Enhancing Urban Walkability Through Integrated Data Analysis and Social Media Insight

Denis Dimitrov, Dessislava Petrova-Antonova

Abstract

Enhancing walkability within urban environments is crucial for promoting walkability behaviours, improving public health, and fostering sustainable cities. Walkability reflects how friendly an area is for walking, which may reflect on the public perception, even on social media. Traditionally perceived walkability has been measured through surveys, which may be resourceful and time consuming. Advances in Geographic Information Systems (GIS), the ease of access to data and machine learning techniques now enables a more objective measurement of walkability. Challenges remain when trying to evaluate objective measurement of satisfaction from social media posts related to political parties and city infrastructure, effectively analysing them and ensuring enough data availability. This study develops integrative models to objectively assess walkability, a the neighbourhood level by incorporating multiple environments and social factors. Drawing upon empirical studies that explore relationship between walking and environmental attributes, the model identifies key characteristics influencing walkability: proximity to points of interest (POIs), pedestrian network accessibility, building density, population density and public sentiment towards walkability-related infrastructure. The model uniquely integrates sentiment analysis of social media data to capture public perception and combines it with questionnaire responses to enhance the understanding of walkability, both from objective and subjective perspective. As a case study the model is applied to the city of Sofia, Bulgaria. The model provides a benchmark for analysis applying it for environment quality and also offer valuable the policy making and urban planning of Sofia. The outcomes include generation of a heat map of Sofia, illustrating spatial variations in walkability across different neighbourhoods. The visualisations aids in identifying areas requiring infrastructure improvements.

Keywords: Walkability, Natural Language Processing, Social Media Analysis, Urban Planning, Public Health, Pedestrian Network, Accessibility, Sentiment Analysis, GIS, Spatial Analysis, Machine Learning

Contents

1 Introduction

1.0.1 Background and literature review

Walkability plays a crucial role in ensuring the quality of urban life, influencing public health environmental sustainability and social equality. Several studies have established that high walkability levels promote physical activity, reduce traffic congestion, and improve social interactions in the cities [6], In the context of rapid urbanization and the push for sustainable cities, walkability indices have emerged as valuable tools for analysing and urban planning. One significant area of research in walkability involves the use of land use and land cover (LULC) mapping, which provides critical insights for urban planning, [3] emphasise the utility of LULC maps in mart city planning and sustainable development, which also highlight the use of machine learning algorithm in the accuracy of these maps. Their study on the city of Melbourne demonstrates how machine learning algorithms, including Random Forest and Support Vector Machines, outperform traditional approaches in LULC classification. The results of their research contribute to urban resilience and the planning of green spaces, underscoring the importance of precise spatial data for the walkability indices.

Natural language processing (NLP) has increasingly seen use in the field of urban research, particularly for analysing public perceptions through social media data. [5] presents a systemic review of how NLP techniques are applied urban studies. This study highlights how unstructured text data from social media, emails, and web pages provide valuable insights into the behaviour, attitudes and perceptions of people, which van enhance the traditional data sources. By integrating sentiment analysis into walkability indices, urban researchers can capture the citizens' subjective experiences, adding depth to objective spatial measures.

Complementing these spatial and text-based approaches, research has also focused on the temporal aspects of walkability. [2] Comfort and Time-Based Walkability Index Design (CTBWI) studies explore how time constraints and comfort levels affect pedestrian behaviours. This concept shifts the focus from static to dynamic evaluation of pedestrian comfort across different times of the day, which provides a more holistic understanding of walkability. Similarly, the research by social media sentiment analysis [4] has shown potential for mining real live lives experiences in real-time, further pushing for walkability indices in with live data. Another key theme in literature is the use of big data, especially regarding environment factors like air quality and temperature, which also impact walkability. [7] The integration of various urban data sources - geospatial information, land use maps, and social media sentiment - offers a more comprehensive framework for assessing urban walkability. This body of research supports the arguments that the design effective walkability indices, it is necessary to integrate objective spatial measures and subjective perception of the urban environment. The combination of machine learning techniques, geospatial data, and NLP opens new avenues for enhancing the accuracy and relevance to walkability indices for policymaking.

1.0.2 Knowledge Gap

Despite the acknowledged importance of subjective perception of walkability, there is lack of comprehensive models that integrate both quality spatial data and qualitative sentiment analysis. Existing studies often treat physical infrastructure and social perceptions separately, leading to incomplete assessments. This study aims to fill this gap by developing a walkability index that combines the

available spatial data about the city of Sofia, questionnaire given to its residents and a social media sentiment analysis in Bulgarian, and we address the linguistic limitations in the previous research.

2 Methods

[1] This section outlines the methodological approach adopted to predict the walkability index using machine learning techniques. The process involves data collection from various sources, preprocessing and feature extraction, model development using Neural Networks, Random Forest and Support Vector Machine (SVM), and also the evaluation of these models.

2.1 Research question

How can the integration of spatial data, residents' perception from questionnaires, and social media sentiment analysis improve the assessment of urban walkability in Sofia, Bulgaria?

2.2 Data Collection

2.2.1 Spatial data

- Points of Interest (POIs): Shapefiles containing locations of amenities and services. Represent the locations that could influence walkability due to their attractiveness or necessity. Gathered by the GATE Institute, "Future Cities".
- Pedestrian Network: Shapefile detailing pedestrian pathways and sidewalks. Gathered by the GATE Institute, "Future cities"
- Residential Buildings: shape file of all residential buildings in Sofia. Gathered by the GATE Institute, "Future cities".
- Land Use Data: GeoPackage containing land use classification and population data. Urban Atlas Land Cover/Land Use 2018 (vector), Europe, 6-yearly; https://land.copernicus.eu/en/products/urban-atlas/urban-atlas-2018

2.2.2 Questionnaire Data

- Survey Instrument: A Questionnaire in Bulgarian constructed and distributed by the GATE Institute - the "15-minute city" team and Bryan Monticelli; It measures the residents' perceptions of walkability factors such as sidewalk conditions, safety and amenities.
- The key questions are: "What is the condition of the sidewalks around the residential building for which you are filling out the questionnaire?" "Are the sidewalks wide enough?" "Are there obstacles on the sidewalks (objects, tree roots, etc.)?" "Are there parked cars on the sidewalks?" "Do you feel safe while walking in the neighborhood?" "Are there shaded areas, seating and recreation spots, waste bins, etc.?" "Are there pedestrian crossings, overpasses, or underpasses where they are needed?"

  The questionnaire was distributed amongst the instate and on social media.

2.2.3 Social Media Data

- Platform: Facebook
- CSV file with posts about the urban living, election news and city infrastructure in the city of Sofia. These posts were gathered by using Boolean search with the keywords: "pedestrian", "accessibility", "transport", "public transport", "connectivity", "stop", "walk", "infrastructure", "sidewalk", "Sofia", "center", "neighborhood"; etc. The results were manually filtered, and only a portion was selected to be used in this study.
- Data Fields: Message contents, user interaction and other metadata.

[1] https://github.com/dennisthekhan/Walkability-Index-with-Social-Media

- Language: Bulgarian.

2.3 Data Processing

2.3.1 Sentiment analysis

Utalization of the *nlptown/bert-base-multilingual-uncased-sentiment* multilingual BERT model to perform sentiment analysis on the Bulgarian text. In the proceeding steps we cleaned and preprocessed the text data, Applied the analysis model to obtain scores ranging from 1 (negative) to 5 (positive) and then mapped the scores to numerical values for integration.

$$\text{Sentiment Score}(r) = \frac{1}{N_r} \sum_{k=1}^{N_r} s_k$$

where:

- r: Cadastre region.
- N_r: Number of social media posts in region r.
- s_k: Sentiment score of post k.

2.3.2 Spatial Feature Calculation

First the distance metrics are calculated - calculating the centroid of each residential building, measuring the distance from each building centroid to the nearest POI, and measuring the distance from each building centroid to the pedestrian network. For building density the number of buildings per region is calculated, and for pedestrian density we use land use data by dividing the population by area. This serves as a proxy for population density and urban form, influencing walkability.

- Distance to nearest POI

$$\text{Distance}_{\text{POI}}(b) = \min_{\forall p \in \text{POIs}} \left\{ \sqrt{(x_b - x_p)^2 + (y_b - y_p)^2} \right\}$$

- Building Density

$$\text{Building Density}(r) = \frac{N_{\text{buildings}}(r)}{A(r)}$$

  where:

  - r: Cadastre region.
  - $N_{\text{buildings}}(r)$: Number of residential buildings in region r.
  - $A(r)$: Area of region r.

- Population Density

$$\text{Population Density}(l) = \frac{P(l)}{A(l)}$$

  where:

  - l: Land use unit.
  - $P(l)$: Population in land use unit l.
  - $A(l)$: Area of land use unit l.

- Entropy Index

$$\text{Entropy Index}(r) = -\frac{\sum_{i=1}^{N} p_i \ln(p_i)}{\ln(N)}$$

  where:

- r: Region being evaluated.
- p_i: Proportion of land use type i in region r.
- N: Total number of land use categories.

2.3.3 Questionnaire Data Encoding

First the categorical responses are encoded numerically by using label encoding. The encoded responses are averaged to create a walkability sore for each respondent. Adresses from the questionnaire and extracted locations from social media posts were geocoded to obtain latitude and longtitute coordinates using the Nominatim API (OpenStreetMap geocoding service). Adresses provided by the responders were converted into coordinates.Locations were concatenated with "Sofia,Bulgaria", to improve accuracy.A rate limit was implemented to comply with the API usage policies. Then only posts with successfully geocoded locations were retained for further analysis. Locations mentioned in social media posts were extracted using a multilingual language model *(xx-ent-wiki-sm)* end geocoded. The responses were mapped to regions using the "cadregion" property.

2.4 Data Integration

The spatial, the social media sentiment score and the questionnaire data is merged on the common key of "cadregion". The feature matrix is constructed by combining the features - disttance to POIs, distance to pedestrian network, building density, population density, average social media sentiment, etc. The composite walkability score from the questionnaire is used as a target variable.

$$\text{Walkability Score}_j = \frac{1}{M} \sum_{i=1}^{M} r_{ij}$$

where:

- j: Respondent index.
- M: Number of questionnaire items.
- r_{ij}: Encoded response of respondent j to question i.

2.5 Model Development

Normalization of standard scaling (StandardScaler from scikit-learn) is applied to the feature matrix. It ensures that all features contribute equally to the model.

$$X_{\text{scaled}} = \frac{X - \mu_X}{\sigma_X}$$

where:

- X: Original feature vector.
- μ_X: Mean of the feature X.
- σ_X: Standard deviation of the feature X.
- X_{scaled}: Standardized feature vector.

The neural network architecture is constructed of an input, hidden and output layer. It utilises dropout and batch normalisation to prevent overfitting.

The input layer corresponds to the number of features. There are two hidden layers with 64 and 32 neurone, respectively, using ReLU activation functions. The output layer is a single neutron with linear activation for regression output.

- First Hidden Layer

$$h_1 = \text{ReLU}(W_1 X_{\text{scaled}} + b_1)$$

  where:

- h_1: Output of the first hidden layer.
- ReLU: Rectified Linear Unit activation function.
- W_1: Weight matrix for the first hidden layer.
- X_{scaled}: Standardized input features.
- b_1: Bias vector for the first hidden layer.

- Second Hidden Layer

$$h_2 = \text{ReLU}(W_2 h_1 + b_2)$$

- Output Layer

$$\hat{y} = W_3 h_2 + b_3$$

where:

- $\hat{y}$: Predicted output.
- W_3: Weight matrix for the output layer.
- h_2: Output of the second hidden layer.
- b_3: Bias term for the output layer.

To prevent overfitting L2 regularisation and dropout layers (of 30The data is split into training and validation sets using a n 80-20 ratio, to evaluate the model on unseen data and prevent overfitting. The model is measured using mean squared error and mean absolute error metrics.

Loss function

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:

- n: Number of samples.
- y_i: Actual value for sample i.
- $\hat{y}_i$: Predicted value for sample i.

Regularisation (Total Loss Function)

$$\text{Loss}_{\text{total}} = \text{MSE} + \lambda \sum_{k} \|W_k\|_2^2$$

where:

- $\text{Loss}_{\text{total}}$: Total loss function including regularization.
- MSE: Mean Squared Error loss.
- λ: Regularization parameter.
- W_k: Weight matrix of layer k.
- $\|W_k\|_2^2$: Squared L2 norm of the weights in layer k.

2.5.1 Random Forest

A random forest algorithm was also implemented which consists of an ensemble of decision threes (RandomforestRegressor).Their prediction are aggregated. The parameters used In this experiment were: number of threes - 100 (n-estimators), and Random State of 42 for reproducibility. The model was trained on a training set without a need for feature scaling. Each tree is trained on a bootstrap sample of the training data. At each split a random subset of features is considered. The prediction is an average of the predictions from each individual tree:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X)$$

where:

- $\hat{y}$: Predicted output.
- T: Total number of trees in the forest.
- $f_t(X)$: Prediction from tree t for input X.

2.5.2 Support Vector Machine Regressor

SVM aims to find a function that derives from the actual observed values by a value no grater than \epsilon for all training data, and at the same time is as flat as possible. Radial Basis Function (RBF) Kernel was used with a kernel coefficient :

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

where:

- $K(x_i, x_j)$: Kernel function between samples x_i and x_j.
- γ: Kernel coefficient.
- x_i, x_j: Input feature vectors.

For the regularisation parameter (C) we have 1.0, which controls the trade off between the flatness of the function and the amount up to which deviations are tolerated; for Epsilon 0.1 was used, whisch specifies the width of the insensitive zone.

For the optimisation objective, we try to minimise:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$\begin{cases} y_i - (w^T \phi(x_i) + b) \leq \epsilon + \xi_i, \\ (w^T \phi(x_i) + b) - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0. \end{cases}$$

where:

- w: Weight vector.
- b: Bias term.
- ξ_i, ξ_i^*: Slack variables for sample i.
- C: Regularization parameter.
- y_i: Actual target value for sample i.
- $\phi(x_i)$: Feature mapping function for sample i.
- ϵ: Epsilon-insensitive loss parameter.
- n: Number of samples.

2.5.3 Model evaluation

The performance was assessed using the testing set to ensure that the models generalize well on unseen data. Two primary metrics were used:

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2$$

Measures the average squared difference between the predicted and the actually value.

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |y_i - \hat{y}_i|$$

where:

- n_{test}: Number of samples in the testing set.
- y_i: Actual value for sample i in the testing set.
- $\hat{y}_i$: Predicted value for sample i in the testing set.

Measures the average absolute difference between the predicted and the actual value, which provides a more importable measure of average error in the same units as the target value.

3 Results

3.1 Heatmaps

3.1.1 Baseline Walkability index

This Figure 1, is a representation of a walkability scorers across various regions of Sofia. It only uses the spacial data, described above.

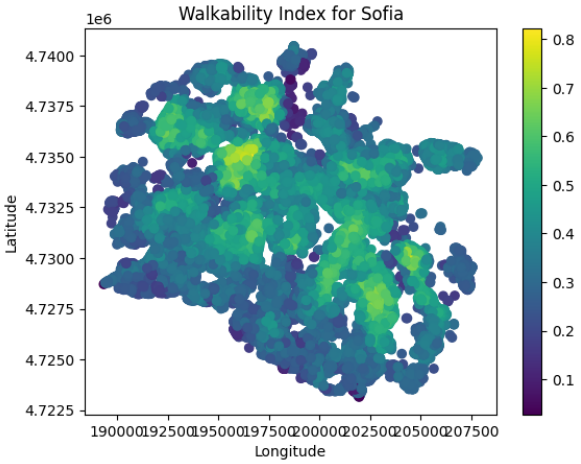

Figure 1: Baseline Spatial Heatmap

This heat map uses a diverging colormap ranging from purple - low walkability, to yellow - high walkability. It is visible that the most walkable regions, indicated by bright yellow, are concentrated in a few clusters, suggesting urban cores with high density and pedestrian infrastructure. Darker areas, are located on the outside of the city area and could correspond to areas with limited pedestrian resources and space amenities. These could also indicated industrial or suburban areas. This index can serve as a baseline, against which to compare the mode detailed and adjusted indices, constructed in this study. Based on the areas with low walkability, city planners and researchers may prioritise these areas for interventions.

3.1.2 Prediction of Neural Network

This Figure 2 is generated using the Neural Network method Here a color gradient of red and blue

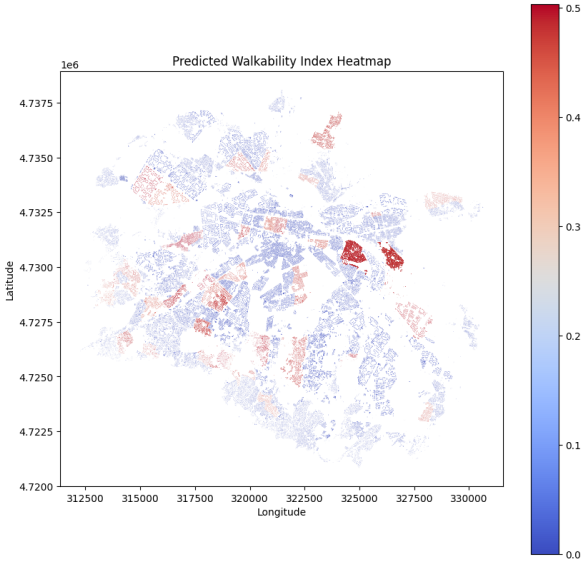

Figure 2: Neural Network Heatmap

is used where red indicates lower walkability and, blue - general higher. High-Walkabilty cluster are again visible on the right side of the map. This prediction is based on a number of features: Distance to POIs - Shorted distance to parks, shops, schools, etc. typically indicates positive contribution to walkability;, Distance to Pedestrian Network - regions closer to pedestrian network are considered more walkable, because of the ease of access; Population Density - regions with a higher population density are considered more well-managed, often have more interconnecting streets and amenities; Sentiment Score - regions with more positive social media impressions are considered with better scores; and Perception Features - the model includes user-reported data, such as sidewalk conditions, presence of obstacles, feeling of safety while walking, etc.

The neural network used for this index includes a multiple hidden layers, which capture non-linear relationships between the input walkability scores. Batch normalisation, which ensures faster and more stable training process. The output layer uses a sigmoid activation which output a walkability index in the range of 0 to 1. The results separate the higher and lower walkability regions clearly, but the median values are not represented well and clearly for the inner city regions.

3.1.3 Prediction of Regional Neural Network

This Figure 3 is a result from a Neural networks, where the sofial media sentiments were separated into regions. his heat map represents the results of the neural network when using regional information to separate the accumulated data from social media and to map it to a designated region. There is a separation into blue and green regions, which could be a result from separation by the Proximity to POIs, which indicated easier access to services. Positive and negative perception from social media may influence the predicted scores with regions showing lower walkability in general for ares with negative perception. The regions in the centre of the city are more walkable in general so this heat map, indicating the opposite means that the user responses and Facebook posts play a significant role in its evaluation.The red regions high need additional services of correction of the pedestrian network. There needs to be further research into the the contents of the post, which were selected manually but may not be entirely related to urban planning. There are more balanced areas, which may suggest not exceptional but sufficient infrastructure and enough services, related to connectivity. This model may be in need of real pedestrian traffic data. This model provides a more well-balanced index on the outskirts to the city, but a lot of outlier negative results in the city centre, which may be a result of the activity of its residents on social media, mainly by political party posts, which were used in the process.

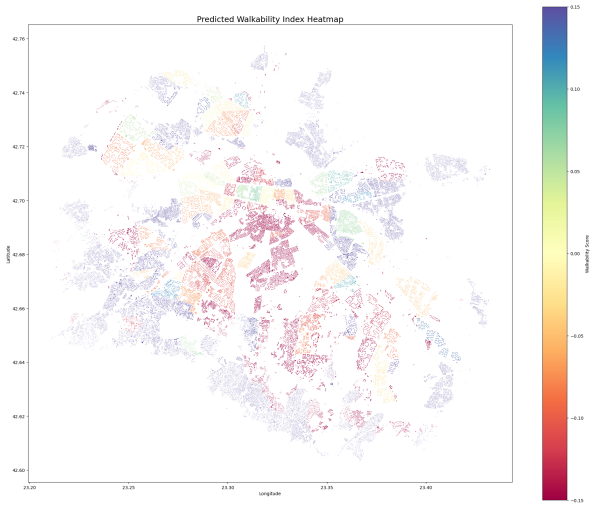

Figure 3: Regional Heatmap

3.1.4 Comparison

These three versions display variations in their indices underlying key differences in the interpretation of the features.

In the baseline first version of the index the "high" areas are mainly centred in clusters, and there is a sharp line between the high and low regions, which indicated that the a more drastic values are obtained from using the spatial data and there is a lot of importance in the proximity to POIs. This version may see the best use for a grater area analysis, suited for urban planning. This heat map represents a binary version of walkability.

In the second version of the index shows less clusters, which may represent an even representation of walkability, but these "red" regions are also located in the further regions. The social media sentiment is spread more, and it indicates more moderate scores.

In the third version the high walkability areas are spread more evenly with fewer clusters with less distance between high and low walkability areas. This could also suggest lower threshold for the low-walkability areas, which are lees concentrated. The moderate (green/purple) areas are spread more evenly. This version focuses the sentiment on specific areas which are discussed more prevalently, which indicated that a validation is needed. This model is complex and multi-dimensional, resulting in a detailed breakdown of walkability.

4 Models comparison

Model	Mean Squared Error	Mean Absolute Error
Random Forest Regressor	0.0406	0.1988
Support Vector Regressor	0.0602	0.2206
Neural Network	1.0478	1.0005

Table 1: Model Performance Metrics

4.0.1 Analysis

- **Random Forest Regressor:** Achieved the lowest MSE and MAE, indicating the best predictive accuracy among the models. The ensemble nature of Random Forest likely captured nonlinear relationships and interactions between features effectively.
- **Support Vector Regressor:** Performed moderately well, with slightly higher errors than the Random Forest model. The RBF kernel enabled modeling of nonlinear patterns, but the model may have been limited by parameter settings or data characteristics.

- **Neural Network:**

  Exhibited significantly higher MSE and MAE, indicating poor predictive performance. Possible reasons include:

  Overfitting - Despite regularization techniques, the model may have overfitted to the training data. *Data Size* - Neural networks typically require large datasets to generalize well. The small sample size may have hindered effective learning. *Hyperparameters* - The chosen architecture and hyperparameters may not have been optimal for the dataset.

4.0.2 Feature Importances

Feature	**Importance Score**
Sentiment Score	Highest
Building Density	High
Population Density	Moderate
Distance to Pedestrian Network	Lower
Distance to POI	Lowest

Table 2: Feature Importance Scores

4.0.3 Interpretation

Random Forrest Regressor achieves the lowest MSE and MAE, indicating the best predictive accuracy. The ensemble nature of this models likely captures the nonlinear relationships and interaction between the features effectively.

Support Vector Regressor performed moderately well but with slightly higher error. The RBF kernel enabled modelling of the nonlinear patterns , but the model might have been limited by the characteristics and the nature of the data.

The neural network on the other hand exhibits significantly higher MSE and MAE, which indicates a poor performance. The model might have been overfitted on the training data. Models of this kind usually require large datasets, with our data size, the learning may have been hindered. The architecture and hyper parameters need to be changed future implementations of the study.

The most significant predictor, is the sentiment score, which indicates that public perception heavily influences perceived walkability. Building and Population densities with higher values are associated with increased accessibility. The distance measures are less influential, possibly due the the scale of variation or multicollinearity with other features.

4.0.4 Conslusion

In conclusion random forests is the preferred method for predicting walkability due to its superior performance. The limitations include the number of questionnaire responses (n 96), which may constrain the models, particularly the neural network. The geocoding inaccuracies and errors with the API responses can also lead to inaccuracies and to insufficient and bias data. For future development more types of data should be considered and integrated in the model (e.g., crime rates, detailed land use mix), if whey can be obtained. More questionnaire responses should be collected to strengthen the datas set. Additional hypertuning of the parameters of the models especially the neural network should be conducted. Additional cross-validation techniques can be used to obtain more robust estimates.

The methodological approach integrates multi-source data and employs advanced machine learning techniques to predict the walkability index. The Random Forest method demonstrated superior performance, effectively capturing the relationships between the built spatial environment, public sentiment in Facebook social media, and perceived walkability by the residents of the neighbourhoods in Sofia . This study underscores the importance of combining objective spatial data with subjective human perceptions to holistically assess walkability. Problems arise in the data collection process, which further complicate the selection of traditional machine learning methods.

References

[1] Razmik Agampatian. Using gis to measure walkability: A case study in new york city, 2014.

[2] Tarek Al Shammas and Francisco Escobar. Comfort and time-based walkability index design: A gis-based proposal. *International Journal of Environmental Research and Public Health*, 16(16):2850, 2019.

[3] Jagannath Aryal, Chiranjibi Sitaula, and Alejandro C. Frery. Land use and land cover (lulc) performance modeling using machine learning algorithms: a case study of the city of melbourne, australia. *Scientific Reports*, 13(13510), 2023.

[4] Christian Berzi, Andrea Gorrini, and Giuseppe Vizzari. Mining the social media data for a bottom-up evaluation of walkability. *arXiv preprint arXiv:1712.04309*, 2017.

[5] Meng Cai. Natural language processing for urban research: A systematic review. *Heliyon*, 7(7):e06322, 2021.

[6] Kimihiro Hino, Hiroki Baba, Hongjik Kim, and Chihiro Shimizu. Validation of a japanese walkability index using large-scale step count data of yokohama citizens. *Cities*, 123:103614, 2022.

[7] Xuanbei Lu. The development and deployment of walkability assessment models for built environments, 2017.

[7] [6] [2] [1] [4] [3] [5]