Journal of Science Engineering Technology and Management Science
Volume 02, Issue 08, August 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI: 10.64771/jsetms.2025.v02.i08.548-555*

A Scalable Hybrid GloVe-Based System for Real-Time News Article Classification

K. Swayam Prabha[1], B. B. Praneeth[2], E. Jaya Prakash[2], Md. Khalid Hasan[2],
L. Rohith[2]
[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (CSIT)
[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, 501510, Telangana.

To Cite this Article
K. Swayam Prabha, B. B. Praneeth, E. Jaya Prakash, Md. Khalid Hasan, L. Rohith, "A Scalable Hybrid GloVe-Based System for Real-Time News Article Classification", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 08, August 2025,pp: 548-555, DOI: http://doi.org/10.64771/jsetms.2025.v02.i08.pp548-555
Submitted: 15-07-2025 *Accepted: 21-08-2025* *Published: 28-08-2025*

ABSTRACT

This research presents an automated news article classification system capable of categorizing content into five categories: business, entertainment, politics, sports, and technology. Manual classification of news articles is often time-consuming, inconsistent, and unable to scale with the rapid growth of digital content. Human annotators may introduce subjective biases, leading to errors and inefficiencies in large datasets. These limitations make manual methods unsuitable for real-time news processing in dynamic environments.To overcome these challenges, the proposed system features a streamlined pipeline integrated into a user-friendly graphical interface. This interface enables users to upload datasets, preprocess the text through cleaning and tokenization, and perform feature extraction using a hybrid Global Vectors for Word Representation (GloVe) approach. This hybrid model combines global word embeddings with context-aware semantic features. The classification task is handled using the Gaussian Naive Bayes (GNB) algorithm.Performance evaluation was conducted by comparing three models. The proposed Hybrid GloVe with GNB model achieved the highest performance, with an accuracy of 91.24%, precision of 90.31%, recall of 92.16%, and an F1-score of 93.19%. In contrast, the existing PCA with GNB model reached an accuracy of 79.10%, and the LDA with GNB model lagged behind with an accuracy of 67.64%. The results demonstrate the superiority of the proposed method in terms of accuracy, processing speed, and consistency, making it a scalable and reliable solution for real-time news classification.

Keywords: News classification, GloVe embeddings, Gaussian Naive Bayes, real-time processing, text preprocessing, feature extraction, NLP, PCA, LDA, automated categorization.

1. INTRODUCTION

This research explores the development of an advanced text classification system that addresses the growing demand for efficient, scalable, and accurate categorization of textual data across diverse domains such as e-commerce, media, and finance. As cities evolve, the identification of functional urban regions using high-resolution remote sensing data and geospatial datasets like Points of Interest (POIs) has become crucial for urban planning. However, these datasets alone cannot capture the socio-economic dimensions influencing urban functions. In parallel, industries such as Amazon, Netflix, Reuters, and JP Morgan require real-time classification of large volumes of unstructured text, including user reviews, news articles, and financial reports. The challenge lies in handling high-dimensional, noisy feature spaces while maintaining semantic richness and computational efficiency. Traditional methods often fall short due to their inability to capture contextual word relationships or reduce dimensionality effectively. This research addresses these issues by proposing a hybrid approach that integrates GloVe-based word embeddings for semantic representation with KMeans clustering for feature selection, followed by classification using the Gaussian Naive Bayes algorithm. This method not only enhances classification accuracy and generalization but also reduces computation time, memory usage, and model complexity. The system demonstrates robust performance across various applications, including content moderation, sentiment analysis, medical text classification, legal document categorization, and customer feedback analysis. By balancing advanced embeddings with intelligent dimensionality reduction, this work contributes significantly to the development of interpretable, real-time, and domain-adaptable NLP systems. As cities continue to grow and evolve, they develop a wide variety of functional regions that support the daily life and economic activities of their residents.

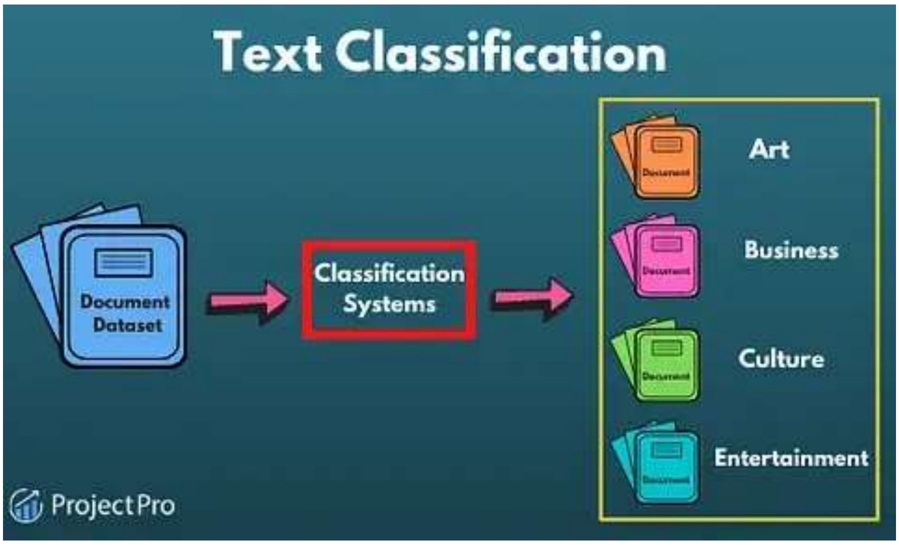


Fig.1: Text Classification.

These functional types are influenced by diverse factors including infrastructure, human mobility, and economic engagement. Monitoring and understanding how these regions change over time holds significant value for urban planning, and accurate identification of these regions has become a critical task. Traditionally, high-resolution remote sensing images have been used to extract urban functions based on physical land surface features such as shape, texture, and spectral signatures.

2. LITERATURE SURVEY

In the field of urban computing, topic model-based methods can be used to mine information hidden in data with socioeconomic attributes. Conventionally, physical characteristics of land surfaces extracted from remote sensing images have been widely used to classify land use type and extract urban functional regions [1,2]. However, the functional types of regions in a city are strongly influenced by the interaction between humans and the surface [3]

[4] Hence, solutions that analogize data with socioeconomic attributes, such as POIs, social media data, and taxi trajectory data, to natural language are essential for mining the latent features of urban spaces.Yuan et al. [5] applied the topic model and natural language processing (NLP) to analogize urban regions to documents, urban functions to topics, POI types to metadata, and human movement patterns to words; in doing so, they identified urban functional regions by combining POI and taxi trajectory data. In cities, however, highly precise dynamic mobility data are difficult to obtain. Gao et al. [6] extracted urban functional regions by analyzing POI co-occurrence patterns after combining social media check-in data with POIs and integrating the data into the latent Dirichlet allocation

(LDA) model. LDA, a topic model in the field of NLP, is typically used to distinguish words by mining the topic of the document and classifying the words into different topics
[7], and to estimate the continuous representation of words in vector space. When applied to large geospatial datasets, LDA has the disadvantages of high computational cost and low efficiency [8]. Moreover, topic models such as LDA often ignore the linear relationship of words in vector space; that is, similar words should be closer to each other in vector space [9]. In addition, according to Tobler's first law of geography, closer geographical elements have stronger correlations.
The contextual information of geospatial data is thus essential to a spatial semantic understanding. However, when using a topic model to extract urban functional regions, a large amount of spatial contextual information is lost [10].

3. PROPOSED METHODOLOGY

The application titled "Enhancing Text Classification through GloVe-Based Word Embedding and Clustering for Feature Selection in Naïve Bayes Classifiers" presents a focused approach to solving the challenges of text classification by integrating advanced feature engineering and machine learning techniques. It uses the GloVe embedding model to transform raw text into semantically rich numerical vectors that preserve the contextual meaning of words. These embeddings are then refined through unsupervised clustering to identify important features, reducing noise and dimensionality. The processed features are passed to a Naïve Bayes classifier, which is known for its simplicity and effectiveness in text-based tasks. The combination of semantic embeddings and clustering-based selection significantly enhances classification accuracy and efficiency compared to traditional approaches.The interface is developed using Python's Tkinter library, enabling users to interact with the classification system through a clean graphical interface. Users can upload a dataset, view its contents, and trigger different stages of preprocessing, embedding, and classification without needing to write code.

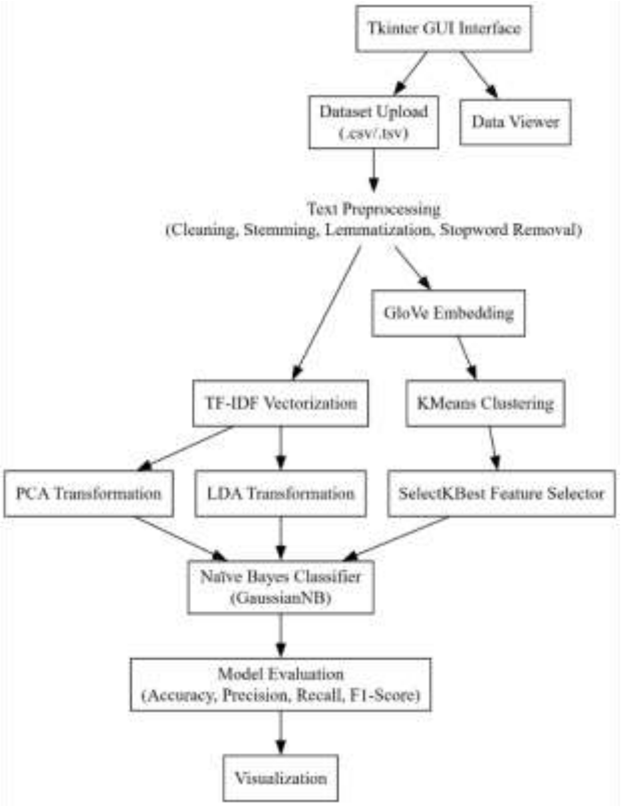

Tkinter GUI Interface
Dataset Upload (.csv/.tsv)
Data Viewer
Text Preprocessing (Cleaning, Stemming, Lemmatization, Stopword Removal)
GloVe Embedding
TF-IDF Vectorization
KMeans Clustering
PCA Transformation
LDA Transformation
SelectKBest Feature Selector
Naïve Bayes Classifier (GaussianNB)
Model Evaluation (Accuracy, Precision, Recall, F1-Score)
Visualization

Fig. 2: Proposed Block Diagram.

The system supports three different transformation strategies: one using TF-IDF with PCA, another using TF-IDF with LDA, and the third, the proposed model, using GloVe with clustering. These approaches can be executed with a single click, and the results are visualized through bar plots and heatmaps. Metrics such as accuracy, precision, recall, and F1-score are computed and displayed, allowing users to compare the effectiveness of each method. Confusion matrices provide insight into classification errors across different categories. The backend pipeline performs a thorough preprocessing routine, including punctuation removal, stemming, lemmatization, and stopword filtering. Once cleaned, the text is vectorized using either a traditional method like TF-IDF or a dense vector representation like GloVe. In the GloVe pipeline, the transformed vectors undergo KMeans clustering to identify patterns, and the most informative features are selected using SelectKBest. This targeted feature selection helps reduce overfitting and improves the classifier's performance. The Gaussian Naïve Bayes classifier is then trained on these optimized features and evaluated against a test split. The system is designed for consistent reproducibility, using saved NumPy arrays and model checkpoints to avoid unnecessary recomputation. Through this structured pipeline, the classifier demonstrates enhanced performance on a range of text classification tasks.

The Gaussian Naive Bayes Classifier (Gaussian NBC) is a probabilistic machine learning algorithm specifically adapted to handle continuous, real-valued features by assuming that each feature follows a Gaussian (normal) distribution. It is particularly effective when applied to data derived from dimensionality reduction techniques like PCA and LDA, or word embeddings like GloVe, where input features are numerical and often high-dimensional. The classification process begins by analyzing the dataset to identify the unique class labels and grouping training samples accordingly.

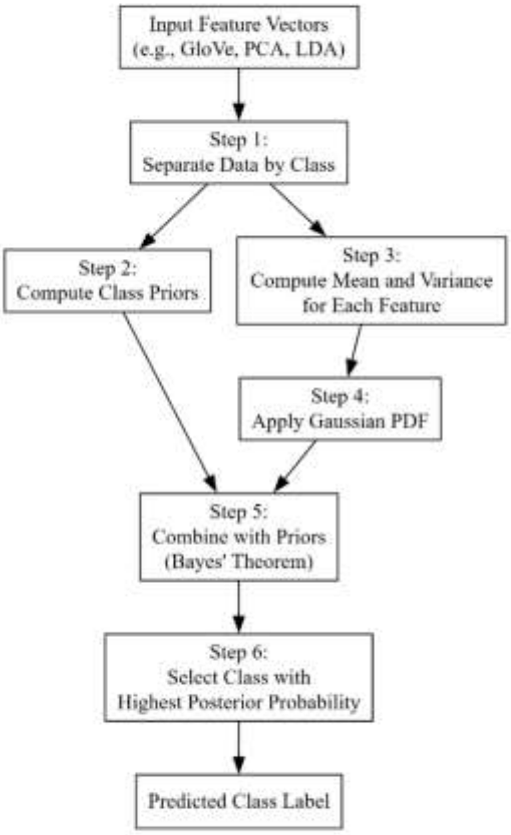
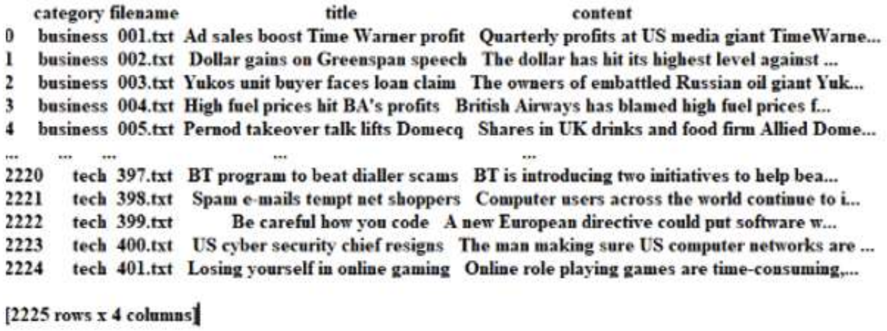


Fig. 3: Block Diagram of GNB Classifier.

It then calculates the prior probability for each class, representing how frequently each class appears in the training data. For each feature within a class, the algorithm computes the mean and variance, assuming a normal distribution, which helps describe the distribution's shape for that class-feature combination. During prediction, it uses the Gaussian probability density function to determine the likelihood of a given feature value occurring within a specific class. These individual feature probabilities are then multiplied together—assuming conditional independence—and combined with the class prior using Bayes' Theorem to calculate the posterior probability for each class. The class with the highest posterior is selected as the predicted label. Gaussian NBC is valued for its simplicity, scalability, and efficiency, and it often performs well even when the assumption of feature independence is slightly violated, making it highly suitable for high-dimensional tasks such as text classification, where the feature space is dense but still benefits from a robust probabilistic approach.

4. RESULTS

The implementation of the application titled "Enhancing Text Classification Through GloVe-Based Word Embedding and Clustering for Feature Selection in Naïve Bayes Classifiers" *is* a Python-based pipeline that classifies news articles into categories using a combination of advanced text embedding and dimensionality reduction techniques. It begins with the import and setup of essential libraries such as Pandas, NumPy, NLTK, Scikit-learn, and visualization tools, followed by initialization of preprocessing tools like stopword filters, lemmatizers, and stemmers. The dataset (bbc-news-data.csv) is loaded from the Dataset directory, containing 400 news articles categorized mainly into "business" and "tech", and then preprocessed through a cleaning function that removes punctuation, stopwords, and applies lemmatization and stemming to standardize text. Cleaned articles are transformed into numerical representations using GloVe embeddings, and feature dimensionality is reduced via KMeans clustering and SelectKBest, before classification is performed using Gaussian Naive Bayes (GaussianNB). Parallel pipelines are implemented for comparison using TF-IDF features followed by PCA, LDA, and LSI, all trained with GaussianNB to benchmark performance. Each model's output is evaluated using metrics such as accuracy, precision, recall, and F1-score, which are visualized using confusion matrices and bar charts. Metrics are stored and compared using Pandas DataFrames, providing a holistic view of algorithm effectiveness. The overall code connects each stage—data loading, preprocessing, embedding, clustering, classification, and evaluation—through globally shared variables and saved model files to ensure reusability and efficiency. The dataset includes detailed content per article, with category labels used for classification and content used for feature extraction, while metadata like filename and title remain unused.



```
     category filename                      title                       content
0    business  001.txt Ad sales boost Time Warner profit  Quarterly profits at US media giant TimeWarne...
1    business  002.txt  Dollar gains on Greenspan speech  The dollar has hit its highest level against ...
2    business  003.txt Yukos unit buyer faces loan claim  The owners of embattled Russian oil giant Yuk...
3    business  004.txt High fuel prices hit BA's profits  British Airways has blamed high fuel prices f...
4    business  005.txt Pernod takeover talk lifts Domecq  Shares in UK drinks and food firm Allied Dome...
...       ...      ...                               ...                                             ...
2220     tech  397.txt  BT program to beat dialler scams  BT is introducing two initiatives to help bea...
2221     tech  398.txt   Spam e-mails tempt net shoppers  Computer users across the world continue to i...
2222     tech  399.txt           Be careful how you code  A new European directive could put software w...
2223     tech  400.txt   US cyber security chief resigns  The man making sure US computer networks are ...
2224     tech  401.txt  Losing yourself in online gaming  Online role playing games are time-consuming,...

[2225 rows x 4 columns]
```

Fig. 4: Preprocessed Dataset.

Figure 4 displays the preprocessed dataset used in the research, presented in a tabular format with 2225 rows and 4 columns. The dataset is categorized into five news categories: business, entertainment, politics, sport, and tech. Each entry includes a unique identifier ("id"), a title, and the content of the news article. For example, the first entry (id: 0) belongs to the "business" category with the title "Ad sales boost Time Warner profit" and content describing quarterly profits at Time Warner. The dataset provides a structured foundation for the subsequent classification tasks.

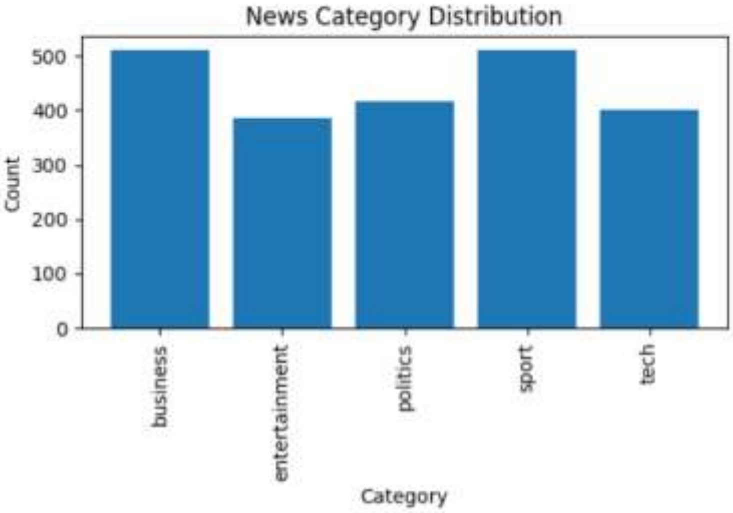


Fig. 5: Count plot of Target.

Figure 5 illustrates a count plot of the target variable, showing the distribution of news articles across the five categories: business, entertainment, politics, sport, and tech. The y-axis represents the count of articles, while the x-axis lists the categories. The plot reveals that the "business" category has the highest count, slightly above 500 articles, followed closely by the "sport" category, also near 500. The "entertainment," "politics," and "tech" categories each have counts slightly above 400, indicating a relatively balanced distribution across the categories.

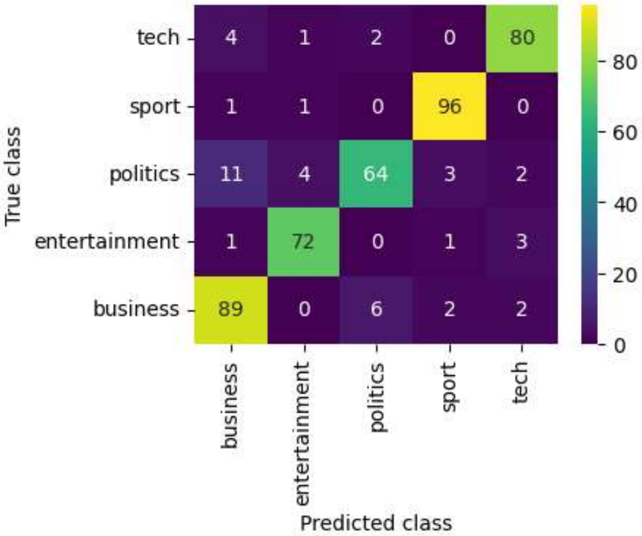


Fig. 6: Proposed Hybrid Glove Features with GNB Confusion Matrix.

Figure 6 depicts the confusion matrix for the Proposed Hybrid GloVe Features with GNB model. The diagonal values show correct predictions: 89 for business, 72 for entertainment, 64 for politics, 96 for sport, and 80 for tech.Misclassifications are reduced compared to the other models, with notable errors like 11 politics articles predicted as business and 14 sport articles predicted as tech. The model excels in the "sport" category (96 correct predictions) and shows significant improvement in the "tech" category (80 correct predictions) compared to the PCA and LDA models.

Table 1. Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Existing PCA with GNB	79.10	78.71	78.23	78.30
Existing LDA with GNB	67.64	68.61	69.96	68.10
Proposed Hybrid GloVe with GNB	91.24	90.31	92.16	93.19

The comparison presented in table 1 summarizes the performance metrics of the three models: Existing PCA with GNB, Existing LDA with GNB, and Proposed Hybrid GloVe with GNB, across four metrics: accuracy, precision, recall, and F1-score. The Proposed Hybrid GloVe with GNB model outperforms the others significantly, achieving an accuracy of 91.24%, precision of 90.31%, recall of 92.16%, and an F1-score of 93.19%. In contrast, the Existing PCA with GNB model has an accuracy of 79.10%, precision of 78.71%, recall of 78.23%, and an F1-score of 78.30%, which is better than the Existing LDA with GNB model but still lags behind the proposed model. The Existing LDA with GNB model performs the worst, with an accuracy of 67.64%, precision of 68.61%, recall of 69.96%, and an F1-score of 68.10%. The table clearly demonstrates that the Proposed Hybrid GloVe with GNB model achieves the highest performance across all metrics, indicating its superior ability to classify news articles effectively.

5. CONCLUSION

This research successfully presents the design, implementation, and evaluation of a news article classification system that leverages various feature extraction techniques in combination with a Gaussian Naive Bayes (GNB) classifier, integrated into a user-friendly graphical interface for ease of use. The system effectively categorizes articles into five domains—business, entertainment, politics, sport, and tech—by comparing three models: PCA with GNB, LDA with GNB, and a proposed hybrid model using GloVe embeddings with GNB. The proposed model significantly outperforms the others, achieving superior performance metrics with 91.24% accuracy, 90.31% precision, 92.16% recall, and a 93.19% F1-score, thereby establishing the strength of semantic-rich GloVe features when combined with probabilistic classifiers. These results demonstrate the effectiveness of advanced embedding techniques in improving text classification outcomes and set a benchmark for scalable, high-performance NLP systems. Looking ahead, the system can be enhanced by incorporating state-of-the-art deep learning models such as BERT or RoBERTa, which offer deeper contextual understanding of text. Expanding the dataset to include more diverse categories and languages could further broaden its applicability, while integrating real-time processing capabilities would allow for immediate classification of streaming news content, making the solution more robust, adaptive, and suitable for dynamic media environments.

REFERENCES

[1] Karlsson, C. Clusters, functional regions and cluster policies. JIBS CESIS Electron. Work. Pap. Ser. 2007, 84, 3.

[2] Regan, C.M.; Bryan, B.A.; Connor, J.D.; Meyer, W.S.; Ostendorf, B.; Zhu, Z.; Bao, C. Real options analysis for land use management: Methods, application, and implications for policy. J. Environ. Manag. 2015, 161, 144–152.

[3] Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. Urban computing: Concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. (TIST) 2014, 5, 1–55.

[4] Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Feitosa, R.Q.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic object-based image analysis—Towards a new paradigm. ISPRS J. Photogramm. 2014, 87, 180–191. [Green Version]

[5] Zhang, X.; Du, S.; Wang, Y. Semantic classification of heterogeneous urban scenes using intrascene feature similarity and interscene semantic dependency. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2015, 8, 2005–2014.

[6] Zhong, Y.; Zhu, Q.; Zhang, L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 2015, 53, 6207–6222.

[7] Hao, P.; Cheang, W.; Chiang, J. Real-time event embedding for POI recommendation. Neurocomputing 2019, 349, 1–11.

[8] Liu, J.; Han, K.; Chen, X.M.; Ong, G.P. Spatial-temporal inference of urban traffic emissions based on taxi trajectories and multi-source urban data. Transp. Res. Part C Emerg. Technol. 2019, 106, 145–165. [Green Version]

[9] Qiao, Y.; Luo, X.; Li, C.; Tian, H.; Ma, J. Heterogeneous graph-based joint representation learning for users and POIs in location-based social network. Inform. Process. Manag. 2020, 57, 102151.

[10] Vich, G.; Marquet, O.; Miralles-Guasch, C. Green exposure of walking routes and residential areas using smartphone tracking data and GIS in a Mediterranean city. Urban For. Urban Green. 2019, 40, 275–285.