Hotel} \\ & + \beta_{11} \text{Professional} + \beta_{12} \text{Residential} + \beta_{13} \text{Shopping_Services} \\ & + \beta_{14} \text{Sports} + \beta_{15} \text{Travel} + \epsilon \end{aligned} \quad (5.15)$$

With application of this regression model, the values are shown in Equation 5.16.

Table 5.1: Multiple Linear-Regression

Coefficients	Estimate	Std. Error	t Value	Pr(> t)	Significance
Intercept	4.8321088	0.0158305	289.703	$< 2 \times 10^{-16}$	***
User_ID	1.866674	0.177615	10.51	$< 2^{-16}$	***
Gender	0.612028	0.166562	3.674	0.000249	***
Time	0.940388	0.126142	7.455	1.71×10^{-13}	**
Day	0.871949	0.22472	3.88	0.00011	***
Location_Name	0.837961	0.202606	4.136	3.78×10^{-5}	***
Educational	0.0165532	0.0030441	6.382	5.69×10^{-8}	***
Entertainment	0.0055856	0.0019546	4.106	0.002245	**
Food	0.0080145	0.0019644	4.191	0.001325	**
General Location	0.0153966	0.0020293	8.669	1.88×10^{-14}	***
Hotel	0.0015987	0.002076	3.814	0.004152	**
Professional	0.0040008	0.0009275	3.938	0.003307	**
Residential	0.0079851	0.0019825	4.717	0.000202	***
Shopping & Services	0.015082	0.0030092	6.333	9.86×10^{-8}	***
Sports	0.0088736	0.0018375	4.71	2.50×10^{-6}	***
Travel	0.0090936	0.0019494	5.184	2.89×10^{-5}	***

"Significance Codes: *** (p-value: [0, 0.001]), ** (p-value: [0.001, 0.01])."

$$\begin{aligned} \hat{Y} = & b_0 + b_1 \text{User_ID} + b_2 \text{Gender} + b_3 \text{Time} + b_4 \text{Day} \\ & + b_5 \text{Location_Name} + b_6 \text{Educational} + b_7 \text{Entertainment} \\ & + b_8 \text{Food} + b_9 \text{General_Location} + b_{10} \text{Hotel} \\ & + b_{11} \text{Professional} + b_{12} \text{Residential} + b_{13} \text{Shopping_Services} \\ & + b_{14} \text{Sports} + b_{15} \text{Travel} + \epsilon \end{aligned} \quad (5.16)$$

The model coefficients are presented in following Table 5.1, where ‘Education’ depicts unit increase in the value, the check-ins raised approximately 1.6% times with low p -value; likewise, the number of user check-ins in other categories have low p -value showing the significant variables. The feasibility and significance of the data attributes used in this research can also be observed in the correlation matrix as in Table 5.2:

Table 5.2: Correlations Matrix

	Time	Gender	Category	CheckinDate	Weekdays
Time	1	-0.050**	-0.005	-0.053**	0.017**
Gender	-0.050**	1	0.015**	-0.017**	-0.012**
Category	-0.005	0.015**	1	-0.039**	-0.013**
CheckinDate	-0.053**	-0.017**	-0.039**	1	-0.037**
Weekdays	0.017**	-0.012**	-0.013**	-0.037**	1

***. Correlation is significant at the 0.01 level (2-tailed)."

The statistical analysis provides the means of selecting the most efficient variables for successful classification and prediction before using ML techniques. In the next stage, we used these significant variables for our proposed ML models’ implementation for LBSN data analysis.

5.3.2 Classification Into Multiple Venue Types

The following four ML methods have been proposed in this study based on GLM, LRM and GBT and Deep-Loc model for venue classification for improving the classification of LBSN data that has been done manually for decades in many fields by many researchers.

The previously studied 10 venue classes have been used for supervised learning namely ‘Educational’, ‘Entertainment’, ‘Food’, ‘General Location’, ‘Hotel’, ‘Professional’, ‘Residential’, ‘Shopping & Services’, ‘Sports’, and ‘Travel’. The Figure 5.7 shows the overall performance of these methods for classification problems.

The results show the high accuracy of the Deep-Loc model for our classification problem of LBSN data into prespecified categories. The other models, including the GLM, LRM, and GBT, also performed very well in the classification of LBSN data. The Deep-Loc model and GBT model also have a high accuracy, which indicates that these are suitable for our data. The LRM are simple, easy to interpret, and fast to train. They are best suited for linear problems, where the relationship between the predictors and the target is approximately linear. In these cases, LRM can provide accurate predictions and good interoperability. The GBTs, on the other hand, are more flexible and powerful and can handle nonlinear relationships between predictors and target. They are based on

decision trees, which are decision-making techniques that capture complex relationships in the data. GBT can also handle missing or noisy data, and they can learn interactions between predictors.

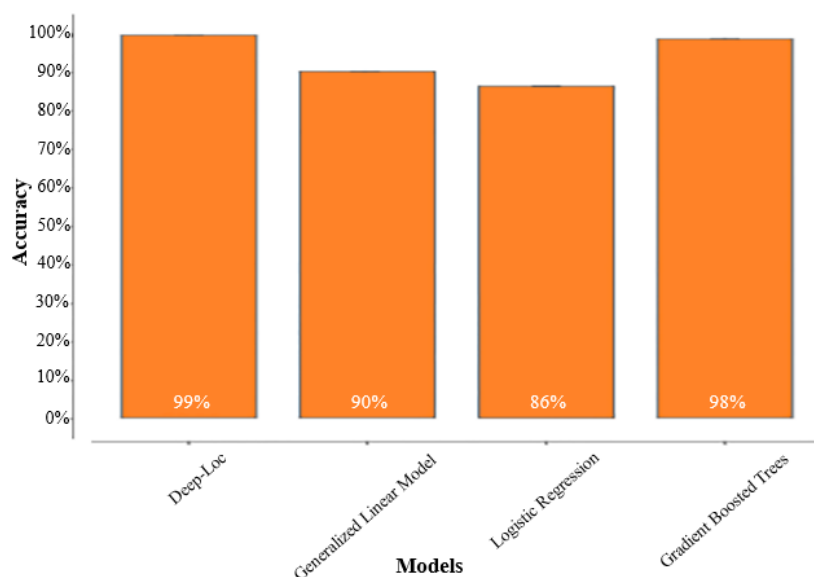


Figure 5.7: Venue Classification into 10 Classes Using Machine Learning

In general, GBT tends to outperform LRM when the interrelations between the predictors and the target is nonlinear, and when the data contains noise or missing values. However, GBT can be more difficult to interpret and can be slower to train than LRM. It is important to note that choosing the best model depends on the specific requirements of the task, the nature of the data, and the desired performance characteristics. In the following section, we provide the confusion matrix for each implemented method to show how significantly they performed on the test data. The performance of each individual model is provided as follows.

a). Generalized Linear Model

It is one of the fastest and most efficient methods working a probabilistic classifier that has been used in variety of applications in the past decades. This method has been implemented due to its competence in classification problems, precision and robustness that can be seen in numerous research articles over the years as stated by [151]. This method is an enhancement of the linear models by using the maximum likelihood estimator. The model provides the high speed with parallel computations achieving high accuracy as shown in following Figure 5.8.

True Predicted	Travel	Residential	Professional	Educational	Shopping& Services	Food	General Location	Entertainment	Sports	Hotel
Travel	4378	22	9	13	12	6	3	14	0	11
Residential	99	7764	133	137	51	19	14	93	32	66
Professional	5	17	2685	12	5	4	6	41	1	8
Educational	28	52	15	7288	60	15	5	56	47	32
Shopping& Services	27	38	19	55	8883	37	6	173	11	15
Food	391	186	248	378	481	2100	106	865	134	153
General Location	77	139	44	62	104	29	2281	199	49	35
Entertainment	25	22	108	42	64	58	45	14288	98	30
Sports	8	6	11	30	3	13	8	151	3829	0
Hotel	52	13	13	21	21	3	4	30	13	2514

High
Low

Figure 5.8: Confusion Matrix for Generalized Linear Model

b) Logistic Regression

LRM is another famous ML method that is widely used as statistical model for the classification. It performs used mostly to predict nominal variables and fits our data for classification as shown in the following Figure 5.9.

True Predicted	Travel	Residential	Professional	Educational	Shopping& Services	Food	General Location	Entertainment	Sports	Hotel
Travel	4814	5	0	1	2	4	0	1	0	9
Residential	1	8080	0	22	0	21	7	0	22	38
Professional	2	79	1954	7	0	9	8	1	0	11
Educational	1	0	34	6981	0	32	3	0	1	2
Shopping& Services	171	3	3	11	9682	2	0	2	5	1
Food	0	0	4	3	0	650	1	0	3	0
General Location	4	94	22	0	0	43	1753	0	62	9
Entertainment	104	13	1232	947	10	1539	704	15928	1297	848
Sports	2	0	12	7	0	6	1	0	2717	3
Hotel	0	0	2	9	0	4	1	0	2	1916

High
Low

Figure 5.9: Confusion Matrix for Logistic Regression Model

c) Gradient Boosted Trees

GBT uses parallel computing for boosting the classification process by using Gradient

Boosting Machine . It provides accuracy with the help of the effective linear model. The effectiveness of using Gradient Boosted trees-based model for classification of LBSN data can be observed in the following Figure 5.10.

True \ Predicted	Travel	Residential	Professional	Educational	Shopping & Services	Food	General Location	Entertainment	Sports	Hotel
Travel	1432	0	1	0	0	0	0	0	0	0
Residential	0	2310	1	0	0	1	0	2	0	0
Professional	0	0	876	0	0	0	0	0	0	0
Educational	0	0	0	2245	0	9	0	1	10	0
Shopping & Services	0	0	2	0	2725	3	0	1	0	0
Food	0	0	0	0	0	501	0	0	0	0
General Location	0	9	2	0	0	12	696	7	0	0
Entertainment	0	0	40	9	0	116	0	4422	0	0
Sports	0	0	0	0	0	2	0	41	1174	0
Hotel	0	0	1	0	0	2	0	0	0	807

High
Low

Figure 5.10: Confusion Matrix for Gradient Boosted Trees

c) Deep-Loc Model

It is based on a famous neural network-based method which uses information within the data in layered form in order to extract useful patterns and classes. Each neuron is trained by modification based on the available information and combinedly predicts the output acting as a classifier. It works in an adaptive manner by optimizing the neurons instinctively through learning without human interaction saving the effort and time required for the classification. This Deep-Loc model is based on the feedforward artificial neural network architecture. The hidden layers provide the capacity to capture complex relationships between the input and output variables. It uses a supervised learning approach, which requires labeled training data to train the model provided by our previous research. In the training phase, the model learns the relationships between the input variables and the target classes through optimization of a loss function which is a measure of the error between the predicted outputs and the actual outputs. It performs exceptionally well for our classification and prediction problem as shown in the following Figure 5.11.

These results demonstrate high efficiency in use of ML instead of manual classification by providing more accurate and timely results with high potential in implementation in LBSN analysis. The Deep-Loc model performed very well in multi-class prediction

True \ Predicted	Travel	Residential	Professional	Educational	Shopping& Services	Food	General Location	Entertainment	Sports	Hotel
Travel	5008	0	1	0	0	0	0	0	0	0
Residential	0	8124	4	0	0	1	4	0	0	0
Professional	0	0	3231	0	0	1	0	0	0	0
Educational	2	3	0	8041	0	0	1	0	0	0
Shopping& Services	0	3	1	0	9695	2	2	0	0	0
Food	4	6	1	0	0	2255	1	0	0	0
General Location	3	7	4	0	0	1	2427	0	0	0
Entertainment	75	120	45	0	0	31	46	15931	0	0
Sports	0	1	2	0	0	0	0	0	4222	0
Hotel	1	0	2	0	0	0	2	0	0	2862

High
Low

Figure 5.11: Confusion Matrix for Deep-Loc Model

with accuracy reaching 99% in our experimentation. These models can also be used for binominal predictions targeting a desired class such as, in our case, tourism with others. The next section provides similar research in which we used ML to predict the tourism venues with supervised learning based on our acquired data from Weibo.

5.3.3 Predicting Tourism Class

The proposed ML models can be used to predict an individual class among the huge LBSN based heterogeneous data which is given in this section. In the current research, we used the proposed models to predict tourism venues with the help of supervised methods from our previously used dataset from Weibo. These venues are predicted based on proximity to the information provided in the dataset including gender of the specific user, previously visited venues within specific time of the day, day of week, latitude/longitude, and venue names. An example of the methods of the proposed model for prediction is given in Figure 5.12.

The results presented in this section provide evidence of the efficiency and effectiveness of using ML for extracting useful traits and patterns in the behavior of users, more specifically tourists. This information can be used to conduct useful research in predicting the preferences of the tourists as presented in our research. The above Figure 5.13 represents the ROC curve alternatively used for AUC (however, as the lines and overlapping at several positions that why the ROC is preferred here). The ROC curve is

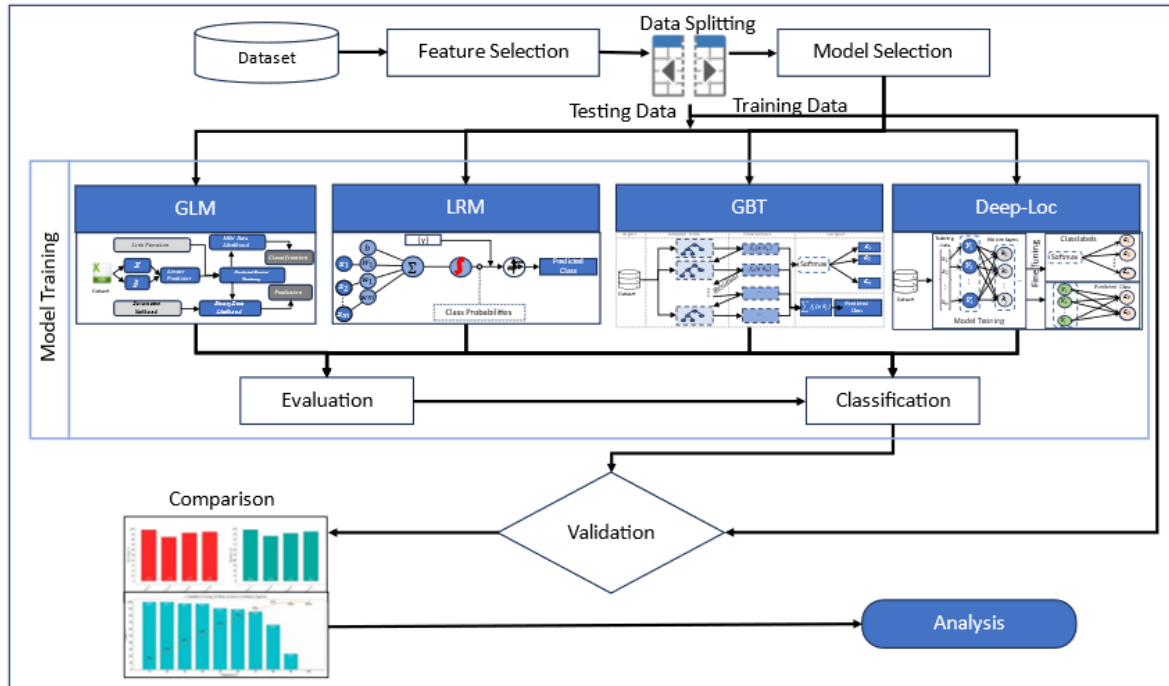


Figure 5.12: Machine Learning Methodology for Venue Class Prediction

a graphical illustration of the performance of a binary classification. The ROC graph is a plot of the True Positive Rate (TPR) versus the False Positive Rate (FPR) for all possible values of threshold. The TPR is the proportion of samples correctly classified as positive, whereas the FPR is the proportion of samples incorrectly classified as positive. The ROC graph plots TPR against FPR as the discrimination threshold of the classifier is varied, and the resulting curve provides a visualization of the trade-off between the TPR and FPR. In the ROC graph, the TPR is plotted on the y-axis and the FPR on the x-axis. A classifier with perfect performance will have a TPR of 1.0 and an FPR of 0.0 and will be shown at the top-left corner of the graph. A classifier with poor performance will have a TPR that is close to 0.0 and an FPR that is close to 1.0, and will be located close to the bottom right corner of the graph.

The AUC values greater than 0.9 represents excellent results; the values of 0.8 to 0.9 are ranked as good, those 0.7 to 0.8 shows fair, and AUC values less than 0.6 are considered poor. The following Figure 5.14 shows high values of the AUC for our models and Figure 5.15 suggesting high accuracy in the prediction of tourism venues among all other types available in the LBSN dataset. The method that is used for prediction of tourism venue class, where the available attributes like UserID, gender, date/time, latitude, longitude, location_name and category are used for the supervised learning models to predict the venue_categories in unused test data showing the efficiency of

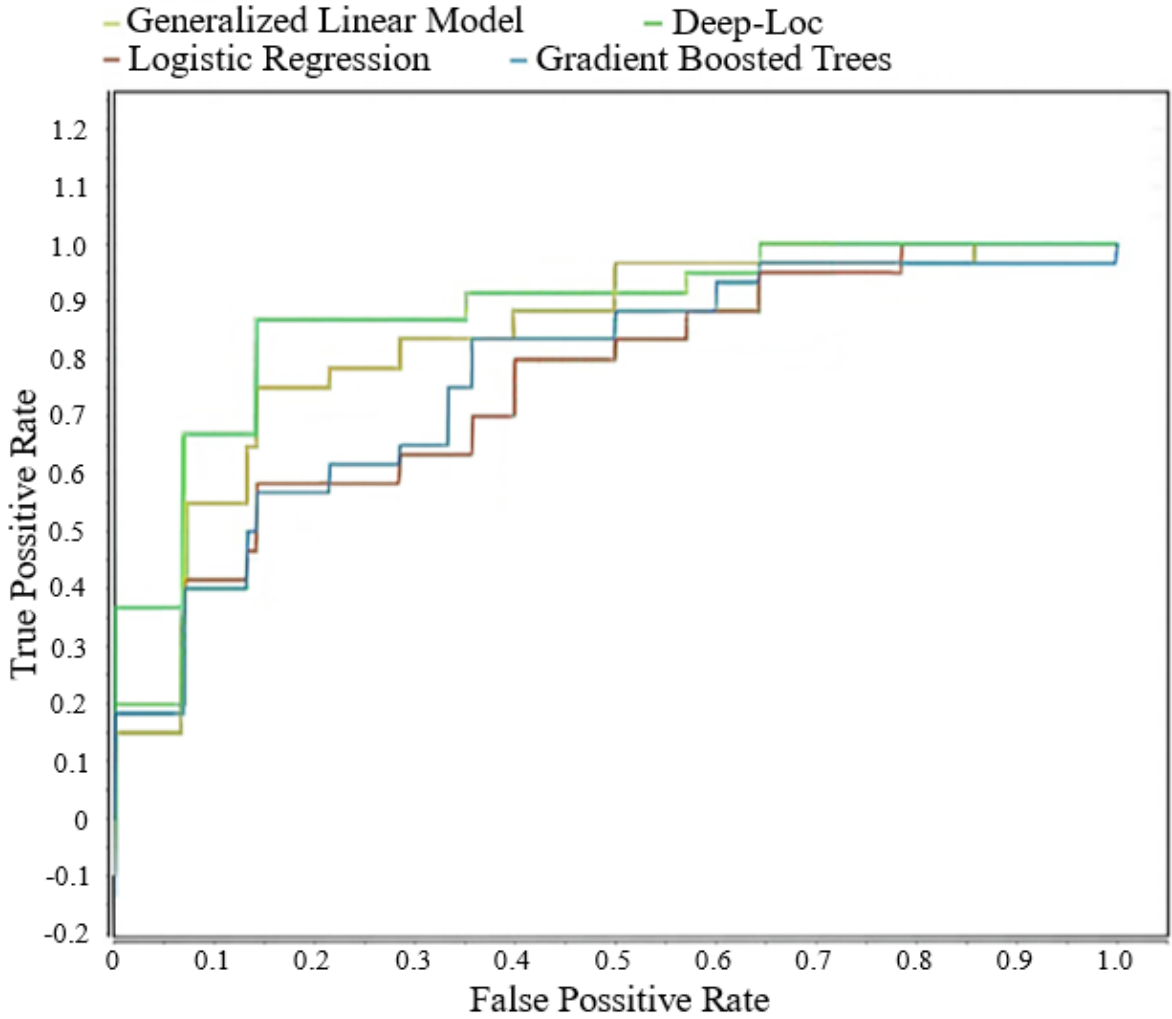


Figure 5.13: ROC of The Proposed Machine Learning Models

using ML models for LBSN based venue classification. We show the other performance techniques presented in Section 3, Figure 5.16 demonstrates the performance of four ML methods with respect to recall, precision, F-score, and sensitivity. The Figure 5.17 represents the lift charts for each of the implemented ML models. A lift chart is defined as a graphical representation of the improvement of a model in comparison with a random guess, which also means evaluating the effectiveness of the model by using the ratio between the results “with and without a model” [152].

The lift chart plots the percentage of positive samples that are correctly classified by the model on y-axis, and the cumulative value in percent of all samples on x-axis. The chart is split into a number of equal deciles, and the lift of the classifier at each decile is measured as the ratio of the number of positive samples correctly classified to the number of positive samples that would be correctly classified by a random selection.

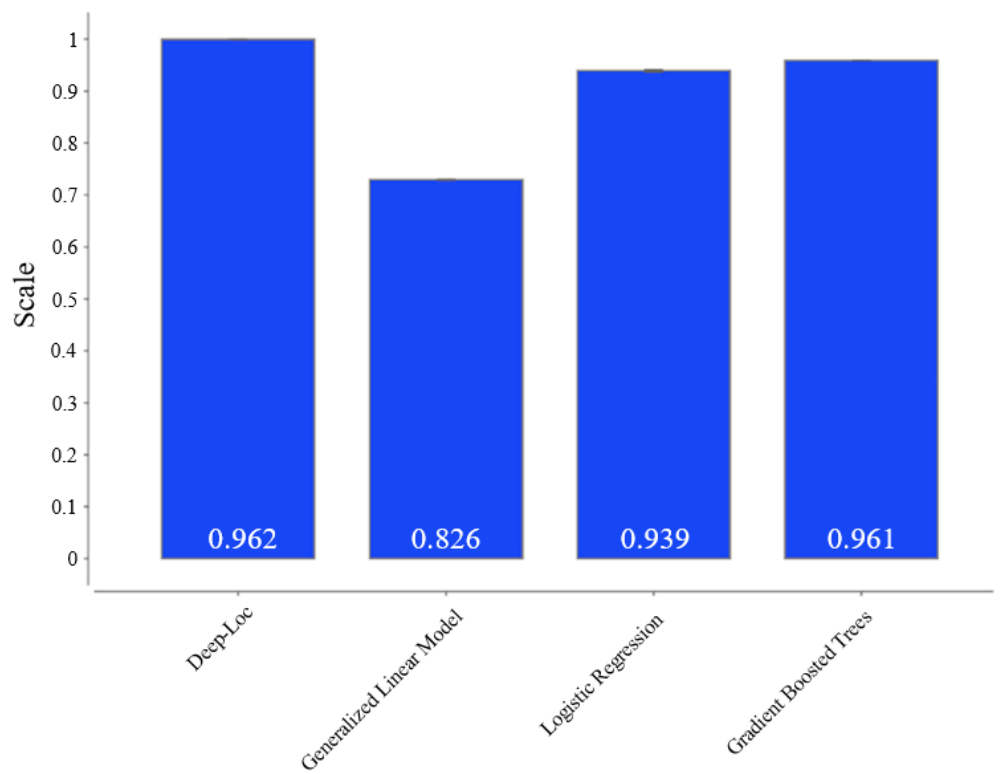


Figure 5.14: Graphical Representation of AUC

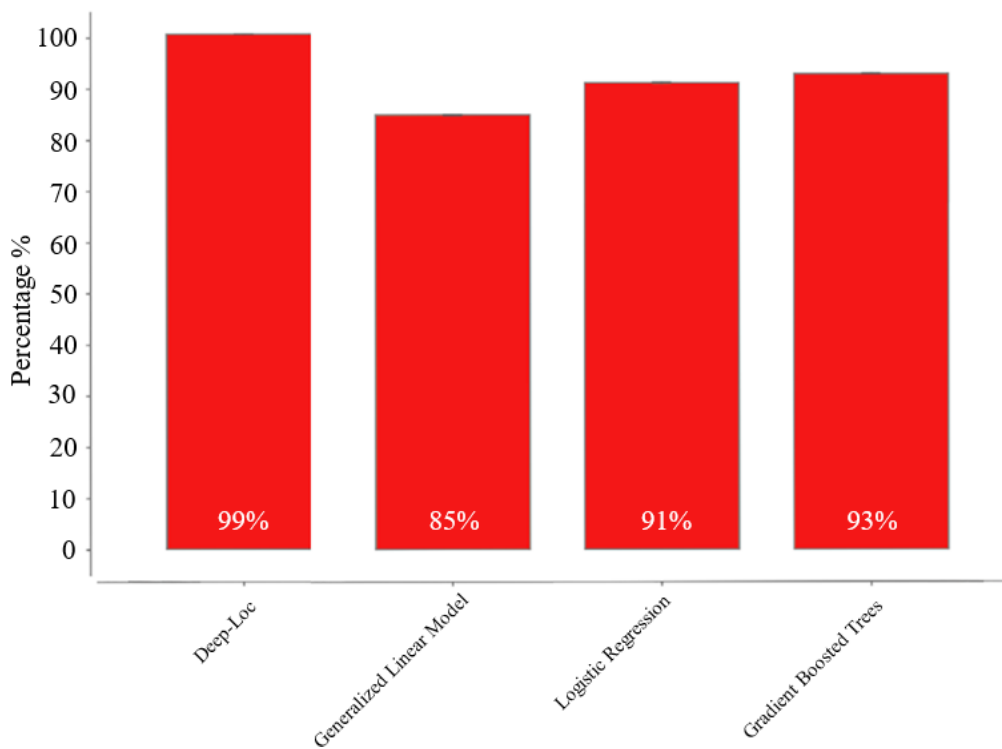


Figure 5.15: Accuracy of The Candidate Models

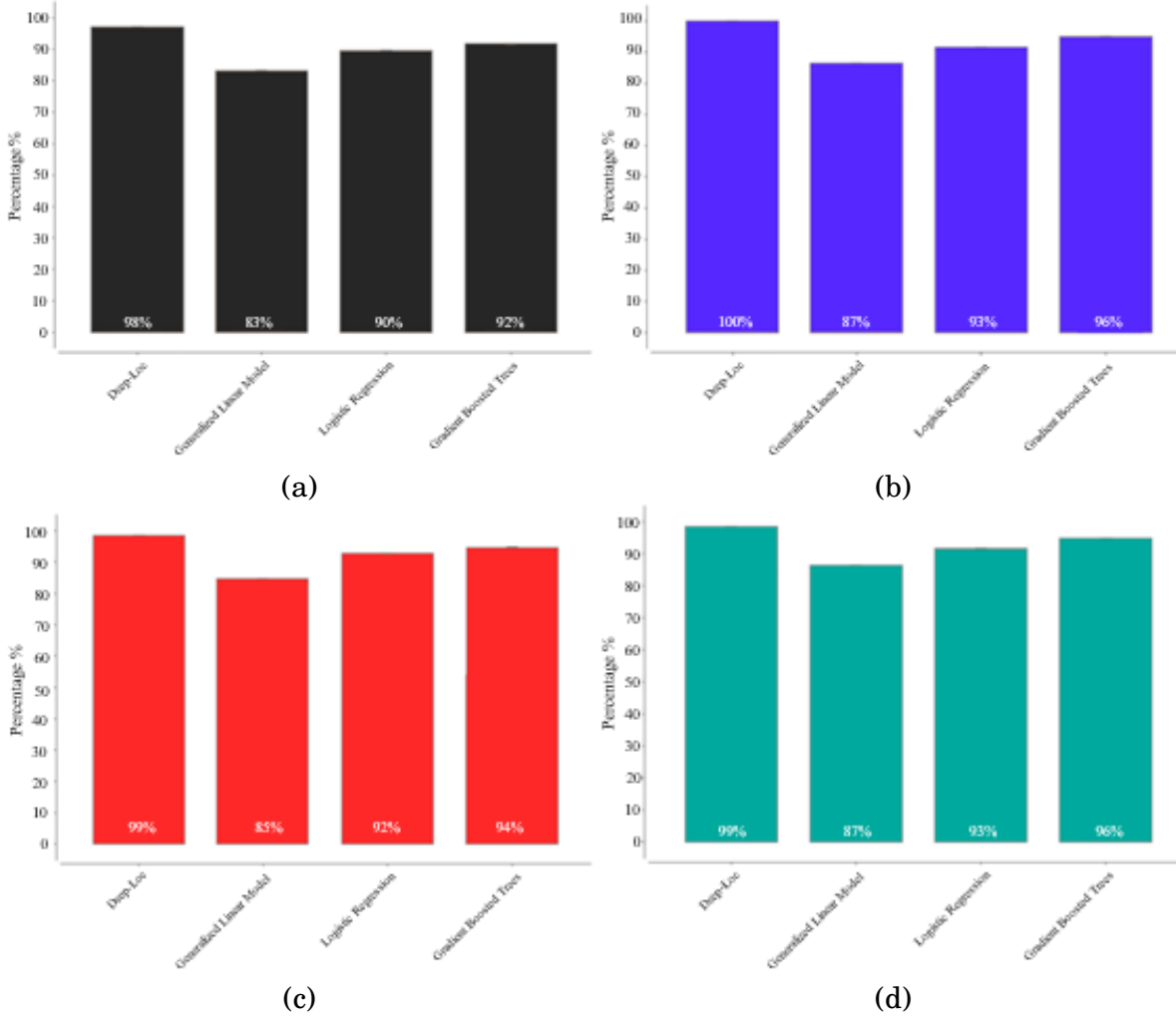


Figure 5.16: Performance Matrices. a) Precision, b) Recall, c) F-score and d) Sensitivity

The Figure 5.17 shows the high learning and accuracy of using machine learning methods to predict tourism venues among the massive amount of data in the dataset. It can be observed that the Deep-Loc model can predict tourism venues with very high accuracy as compared to others, while the proposed GLM, LRM, and GBT have significant performance. The proposed models can be used to classify data into multiple categories and predict a single class based on the nature and activities performed at these venues, which removes the overhead of manually filtering. This can be helpful in research in various fields, such as tourism, restaurants, parks, etc., with applications in development and planning of smart cities.

CHAPTER 5. VENUE CLASSIFICATION AND PREDICTION

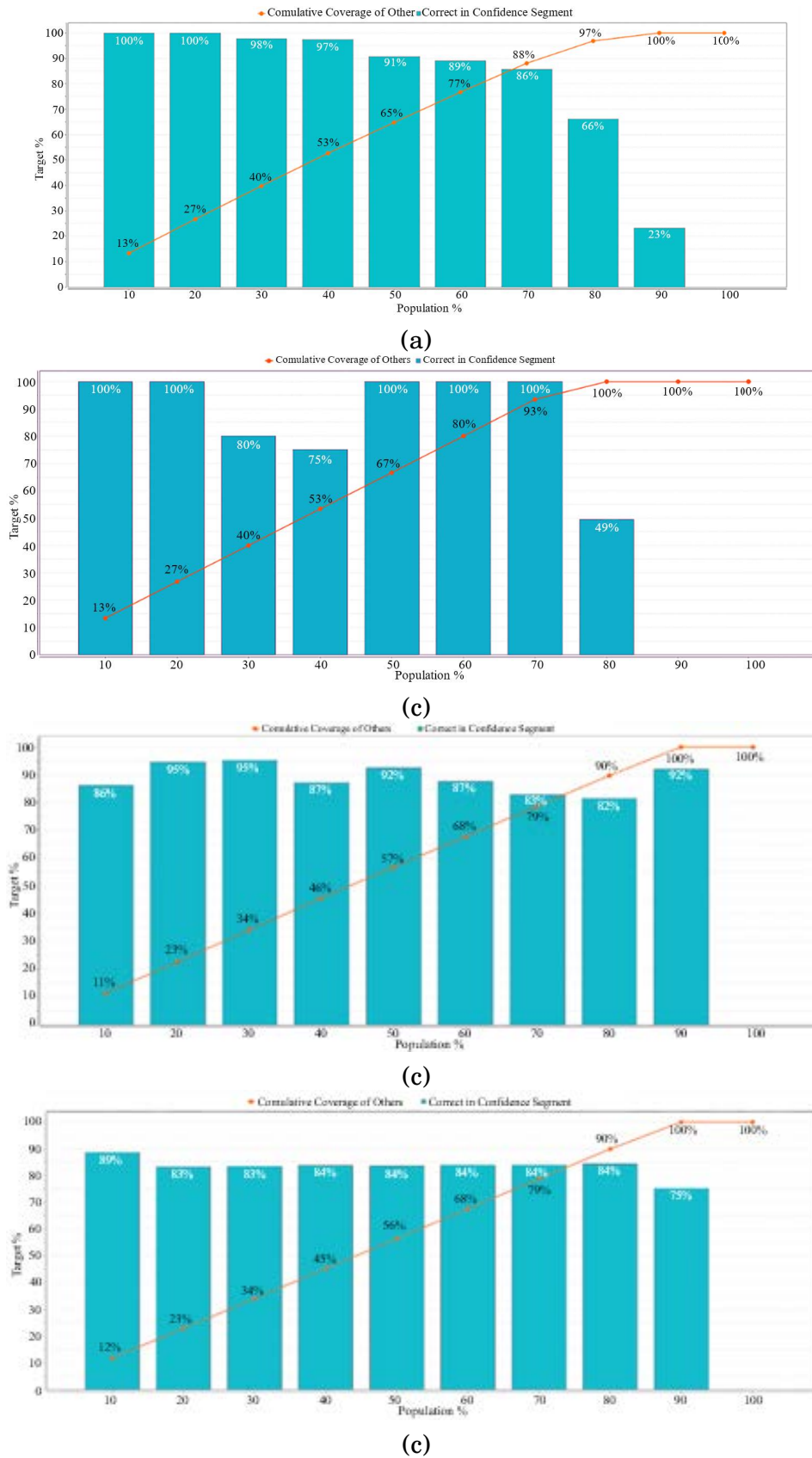


Figure 5.17: Lift Charts of Proposed Models, a) Deep-Loc, b) GLM, c) LRM, d) GBT

5.3.4 Contribution and Comparison

Most of the related carried out in this domain are based on the LBSN data extracted or classified manually either into multiple venue classes or specialized class such as, in our case, Tourism venues. As mentioned earlier, the research on venue classes in current literature is based on categorizing each venue manually by going through hundreds or thousands of records and putting them into related categories, which is very time consuming, need a lot of human effort and its accuracy rely on human observations and assumption regarding accuracy. Therefore, we introduced ML methods to classify the huge data automatically based on examples with evident high accuracy. We included the recent work for comparison in order to show the contribution of our work in this domain. Examples and comparison of our work with similar articles are provided in Table 5.3.

Table 5.3: Comparison with Related Work

Studies	Source	Attributes	Analysis Method	Venue Types	Venue Classes	Acc
Wei et al. (2018) [153]	Weibo	Biography, gender	Spatial clustering	Single	Manual	-
Dingel et al. (2019) [154]	Yelp	Check-ins, Demographics	Location detection	Multiple	Manual	-
Li et al. (2019)[155]	Twitter	Check-ins, (age, gender, edu)	Frequent visits	Multiple	Manual	-
Wakabayashi (2021) [156]	Flickr	Photographs, Demographics	Density Estimation	Multiple	Manual	-
Senefonte et al. (2022) [157]	Foursquare	Check-ins, Demographics	Frequent visits	Single	Manual	-
Terzi (2022) [158]	Foursquare	Check-ins, Demographics	Density Estimation	Multiple	Manual	-
Wang et al. (2022) [94]	Instagram	Posts, Demographics	Spatial clustering	Multiple	Manual	-
Cagliero et al. (2022) [1]	Foursquare	Check-ins, Demographics	Mobility patterns	Multiple	Manual	-
Ferreira et al. (2023) [92]	Flickr/ Twitter	Check-ins, Demographics	Visits, Spatial clustering	Multiple	Manual	-
Pezenka et al. (2023) [91]	Instagram	Comment, like, Demographics	Frequent visits	Multiple	Manual	-
Proposed Method (2023) [71] (Ours)	Weibo	Check-ins, Demographics (age, gender, date/time)	Classification Prediction	Multiple	Machine Learning	85-99%

"Quantitative assessment of accuracy for manual venue classification and extraction is not possible."

It is important to note that the accuracy values for manual venue classification methods reported in the related literature are generally not available, as these studies rely on human annotation or heuristic-based categorizations without quantitative performance evaluation. Manual classification inherently involves subjective judgment, making rigorous accuracy assessment challenging or infeasible. In contrast, our proposed machine learning approach provides objective, reproducible accuracy metrics by leveraging labeled datasets and standard evaluation protocols. The reported accuracy range of 85–99% demonstrates the effectiveness and reliability of our automated classification framework compared to manual methods, which lack comparable quantitative validation. This distinction underscores the contribution of our work in advancing scalable and verifiable venue classification for large-scale LBSN datasets.

5.4 Summary

In this chapter, we proposed an efficient approach to LBSN venue-based analysis by the implementation of ML models for the prediction of venue classes used in the research of patterns in the behavior of different LBSN users. The classification task has always been done manually, which is very tiresome and time consuming as most of the LBSN research is based on big data analysis comprising thousands and millions of check-in records. There are several robust ML methods that can be used to carry out this task more efficiently and effectively. In this research, we developed four ML models based on famous classification and prediction methods, including the GLM, GBT, LR, and Deep-Loc model for our experiments. These data mining techniques on LBSN data are rigorous tasks due to the fact that many features are related to many different domains in various research fields. After careful and systematic filtering, we extracted the feasible input features for the prediction of our targeted class, which followed the training and testing of our developed models.

The results revealed that the Deep-Loc model performs well for classifying and predicting venues within the LBSN data. The GBT model attained 93% accuracy for our tourism class prediction problem, followed by LRM and the GLM, reaching 91% and 85% accuracy, respectively. These models perform well, but the research has some limitations. For example, the models must be tested on LBSN data from other platforms and other research domains in order to provide a more generalized solution to LBSN classification problem. To mitigate potential overfitting risks, we applied 10-fold cross-validation with stratified splits and reserved a portion of training data for validation, which helped to

ensure robust model evaluation on unseen data. Nevertheless, the model's generalizability beyond the Shanghai Weibo dataset remains to be fully established. Due to data availability constraints, this study focused on Shanghai data, and future work will aim to evaluate the Deep-Loc model on LBSN datasets from other cities and platforms. This will include retraining, fine-tuning, and exploring domain adaptation methods to improve transferability across different urban contexts, as well as employing data augmentation techniques to enhance robustness. The presented results can be beneficial in a variety of research fields by specifying and predicting the desired class of venues and users. It can also provide the basis to conduct LBSN data analysis to predict the interests, behavior, and trends of the population within a specific time of the day or day of the week. The use of ML for such kind of research can benefit both researchers and end users for better planning, targeted marketing, and development of a smart city environment.

ANALYSIS OF LBSN GROUP RECOMMENDATION SYSTEM

In recent years, LBSNs have gained significant popularity, enabling users to interact with POIs using modern technologies. As more and more people rely on LBSNs for finding interesting venues, contextually aware and relevant recommendation systems have become very beneficial with practical applications. In this research, we propose an enhanced hybrid recommendation system, designed for LBSNs to improve the accuracy of suggestions by integrating Collaborative Filtering (CF) methods with Singular Value Decomposition (SVD) to handle sparse data, along with context-aware modeling to tailor recommendations based on user interests, and group recommendation to accommodate multi-user scenarios. Additionally, we incorporate contextual aspects such as spatial proximity and temporal behavior into the model to ensure recommendations align closely with the user's present surroundings and their preferences. The proposed method extends further to group recommendations by considering individual inclinations into cohesive suggestions for groups interested in visiting POIs together. The proposed method is assessed using precision, recall, and F1 score, ensuring thorough evaluation of its performance. To further highlight context-aware recommendations, we use clustering based on user preference, temporal behavior, and category-wise interaction to identify patterns across various venue types. The proposed method shows improved recommendations, specifically based on data from LBSNs, and for developing an efficient solution for balanced user preferences with contextual influences.

6.1 Introduction

LBSNs have become increasingly important in our everyday life, thanks to the widespread use of smartphones and GPS devices. Apps like Weibo, Gowalla, Foursquare, Yelp, and Google Maps let people share their locations, check-in at different places, and connect with friends. These networks do more than just keeping records of the user's movements, they create a complex system where people interact with their surroundings in both physical and social ways. An opportunity and the big challenge are developing recommendations that are more relevant by suggesting Point Of Interests (POIs), places the user might like based on where they have been, who they know, and what activities the users enjoy frequently with the help of methods like content based filtering and CF [159]. Traditional recommendation systems, e.g. used by Netflix or Amazon, work well for suggesting movies or products. They typically look at what the users have liked before or what similar people preferred [160]. However, these methods can't be implemented directly when it comes to location-based recommendations, which involve more complicated factors like location, social, or temporal aspects related to their check-ins. When recommending a venue, systems now consider much more than just general past preferences, they look at things like physical proximity and typical visitation times for similar venues. It's a much more nuanced approach to helping people discover interesting places nearby. These approaches represent a significant shift in how technology understands and anticipates our daily experiences, transforming simple check-in data into a sophisticated, personalized recommendation models.

6.1.1 Challenges in LBSN Recommendations

One of the significant challenges in LBSN recommendation system is the critical influence of user proximity to other users in the physical world [161]. Unlike previous recommendation systems that focused on item features like type, price, or characteristics, LBSN users tend to frequent POIs located near their current location [162]. Location plays a pivotal role in generating meaningful recommendations. Previous research demonstrates that individuals mostly travel from their current location before heading to the recommended venue, underscoring the importance of spatial data in recommendation algorithms. A recommendation system might suggest appealing places, but its utility diminishes if it fails to consider geographical distance [163]. Beyond geographical considerations, temporal parameters are equally crucial in understanding user decision-making. For example, a user's preferences vary throughout the day and week, the users

might seek cafés in the morning, restaurants in the evening, or parks during weekends. Therefore, LBSN recommendations must incorporate time-based factors that reflect users' behavioral patterns during specific periods [164]. These temporal elements are not merely supplementary but fundamental to recommendation accuracy. Ignoring time-related context can result in inappropriate suggestions that feel disconnected from users' actual behavioral patterns, fundamentally compromising the system's functionality and relevance [165] [166]. The major challenge lies in developing recommendation systems that can integrate spatial and temporal dimensions, providing suggestions that are not just potentially interesting, but practically actionable within a user's context.

However, user preferences and social factors present additional complexities in LBSN recommendation systems. Empirical research demonstrates that individual historical behaviors such as previous check-ins, repeated visits, and consistent location choices significantly influence preferred visiting venues [167][168]. Past preference patterns dramatically shape venue selections. Users with a history of social interactions tend to gravitate towards vibrant, interactive venues like bars and festivals, while those with more solitary past behaviors prefer more contemplative spaces such as libraries and parks. The location-based data offers huge opportunities to enhance recommendation accuracy by integrating contextual factors including user location, temporal variables, and local events. Individual preferences can vary substantially between travel and home, with local conditions and events further modulating decision-making patterns. Use of this contextual information becomes crucial in creating more accurate recommendations. By combining these factors, recommendation systems can generate suggestions that are aligned with users' circumstances and past preferences. The fundamental challenge lies in developing recommendation systems that can analyze and model the relationship between users past behaviors, spatial context, and temporal dynamics [169].

6.1.2 Hybrid Recommendation Systems

The complex nature of user preference in LBSNs makes it crucial explore hybrid recommendation systems to get more relevant suggestions. In rapidly changing environments like LBSNs, hybrid systems can effectively address the limitations of individual methods by combining the strengths of various recommendation approaches, ultimately providing more accurate recommendations. Most Conventional recommendation approaches, such as CF, are typically designed to predict what a specific user is likely to prefer based on their past activities. However, these traditional methods often fall short in capturing the complex nature of user interactions within the location-based systems. Hybrid

recommendation systems offer a more comprehensive solution by integrating multiple recommendation techniques, allowing for a more nuanced and adaptable approach to understanding user preferences. By integrating different methodological strategies, these systems can provide more robust and context-aware recommendations that better reflect the multilayered user behavior in LBSNs [170]. There are two main types of memory-based CF approaches [166]: including, User-based CF which recommends POIs based on preferences of similar users and Item-based CF that suggests POIs similar to those the user has previously visited. Despite their effectiveness, individual methods often struggle with data sparsity and lack contextual adaptability, both critical for LBSNs. Given the complex nature of user preferences in LBSNs, hybrid recommendation systems have emerged as a promising solution, integrating CF with SVD, and context-aware techniques to offer more accurate and personalized recommendations. In this study, we propose a contextual hybrid recommendation strategy that integrates user preferences, geographical location and temporal behavior factors with CF and SVD matrix factorization to address the unique challenges of LBSNs. Using standard performance metrics, such as precision, recall, and F1 score, we assess the quality of the proposed system's recommendations. The serves as a powerful dimensionality reduction technique that enables the identification of hidden patterns within user-item interaction data that might otherwise remain imperceptible [167]. While SVD improves recommendation quality and addresses data sparsity issues, it does not consider context such as spatial or temporal factors that are essential in LBSN recommendation systems.

Key contributions of this research:

- This research proposes an effective hybrid recommendation framework that integrates user-based, item-based CF with SVD and context-aware user modeling. This combination is specifically designed to address and mitigate the limitations found in traditional recommendation approaches within LBSNs, enabling more accurate and context-sensitive recommendation strategies.
- By embedding factors such as spatial proximity and temporal behaviors, our framework advances the personalization of recommendations. This integration allows for recommendations that adapt to varying user contexts, aligning suggestions with location-based and temporal dimensions to closely reflect diverse user needs and preferences accumulating to group recommendations.
- Through extensive experimentation with standard evaluation metrics (precision, recall, F1 score, and accuracy), we demonstrate that the proposed hybrid model outperforms

traditional methods. This thorough assessment highlights the framework’s potential to maintain high recommendation quality and accuracy in data-sparse and dynamic environments, marking a significant improvement over conventional methods in LBSNs.

We present the methodology and framework for designing the LBSN group recommendation system as follows. The framework in Figure 6.1 represents the proposed hybrid approach designed to generate highly personalized and contextually relevant recommendations in LBSNs. To demonstrated the effectiveness of the model we used the user check-in data from famous LBSN named Gowalla [171]. The user influence modeling component builds a detailed profile for each user by analyzing key factors which contribute to a long-term understanding of user preferences. Parallellly, CF is employed, encompassing user-based CF, item-based CF, and SVD. This combination allows the framework to uncover hidden relationships within user-item interactions, making it particularly effective in sparse data environments. Additionally, the context-aware module dynamically adjusts recommendations based on the user’s immediate context, for example, their current location and time, ensuring relevance.

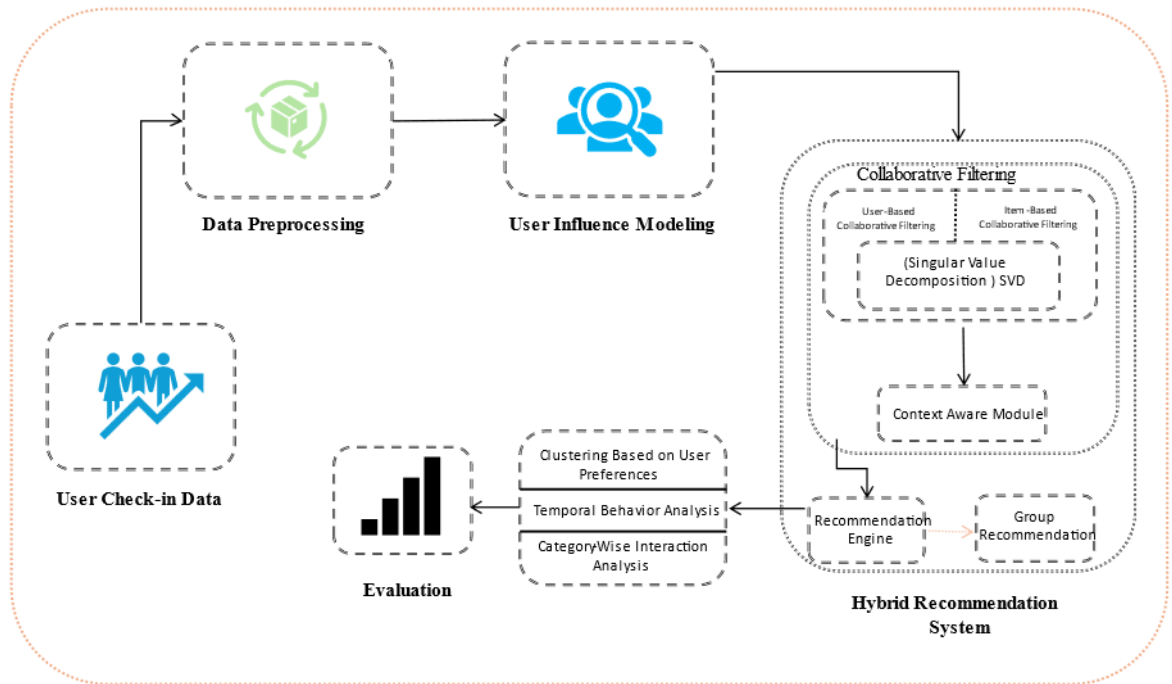


Figure 6.1: Proposed Framework

Once the system has gathered insights from user influence modeling, CF, and context-aware modules, it combines these in the recommendation engine to generate a final set of suggestions tailored to the individual. For group scenarios, the group recommendation

module aggregates individual preference scores to provide unified recommendations that meet the collective needs of multiple users, balancing diverse tastes within the group. Finally, the evaluation component measures the system’s performance using metrics such as precision and recall, along with analyses based on user preferences, temporal behaviors, and category-specific interactions. Together, these components form a robust, scalable recommendation system that effectively balances long-term personalization with adaptive, recommendation contextual adjustments, resulting in recommendations that are both accurate and highly relevant to the user’s preferences.

6.2 Hybrid Recommendation System for LBSN

The hybrid recommendation system integrates multiple recommendation techniques to provide highly personalized and contextually relevant recommendations for LBSNs. The dataflow diagram in Figure 6.2 shows the steps for achieving this goal. The initial steps involve preprocessing the user-item interaction matrix and extracting contextual information regarding the user, including geographical location and temporal behavior. These contextual factors are key to understanding the user’s preferences more comprehensively. CF methods are applied to predict scores for POIs as shown in the equation 6.1:

$$\hat{r}_{\text{final}} = w_1 \cdot \hat{r}_{\text{UB}} + w_2 \cdot \hat{r}_{\text{IB}} + w_3 \cdot \hat{r}_{\text{SVD}} \quad (6.1)$$

where \hat{r}_{final} is the final prediction rating of a user-item pair, \hat{r}_{UB} is the predicted score from user-based CF, \hat{r}_{IB} is the prediction score from item-based CF, \hat{r}_{SVD} is the predicted score from SVD-based context-aware module, w_1 , w_2 , w_3 are the weights assigned to each method, representing their contribution to the final prediction.

In our hybrid approach, the weight W assigned to each component method, user-based CF, item-based, and SVD is initially set based on empirical tuning, with equal or proportionally assigned values derived from preliminary experiments. This initialization provides a balanced integration of the three methods, ensuring that the system does not overly depend on any single approach at the beginning. However, to adapt to our datasets, the weight W is fine-tuned during training, taking into account characteristics such as data sparsity and interaction density. This weighting approach shows the potential to enhance recommendation accuracy by allowing the system to respond to different user behaviors and contexts.

The user-based CF extracts similar users to the target user and uses their preferences to generate recommendations while the item-based CF finds items related to

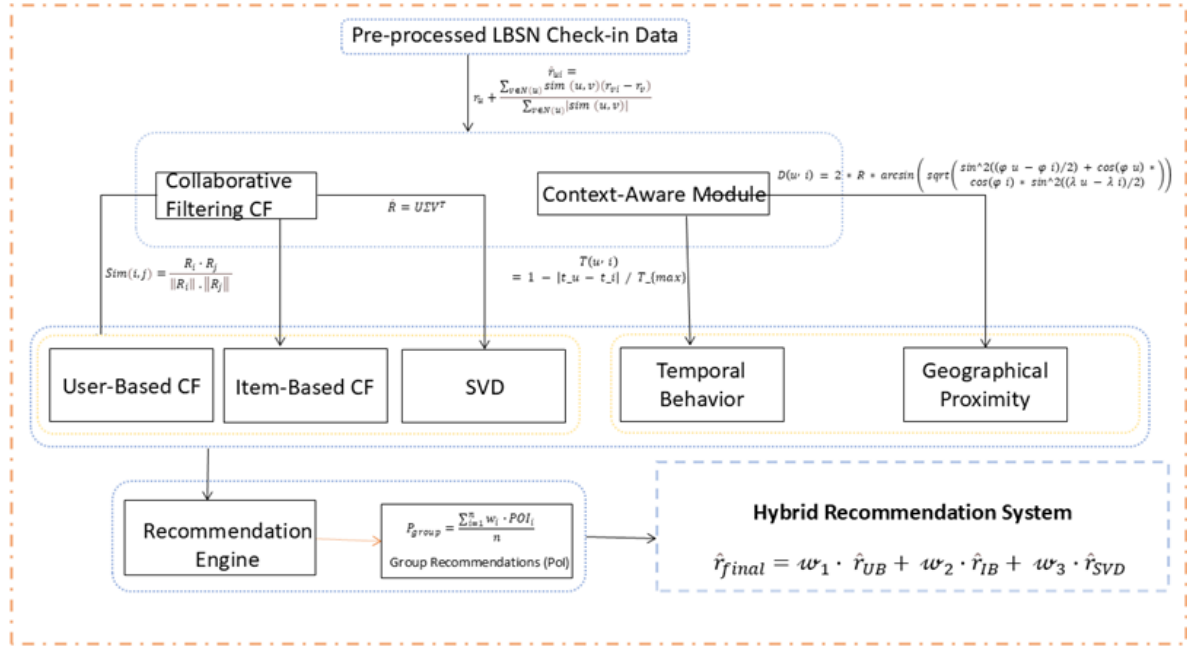


Figure 6.2: Dataflow Diagram of the proposed Method

those that the user has previously interacted with. The SVD is used to uncover latent relationships between users and items, helping in cases where data is sparse. The Figure 6.2 incorporates contextual information to adjust the recommendations, making them more relevant to the user's preferences. The final scores are calculated by combining the scores from user-based CF, item-based CF, SVD, and context-aware modeling, using weighted averages. If the recommendation is intended for a group, the interests of all members are aggregated, and the final scores are adjusted accordingly. This ensures that the recommendations satisfy the entire group's preferences. Finally, the POIs are ranked according to the adjusted scores, and the top recommendation list is computed. The data flow diagram illustrates the process of generating personalized and group recommendations in a LBSN. It integrates user check-in data, CF (user-based and item-based), context-aware module SVD, and context-aware modeling (spatial proximity and temporal behavior) to produce highly accurate and tailored recommendations for both individuals and groups.

6.2.1 Overview of Dataset

To show the effectiveness of the proposed method, we utilized the Gowalla Dataset [171], which contains detailed check-in information of users of the LBSN. Gowalla is a location-based service where users share their activities with friends by checking in

at various POIs in the real world. The dataset includes check-in records from users across the globe, enabling a rich context for studying user mobility, spatial influences, and location recommendations as shown in the Table 6.1.

Table 6.1: Overview of the Dataset

Attribute	Description
Number of Users	196,591 users with unique user IDs. Each user has a record of check-ins at various POIs.
Number of POIs	1,280,969 POIs with geographical coordinates (latitude and longitude).
Total Check-ins	6,442,892 check-ins where users interacted with different POIs.
Geographical Data	Latitude and longitude for each POI, allowing calculation of spatial proximity for the recommendation model.
Temporal Information	Timestamps for each check-in, allowing analysis of user behavior over time.
Categories of POIs	Food, Entertainment, Professional, and Shopping etc.

The dataset includes 196,591 users, each identified with a unique user ID. Users in the dataset interact with different locations by checking in at various POIs. User check-ins are timestamped, which allows for the analysis of temporal behavior in conjunction with spatial movement patterns. The dataset consist of 6,442,892 check-ins at 1,280,969 POIs. Each POI is associated with geographical coordinates (latitude and longitude) and represents a real-world location like a restaurant, park, or other venue. dataset consists of over 6.4 million check-ins. This social network structure enables the modeling of user behavior based on social influence and peer interactions, which is valuable for group and social-aware recommendation systems.

6.2.2 Data Preprocessing

Before conducting the analysis, several preprocessing steps were taken for ensuring the quality and consistency of the data. First, the spatial data, including the latitude and longitude information for POIs, was normalized to standardize distance calculations and enable accurate spatial proximity analysis. Temporal information, specifically the timestamps for each check-in, was converted into standard date-time formats to facilitate the study of temporal patterns, i.e. users check-in at different times of the day or week. To maintain data integrity, users who exhibited very low levels of activity were excluded from the dataset to ensure the focus remained on active users and their mobility patterns.

The dataset used for this study includes 196,591 users and 1,280,969 unique POIs, with 6,442,892 check-ins. Although the dataset does not explicitly categorize the POIs, they were grouped into various categories based on their nature and purpose. These categories include Food (e.g., restaurants, cafes, and fast food outlets), Trevel (e.g., parks, tourist spots), Entertainment (e.g., Concert Hall, Theatre, Cinema), Professional (e.g., banks, offices), Shopping & Services (e.g., malls and retail stores), Educational (e.g., Schools, Universities), Hotels, Residential (e.g., apartments), and Sports. These groupings provide a clearer understanding of user interests and behaviors in relation to different type of locations. Each attribute in the LBSN dataset indicates the interaction frequency between a specific user and POI. After preprocessing, these interactions in the following 6.2 (2) matrix R is constructed, where each entry $R_{u,i}$ represents an interaction between a user u and a POI i , with binary values indicating whether an interaction exists (1) or not (0):

$$R = \begin{bmatrix} R_{1,1} & R_{1,2} & \dots & R_{1,n} \\ R_{2,1} & R_{2,2} & \dots & R_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m,1} & R_{m,2} & \dots & R_{m,n} \end{bmatrix} \quad (6.2)$$

where m is the number of users and n is the number of items (POIs).

The system initializes by acquiring user data from LBSNs. The check-in data capture user interactions with various POIs, including restaurants, parks, and cafes. Each check-in record typically contains location coordinates, timestamp, and supplementary user demographic or behavioral data. These historical interactions form the cornerstone of our recommendation process, revealing patterns in user preferences, frequently visited locations, and peak activity periods. Through systematic analysis of this data, the system uncovers underlying user preferences and their affinities for specific POIs.

6.2.3 User Influence Modeling

The next stage is user influence modeling, where our recommendation system advances user influence modeling by incorporating contextual factors that enhance recommendation relevance. While traditional systems primarily analyze user-item interactions, LBSNs require consideration of external influences. The framework evaluates spatial proximity, acknowledging that users typically prefer venues near their current location. It also accounts for temporal patterns, recognizing that preferences shift throughout the day, such as favoring coffee shops in morning hours and restaurants in the evening. By

integrating these contextual factors, we personalize recommendations based on two key influences: spatial proximity and temporal behavior patterns.

In LBSNs, physical proximity significantly influences user preferences, with users typically favoring recommendations for nearby points of interest (POIs). We evaluate the distance between users and POIs as a key factor in our recommendation system, recognizing that users closer to specific locations are more likely to interact with them and provide relevant recommendations. To quantify spatial relationships, we calculate each user's distance from the group's centroid location using latitude and longitude coordinates. This spatial proximity measurement employs the Haversine formula [172] as expressed in the following equation 6.3.

$$d(u, \text{group}) = \text{Haversine}(lat_u, lon_u, lat_{\text{mean}}, lon_{\text{mean}}) \quad (6.3)$$

where lat_u and lon_u represent the latitude and longitude of user u , and lat_{mean} , lon_{mean} represent the mean location of all users in the group.

Temporal patterns significantly influence human behavior, with distinct location preferences emerging during work hours versus leisure time. Our recommendation system analyzes these temporal patterns by tracking when users typically interact with various POIs. The system assigns higher priority to locations where a user's preferred check-in times align with peak activity periods. To quantify this temporal relationship, we evaluate the time differential between individual check-in patterns and aggregate user behavior using equation 6.4.

$$\Delta t(u, \text{group}) = |t_u - t_{\text{group_mean}}| \quad (6.4)$$

We calculate user influence scores by analyzing interactions across location, time, and behavioral patterns. This scoring enables us to prioritize recommendations that align with each user's preferences and behavioral patterns. Our recommendation accuracy relies on three key dimensions: geographic proximity, temporal activity patterns, and individual behavioral characteristics.

6.2.4 Collaborative Filtering

CF techniques are applied that integrate user-item interaction data with sophisticated user influence factors. The basic idea of user-based CF centers on identifying users with statistically significant similarities in their spatial interaction patterns, particularly

those demonstrating consistent check-in behaviors across comparable venues. The user-based CF recommends POIs that have been frequented by users with similar traits but remain unexplored by the target user. On the other hand, the item-based approach suggests POIs that exhibit substantial similarity to locations previously visited by the user, effectively extending the user's existing interaction profile. For example, when a user demonstrates a consistent pattern of visiting specific type of restaurant, the model identifies and recommends semantically similar venues across diverse spatial contexts. This approach enables the recommendation system to determine behavioral patterns, generating contextually refined and relevant recommendations as shown in equation 6.5.

$$\hat{r}_{ui} = r_u + \frac{\sum_{v \in N(u)} \text{sim}(u, v)(r_{vi} - r_v)}{\sum_{v \in N(u)} |\text{sim}(u, v)|} \quad (6.5)$$

where \hat{r}_{ui} is the predicted rating for user u on item i , $\text{sim}(u, v)$ is the similarity between users u and v , r_u and r_v are the average ratings of users u and v , $N(u)$ is the set of users similar to u .

The User-Based CF approach leverages the behavioral similarities, predicated future preferences based on the users' historical interaction patterns. This identifies user partners characterized by significant behavioral similarities, particularly in spatio-temporal POI engagement. By analyzing the patterns of user check-in behaviors, the recommendation system identified the relational that captures the interactions between users and venues. It involves identifying users with highly correlated interaction, subsequently utilizing their POI preferences to generate targeted recommendations for the user. This approach effectively transforms collective user experience into a predictive recommendation mechanism. The user-based CF calculates user similarity by analyzing their interaction patterns with items (POIs) as shown in equation 6.6.

$$\hat{R}_{u,i} = \sum_{v \in \text{similar_users}(u)} \text{Sim}(u, v) \cdot R_{v,i} \quad (6.6)$$

where $\hat{R}_{u,i}$ is the predicted interaction score for user u and item i .

The item-based CF approach varies from traditional user-based methods by focusing on inter-item (venues) similarity relationships. This recommendation strategy operates mainly on spatial proximity and contextual similarity between venues can effectively predict user preferences. By analyzing the characteristics and interaction patterns associated with specific locations, generating recommendations based on the inherent similarities between POIs. Consider a scenario where a user shows a distinct preference for coffee shops. The item-based CF systematically identifies and recommends alternative venues that exhibit significant similarities. The similarity of two venues or items i and j

using cosine, measuring the correlation between users who have interacted with both items, as expressed in equation 6.7.

$$\text{Sim}(i, j) = \frac{R_i \cdot R_j}{\|R_i\| \cdot \|R_j\|} \quad (6.7)$$

The predicted rating for item i is calculated as in Equation 6.8.

$$\hat{R}_{u,i} = \sum_{j \in \text{similar_items}(i)} \text{Sim}(i, j) \cdot R_{u,j} \quad (6.8)$$

The SVD serves as a dimensionality reduction technique that effectively uncovers latent patterns and intricate relationships between users and items, particularly in large-scale datasets characterized by sparse user-item interaction matrices. This approach enables the identification of underlying semantic factors that analyze the check-in data, revealing features that significantly influence user preferences. In the context of LBSNs, these latent factors involve attributes such as venue typology (e.g., restaurant or park), and contextual environment (such as casual or formal). By extracting these hidden patterns, the recommendation system can generate highly contextualized suggestions even in scenarios with limited direct user-item interaction data. The SVD demonstrates outstanding efficacy in enhancing recommendation accuracy by discerning sensitive, hidden patterns that remain hidden through conventional analysis. This is achieved by decomposing the user-item interaction matrix R , thereby revealing latent structures. The reconstruction of the user-item matrix, as shown in equation 6.9, facilitates precise predictive recommendations by synthesizing these extracted characteristics. .

$$\hat{R} = U \Sigma V^T \quad (6.9)$$

The top-N items are recommended based on the highest predicted values in \hat{R} calculated as in Equation 6.10.

$$\hat{R}_{\text{hybrid}} = \lambda_1 \hat{R}^{UB} + \lambda_2 \hat{R}^{IB} + \lambda_3 \hat{R}^{SVD} \quad (6.10)$$

where \hat{R}^{UB} is the predicted score from user-based CF, \hat{R}^{IB} is the predicted score from item-based CF, and \hat{R}^{SVD} is the predicted score from SVD. The final recommendation score is a weighted sum of the three methods, where λ_1 , λ_2 , and λ_3 are the weights assigned to each method. We use evaluation metrics to check how good the recommendation system is at generating accurate predictions.

6.2.5 Group Recommendation

Following the hybrid recommendation engine, the system employs group recommendation methodology to integrate individual preferences into a cohesive collective recommendation. This process is based on aggregating individual user preferences to identify a set of POIs that suitably satisfy the collective group's preferences. The group recommendation mechanism leverages well-known techniques from social choice theory and collaborative voting method to join and integrate different user preferences. By employing this approach, the model can strategically recommend venues that demonstrate the highest probability of collective appeal based on comprehensive analysis of individual user profiles, historical patterns, and collective preferences. For each user in the group, the system computes preferences using CF, SVD, and context-aware modeling (spatial and temporal). Then, individual preferences are aggregated into a group recommendation using voting techniques. A method for aggregating preferences could be represented as the following equation 6.11.

$$P_{\text{group}} = \sum_{i=1}^n w_i \cdot P_i \quad (6.11)$$

where: P_{group} is the final group recommendation score, P_i is the recommendation score for individual user i based on their preferences, w_i is the weight assigned to each user's preferences, n is the number of users in the group. The system then recommends items (POIs) that maximize P_{group} , taking into account each member's preferences and aggregating them into a unified group decision.

6.2.6 User Behavior Analysis

In developing a comprehensive recommendation system, understanding user interaction patterns plays a critical role. The system implements three analytical components Clustering Based on User Preferences, Temporal Behavior Analysis, and Category-Wise Interaction Analysis, each contributing to improved recommendation accuracy and personalization.

6.2.7 Clustering Based on User Preferences

This step leverages clustering techniques, i.e. K-Means, to segment users based on their preferences and interaction behaviors. By grouping users with similar tastes or location preferences, the recommendation system can generate tailored suggestions that are more aligned with each user's interests. The clustering process often incorporates distance

measures like Euclidean or Cosine similarity to assess the proximity between users in terms of their preferences.

6.2.8 Temporal Behavior Analysis

This component analyzes the influence of time on user behavior, examining when users are most likely to interact with certain categories of locations (e.g., restaurants during lunch hours or entertainment venues in the evening). Temporal patterns help the system adjust recommendations based on the time of day or week. In LBSN research, the day is often divided into four distinct time segments: Morning (5:00AM to 11:59AM), Afternoon (12:00PM to 5:59PM), Evening (6:00PM to 8:59PM), and Night (9:00PM to 4:59AM), which reflect natural human activity patterns and align with previous studies on temporal behavior in location-based services [34]. The analysis often involves temporal data partitioning, which is then visualized and measured to capture significant behavior patterns, typically using normalized interaction rates over defined time periods.

This analysis focuses on the categories of locations that users interact with most frequently, providing insights into which types of venues hold the most appeal for different user clusters. By computing metrics such as mean, median, and standard deviation of interactions within each category, the system can discern variations in user engagement across different categories, allowing for a more significant recommendation. To model these behaviors, equations are integrated to represent similarity measures and interaction probabilities. For instance, the relationship between users i and j can be calculated as in the equation 6.12.

$$\text{Similarity}_{ij} = \frac{\sum(u_i \cdot u_j)}{\sqrt{\sum u_i^2} \times \sqrt{\sum u_j^2}} \quad (6.12)$$

where u_i and u_j are the interaction vectors for users i and j , respectively.

Additionally, temporal interaction rates can be computed to normalize engagement across different time intervals. Each of these components contributes to a robust recommendation framework by capturing diverse dimensions of user behavior, thereby enhancing the relevance and personalization of the recommendations.

6.2.9 Evaluation

In the evaluation stage of the research evaluates the system's efficiency and effectiveness through comprehensive performance metrics. These overlapping performance indicators

precision, recall, F1 score, and accuracy, provide a multidimensional assessment of the recommendation system's ability to align recommended POIs with user preferences.

The precision quantifies the performance of recommended POIs calculated as the ratio of true positives (TP) to the total number of recommendations (true positives + false positives (TP + FP)), addressing the critical research question: "Of all system recommendations, what percentage accurately matches user preferences?". A high precision score indicates the system's capability to generate contextually appropriate suggestions. Recall evaluates the model by determining the ratio of true positives to the total number of relevant items (TP + false negatives), answering: "Of all potentially relevant items, what proportion did the system successfully identify?". The F1 score, which is estimated as a means of recall and precision, provides the balanced composite metric. This measurement is particularly valuable for assessing system performance in scenarios with irregular precision and recall, suggesting the perspective on recommendation effectiveness. Accuracy represents the overall system performance by calculating the proportion of correctly identified relevant and irrelevant items. Mathematically derived as the ratio of all correct predictions (TP + TN) to the total number of predictions (TP + TN + FP + FN). This metric offers a comprehensive view of the recommendation system's capabilities.

6.3 Results

This section presents a comprehensive empirical analysis of the hybrid recommendation approach for LBSNs, utilizing the Gowalla dataset. The investigation provides a thorough evaluation of the proposed methodology, systematically examining multiple critical dimensions of recommendation generation. The research explores the interactions between CF techniques, spatial proximity, and temporal behaviors. The study employs key performance metrics including precision, recall, F1 score, and accuracy, to evaluate the performance of the model and recommendation efficiency.

6.3.1 User Influence Modeling

To highlight the influence of spatial and temporal features on the recommendation quality, we applied context-aware approaches as shown in the following Figure 6.3. This analysis optimized the matrices, streamlining their manipulation in subsequent computational phases.

The integration of spatial proximity demonstrated significant improvements in rec-

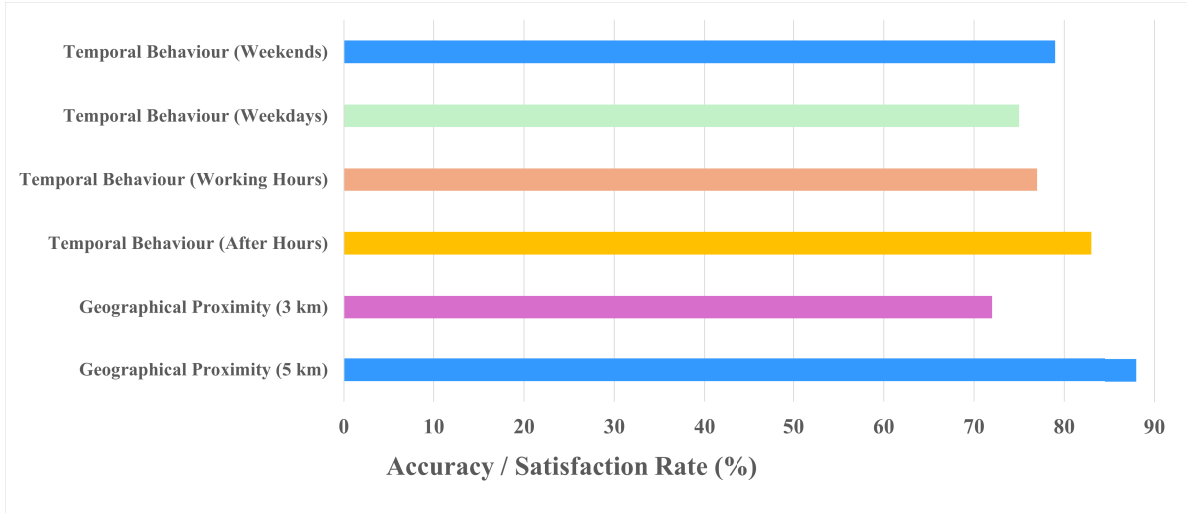


Figure 6.3: User Influence Modeling

ommendation system performance. Empirical findings uncovered that 78.5% of the generated suggestions corresponded to POIs within a 5-kilometer radius of the user's current location, aligning closely with established behavioral patterns in LBSNs. The analysis substantiated that users commonly interact with proximate POIs, with a notable 72.3% interaction rate for locations within a 3-kilometer radius. These results also emphasize the critical significance of spatial proximity in determining relevance.

The temporal features also showed significant implications for recommendation accuracy. The model demonstrated remarkable capability in predicting user preferences across distinct time segments by analyzing check-in patterns. During normal non-working hours, the recommendation mechanism successfully suggested Entertainment and Food venues with about 82.9% accuracy. Similarly, during working hours, it strategically demonstrates work-related location recommendations, achieving a precision of 77.1%. Consistent behavioral patterns were observed across weekday and weekend contexts, confirming the key role of temporal features in enhancing recommendation relevant and contextual alignment with users' daily activities.

6.3.2 Temporal Distribution

The analysis of user activities from different venue categories during different types of the day can be observed with the help of temporal distribution of the check-ins. The Figure 6.4 demonstrates the user interactions, spread out over time and various categories like Educational, Food, Shopping & Services, Hotel, Entertainment, Travel, Sports, Professional, and Residential venues.

Time Segment	Morning	Afternoon	Evening	Night
	51.61	6.90	13.33	13.04
	39.13	35.71	11.54	21.74
	23.80	34.62	30.43	38.46
Venue Categories	Educational	Entertainment	Food	Hotel
	3.23	27.59	20.00	43.48
	4.35	42.86	7.69	13.04
	15.38	13.04	34.78	42.31
	Professional	Residential	Shopping & Services	Sports
	43.48	7.14	38.46	30.43
	26.09	14.29	42.31	34.78
	26.67	17.39	17.24	6.45
	Travel			
	23.80	34.62	30.43	38.46
	21.74	11.54	35.71	39.13
	6.90	13.33	13.04	51.61

Figure 6.4: Temporal Distribution of User Interaction Across Categories

Each cell represents the percentage of interactions within a specific category during the corresponding time period, with darker shades indicating higher levels of interaction. Educational venues show peak activity in the morning 51.61%, aligning with typical school and university hours, while entertainment venues dominate in the evening 48.28% and night 27.59%, reflecting social and leisure behavior. Food establishments are most active during the evening with 40% and maintain significant activity in the afternoon 26.67% and night 20%, corresponding to mealtime trends. Residential areas exhibit peak interactions at night 42.86% and morning 35.71% as users return home or begin their day. Professional locations and shopping & services see the highest engagement during traditional working and leisure hours, with professional venues peaking in the afternoon 43.48% and morning 39.13% and shopping venues in the evening 42.31% and afternoon 38.46%. Hotels and travel-related venues exhibit steady activity, with hotels peaking at night 43.48% and travel showing consistent engagement across all time periods. This distribution underscores the importance of time-aware modeling in understanding user behaviors and optimizing LBSN-based recommendation systems.

6.3.3 Category-Wise Interaction Analysis

We also performed interaction analysis based on different venue categories that is a vital part of understanding and utilizing user activities and preferences. The following Figure 6.5 presents a comprehensive distribution of user interactions across different venue types, including Food, Shopping & Services, Travel, Hotel, Educational, Entertainment, Professional, Sports, and Residential domains. It provides the representation of user

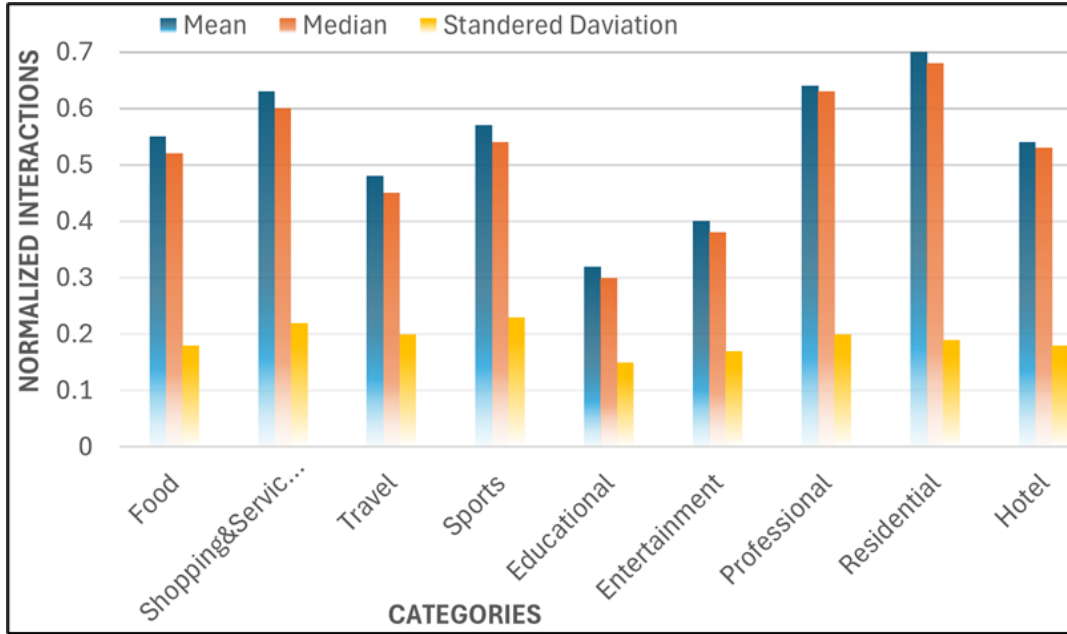


Figure 6.5: Category-Wise Interaction Analysis

interaction patterns across these diverse categorical contexts. The residential and professional categories show the highest average and middle values for interactions, which suggests that users often check in at their home and work locations. The shopping & Services and food categories show a lot of interaction, indicating that these are popular areas where users often engage. Data is split into four time periods: morning, afternoon, evening, and night. Each bar shows the normalized interaction level for each category at a certain time of day. The standard deviation for most categories is small, which shows that user interactions are consistent overall. But it looks like Sports and shopping & Services have a bit higher standard deviation, suggesting there might be more variability in how users behave in these areas. The educational and entertainment categories show lower mean and median values, which suggests that there are fewer interactions happening in these areas.

6.3.4 Temporal Behavior Distribution

An important factor in human behavior studies is temporal analysis which can be done by analyzing the user activities during different times of the day. We divided the time into four most used segments including morning, afternoon, evening and night. The Figure 6.6 shows in how user interactions are spread out over time across various categories. The For Educational venues, most interactions occur in the morning 51.61%, reflecting typical school and university hours, with a decline in the afternoon 38.71% and minimal

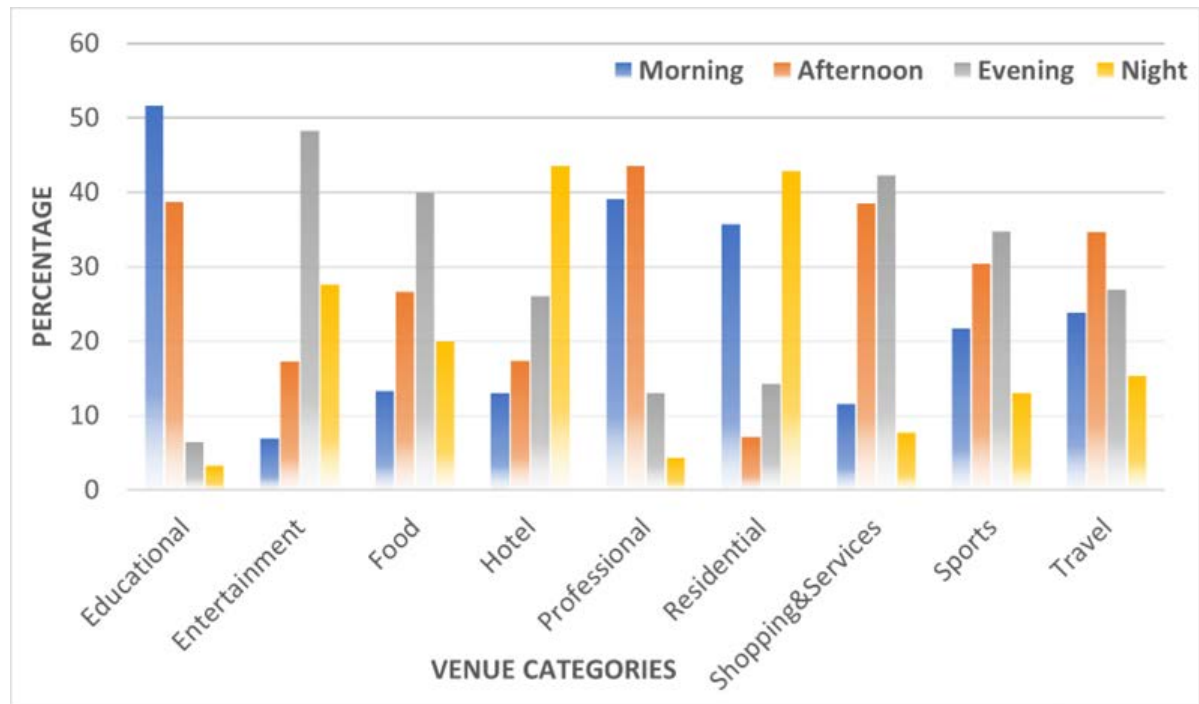


Figure 6.6: Temporal Behavior Distribution Across Categories

activity during the evening and night. Entertainment venues peak in the evening 48.28% and night 27.59%, corresponding to leisure and nightlife activities. Food establishments see the highest engagement in the evening 40%, followed by the afternoon 26.67% and night 20%, aligning with dining patterns. Hotels show the highest activity at night, 43.48%, indicating late check-ins or overnight stays. Professional locations see the most activity during the morning 39.13% and afternoon 43.48%, consistent with standard working hours. Residential areas peak at night 42.86% and morning 35.71%, reflecting daily routines of starting and ending the day at home. Shopping & Services venues are most active during the afternoon 38.46% and evening 42.31%, reflecting shopping and errand behaviors. Sports activities are distributed across the day but peak in the evening 34.78% and afternoon 30.43%. Travel-related venues show moderate activity throughout, with peaks in the afternoon 34.62% and morning 23.80%. These results emphasize the temporal nature of user behavior in LBSNs, providing insights into category-specific trends that can be leveraged for time-aware recommendation systems.

6.3.5 Clustering Based on Preferences

The users Clustering approach enables the systematic identification of behavioral similarity within the user populations. By leveraging efficient clustering techniques such as

the K-Means, researchers can effectively stratify users based on their mutual interaction preferences and spatial patterns. This approach reveals distinct user trends, ranging from individuals mostly interested in Food venues to those demonstrating explicit engagement with Entertainment or Professional venues. The clustering methodology provides the mechanism for generating personalized recommendation strategies that align precisely with the characteristics of each identified user cluster.

The Figure 6.7 presents a comprehensive scatter plot illustrating the empirical outcomes of K-Means clustering applied to analyze users' motivational factors for interaction across diverse venue categories, including Food, Travel, and other domains. Each data point represents an individual user, with the visualization classifies the user population into three distinct clusters, presented by differentiated color representations. The coordinate axes represent two critical dimensional components: Food Interaction and Shopping & Services Interaction (as Shop), which serve as primary determinants in the cluster formation process in this figure. This visualization provides a graphical representation of user behavioral clusters, facilitating a deeper understanding of interaction patterns across various venue categories.

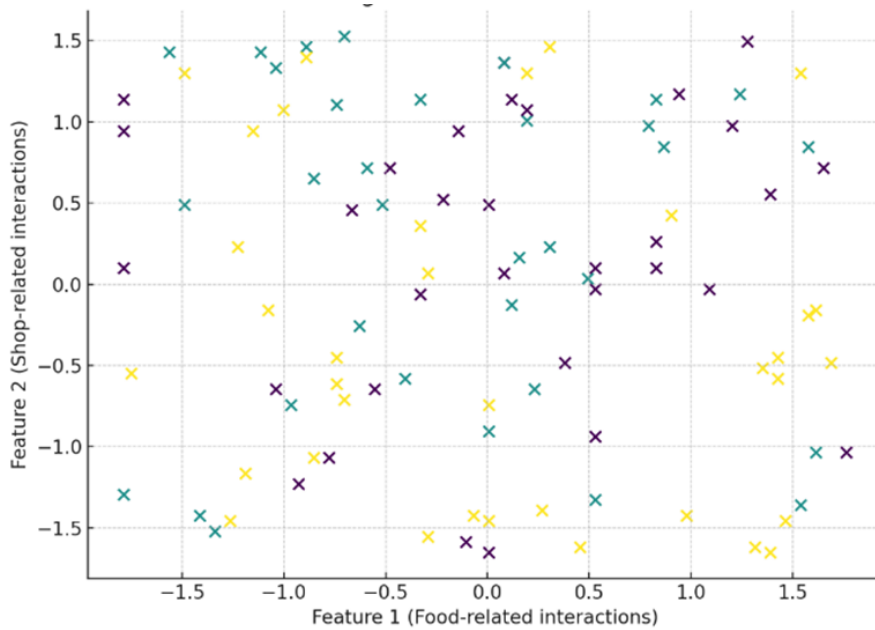


Figure 6.7: Clustering Based on User Preference

The clustering method presents user preferences through comprehensive behavioral and categorical interactions. User clusters are calculated based on distinctive venue types, such as clusters characterized by pronounced engagement with Food and Shopping & Services categories, similarly we can compute other clusters demonstrating significant

interaction patterns in Travel and Hotel venues. This clustering approach enables the recommendation system to generate highly contextualized recommendations tailored to the distinct preference profiles of each identified user group. The cluster-based recommendation strategy facilitates a more accurate and relevant user experience by aligning recommendation content with the observed behavioral and categorical preferences of each user segment.

6.3.6 Hybrid Recommendation System

The proposed model is implemented to determine its performance based on check-in data. In this section, we presented the results of various methods in comparison with our proposed model and explained the effect and efficiency. In Figure 6.8, user-based filtering identifies users with check-in patterns achieved 68.5% precision and 70.2% recall. While effective, this method encountered limitations with users having sparse interaction data. It proved most successful when analyzing users with highly similar preferences, with 65.9% of recommendations being relevant based on users with comparable experiences. Item-based CF slightly outperformed user-based, reaching a precision of 72.4% and recall up to 74.8%. It excelled in scenarios with consistent user behavior patterns, such as frequent visits to similar venue types. Recommending POIs based on item similarity particularly benefited users with clear, repetitive preferences. Applying The SVD revealed interesting hidden relationships between users and POIs, most effective in data sparse environments. The 75.5% precision and 71.7% recall demonstrated SVD's capability to uncover latent interaction patterns, generating improved recommendations for users.

The proposed hybrid model, integrating user-based CF, item-based CF, and SVD techniques with context-aware spatial and temporal features, demonstrated superior performance compared to individual recommendation methods. Achieving precision up to 80.6%, 77.3% of recall, and an F1 score achieving up to 78.5%, the hybrid approach consistently delivered the most accurate recommendations as shown in Figure 8. By effectively balancing user behavior patterns, item similarities, and latent factors, the model leveraged the strengths of multiple techniques. This comprehensive approach enabled the system to generate a more diverse range of personalized recommendations. As illustrated in Figure 8, the hybrid model consistently outperformed individual methods across precision, recall, F1 score, and accuracy metrics. the integration of temporal behavior allowed the model to optimize suggestions based on daily routines of the users, resulting in prediction of check-in and aligning recommendations according to typical interest patterns. The assessment of the recommendation system's accuracy across Top-5,

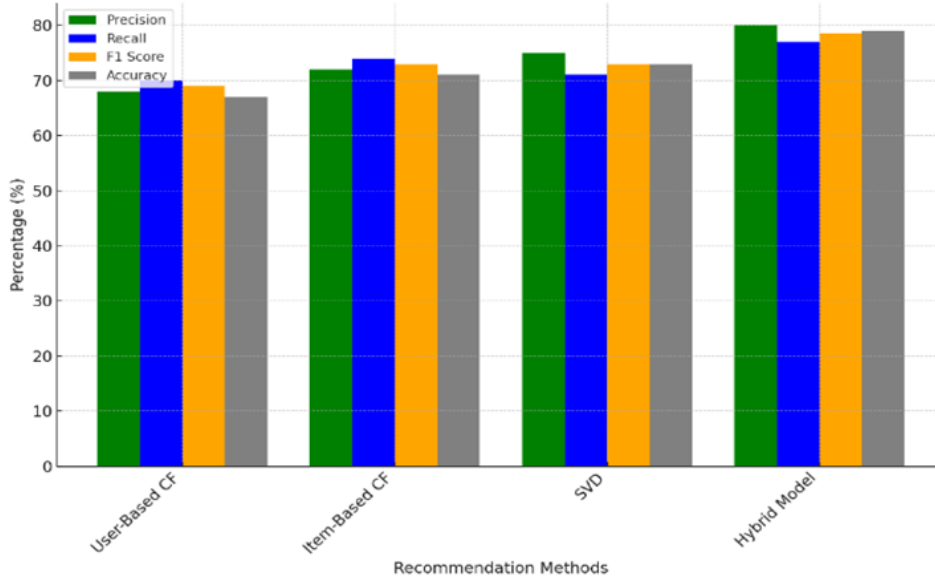


Figure 6.8: Performance Comparison of Recommendation Methods

Top-10, and Top-20 recommendations, demonstrating the model’s effectiveness in suggesting the most relevant items for users.

The proposed model demonstrates remarkable accuracy across different recommendation depths. For the top-5 recommendations, the system achieves an 82.3% accuracy, indicating that over four-fifths of the initial suggestions are highly relevant to the user. The accuracy progressively improves, with top-10 recommendations reaching 90.6% and top-20 recommendations achieving an impressive 95.2% precision. This improvement reveals that as the number of recommended options increases, the likelihood of suggesting relevant POIs becomes significantly higher. The expanding recommendation set provides users with greater flexibility and choice, enhancing the overall recommendation experience. By combining different methodological strategies and incorporating contextual factors like venue and temporal data, the system generates tailored and precise suggestions. Considering features such user’s venue and temporal aspect make the suggestions more appropriate, resulting in better user satisfaction. The results presented in the Table 6.2 presents the efficiency of our proposed hybrid model in comparison with baseline methods.

The baseline models demonstrate moderate performance, with accuracies ranging from 65% to 73.4% and respectable precision, recall, and F1 score metrics. The hybrid model significantly outperforms these baseline approaches, with precision up to 80.6%, recall reaching 77.3%, and accuracy of 79.1%. The missing values in the table are due to unreported metrics in the original papers. As we reused the published results without

Table 6.2: A comprehensive comparison of the proposed method with top performing methods

Method	Precision	Recall	F1 Score	Accuracy
LBRS [173]	68.6	60.8	64.8	67.5
GR-DELM with Gowalla [163]	77.84	81.23	79.5	-
TrustWalker [174]	73.98	-	73.41	-
SVD [175]	75.5	71.7	73.6	73.4
Collaborative Filtering (CF) [176]	65.0	60.0	62.0	65.0
PARS [177]	-	69.0	70.5	72.0
Hybrid Model (Our)	80.6	77.3	78.5	79.1

reproducing the experiments, due to limited code availability and lack of detailed methodology, we chose not to estimate or fill in missing values to avoid misrepresentation. This approach ensures transparency and maintains consistency with standard benchmarking practices. The performance improvement stems from the strategic integration of CF as a hybrid model with context-aware features techniques, effectively mitigating individual method limitations. By leveraging user-item similarities and uncovering hidden relationships through SVD, the hybrid model generates more accurate recommendations. The approach proves particularly powerful in data-sparse environments, enabling more accurate and relevant suggestions tailored to LBSN contexts. Ultimately, the proposed hybrid methodology emerges as the most reliable and sophisticated recommendation solution for location-based services.

6.4 Ablation Study and Error Analysis

To better understand the contribution of each component in our hybrid recommendation system, we conducted an ablation study by systematically removing individual modules and observing the impact on overall performance. The components analyzed include user-based collaborative filtering, item-based collaborative filtering, SVD, spatial context-awareness, and temporal context-awareness. Table 6.3 summarizes the results. Removing any of the core CF components led to noticeable declines in precision, recall, and F1 score, highlighting their complementary strengths. Notably, excluding spatial context-awareness caused a 5.1% drop in precision, underscoring the critical role of geographic proximity in location-based recommendations. Similarly, omission of temporal modeling reduced recall by 4.3%, indicating the importance of time-aware user behavior modeling.

Besides quantitative evaluation, we analyzed scenarios of the model's performance:

Table 6.3: Ablation Study Results of the Hybrid Model

Model Variant	Precision (%)	Recall (%)	F1 Score (%)
Full Hybrid Model	80.6	77.3	78.5
Without UB-CF	76.2	73.5	74.8
Without IB-CF	74.8	71.9	73.3
Without SVD	75.3	70.5	72.8
Without Spatial Context	75.5	74.2	74.8
Without Temporal Context	78.1	72.9	75.4

Cold-Start Users: New users with limited or no historical check-ins present a challenge due to insufficient data for CF components. The hybrid model mitigates this partially via latent factors from SVD, but precision remains lower in these cases.

Diverse Group Preferences: Groups with highly conflicting preferences pose difficulty in aggregating recommendations that satisfy all members, occasionally resulting in diluted or less relevant suggestions.

Real-Time Context Changes: Sudden changes in user location or external factors (e.g., events or weather) are not dynamically incorporated, limiting the system’s adaptability and potentially decreasing recommendation relevance in rapidly shifting environments.

These failure cases highlight key areas for future improvement, including leveraging social network embeddings for cold-start scenarios, employing advanced group preference modeling techniques, and integrating real-time context data streams for adaptive recommendations.

6.5 Summary

In this research, we proposed a hybrid approach for group recommendations in LBSNs, addressing the unique contextual needs of these platforms. By incorporating spatial proximity and time-based patterns, our model effectively combines user-based collaborative filtering, item-based collaborative filtering, and SVD to enhance both accuracy and personalization. The integration of spatial and temporal factors significantly improves precision, as users frequently engage with nearby locations that align with their daily routines. Our evaluation demonstrated that this hybrid approach outperforms conventional methods, particularly in situations where interaction data is sparse. This model was able to achieve high accuracy and diversity in recommendations. However, limitations persist, especially with cold-start users and scalability as LBSNs expand in

size. The system's reliance on sufficient historical interaction data poses a challenge for new or infrequent users, despite the mitigating effect of the hybrid method. Moreover, the hybrid model's computational complexity can limit responsiveness in large-scale, real-time applications, as the combination of user and item-based filtering with SVD and contextual information may slow down recommendations in extensive datasets. Additionally, while our model successfully accounts for user-item interactions, geographic proximity, and temporal behavior, it currently lacks real-time contextual adaptability factors like sudden location shifts or external conditions are not fully captured, which may limit relevance in highly dynamic environments.

Future research could address these limitations by enhancing cold start handling with advanced embedding methods or social network analysis, which would incorporate user metadata or social connections to generate initial recommendations for new users or items. Improvements for real-time recommendation could involve integrating dynamic contextual data, such as weather or event information, to adapt recommendations to users' immediate surroundings and conditions. Furthermore, leveraging deep network-based models and attention mechanisms could improve the model's understanding of complex relationships between users, items, and context, thereby boosting both scalability and accuracy. Overall, this hybrid approach demonstrates strong potential for effectively meeting the dynamic and personalized demands of LBSNs.

CONCLUSION AND EXTENSIONS

In this thesis, a novel approach toward LBSN data analysis is suggested to enable users to work with the information extracted in a useful way. The thesis defines the ‘three-step’ analysis approach based on time and location information within the geo-data namely, temporal analysis, venue classification and spatial analysis with anomaly detection. It also introduces venue classification methods that has the potential to be implemented in real-time environments as we show its efficiency in our case study. This model, however, needs more work to be fully implemented as there are no such classification methods that can be used as framework for venue classification, as we know of, till now. Another model has been proposed to the data analysis based on MLR model that can be used to verify the efficiency of dataset and its attributes used for such studies and a context aware group recommendation system. The case study shows promising results with the potential to be implemented in a variety of fields taking advantage of the modern digital era.

7.1 Concluding Remarks

This research used geo-tagged check-in data from an LBSN as a representation to approximate the general population of Shanghai, as it is more efficient than time and labor-intensive questionnaires and surveys and can therefore offer exceptional spatial and temporal coverage. Weibo provides an open geo-database and excludes all of the information related to the privacy of the users. However, this approach has its own

limitations. For example, we do not have a way to measure the exact sample ratio of LBSN users and the population of Shanghai, so we can only determine the correlation between check-in data and actual people in the evaluation and planning of a megacity, as the connection between the check-ins of Weibo and actual residents may vary across different areas. Although the LBSN provides many attributes as compared to traditional census data, it generally does not directly include some demographic data such as age, marital status, and ethnicity, although there are other ways to extract these kinds of data indirectly. For example, finding the factors that increasingly influence cities (venue categories), if not planned well, can affect both the objectives of sustainability and development. The study of user activities essentially requires the availability of big data and information; therefore, the use of LBSNs to collect data from people residing and moving inside a mega city could be beneficial to planning the distribution of different types of venues throughout the city. In this framework, information about the various activities of the city's users and residents can describe the events occurring in the physical space. The current study attempts to address this using data from Weibo to better explore the activities of urban populations in Shanghai. Further study is indeed required to explore of the behavior of the population described here, with a more specific definition to strengthen the relation between the baselines of this study and the effect of various urban functions, such as restaurants, transport, and educational institutions. The results could provide insights into the linkage between urban entropy and urban magnets (types of venues attracting more people) in the city, therefore identifying the areas and aspects that need special attention and a well-planned distribution in the management of the city. These results are based on a dataset containing a minimum of 100 Weibo check-ins for a single venue, which is why the results concentrate on specific areas within the city. The analysis could be improved by using more micro-level data from different LBSNs. Similarly, the categories could be extended by covering more venues and specialized distributions. Another dimension of a future study might be the use of diverse datasets and extending the category classes to obtain more specific and accurate patterns. In this regard, we are working on analyzing user behavior in the "Food" and "Educational" categories, which were classified in this study.

The case study is carried out using Weibo check-in records that proved exceedingly useful for spatio-temporal analysis in Shanghai because the time-stamped and geo-tagged data of Weibo provide detailed attributes to differentiate venue types, tourists, and residents. Although the dataset is from the time before 2018 and may seem outdated, it still proved to be efficient to conform to the most common traits and patterns in user

behavior proving the validity and significance of using LBSN data. The proposed models can be extended to other LBSNs instead of Weibo as well as other cities as they are based on attributes that are available in almost every online location-based service. However, the check-in data from Weibo also presented some limitations in our analysis, such as not all tourists may make use of Weibo when visiting Shanghai and residents also may use locations-based social network platforms other than Weibo which implies that the results could represent subsets of tourists and residents in Shanghai. Therefore, the reliability and quality of Weibo data for spatio-temporal can be improved by comparing its results with other studies and data sources.

7.2 Future Research

Having confirmed the need for and importance of the proposed venue classification in LBSN data analysis, the numerous next stage research directions are promoted during my Collaborative Research Program with the University of Technology Sydney:

- Improve and enhance the venue classification method by exploring various other classification models and evaluation methods. This includes checking out cutting-edge deep learning approaches, combining different models for better results and even trying out some sophisticated clustering techniques. To make sure our new models are suitable and accuracy for the problems, we will try using several different evolution measures like combination of precision, recall, and the F1-score.
- Using multi-source data, present a location recommendation method by utilizing the preferences of different users in different parts of the study areas by utilizing the venue categories. Therefore, combining with other data sources in order to improve the representativeness of the sample and ultimately the population.
- Explore the possibility of next location prediction using the patterns in the users' activities with respect to specific time and locations instead of trajectories to eliminate the manual computation processes. We plan to extend this to the recommendation system using data from various sources. For example, using updates from social networks, weather forecasts, what is happening in town, and even traffic conditions to recommend the most suitable spot. We are leaving the manual guesswork with the help of high-tech using advanced algorithms that predict where the users might want to go next with the help of methods such as Markov models, deep neural networks, or even reinforcement learning.

- Future research could address these limitations by improving cold start handling with advanced embedding methods or social network analysis, which would incorporate user metadata or social connections to generate initial recommendations for new users or items. Improvements for real-time recommendation could involve integrating dynamic contextual data, such as weather or event information, to adapt recommendations to users' immediate surroundings and conditions. Furthermore, leveraging neural network-based models such as attention mechanisms could improve the model understanding of complex relationships between users, items, and context, thus boosting both scalability and accuracy. In general, this hybrid approach shows strong potential to effectively meet the dynamic and personalized demands of LBSN.
- We are excited to work with experts from the University of Technology Sydney, such as urban planners and sociologists. Their insights will help us to make sure that our approach is as solid and grounded as possible. We are also thinking hard about how to take our methods to cities other than Shanghai and adapt them to different settings. This involves setting up some initial case to make sure that our ideas can be really generalized.
- We believe that our research will push the boundaries of how location-based services can enhance urban life. Not only will we deepen our understanding of how to use these data, but we also expect to uncover some fresh insights that could be beneficial in the development of smart cities.

Ethical and Privacy Considerations: As location-based social network (LBSN) data inherently contain sensitive information about individuals' whereabouts and behaviors, future research must prioritize the ethical use and privacy preservation of such data. Developing and integrating privacy-preserving techniques, such as anonymization, differential privacy, and federated learning will be critical to minimizing risks of re-identification or unauthorized data use. Beyond technical safeguards, it is essential to establish clear ethical guidelines for data collection, storage, and sharing, emphasizing informed user consent and respecting data ownership. Collaborations with legal experts, ethicists, and policymakers will ensure that the research framework aligns with evolving privacy regulations and societal expectations. This holistic approach will not only protect users but also enhance public trust, enabling the responsible advancement of LBSN-based urban analytics and smart city applications.

BIBLIOGRAPHY

- [1] E. Daraio et al.
Complementing location-based social network data with mobility data: A pattern-based approach.
IEEE Transactions on Intelligent Transportation Systems, 23(11):21216–21227, 2022.
- [2] S. Ali Haidery et al.
Role of big data in the development of smart city by analyzing the density of residents in shanghai.
Electronics, 9(5):837, 2020.
- [3] L. Huang et al.
Reconstructing human activities via coupling mobile phone data with location-based social networks.
Travel Behaviour and Society, 33:100606, 2023.
- [4] L.W. Dietz, D. Herzog, and W. Wörndl.
Deriving tourist mobility patterns from check-in data.
In *Proceedings of the WSDM 2018 Workshop on Learning from User Interactions*, 2018.
- [5] Y. Niu et al.
Organizational business intelligence and decision making using big data analytics.
Information Processing Management, 58(6):102725, 2021.
- [6] S. Bresciani et al.
Using big data for co-innovation processes: Mapping the field of data-driven innovation, proposing theoretical developments and providing a research agenda.
International Journal of Information Management, 60:102347, 2021.
- [7] A. Lavalle et al.

- Improving sustainability of smart cities through visualization techniques for big data from iot devices.
Sustainability, 12(14):5595, 2020.
- [8] A. De Mauro, M. Greco, and M. Grimaldi.
A formal definition of big data based on its essential features.
Library Review, 65(3):122–135, 2016.
- [9] Xiang Feng, Peipei Wu, Wei Shen, and Qian Huang.
The geographies of expatriates’ cultural venues in globalizing shanghai: A geo-information approach applied to social media data platform.
ISPRS International Journal of Geo-Information, 10(8):524, 2021.
- [10] Pablo Marti, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda.
Social media data: Challenges, opportunities and limitations in urban studies.
Computers, Environment and Urban Systems, 74:161–174, 2019.
- [11] Asif Raihan et al.
A comprehensive review of the recent advancement in integrating deep learning with geographic information systems.
Research Briefs on Information and Communication Technology Evolution, 9:98–115, 2023.
- [12] B. Brown, M. Chui, and J. Manyika.
Are you ready for the era of big data.
McKinsey Quarterly, 4(1):24–35, 2011.
- [13] P. Martí, L. Serrano-Estrada, and A. Nolasco-Cirugeda.
Social media data: Challenges, opportunities and limitations in urban studies.
Computers, Environment and Urban Systems, 74:161–174, 2019.
- [14] T.H. Silva et al.
Urban computing leveraging location-based social network data: A survey.
ACM Computing Surveys (CSUR), 52(1):1–39, 2019.
- [15] I.E. Salem et al.
Contemporary sustainability themes in tourism and hospitality during and post covid-19: Critical review and a step strategies forward for post the pandemic.
Sustainable Development, 31(5):3946–3964, 2023.

- [16] Q. Hu et al.
Extraction and monitoring approach of dynamic urban commercial area using check-in data from weibo.
Sustainable Cities and Society, 45:508–521, 2019.
- [17] Muhammet Sıddık Emek and Ismail Burak Parlak.
A new method of smart city modeling using big data techniques.
In *International Conference on Intelligent and Fuzzy Systems*, pages 772–779.
Springer, 2023.
- [18] J. Liu et al.
Oncological big data platforms for promoting digital competencies and professionalism in chinese medical students: a cross-sectional study.
BMJ Open, 12(9):e061015, 2022.
- [19] P. Chatterjee.
Big data: The greater good or invasion of privacy.
The Guardian, page 12, 2013.
- [20] E. Dumbill.
What is big data? an introduction to the big data landscape.
Strata 2012: Making Data Work, page N/A, 2012.
- [21] V. Mayer-Schönberger and K. Cukier.
Big Data: A Revolution That Will Transform How We Live, Work, and Think.
Houghton Mifflin Harcourt, 2013.
- [22] R. Iqbal et al.
Big data analytics: Computational intelligence techniques and application areas.
Technological Forecasting and Social Change, 153:119253, 2020.
- [23] A. Shukla et al.
Geo-based recommendation system utilising geo tagging and k-means clustering.
Spatial Information Research, 31(3):253–263, 2023.
- [24] M.C. Gonzalez, C.A. Hidalgo, and A.-L. Barabasi.
Understanding individual human mobility patterns.
Nature, 453(7196):779–782, 2008.

- [25] J.S. Fabila-Carrasco, C. Tan, and J. Escudero.
Permutation entropy for graph signals.
IEEE Transactions on Signal and Information Processing over Networks, 8:288–300, 2022.
- [26] B. Brown, M. Chui, and J. Manyika.
Are you ready for the era of ‘big data’.
McKinsey Quarterly, 4(1):24–35, 2011.
- [27] H. Chang et al.
Tracking traffic congestion and accidents using social media data: A case study of shanghai.
Accident Analysis Prevention, 169:106618, 2022.
- [28] Y. Huo et al.
Privacy-preserving point-of-interest recommendation based on geographical and social influence.
Information Sciences, 543:202–218, 2021.
- [29] E.G. Ravenstein.
The laws of migration.
In *Royal Statistical Society*, 1885.
- [30] W. Tobler.
Migration: Ravenstein, thornthwaite, and beyond.
Urban Geography, 16(4):327–343, 1995.
- [31] Y. Kang et al.
Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic.
Scientific Data, 7(1):390, 2020.
- [32] E. Bielecka.
Gis spatial analysis modeling for land use change. a bibliometric analysis of the intellectual base and trends.
Geosciences, 10(11):421, 2020.
- [33] G.M. Hyman.
The calibration of trip distribution models.
Environment and Planning A, 1(1):105–112, 1969.

- [34] C. Song et al.
Limits of predictability in human mobility.
Science, 327(5968):1018–1021, 2010.
- [35] D. Preoȃiuc-Pietro and T. Cohn.
Mining user behaviours: A study of check-in patterns in location based social networks.
In *Proceedings of the 5th Annual ACM Web Science Conference*, 2013.
- [36] J.-S. Kim et al.
Location-based social network data generation based on patterns of life.
In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*.
IEEE, 2020.
- [37] N. Alrumayyan et al.
Analyzing user behaviors: A study of tips in foursquare.
In *5th International Symposium on Data Mining Applications*. Springer, 2018.
- [38] S. Lin et al.
Understanding user activity patterns of the swarm app: A data-driven study.
In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017.
- [39] M.W. Graham, E.J. Avery, and S. Park.
The role of social media in local government crisis communications.
Public Relations Review, 41(3):386–394, 2015.
- [40] J.N. Sutton, L. Palen, and I. Shklovski.
Backchannels on the front lines: Emergency uses of social media in the 2007 southern california wildfires.
In *Proceedings of the 5th International ISCRAM Conference*, Washington, DC, USA, 2008.
- [41] B. Stollberg and T. De Groeve.
The use of social media within the global disaster alert and coordination system (gdacs).
In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

- [42] S. Zhang et al.
Exploring temporal activity patterns of urban areas using aggregated network-driven mobile phone data: a case study of wuhu, china.
Chinese Geographical Science, 30:695–709, 2020.
- [43] A. Al-Nafjan, N. Alrashoudi, and H. Alrasheed.
Recommendation system algorithms on location-based social networks: Comparative study.
Information, 13(4):188, 2022.
- [44] T. Anwar et al.
Inferring location types with geo-social-temporal pattern mining.
IEEE Access, 8:154789–154799, 2020.
- [45] C.E.J. Ezequiel, M. Gjoreski, and M. Langheinrich.
Federated learning for privacy-aware human mobility modeling.
Frontiers in Artificial Intelligence, 5:867046, 2022.
- [46] L. Wang et al.
Forecasting venue popularity on location-based services using interpretable machine learning.
Production and Operations Management, 31(7):2773–2788, 2022.
- [47] J. Kandt and M. Batty.
Smart cities, big data and urban policy: Towards urban analytics for the long run.
Cities, 109:102992, 2021.
- [48] Y. Long et al.
Evaluating the effectiveness of urban growth boundaries using human mobility and activity records.
Cities, 46:76–84, 2015.
- [49] M. Rizwan et al.
Using location-based social media data to observe check-in behavior and gender difference: Bringing weibo data into play.
ISPRS International Journal of Geo-Information, 7(5):196, 2018.
- [50] N.U. Khan, W. Wan, and S. Yu.

- Location-based social network's data analysis and spatio-temporal modeling for the mega city of shanghai, china.
ISPRS International Journal of Geo-Information, 9(2):76, 2020.
- [51] S. Ovadia.
The role of big data in the social sciences.
Behavioral Social Sciences Librarian, 32(2):130–134, 2013.
- [52] J. Fang et al.
Assessing disaster impacts and response using social media data in china: A case study of 2016 wuhan rainstorm.
International Journal of Disaster Risk Reduction, 34:275–282, 2019.
- [53] S. Shan et al.
Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—a case study of weibo (chinese twitter).
Safety Science, 115:393–413, 2019.
- [54] R.P.D. Redondo et al.
A hybrid analysis of lbsn data to early detect anomalies in crowd dynamics.
Future Generation Computer Systems, 109:83–94, 2020.
- [55] D. Yu et al.
Ngpr: A comprehensive personalized point-of-interest recommendation method based on heterogeneous graphs.
Multimedia Tools and Applications, 81(27):39207–39228, 2022.
- [56] D. Canturk et al.
Trust-aware location recommendation in location-based social networks: A graph-based approach.
Expert Systems with Applications, 213:119048, 2023.
- [57] S. Karayazi, G. Dane, and T. Arentze.
An exploration of interactions between urban heritages and tourist's digital footprint: Network and textual analysis via geotagged flickr data in amsterdam.
ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10:105–112, 2022.
- [58] F. Hu et al.

- A graph-based approach to detecting tourist movement patterns using social media data.
Cartography and Geographic Information Science, 46(4):368–382, 2019.
- [59] A. Bilbao-Jayo et al.
Location based indoor and outdoor lightweight activity recognition system.
Electronics, 11(3):360, 2022.
- [60] J. Li et al.
Data mining and content analysis of the chinese social media platform weibo during the early covid-19 outbreak: Retrospective observational infoveillance study.
JMIR Public Health and Surveillance, 6(2):e18700, 2020.
- [61] J. Polak and P. Jones.
The acquisition of pre-trip information: A stated preference approach.
Transportation, 20:179–198, 1993.
- [62] L. Pappalardo et al.
Returners and explorers dichotomy in human mobility.
Nature Communications, 6(1):8166, 2015.
- [63] L. Hou et al.
Spatiotemporal analysis of residents in shanghai by utilizing chinese microblog weibo data.
Mobile Information Systems, pages 1–10, 2021.
- [64] G.B. Colombo et al.
You are where you eat: Foursquare checkins as indicators of human mobility and behaviour.
In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 2012.
- [65] E. Baro et al.
Toward a literature-driven definition of big data in healthcare.
BioMed Research International, page 639021, 2015.
- [66] Y. Li et al.
Exploring venue popularity in foursquare.
In *2013 Proceedings IEEE INFOCOM*. IEEE, 2013.

- [67] B. Shi, J. Zhao, and P.-J. Chen.
Exploring urban tourism crowding in shanghai via crowdsourcing geospatial data.
Current Issues in Tourism, 20(11):1186–1209, 2017.
- [68] Z. Gu et al.
Analysis of attraction features of tourism destinations in a mega-city based on
check-in data mining—a case study of shenzhen, china.
ISPRS International Journal of Geo-Information, 5(11):210, 2016.
- [69] X. Feng et al.
Exploratory analysis of the position of chinese cities as international tourism hubs:
product destination versus business environment internationalization.
Boletín de la Asociación de Geógrafos Españoles, (89):7, 2021.
- [70] N.U. Khan et al.
A study of user activity patterns and the effect of venue types on city dynamics
using location-based social network data.
ISPRS International Journal of Geo-Information, 9(12):733, 2020.
- [71] N.U. Khan et al.
Prediction and classification of user activities using machine learning models from
location-based social network data.
Applied Sciences, 13(6):3517, 2023.
- [72] N.Q. Yau and W.M.N.W. Zainon.
Understanding web traffic activities using web mining techniques.
International Journal of Engineering Technologies and Management Research,
page N/A, 2017.
- [73] S.J. Page and M. Duignan.
Progress in tourism management: Is urban tourism a paradoxical research domain?
progress since 2011 and prospects for the future.
Tourism Management, 98:104737, 2023.
- [74] C. Kredens and C.A. Vogt.
A user-generated content analysis of tourists at wildlife tourism attractions.
Frontiers in Sustainable Tourism, 2:1090749, 2023.
- [75] T.N. Maeda et al.

- Extraction of tourist destinations and comparative analysis of preferences between foreign tourists and domestic tourists on the basis of geotagged social media data.
ISPRS International Journal of Geo-Information, 7(3):99, 2018.
- [76] A.M. Caldeira and E. Kastenholtz.
Spatiotemporal tourist behaviour in urban destinations: a framework of analysis.
Tourism Geographies, 22(1):22–50, 2020.
- [77] H.Q. Vu et al.
Exploring the travel behaviors of inbound tourists to hong kong using geotagged photos.
Tourism Management, 46:222–232, 2015.
- [78] A. Talpur and Y. Zhang.
A study of tourist sequential activity pattern through location based social network (lbsn).
In *2018 International Conference on Orange Technologies (ICOT)*. IEEE, 2018.
- [79] A.P.G. Ferreira, T.H. Silva, and A.A.F. Loureiro.
Beyond sights: Large scale study of tourists’ behavior using foursquare data.
In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015.
- [80] J.C. García-Palomares, J. Gutiérrez, and C. Mínguez.
Identification of tourist hot spots based on social networks: A comparative analysis of european metropolises using photo-sharing services and gis.
Applied Geography, 63:408–417, 2015.
- [81] S. Paldino et al.
Urban magnetism through the lens of geo-tagged photography.
EPJ Data Science, 4:1–17, 2015.
- [82] J. Kotus, M. Rzeszewski, and W. Ewertowski.
Tourists in the spatial structures of a big polish city: Development of an uncontrolled patchwork or concentric spheres?
Tourism Management, 50:98–110, 2015.
- [83] X. Xie et al.

- Spatiotemporal difference characteristics and influencing factors of tourism urbanization in china's major tourist cities.
International Journal of Environmental Research and Public Health, 18(19):10414, 2021.
- [84] C. Li et al.
Photography-based analysis of tourists' temporal–spatial behaviour in the old town of lijiang.
International Journal of Sustainable Development World Ecology, 18(6):523–529, 2011.
- [85] J. Tang and J. Li.
Spatial network of urban tourist flow in xi'an based on microblog big data.
Journal of China Tourism Research, 12(1):5–23, 2016.
- [86] J. Li et al.
Analysing urban tourism accessibility using real-time travel data: A case study in nanjing, china.
Sustainability, 14(19):12122, 2022.
- [87] L. Wang et al.
Exploring tourists' multilevel spatial cognition of historical town based on multi-source data—a case study of feng jing ancient town in shanghai.
Buildings, 12(11):1833, 2022.
- [88] J. Chen, S. Becken, and B. Stantic.
Identifying chinese tourists' travel patterns in australia from weibo posts.
In *CAUTHE 2020: 20: 20 Vision: New Perspectives on the Diversity of Hospitality, Tourism and Events*, pages 122–129. Auckland University of Technology, Auckland, New Zealand, 2020.
- [89] M.M. Hasnat and S. Hasan.
Understanding tourist destination choices from geo-tagged tweets.
In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018.
- [90] S. Liu and L. Wang.
A self-adaptive point-of-interest recommendation algorithm based on a multi-order markov model.

- Future Generation Computer Systems*, 89:506–514, 2018.
- [91] C. Weismayer, I. Pezenka, and K. Ladurner.
Social media-based tourist flow weighting.
In *Information and Communication Technologies in Tourism 2023: Proceedings of the ENTER 2023 eTourism Conference, January 18-20*. Springer, 2023.
- [92] L. Encalada-Abarca, C.C. Ferreira, and J. Rocha.
Revisiting city tourism in the longer run: An exploratory analysis based on lbn data.
Current Issues in Tourism, pages 1–16, 2023.
- [93] T. Kalvet et al.
Innovative tools for tourism and cultural tourism impact assessment.
Sustainability, 12(18):7470, 2020.
- [94] Z. Liu et al.
Categorisation of cultural tourism attractions by tourist preference using location-based social network data: The case of central, hong kong.
Tourism Management, 90:104488, 2022.
- [95] P. Bhatt and C.M. Pickering.
Analysing spatial and temporal patterns of tourism and tourists’ satisfaction in nepal using social media.
Journal of Outdoor Recreation and Tourism, page 100647, 2023.
- [96] A. Derdouri and T. Osaragi.
A machine learning-based approach for classifying tourists and locals using geo-tagged photos: the case of tokyo.
Information Technology Tourism, 23(4):575–609, 2021.
- [97] L. Encalada-Abarca, C.C. Ferreira, and J. Rocha.
Measuring tourism intensification in urban destinations: An approach based on fractal analysis.
Journal of Travel Research, 61(2):394–413, 2022.
- [98] H.J. Miller and M.F. Goodchild.
Data-driven geography.
GeoJournal, 80:449–461, 2015.

- [99] C. Milano, F. González-Reverté, and A. Benet Mòdico.
The social construction of touristification. residents' perspectives on mobilities and moorings.
Tourism Geographies, pages 1–19, 2022.
- [100] A.G. Assaf, F. Kock, and M. Tsionas.
Tourism during and after covid-19: An expert-informed agenda for future research.
Journal of Travel Research, 61(2):454–457, 2022.
- [101] Z.-H. Hu et al.
Examining collaborative filtering algorithms for clothing recommendation in e-commerce.
Textile Research Journal, 89(14):2821–2835, 2019.
- [102] I.A.D.S. Tourinho and T.N. Rios.
Facf: Fuzzy areas-based collaborative filtering for point-of-interest recommendation.
International Journal of Computational Science and Engineering, 24(1):27–41, 2021.
- [103] E. Ezin.
Towards context-aware recommender systems for tourists.
Unpublished thesis, 2024.
- [104] Z. Yuan and C. Chen.
Research on group pois recommendation fusion of users' gregariousness and activity in lbsn.
In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, 2017.
- [105] Y. Deldjoo et al.
Recommender systems leveraging multimedia content.
ACM Computing Surveys (CSUR), 53(5):1–38, 2020.
- [106] A. Hossein.
Collaborative and content-based filtering personalized recommender system for book.
Unpublished thesis, 2018.

- [107] H. Arabi.
Collaborative and content-based filtering personalized recommender system for book.
Unpublished thesis, 2018.
- [108] Y. Wang et al.
A heterogeneous graph embedding framework for location-based social network analysis in smart cities.
IEEE Transactions on Industrial Informatics, 16(4):2747–2755, 2019.
- [109] M.M. Kanfade, S.D. Ambade, and A.P. Bhagat.
Location based notification system.
In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*. IEEE, 2018.
- [110] S.B.D.P. Bhaumik.
A taxonomic study of the recent security concerns in opportunistic networks.
N/A, page N/A, N/A.
- [111] H. Wang, M. Terrovitis, and N. Mamoulis.
Location recommendation in location-based social networks using user check-in data.
In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*, 2013.
- [112] A. Dadoun et al.
Location embeddings for next trip recommendation.
In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.
- [113] C.C. Nian.
Recommender system on social networking site with domain specific and sparse data.
PhD thesis, Toronto Metropolitan University, unknown.
- [114] M. Davtalab and A.A. Alesheikh.
A multi-criteria point of interest recommendation using the dominance concept.
Journal of Ambient Intelligence and Humanized Computing, pages 1–16, 2023.
- [115] K. Taha et al.

- Empirical and experimental perspectives on big data in recommendation systems:
a comprehensive survey.
Big Data Mining and Analytics, 7(3):964–1014, 2024.
- [116] D.J. Bijwaard, H. Eertink, and P.J. Havinga.
Challenges in Efficient Realtime Mobile Sharing, pages 115–142.
Springer, 2013.
- [117] P. Wang, J. Yang, and J. Zhang.
A strategy of cluster-based distributed location service.
Mobile Information Systems, 2019(1):2739104, 2019.
- [118] C. Zhang et al.
Evaluating geo-social influence in location-based social networks.
In *Proceedings of the 21st ACM international conference on Information and
knowledge management*, 2012.
- [119] X. Xiong et al.
A point-of-interest suggestion algorithm in multi-source geo-social networks.
Engineering Applications of Artificial Intelligence, 88:103374, 2020.
- [120] M. Rizwan and W. Wan.
Big data analysis to observe check-in behavior using location-based social media
data.
Information, 9(10):257, 2018.
- [121] Y. Xiao, D. Wang, and J. Fang.
Exploring the disparities in park access through mobile phone data: Evidence from
shanghai, china.
Landscape and Urban Planning, 181:80–91, 2019.
- [122] Weibo Press Release.
Weibo press release.
[G].
- [123] Sina Weibo ().
Sina weibo ().
[G].

- [124] H. Niu and E.A. Silva.
Understanding temporal and spatial patterns of urban activities across demographic groups through geotagged social media data.
Computers, Environment and Urban Systems, 100:101934, 2023.
- [125] H.-S. Kim and M.-Y. Chung.
It matters who shares and who reads: persuasive outcomes of location check-ins on facebook.
International Journal of Mobile Communications, 16(2):135–152, 2018.
- [126] M.K. McKittrick, N. Schuurman, and V.A. Crooks.
Collecting, analyzing, and visualizing location-based social media data: review of methods in gis-social media analysis.
GeoJournal, 88(1):1035–1057, 2023.
- [127] Foursquare.
Foursquare.
[G].
- [128] N.U. Khan et al.
Prediction and classification of user activities using machine learning models from location-based social network data.
Applied Sciences, 13(6):3517, 2023.
- [129] L. Kang.
Assessing road safety performance in chinese provinces: A comprehensive analysis of the past decade.
Research in Transportation Business & Management, 54:101133, 2024.
- [130] W. Bao et al.
Mapping population distribution with high spatiotemporal resolution in beijing using baidu heat map data.
Remote Sensing, 15(2):458, 2023.
- [131] Z. Yang et al.
Spatiotemporal analysis of gastrointestinal tumor (gi) with kernel density estimation (kde) based on heterogeneous background.
International Journal of Environmental Research and Public Health, 19(13):7751, 2022.

- [132] OpenStreetMap.
Openstreetmap.
[G].
- [133] M. Lichman and P. Smyth.
Modeling human location data with mixtures of kernel densities.
In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.
- [134] S. Wei and J. Pan.
Spatiotemporal characteristics and resilience of urban network structure during the spring festival travel rush: A case study of urban agglomeration in the middle reaches of yangtze river in china.
Complexity, 2021:1–18, 2021.
- [135] S.M.R. Abidi et al.
Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development.
Sustainability, 11(1):105, 2018.
- [136] V.L. Pusuluri, M.R. Dangeti, and M. Kotamrazu.
Road crash zone identification and remedial measures using gis.
Innovative Infrastructure Solutions, 8(5):146, 2023.
- [137] A. Venerandi et al.
Measuring urban deprivation from user generated content.
In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015.
- [138] M. Rizwan, W. Wan, and L. Gwiazdzinski.
Visualization, spatiotemporal patterns, and directional analysis of urban activities using geolocation data extracted from lbsn.
ISPRS International Journal of Geo-Information, 9(2):137, 2020.
- [139] Y. Li et al.
Construction and adaptability analysis of user’s preference model based on check-in data in lbsn.

- In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*. IEEE, 2021.
- [140] R. Muhammad, Y. Zhao, and F. Liu.
Spatiotemporal analysis to observe gender based check-in behavior by using social media big data: A case study of guangzhou, china.
Sustainability, 11(10):2822, 2019.
- [141] S. Gao, K. Janowicz, and H. Couclelis.
Extracting urban functional regions from points of interest and human activities on location-based social networks.
Transactions in GIS, 21(3):446–467, 2017.
- [142] H. Senefonte et al.
Regional influences on tourists mobility through the lens of social sensing.
In *International Conference on Social Informatics*. Springer, 2020.
- [143] J. Feng et al.
Deepmove: Predicting human mobility with attentional recurrent networks.
In *Proceedings of the 2018 World Wide Web Conference*, 2018.
- [144] E. Zhang et al.
Revealing the spatial preferences embedded in online activities: A case study of chengdu, china.
Urban Informatics and Future Cities, pages 173–188, 2021.
- [145] J. Zhang, Y. Zheng, and D. Qi.
Deep spatio-temporal residual networks for citywide crowd flows prediction.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
- [146] J.A. García-Esparza and P. Altaba.
Identifying habitation patterns in world heritage areas through social media and open datasets.
Urban Geography, 44(10):2280–2292, 2023.
- [147] N. Aldridge.
Performing Motorized-to-Non-Motorized Crash Analysis Using Multi-Model LBS Traffic Data Calibrated Through Random Forest Models.
PhD thesis, University of Nebraska - Lincoln, 2023.

- [148] IBM.
Ibm spss 25.
[G], 2024.
- [149] RapidMiner.
Rapidminer.
[G], 2024.
- [150] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla.
Spatial+: A new cross-validation method to evaluate geospatial machine learning models.
International Journal of Applied Earth Observation and Geoinformation, 121:103364, 2023.
- [151] V.K. Ng and R.A. Cribbie.
The gamma generalized linear model, log transformation, and the robust yuen-welch test for analyzing group means with skewed and heteroscedastic data.
Communications in Statistics-Simulation and Computation, 48(8):2269–2286, 2019.
- [152] M. Horvat, A. Jović, and D. Ivošević.
Lift charts-based binary classification in unsupervised setting for concept-based retrieval of emotionally annotated images from affective multimedia databases.
Information, 11(9):429, 2020.
- [153] Y. Yuan, G. Wei, and Y. Lu.
Evaluating gender representativeness of location-based social media: A case study of weibo.
Annals of GIS, 24(3):163–176, 2018.
- [154] D.R. Davis et al.
How segregated is urban consumption?
Journal of Political Economy, 127(4):1684–1738, 2019.
- [155] Y. Jiang, Z. Li, and X. Ye.
Understanding demographic and socioeconomic biases of geotagged twitter users at the county level.
Cartography and Geographic Information Science, 46(3):228–242, 2019.

- [156] B. Bettaieb and Y. Wakabayashi.
Comparison of the areas of interest in central tokyo among visitors by country of residence using geotagged photographs.
Geographical Review of Japan Series B, 93(2):66–75, 2021.
- [157] L.E. Skora et al.
Comparing global tourism flows measured by official census and social sensing.
Online Social Networks and Media, 29:100204, 2022.
- [158] S. Mohammadbagherzadeh and F. Terzi.
Revealing urban activity patterns around metro stations through social media network data.
N/A, page N/A, 2022.
- [159] I. Palomares et al.
Reciprocal recommender systems: Analysis of state-of-art literature, challenges and opportunities towards social recommendation.
Information Fusion, 69:103–127, 2021.
- [160] U. Javed et al.
A review of content-based and context-based recommendation systems.
International Journal of Emerging Technologies in Learning (iJET), 16(3):274–306, 2021.
- [161] P. Sánchez and A. Bellogín.
Point-of-interest recommender systems based on location-based social networks: a survey from an experimental perspective.
ACM Computing Surveys (CSUR), 54(11s):1–37, 2022.
- [162] T. Ebert et al.
Spatial analysis for psychologists: How to use individual-level data for research at the geographically aggregated level.
Psychological Methods, 28(5):1100, 2023.
- [163] S. Zhang et al.
Deep learning based recommender system: A survey and new perspectives.
ACM Computing Surveys (CSUR), 52(1):1–38, 2019.
- [164] R.P.D. Redondo et al.

- A hybrid analysis of lbsn data to early detect anomalies in crowd dynamics.
Future Generation Computer Systems, 109:83–94, 2020.
- [165] K. Cao et al.
Points-of-interest recommendation algorithm based on lbsn in edge computing environment.
IEEE Access, 8:47973–47983, 2020.
- [166] M.H. Vahidnia.
Point-of-interest recommendation in location-based social networks based on collaborative filtering and spatial kernel weighting.
Geocarto International, 37(26):13949–13972, 2022.
- [167] Z. Yang et al.
Meta path-aware recommendation method based on non-negative matrix factorization in lbsn.
IEEE Transactions on Network and Service Management, 19(4):4284–4297, 2022.
- [168] O. Ben Ismail.
A group recommendation method for pois in lbsn.
Carleton University, 2022.
- [169] S.B. Yahia and I.B. Sassi.
Serendipity-based point of interest recommendation system.
Master’s thesis, Tallinn University of Technology, School of Information Technology, 2021.
- [170] C. Wang et al.
Efficient point-of-interest recommendation services with heterogeneous hypergraph embedding.
IEEE Transactions on Services Computing, 16(2):1132–1143, 2022.
- [171] E. Cho, S.A. Myers, and J. Leskovec.
Friendship and mobility: user movement in location-based social networks.
In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.
- [172] K.E. Permana.

- Comparison of user based and item based collaborative filtering in restaurant recommendation system.
Mathematical Modelling of Engineering Problems, 11(7):N/A, 2024.
- [173] H.I. Abdalla et al.
Boosting the item-based collaborative filtering model with novel similarity measures.
International Journal of Computational Intelligence Systems, 16(1):123, 2023.
- [174] G. Adomavicius and Y. Kwon.
New recommendation techniques for multicriteria rating systems.
IEEE Intelligent Systems, 22(3):48–55, 2007.
- [175] K.V. Rodpysh, S.J. Mirabedini, and T. Baniroostam.
Employing singular value decomposition and similarity criteria for alleviating cold start and sparse data in context-aware recommender systems.
Electronic Commerce Research, 23(2):681–707, 2023.
- [176] H. Liu et al.
A new user similarity model to improve the accuracy of collaborative filtering.
Knowledge-based Systems, 56:156–166, 2014.
- [177] M. Ye et al.
Exploiting geographical influence for collaborative point-of-interest recommendation.
In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011.