Hybrid geo-semantic model for tourist accommodation recommendation based on web scraping and unsupervised evaluation

Modelo geosemántico híbrido para la recomendación de alojamientos turísticos basado en web scraping y evaluación no supervisada

Modelo geo-semântico híbrido para recomendação de acomodações turísticas baseado em Web Scraping e avaliação não supervisionada

Juan Camilo Castañeda Alvarado[1]
José Ignacio Rodríguez Molano[2]
Jorge Enrique Salamanca Céspedes[3]

Received: April 15th, 2026
Approved: July 27th, 2026
Available: September 1st, 2026

How to cite this article:
J. C. Castañeda Alvarado, J. I. Rodríguez Molano, and J. E. Salamanca Céspedes, "Hybrid Geo-Semantic Model for Tourist Accommodation Recommendation Based on Web Scraping and Unsupervised Evaluation," *Revista Ingeniería Solidaria,* vol. 22, no. 3, 2026.
doi: https://doi.org/10.16925/2357-6014.2026.03.05

Research article. https://doi.org/10.16925/2357-6014.2026.03.05

[1] Researcher, M.Sc. in Information Science, Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.
Email: juccastanedaa@udistrital.edu.co
ORCID: https://orcid.org/0009-0009-3120-5627
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0002474984

[2] Professor, Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.
Email: jirodriguezm@udistrital.edu.co
ORCID: https://orcid.org/0000-0003-2581-277X
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000593869

[3] Professor, Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.
Email: jsalamanca@udistrital.edu.co
ORCID: https://orcid.org/0000-0002-9396-2641
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000177407

CC BY

Abstract

Introduction: This article is the result of research on the *Tourist Accommodation Recommendation Model Using Machine Learning with Information Extraction Through the Web Scraping Technique* conducted at the Universidad Distrital Francisco José de Caldas in 2026.

Problem: In tourism cold-start scenarios, the lack of structured data and user history reduces the quality of accommodation recommendations. Web Scraping is required [2], but it introduces inconsistencies and structural variability that hinder accurate recommendations.

Objective: To design and evaluate a hybrid geo-semantic recommender system that integrates content-based similarity between accommodations with geographic proximity under an unsupervised framework using web-scraped data.

Methodology: Data from 13,204 Colombian accommodations were scraped from Booking.com using Playwright, preprocessed, and represented as vectors. Then, cosine similarity and Haversine distance were combined into a weighted hybrid scoring function, evaluated through similarity standard deviation, intra-list similarity, and geographic relevance.

Results: The base model achieved an average cosine similarity of 0.868 and a mean geographic distance of 0.16 km, with 100% of recommendations within the defined geographic constraint radius. In an exploratory comparison against Booking.com, the proposed model returned semantically and spatially consistent alternatives independent of commercial positioning criteria.

Conclusions: The geo-semantic model generates coherent and geographically relevant recommendations from web-scraped data in cold-start contexts where user interaction data are unavailable.

Originality: This work proposes a reproducible geo-semantic recommender model for tourist accommodation in cold-start scenarios, integrating semantically represented web-scraped accommodation data with geographic proximity to generate recommendations.

Limitations: The model does not incorporate user preference data.

Keywords: Recommender Systems; Cosine Similarity; Haversine Distance; Web Scraping; Tourism.

Resumen

Introducción: Este artículo es el resultado de la investigación *Modelo de recomendación de alojamientos turísticos mediante Machine Learning, con extracción de información a través de la técnica Web Scraping*, desarrollada en la Universidad Distrital Francisco José de Caldas en 2026.

Problema: La falta de datos estructurados y del historial de los usuarios dificulta las recomendaciones de alojamientos. Aunque el Web Scraping es una forma de obtención de información [2], este proceso suele generar inconsistencias y variaciones en los datos, lo que produce recomendaciones imprecisas.

Objetivo: Diseñar y evaluar un sistema de recomendación híbrido geo-semántico que integre la similitud entre alojamientos y la proximidad geográfica, utilizando datos obtenidos mediante Web Scraping.

Metodología: Se realizó Web Scraping de 13.204 alojamientos de Colombia en Booking.com, cuya información fue procesada y representada en vectores. Luego, el modelo se evaluó mediante la desviación estándar de la similitud, la similitud intra-lista y la cercanía geográfica.

Resultados: El modelo obtuvo una similitud promedio de 0.868 y una distancia media de 0.16 km. Se comparó con Booking.com, donde el modelo recomendó alojamientos con mayor coherencia semántica y geográfica.

Conclusiones: El modelo geo-semántico genera recomendaciones de alojamientos acordes con sus características y posición geográfica, a partir de datos obtenidos mediante Web Scraping, sin utilizar información de interacción de los usuarios.

Originalidad: Se propone un modelo de recomendación geo-semántico que integra la similitud semántica y la cercanía geográfica, utilizando datos obtenidos mediante Web Scraping en Booking.com para generar recomendaciones de alojamientos en Colombia.

Limitaciones: El modelo no incluye información previa sobre la interacción de los usuarios.

Palabras clave: Sistemas de recomendación; Similitud del coseno; Distancia de Haversine; Web Scraping; Turismo.

Resumo

Introdução: Este artigo é resultado da pesquisa Modelo de recomendação de acomodações turísticas por meio de Machine Learning, com extração de informações através da técnica de Web Scraping, desenvolvida na Universidad Distrital Francisco José de Caldas em 2026.

Problema: Em cenários de cold-start no turismo, a falta de dados estruturados e de histórico dos usuários reduz a qualidade das recomendações de acomodações. Embora o Web Scraping seja utilizado para obtenção de informações, ele introduz inconsistências e variações estruturais que dificultam recomendações precisas.

Objetivo: Projetar e avaliar um sistema híbrido de recomendação geo-semântica que integre a similaridade entre acomodações e a proximidade geográfica, utilizando dados obtidos por Web Scraping.

Metodologia: Foram coletados por Web Scraping dados de 13.204 acomodações na Colômbia a partir do Booking.com, utilizando Playwright. Os dados foram processados e representados como vetores. Em seguida, a similaridade de cosseno e a distância de Haversine foram combinadas em uma função de pontuação híbrida ponderada.

Resultados: O modelo apresentou similaridade média de 0,868 e distância geográfica média de 0,16 km. Em comparação exploratória com o Booking.com, o modelo recomendou acomodações com maior coerência semântica e geográfica.

Conclusões: O modelo geo-semântico gera recomendações coerentes e geograficamente relevantes a partir de dados obtidos por Web Scraping em cenários de cold-start, nos quais não estão disponíveis dados de interação dos usuários.

Originalidade: Propõe-se um modelo de recomendação geo-semântico que integra similaridade semântica e proximidade geográfica utilizando dados obtidos por Web Scraping.

Limitações: O modelo não incorpora informações sobre as preferências dos usuários.

Palavras-chave: Sistemas de recomendação; similaridade de cosseno; distância de Haversine; Web Scraping; turismo.

1. INTRODUCTION

Digital tourism has experienced a significant transformation in recent years, driven by the growth of online platforms that centralize the supply of tourism services, particularly in the accommodation sector [3]. These platforms allow users to access large volumes of information in real time, including descriptions, reviews, services, images, prices, and geographic location [4], [5]. However, this abundance of information has introduced a structural issue: user cognitive overload [6]. In practice, having more

information does not necessarily lead to better decisions, as the multiplicity of options and attributes makes it difficult to efficiently identify truly relevant alternatives [1], [7].

In response, recommender systems have become a relevant technological solution, enabling the filtering, ranking, and prioritization of options based on similarity or preference criteria. In this field, hybrid models have gained particular attention in recent investigations because of their ability to integrate multiple sources of information and mitigate the limitations of individual approaches such as collaborative filtering or content-based methods [8]. Multiple studies have explored combinations of user profiles, item attributes, contextual data, and semantic techniques to improve recommendation accuracy and relevance [8], [9], [10]. However, in the tourism domain, recommendation relevance depends not only on similarity between accommodation attributes but also on geographic proximity to a reference point or area of interest, introducing a spatial dimension that is not always explicitly incorporated into traditional models.

Even with these advances, significant challenges remain in real-world implementations. One of the most critical is the cold-start problem, where no user interaction history is available, preventing the effective application of approaches based on previous behavior such as bookings, clicks, or historical ratings [11]. Additionally, in many practical scenarios, data do not come from structured repositories or carefully selected training datasets but from automated extraction processes through Web Scraping. This introduces difficulties related to heterogeneity, incompleteness, noise, and structural variability in the collected data. Consequently, there is a need for models capable of operating without explicit user feedback, using data extracted from unstructured sources, integrating both semantic and geographic information, and supporting evaluation schemes where no ground truth is available [12], [13].

Developing efficient recommender systems in tourism not only represents a technical contribution but also supports the democratization of access to useful and actionable information, enabling users to make more informed decisions based on objective criteria such as quality, services, and location [11]. This approach promotes more autonomous, transparent, and potentially more sustainable tourism by reducing dependence on visibility mechanisms driven only by commercial or platform positioning strategies. It also increases the exposure of less visible accommodations that are geographically and semantically relevant, contributing to a more balanced distribution of tourism demand [14].

Although hybrid recommender systems in tourism have been widely studied, most existing approaches rely on user profiles, collaborative signals, or supervised learning strategies. Limited attention has been given to fully unsupervised geo-semantic

recommendation using exclusively web-scraped accommodation data in cold-start scenarios. This gap is especially relevant in tourism environments where labeled relevance data and user interaction histories are unavailable.

In this context, the results can support the optimization of recommendation platforms and improve the user experience in intensive search scenarios [15]. For software developers, the study provides a methodological approach for implementing robust recommender systems in environments with limited, incomplete, or unstructured data. For academia, it contributes to the design and evaluation of unsupervised hybrid models, particularly in the integration of semantic representation, applied geomatics, and descriptive evaluation metrics.

Under this approach, the objective of this article is to design, implement, and evaluate a tourist accommodation recommender system based on an unsupervised hybrid geo-semantic approach that integrates content similarity using cosine similarity and geographic proximity using the Haversine distance. The paper presents the construction of the model from web-scraped data, its methodological and mathematical formalization, the application of dimensionality reduction techniques, and its evaluation through unsupervised metrics of coherence, stability, and proximity. The results are then analyzed and discussed in the context of digital tourism, recommender systems, and engineering applications in data-scarce environments. The main contributions of this study are summarized as follows:

- It proposes an end-to-end geo-semantic recommendation framework for tourist accommodation based on heterogeneous data obtained through Web Scraping.
- It formulates a cold-start item-to-item recommendation strategy that does not require user profiles, historical interactions, bookings, clicks, or explicit relevance labels.
- It integrates semantic similarity and geographic proximity into an explicit hybrid ranking function, allowing control over the balance between accommodation attributes and spatial relevance.
- It evaluates the model using unsupervised internal metrics, including semantic coherence, intra-list similarity, ranking stability, and geographic relevance, avoiding the misuse of supervised metrics in the absence of ground truth.
- It provides empirical evidence from a large-scale dataset of Colombian accommodations extracted from Booking.com, demonstrating the feasibility of geo-semantic recommendation using real-world, noisy, and incomplete tourism data.

1.1. REVIEW OF RELATED LITERATURE

Recent literature on recommender systems shows increasing interest in hybrid approaches that combine multiple sources of information to improve recommendation quality. Studies indexed in leading academic databases, including Scopus and Web of Science, have explored the integration of content-based filtering with semantic and contextual models [16].

In the tourism domain, studies such as Solano-Barliza et al. [17] highlight the importance of incorporating contextual and geographic features into recommender systems, while studies such as Troussas et al. [18] explore the use of semantic representation techniques to improve the quality of recommendations.

In addition, models have been developed that incorporate natural language processing techniques to capture the semantics of accommodation descriptions, as well as approaches that integrate geospatial information to improve recommendation relevance based on location [19]. The following table summarizes the articles identified and analyzed in this research, all related to this topic.

TABLE I. ARTICLES RELATED TO RECOMMENDER SYSTEMS

Recommender System	Destination	Tourism Typology	Attributes
A Two-Tiered Recommender System for Tourism Product Recommendations [20]	27 popular destinations in North America and Europe	General Points of Interest	Constraints, flights, hotels, transportation
Coastal Tourism and Climate-Related Disasters in an Archipelago Country of Indonesia: Tourists' Perspective [21]	Indonesia	Coastal Tourism	User profile, climate
A Tourist Guide with a Recommender System and Social Interaction [22]	Portugal	General Points of Interest	Tourist Location, POI, Age and Gender
A Suggestive Recommendation Method to Make Tourists "Feel like going" [23]	Area de Gion y Kiyomizu Kyoto. (Japan)	General Points of Interest	Location and pictures
A personalized hybrid tourism recommender system [24]	Paris	General Points of Interest	People profile and Ranking
M-Guide: Hybrid Recommender System Tourism in East-Timor [25]	East Timor	Natural Tour, Historic, Religious and Cultural.	People profile and Community Information
Intelligence Traveling Schedule Recommender Based on Commonsense Reasoning Algorithm [26]	China	General Tourism	People profile
Data gathering system for recommender system in tourism [27]	Japan	General Points of Interest	People profile

(continúa)

(viene)

Recommender System	Destination	Tourism Typology	Attributes
Intelligent travel recommendation system by mining attributes from community contributed photos [28]	India	General Tourism	People profile, Face Numbers and Season
A Cocktail Approach for Travel Package Recommendation [29]	General	General Tourism	People profile, Group profile and Season
A Collaborative Awareness Framework for [30]	Malaysia	Ethnic, cultural, Historical, environmental, recreational and business tourism	People profile and tourist Location
A Context-Aware Proactive Recommender System (CAPTRS) for Tourist [31]	Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Roma, Tuscany	General Points of Interest, Gastronomic Tourism	Location, weather, season and date
Comparison of Various Metrics Used in Collaborative Filtering for Recommendation System [32]	Movie lens	Movies	Accuracy, Sensitivity (Recall), Precision, Matthews Correlation Coefficient (MCC)
Re-rank Algorithm for Online Hotel Search [33]	Italia	General Points of Interest	Price and location
Recommendation of Tourist Attractions Based on Slope One Algorithm [34]	Ilvping	Domestic tourism	The features are used to compute similarity between items
Recommending the duration of Stay I Personalized Travel Recommender Systems [1]	Germany	General Points of Interest	Geolocation
A personalized hybrid tourism recommender system [24]	Paris	General Points of Interest	User information, activities, ratings
A hybrid location-based recommender system for trip planning and mobility using geo-tagged media [35]	India	General Points of Interest	Geolocation

Source: Own work.

A recommender system was also identified that can be fed using data collected from mobile devices, which can provide valuable information to the system [31]. In this configuration, the author uses the visitor's location as the primary factor to identify points of interest. However, additional factors are suggested for consideration, including climate, time, user characteristics, and travel season. The model is validated by selecting a specific area in Paris and feeding the system with a set of restaurants, confirming its proper operation.

Another case describes a recommender system that uses a hybrid algorithm, combining content-based filtering and collaborative filtering to generate recommendations based on user profiles and ratings of different points of interest in Paris [24]. The results show that combining both approaches improves accuracy compared to

using a single method. However, the author notes that system behavior may change across regions because of differences in user profiles.

Of the 18 works reviewed, 12 were developed in Asia or Europe, indicating strong research activity in those regions. In contrast, the American continent, with the exception of a single study conducted in the United States, shows limited documented development in this area. This suggests significant opportunities for further research and advancement in the Americas.

Most of these studies were developed using supervised or semi-supervised approaches, which limits their applicability in contexts where labeled data or user interaction are unavailable. A gap remains in the design of fully unsupervised hybrid models that explicitly integrate semantic and geographic components.

Finally, a bibliometric analysis was conducted using the VOSviewer software to identify research trends in tourism related to the keywords "Tourism," "Machine Learning," "Recommender Systems," and "Web Scraping." The analysis shows a substantial number of studies applying machine learning and recommender systems in tourism. However, there is a clear lack of research specifically addressing the use of Web Scraping in this domain, highlighting the need to develop approaches that incorporate Web Scraping to improve data acquisition in the tourism sector.

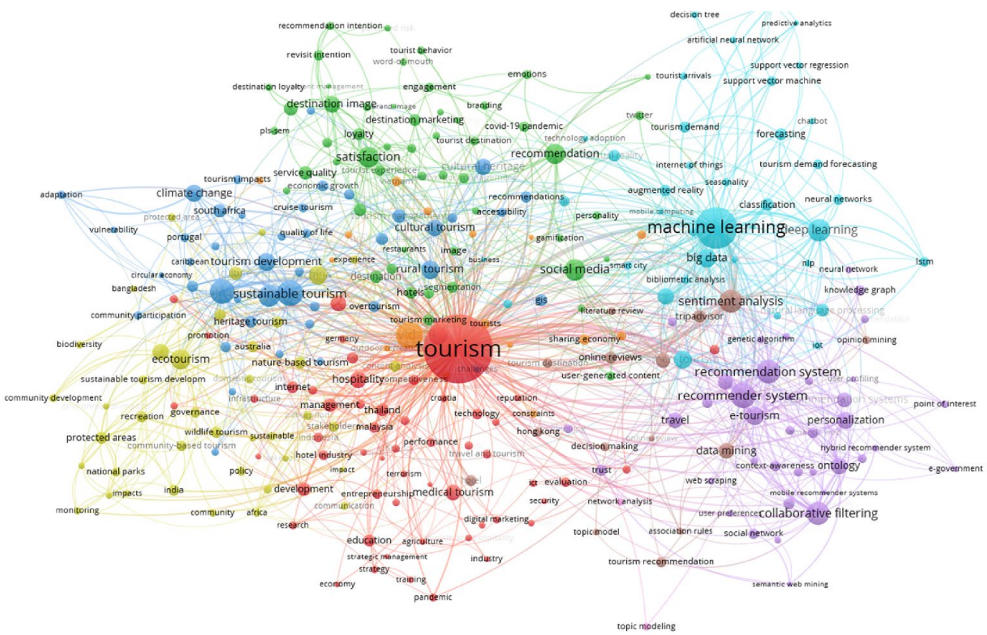

Fig. 1. Bibliometric network analysis using VOSviewer.
Source: Own work.

In summary, prior tourism recommender systems predominantly rely on collaborative, contextual, or user-driven information. While some studies incorporate either semantic similarity or geographic relevance, very few explicitly integrate both dimensions under a fully unsupervised framework, and none of the reviewed studies use exclusively web-scraped accommodation data as the primary source of recommendation knowledge.

2. MATERIALS AND METHODS

This research follows a quantitative, experimental, and evaluative design oriented toward developing and validating a tourist accommodation recommender system under conditions of incomplete user information. The quantitative component is based on representing accommodations as numerical feature vectors, applying standardization and dimensionality reduction techniques, and analyzing objective metrics such as semantic coherence, ranking stability, and geographic proximity. The experimental component involves building two comparable system configurations and calibrating the weighting parameter of the hybrid model. The evaluative component is reflected in the definition of an unsupervised evaluation framework, aligned with the lack of ground truth derived from real user interactions.

The problem addressed corresponds to a cold-start user scenario, where no browsing history, bookings, clicks, or prior ratings are available. Under these conditions, the system cannot depend on traditional collaborative filtering approaches or supervised relevance prediction strategies. The recommendation process is formulated as an item-to-item similarity retrieval problem, where the user's most recent interaction is treated as the query item, and recommendations are generated based on structural and spatial similarity with the remaining accommodations.

From a methodological perspective, the system follows a sequential pipeline: (i) automated data acquisition from Booking.com; (ii) data cleaning, structuring, and feature enrichment; (iii) standardization and vector representation of accommodations; (iv) dimensionality reduction using Principal Component Analysis (PCA); (v) computation of semantic similarity between items; (vi) calculation of geographic distance and its transformation into proximity; (vii) integration of both signals into a hybrid scoring function; and (viii) evaluation of model performance using indirect metrics. This architecture ensures full traceability from data acquisition to the final ranking generation.

PCA was included as an experimental dimensionality reduction configuration to assess whether a compressed representation could preserve recommendation

quality. The final model selection was based on the comparison between the original standardized representation and the PCA-based representation.

Formally, let i_q denote the accommodation queried by the user, and let $I = \{i_1, i_2, \ldots, i_n\}$ be the set of available accommodations. The objective is to define a scoring function $s(i_q, i_j)$ that ranks items according to their relevance with respect to the query item. Relevance is not directly observed as labeled data but is approximated through two latent dimensions: attribute similarity and spatial proximity. The final ranking is therefore obtained as an ordering induced by a composite scoring function.

This approach is well suited to tourism applications, where the selection of an accommodation depends not only on its internal features but also on its relative location within the destination [36]. In other words, two accommodations may be semantically similar in terms of facilities, ratings, and overall profile, but differ significantly in the user's experience if they are located in distant areas. Geographic information is therefore treated not as a descriptive variable but as an explicit component of the recommendation mechanism.

2.1. Data Acquisition and Preprocessing

The primary data source was Booking.com, selected for its extensive coverage, rich descriptive content, and public availability of relevant attributes. The scraping process was conducted for academic purposes, with no attempt to bypass authentication, access restricted data, or extract personally identifiable information. The use of web-scraped data is discussed as a methodological limitation due to platform terms, structural variability, and reproducibility constraints.

Data collection was performed through automated web Scraping using the Playwright library of python, chosen for its ability to interact with dynamic content, simulate real browsing behavior, and handle complex HTML structures across browsers [37]. This approach addresses the limited availability of structured tourism datasets and uses Playwright to ensure stable automated extraction [38].

The data collection process was conducted in multiple extraction stages between June and September 2023, during which several scraping sessions were performed to progressively increase coverage and improve data completeness. An additional update was carried out in December 2024 to validate extraction consistency and verify structural stability of the source platform.

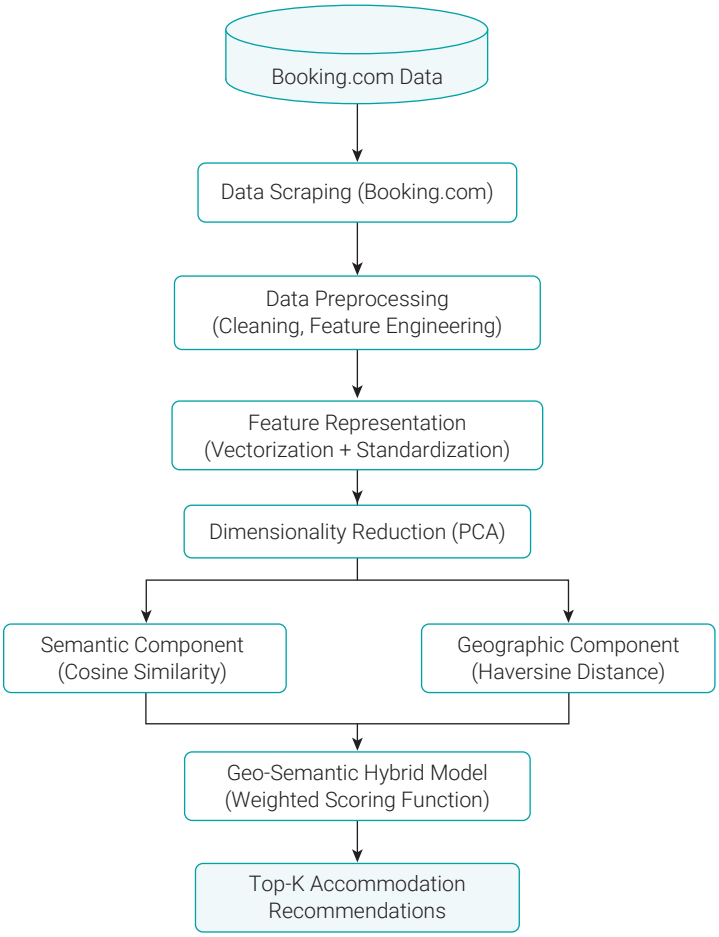

Fig. 2. Recommender System Architecture.
Source: Own work.

The extraction process was carried out in two stages. First, individual accommodation links were collected from result pages segmented by regions and subregions in Colombia. This segmentation was required because of platform limitations on the number of visible results per query, making it necessary to partition the search space by city, region, and, in some cases, star rating. In the second stage, attributes were extracted from each individual URL, including name, address, region, geographic coordinates, list of facilities, ratings by specific category, overall rating, number of reviews, and URL as identifier. The initial Scraping process produced more than nineteen thousand records, which were later refined to remove duplicates, as a result so the final dataset contained approximately 13,000 accommodations.

From a data quality perspective, the Scraping process introduced three main challenges: structural heterogeneity across pages, missing values due to absent

elements in non-standardized listings, and redundancy caused by duplicates from segmented navigation. To address these issues, preprocessing included a semantic restructuring phase of the attributes.

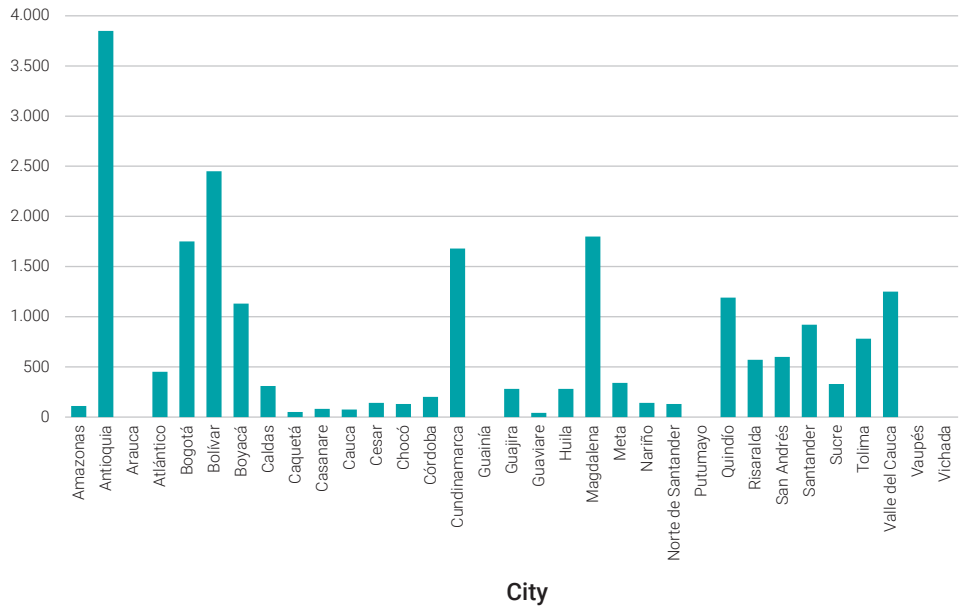

Fig. 3. Number of scraped accommodations by region.
Source: Own work.

The facilities field, originally stored as textual lists, was transformed into a binary feature space using one hot encoding. This converts unstructured descriptions into an algebraically tractable representation. If a facility f is present in accommodation i, the corresponding binary variable takes the value 1; otherwise, it is 0. Formally, if F represents the set of unique facilities, each accommodation is represented by a binary subvector.

Similarly, the composite variable of ratings based on was decomposed into explicit dimensions: Staff, Facilities, Cleanliness, Comfort, Value for Money, Location, and Free WiFi. This transformation increases the granularity of the model by separating general evaluation from specific experience dimensions.

A relevant methodological aspect is the use of meaningful imputation, where missing information is modeled as an explicit variable rather than replaced with artificial values that could distort the dataset structure. In addition, normalization techniques were applied, including Min-Max scaling for ratings and Z-score standardization for continuous variables, ensuring comparability across dimensions.

2.2. Semantic component: cosine similarity in vector spaces

The semantic core of the model represents each accommodation as a vector in a feature space. Let $\mathbf{X}a$ and $\mathbf{X}b$ denote two accommodations, represented as:

$$X_a = (x_{a1}, x_{a1}, \ldots . x_{ap})$$
$$X_b = (x_{b1}, x_{b1}, \ldots . x_{bp})$$

Each dimension encodes an accommodation attribute, such as a binary facility, a scaled rating, a transformed geographic variable, or a principal component in the reduced version of the model. Content-based recommendation is designed to estimate the degree of similarity between such vectors[39]. In particular, Vector representation enables the use of similarity metrics such as cosine similarity, which measures the relationship between accommodations based on their characteristics [40]. When data is represented as vectors in a feature space, different mathematical measures can be applied to quantify similarity between objects. Among these, cosine similarity is widely used because it evaluates the angle between vectors in a multidimensional space[41].

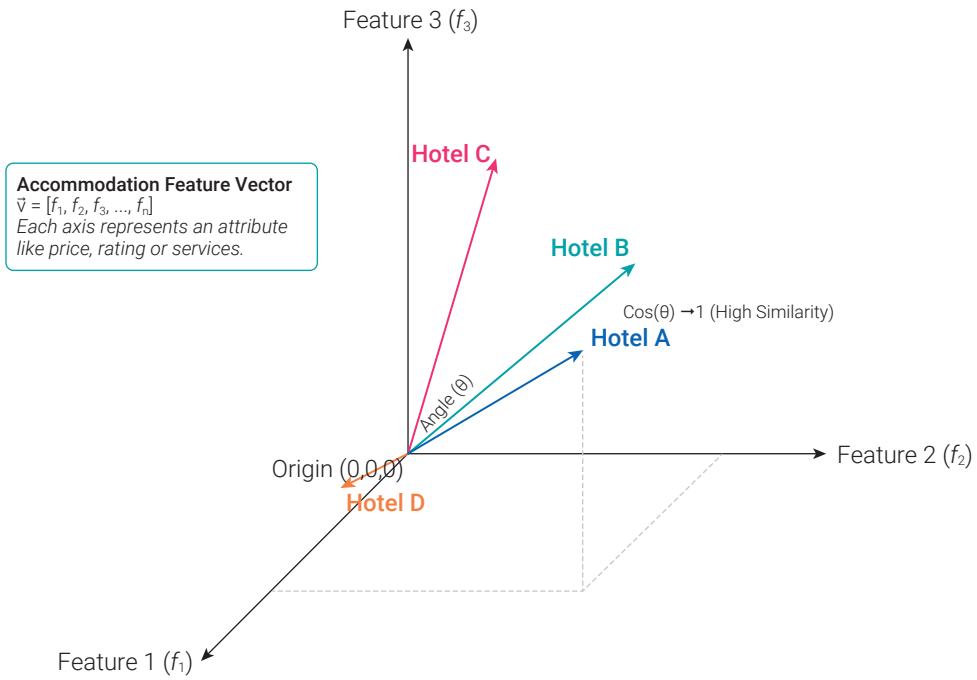

Fig. 4. Vector representation of accommodations.
Source: Own work.

The selected metric is cosine similarity, commonly used in recommender systems and information retrieval when objects are represented in high-dimensional vector spaces. Its formulation is [39]:

$$Sim(A,B) = \frac{A_i * B_i}{//A_i// * //B_i//}$$

This metric captures similarity in terms of orientation, making it particularly suitable for normalized, high-dimensional spaces where vector magnitude may vary without altering semantic structure. Here, $A_i * B_i$ is the dot product and $//A_i//$ is the Euclidean norm of the vector. Cosine similarity approaches 1 when the angle between vectors is small, indicating that both share a similar orientation in the attribute space.

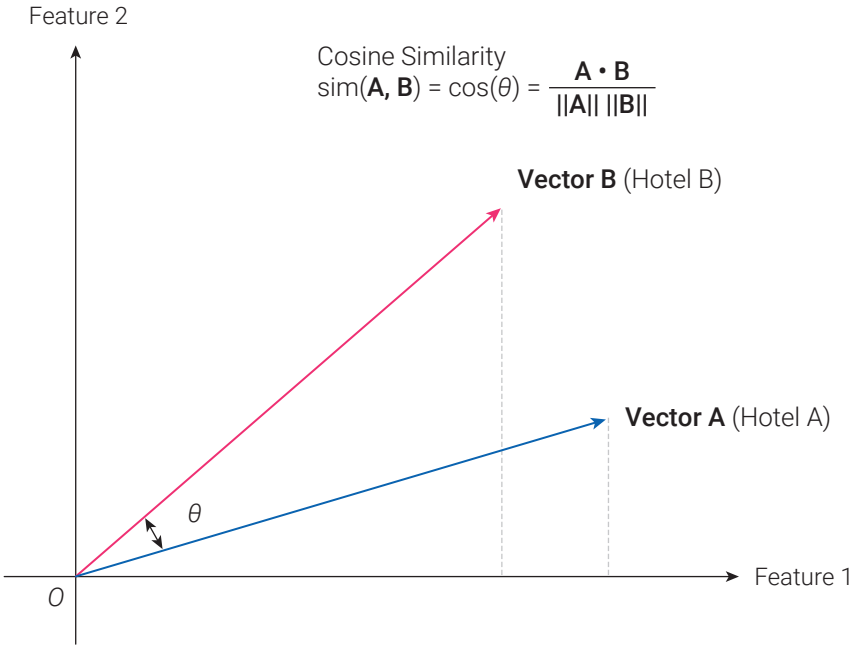

Fig. 5. Cosine similarity in the feature space.
Source: Own work.

The choice of this metric is particularly well suited to the problem. Unlike pure distance measures, cosine similarity emphasizes the relative orientation between vectors rather than their absolute magnitude [18]. This is relevant because two accommodations may differ in the intensity of certain attributes while still sharing a similar structural pattern in terms of services, ratings, and general profile. Cosine similarity

captures similarity in shape, not just metric proximity. This property is especially useful in heterogeneous tourism data, where binary and semi-continuous variables coexist after preprocessing.

2.3. Geographic component: Haversine distance and proximity score

The second dimension of the model captures spatial proximity between accommodations. Let the query accommodation i_q have coordinates $(\phi_q\,,\ \lambda_q)$ and a candidate accommodation i_j have coordinates $(\phi_j\,,\ \lambda_j)$, where ϕ represents latitude y λ longitude. To calculate the Geographic distance is computed using the Haversine formula, which estimates great-circle distance between two points on the Earth's surface [42]:

$$d = 2R * \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Once the distance is obtained, the system does not use it directly as a score. After computing the Haversine distance, the geographic distance is transformed into a normalized proximity score in order to make it compatible with the semantic similarity component. Since cosine similarity increases when two accommodations are more similar, the geographic component must follow the same direction: higher values must indicate greater geographic relevance.

The hybrid model requires both components to follow the same semantic direction, where higher values indicate greater relevance [43]. In addition to the proximity transformation, a geographic constraint radius of 1 km was considered to evaluate whether the recommended accommodations remained within this distance.

A geographic constraint radius of 1 km was adopted as a pragmatic walkability threshold for urban tourism scenarios. This decision was theoretically informed by the "15-minute city" framework, which proposes that essential urban services and destinations should be reachable within approximately 15 minutes on foot. Under typical urban mobility conditions, this corresponds approximately to a radius of 1 km. Although this threshold may vary depending on the rural or metropolitan context, it provides an interpretable and conservative benchmark for evaluating geographic coherence in Colombian tourist accommodation scenarios [44].

2.4. GEO-SEMANTIC HYBRID MODEL

Once semantic similarity and geographic proximity are defined, both signals are integrated into a hybrid scoring function. The model follows a weighted linear combination:

$$Score(i,j) = \propto Sim(i,j) + (1-\propto)Prox(i,j)$$

where:

- $Sim(i,j)$ is the cosine similarity between accommodations.
- $Prox(i,j)$ is the proximity derived from geographic distance
- $\propto$ is a weighting parameter controlling the relative contribution of each component

This formulation is appropriate for several reasons. First, both signals are bounded approximately within the interval [0,1], enabling an interpretable linear combination. Second the weighting structure supports sensitivity analysis: when α=1, the model reduces to a purely content-based approach but when α=0, it becomes strictly geographic. The convex combination also avoids discontinuities and allows controlled analysis of the trade-off between attribute similarity and spatial proximity [45].

The parameter α was calibrated through sensitivity analysis by evaluating different values and observing the behavior of average similarity and mean distance in the recommendations.

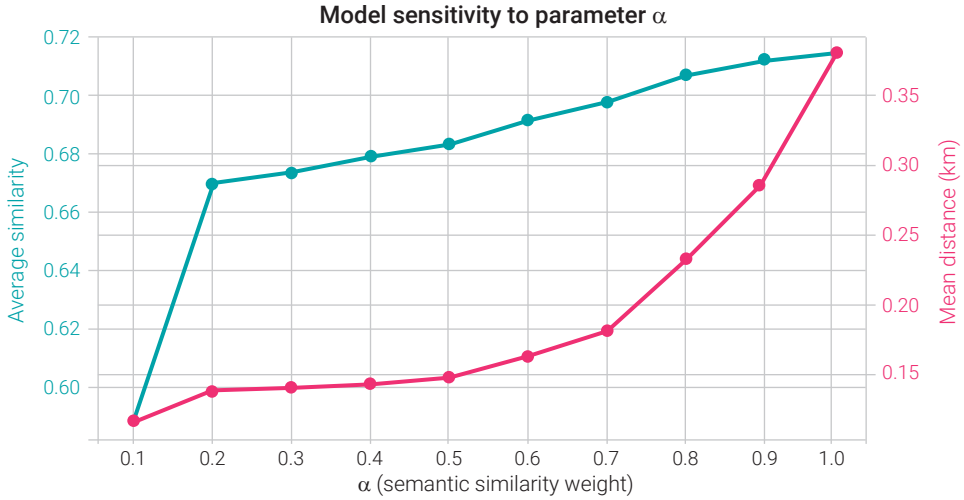

Fig. 6. Sensitivity analysis of parameter α.
Source: Own work.

e-ISSN 2357-6014 / Vol. 22, no. 3 / september-december 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

The selected operating point was α=0.7, as it provides the best empirical balance between both dimensions. Under this configuration, the system maintains high semantic coherence while preserving geographic constraints.

2.5. Unsupervised evaluation framework

The evaluation follows an unsupervised framework because of the lack of ground truth derived from real user interactions. This is a key methodological constraint. Metrics such as Precision@K, Recall@K, F1, RMSE, or MAE require prior knowledge of relevant items or observed user feedback such as clicks, bookings, or ratings. Without these signals, their application is not valid [46], [47], [48].

The evaluation focuses on structural properties of system behavior: semantic coherence, ranking stability, and internal consistency of recommendations. This approach is consistent with cold-start and content-based scenarios, where the goal is not to predict observed user choices but to generate recommendations that are semantically and spatially coherent.

2.5.1. Overlap@K

To evaluate stability across model configurations, Overlap@K is used to measure the proportion of shared items between two Top-*K* recommendation sets. If A and B denote the recommendation sets from two configurations, then [16]:

$$Overlap@k = \frac{/A \cap B/}{K}$$

Its value ranges from 0 to 1. Higher values indicate that both rankings share a large portion of their recommendations, regardless of order.

2.5.2. Similarity Standard Deviation

Semantic coherence depends not only on high average similarity but also on low dispersion within the recommended set. To capture this variability, the standard deviation of cosine similarity is computed [39]:

$$std_{sim} = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(S_j - S)^2}$$

2.5.3. Intra-list similarity

Intra-list similarity evaluates internal consistency within the recommendation set, measuring similarity among the recommended items themselves rather than with the query item [49].

$$ILS = \frac{2}{K(K-1)}\sum_{i<j} sim(i,j)$$

2.6. METHODOLOGICAL AND ETHICAL CONSIDERATIONS

Data collection was performed exclusively using publicly available information from digital platforms, without involving sensitive personal data or user identification. The model is designed to improve access to relevant information for decision-making, aligned with the principles of transparency and responsible data use [2].

3. RESULTS

3.1. CHARACTERIZATION OF THE FINAL DATASET

The Web Scraping process initially collected 19,660 accommodation URLs across multiple regions in Colombia. After cleaning and duplicate removal, the dataset was reduced to 13,204 records, forming the basis for model development. The initial vector representation consisted of 467 descriptive variables, later reduced to 295 principal components based on a 90% cumulative explained variance threshold.

TABLE II. BOOKING WEB SCRAPING RESULTS

Variable	Valor
Total number of accommodations	13,204
Number of original variables	467
Numerical variables	9
Transformed categorical variables	458
Geographic variables (lat, lon)	2
Derived variables (imputation)	> 20
Duplicate records removed	yes

Source: Own work.

From a structural perspective, the dataset integrates three types of variables:

I. Perceived quality variables, such as overall rating and category-based ratings.
II. Binary variables derived from facility expansion.
III. Spatial and accessibility variables, including latitude, longitude, and distance to the airport.

This structure defines a high-dimensional recommendation problem, where system performance depends on preserving both semantic information and geographic context. These descriptive results highlight that the model was not built on a curated or idealized dataset but on a large, heterogeneous, real-world data source with the inherent complexity of web-extracted information. Consequently, model performance should be interpreted within the context of a rich but imperfect dataset, reinforcing the applied validity of the study.

3.2. IMPACT OF DIMENSIONALITY REDUCTION USING PCA

To evaluate the effect of dimensionality reduction on recommendation quality, PCA was analyzed as an ablation experiment. The objective was not to assume that the reduced representation would improve the model, but to determine whether compressing the original feature space preserved the semantic structure required for item-to-item recommendation. For this purpose, two configurations were compared: a base model using the original standardized variables and a PCA-based model using 295 principal components, corresponding to 90% of the cumulative explained variance.

TABLE III. SEMANTIC COHERENCE COMPARISON: BASE MODEL VS. PCA

Metrics	Model Base	Model PCA
Average similarity	0.868	0.521
Median	0.865	0.491
Dispersion (std_sim)	0.025	0.120
Intra-list similarity	0.831	0.344

Source: Own work.

The results show that PCA reduced the dimensionality of the feature space but negatively affected the semantic coherence of the recommendations. The base model achieved an average cosine similarity of 0.868 and a median of 0.865, whereas the PCA-based model decreased to an average similarity of 0.521 and a median of 0.491. Similarly, the standard deviation of similarity increased from 0.025 in the base model to 0.120 in the PCA-based model, indicating greater dispersion and lower stability in the reduced representation. Intra-list similarity also decreased substantially, from 0.831 in the base model to 0.344 in the PCA-based model. These results indicate that the PCA transformation weakened the local semantic relationships relevant for recommendation, even though it retained 90% of the global variance.

Therefore, PCA should be interpreted as a controlled negative finding rather than as an improvement to the model. Although dimensionality reduction provides a more compact representation, it does not necessarily preserve the fine-grained co-occurrence patterns among facilities, ratings, and accommodation attributes that are essential for similarity-based recommendation. Based on these results, the original standardized feature representation was retained as the preferred configuration for the final geo-semantic recommender.

3.3. RESULTS OF THE GEO-SEMANTIC HYBRID MODEL

After integrating the geographic component through Haversine distance and transforming it into a proximity score, the system achieved an average geographic distance of approximately 0.16 km in its recommendations. This confirms strong spatial proximity between the query accommodation and the recommended set. This result should be interpreted alongside the average similarity of 0.868 from the base model, showing that the system recommends accommodations that are both similar and geographically close.

Fig. 7. Geo-semantic hybrid model results.
Source: Own work.

TABLE IV. INTEGRATED RESULTS OF THE HYBRID MODEL

Metric	Value
Average cosine similarity	0.868
Overlap@5	~0.50
Average geographic distance	~0.16 km
α optimal	0.7
Average distance with (α =0.7)	0.18 km
Geographic constraint compliance	100 %

Source: Own work.

Sensitivity analysis of the parameter α shows that increasing the weight of the semantic component raises the average similarity but also leads to a gradual increase in the average geographic distance. In this context, the value α = 0.7 was identified as the most suitable operating point, with an average similarity close to 0.7 and an average distance of 0.18 km, while maintaining 100% of recommendations within the defined geographic radius.

Illustrative case study: Zamia Hostal

As an external exploratory reference, a comparison was conducted between the recommendations generated by the proposed model and the publicly visible suggestions displayed by Booking.com for the same query accommodation. This comparison was not designed as a supervised benchmark, since Booking.com uses proprietary ranking mechanisms and may incorporate commercial, behavioral, availability, pricing, and personalization signals that are not observable in this study. Therefore, the objective of this analysis is not to claim superiority over Booking.com but to examine how the proposed geo-semantic model behaves in relation to the output of a commercial platform.

A specific case is documented using the accommodation Zamia Hostal. For this query, the proposed model generated recommendations such as Sachar Lodging Apartahotel, Casa Mendoza, Hostal Antiguo Campestre, Palonegro Bucaramanga, and Quinta de Cabecera, with similarities ranging from 0.7547 to 0.7921, distances between 0.0968 km and 0.5075 km, and hybrid scores between 0.7435 and 0.8018. As shown in the following figure, Booking.com recommends different accommodations, including "San Lorenzo Bucaramanga," Urban Hostal Bar, and "La Bonita Guesthouse." The accommodation names are presented in Spanish because the study focuses on recommendations within Colombia.

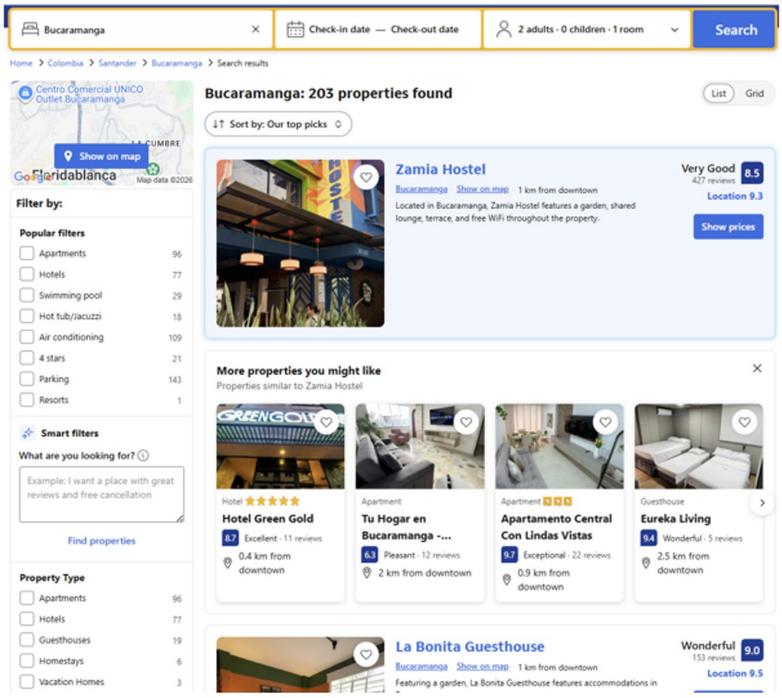

Fig. 8. Booking.com recommendation results.
Source: Screenshot captured by the authors from Booking.com, 2026.

An extended exploratory comparison was performed using 30 query accommodations selected through simple random sampling from the complete dataset, ensuring that each accommodation had the same probability of being included and reducing potential selection bias. For each selected accommodation, the Top-5 recommendations generated by the proposed geo-semantic model were compared with the publicly visible suggestions displayed by Booking.com. The comparison considered cosine similarity, geographic distance, and the proposed hybrid score as descriptive indicators of semantic and spatial consistency.

Across the analyzed cases, the proposed model produced recommendations with cosine similarity values ranging from 0.7547 to 0.7921, geographic distances between 0.0968 km and 0.5075 km, and hybrid scores between 0.7435 and 0.8018. In contrast, the Booking.com suggestions showed cosine similarity values between 0.701 and 0.807, but with larger geographic distances, ranging from 1.31 km to 2.50 km, and lower hybrid scores between 0.580 and 0.695.

TABLE V. EXPLORATORY COMPARISON

Metric	Proposed Model	Booking.com
Number of Accommodations	0.868	30
Cosine Similarity Range	0.75 – 0.7921	0.701 – 0.807
Geographic distance Range (km)	0.09 – 0.50	1.31 – 2.50
Hybrid Score	0.7435 - 0.8018	0.580 – 0.695

Source: Own work.

These results suggest that, under the proposed geo-semantic scoring criteria, the model tends to prioritize geographically closer accommodations while maintaining comparable semantic similarity. However, this comparison should be interpreted as an exploratory external reference rather than a conclusive benchmark, since Booking.com uses proprietary ranking mechanisms and may incorporate commercial, behavioral, pricing, availability, and personalization signals that were not observable in this study.

4. CONCLUSIONS

The results indicate that the proposed geo-semantic hybrid model provides a consistent and effective solution to the problem of recommending tourist accommodations in cold-start scenarios [11]. In summary, the study evaluates whether it is possible

to implement a recommender system based on web data extraction and machine learning, using the user's most recent interaction as the starting point. The findings support this premise: combining vector representation, cosine-based semantic similarity, and spatial proximity through the Haversine distance enables the generation of recommendations that are both semantically coherent and geographically consistent. This outcome aligns with the main objective and confirms the technical feasibility of the approach.

Prior work in tourism has shown the usefulness of hybrid models, particularly those that integrate contextual information, item attributes, or user preferences [39]. However, most of these approaches depend on user profiles, historical interactions, or explicit relevance signals. In contrast, the approach developed here is based on an unsupervised item-to-item similarity framework, operating without ground truth, without observed interactions, and using heterogeneous data obtained through Web Scraping. The contribution lies not only in combining two dimensions but in doing so under these constraints.

A central aspect of the model is the integration of semantic and geographic dimensions within a single scoring function. In tourism, recommendation quality depends not only on similarity in attributes but also on spatial coherence. Post-pandemic studies suggest that relevant options should remain within a threshold of approximately 1 km [44]. The results show that the geographic component significantly reshapes the ranking, guiding semantically similar results toward spatially appropriate alternatives. This behavior is evident in the sensitivity analysis of parameter α: higher values prioritize attribute similarity, while lower values emphasize spatial proximity. The choice of $\alpha = 0.7$ provides a balanced trade-off.

The PCA-based representation was evaluated as an ablation experiment to determine whether dimensionality reduction could preserve the semantic structure of the recommendation space [50]. Although PCA retained 90% of the cumulative explained variance, it reduced average cosine similarity and intra-list similarity while increasing semantic dispersion. Therefore, the original standardized feature representation was retained as the preferred configuration for the final model.

The use of cosine similarity as the core of the semantic component is supported by the results. In a feature space combining binary variables from facilities, scaled ratings, and indicators of missing information, cosine similarity offers clear advantages by focusing on relative orientation rather than absolute magnitude [51]. This makes it suitable for capturing structural similarity patterns between accommodations. Although metrics such as Mahalanobis distance may appear theoretically appealing due to their use of covariance structure [52], they are more sensitive to noise,

multicollinearity, and sparsity associated with scraped data and expanded categorical features. In practice, cosine similarity proves to be more stable and interpretable for this representation.

The exploratory comparison with Booking.com provides a useful external reference for interpreting the behavior of the proposed model. However, this comparison should not be understood as a conclusive benchmark or as evidence of superiority over a commercial platform, since Booking.com relies on proprietary ranking mechanisms and may incorporate variables that were not available in this study, such as pricing, availability, user behavior, popularity, conversion signals, and commercial positioning [6]. Within this limitation, the comparison suggests that the proposed model can generate geographically closer recommendations while maintaining semantic coherence under the defined geo-semantic scoring criteria.

Another significant aspect is the evaluation framework. In the absence of ground truth, applying metrics such as precision, recall, RMSE, or MAE is not appropriate [16], [48]. Therefore, the study instead uses structural metrics such as Overlap@K, intra-list similarity, and the standard deviation of similarity to assess consistency, stability, and internal coherence. This approach is aligned with cold-start and content-based scenarios and is consistent with the literature on the offline evaluation of recommender systems. These metrics, however, only capture internal system properties and do not reflect perceived usefulness or user satisfaction [39].

Regarding the limitations, accommodations located within the same building, block, or highly concentrated urban zones may naturally share similar facilities and ratings, which could partially inflate cosine similarity and intra-list similarity values. Therefore, these metrics should be interpreted as indicators of internal coherence rather than direct evidence of recommendation superiority. The main limitation is the absence of ground truth, which prevents direct validation against real user preferences. The use of Web Scraping introduces additional challenges, including noise, structural variability, and missing values. The geographic representation, based on great-circle distance and a fixed radius, enables a clear and simple integration but does not account for more complex contextual variables [42].

In summary, the results support a clear conclusion: the proposed hybrid model effectively integrates semantic and geomatic dimensions in a cold-start scenario, using data obtained through Web Scraping and operating without user history or labeled relevance data. Tourist accommodation recommendation can be formulated as a combination of structural similarity and spatial proximity, even under limited information conditions. The study also confirms the suitability of unsupervised descriptive

metrics for evaluating this type of system, provided they are interpreted as indicators of coherence, stability, and plausibility.

The main contribution of this research is the design and evaluation of a transparent geo-semantic recommendation framework for tourist accommodation under cold-start and data-scarce conditions. Unlike approaches that depend on user profiles, historical interactions, or supervised relevance labels, the proposed model operates on web-scraped accommodation attributes and generates recommendations through an explicit combination of semantic similarity and spatial proximity. Therefore, the value of the study lies not only in the use of cosine similarity and the Haversine distance, but also in the methodological integration of data acquisition, preprocessing, vector representation, geo-semantic ranking, and unsupervised evaluation within a single reproducible pipeline.

REFERENCES

[1] J. Neidhardt, W. Wörndl, T. Kuflik, D. Goldenberg, and M. Zanker, "Workshop on Recommenders in Tourism (RecTour)," in *Proc. 16th ACM Conference on Recommender Systems (RecSys)*, Seattle, WA, USA, Sep. 2022, pp. 678–679, doi: https://doi.org/10.1145/3523227.3547416.

[2] V. Krotov, L. Johnson, and L. Silva, "Legality and ethics of web scraping," *Communications of the Association for Information Systems*, vol. 47, Art. no. 24, pp. 539–563, 2020, doi: https://doi.org/10.17705/1CAIS.04724.

[3] S. Guardia, "Ensaio sobre destinos turísticos inteligentes," *Journal of Tourism & Development*, no. 27/28, pp. 1515–1528, 2017, doi: https://doi.org/10.34624/rtd.v1i27/28.9897.

[4] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015, doi: https://doi.org/10.1016/j.eij.2015.06.005.

[5] L. Celardo, M. Misuraca, and M. Spano, "Geo-referenced sentiment analysis for tourists' points of interest: The case of Matera European Capital of Culture," *Rivista Italiana di Economia Demografia e Statistica*, vol. 78, no. 2, pp. 103–114, 2024, doi: https://doi.org/10.71014/sieds.v78i2.258.

[6] N. Thi Diem Kieu and B. Trong Tien Bao, "Determinants of tourist intentions to use mobile applications for accommodation booking: Insights from Ho Chi Minh City," *Journal of Tourism & Development*, no. 49, pp. 141–170, 2025, doi: https://doi.org/10.34624/rtd.v49i0.37828.

[7] M. Sumardi, Jufery, Frenky, R. Wongso, and F. A. Luwinda, "'tripBuddy' travel planner with recommendation based on user's browsing behaviour," *Procedia Computer Science*, vol. 116, pp. 326–333, 2017, doi: https://doi.org/10.1016/j.procs.2017.10.084.

[8] T. Bomatpalli, "A survey of travel recommender system," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 3, pp. 356–362, 2019, doi: https://doi.org/10.26438/ijcse/v7i3.356362.

[9] J. Borràs, A. Moreno, and A. Valls, "Intelligent tourism recommender systems: A survey," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7370–7389, 2014, doi: https://doi.org/10.1016/j.eswa.2014.06.007.

[10] M. K. Najafabadi, A. H. Mohamed, and M. N. Mahrin, "A survey on data mining techniques in recommender systems," *Soft Computing*, vol. 23, no. 2, pp. 627–654, 2019, doi: https://doi.org/10.1007/s00500-017-2918-7.

[11] O. A. S. Ibrahim, E. M. G. Younis, E. A. Mohamed, and W. N. Ismail, "Revisiting recommender systems: An investigative survey," *Neural Computing and Applications*, 2025, doi: https://doi.org/10.1007/s00521-024-10828-5.

[12] A. Cavallo, "Scraped data and sticky prices," *Review of Economics and Statistics*, vol. 100, no. 1, pp. 105–119, 2018, doi: https://doi.org/10.1162/REST_a_00652.

[13] R. N. Landers, R. C. Brusso, K. J. Cavanaugh, and A. B. Collmus, "A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research," *Psychological Methods*, vol. 21, no. 4, pp. 475–492, 2016, doi: https://doi.org/10.1037/met0000081.

[14] D. Kaur, S. Khanna, and D. Aulakh, "A collaborative location-based travel recommendation system through enhanced rating prediction for the group of users," *Structural Chemistry*, vol. 24, no. 1, pp. 357–367, 2013, doi: https://doi.org/10.1007/s11224-012-0084-1. **(SD: verificar correspondencia entre revista, título y DOI).**

[15] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015, doi: https://doi.org/10.1016/j.eij.2015.06.005.

[16] E. Zangerle and C. Bauer, "Evaluating recommender systems: Survey and framework," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2023, doi: https://doi.org/10.1145/3556536.

[17] A. Solano-Barliza *et al*., "Recommender systems applied to the tourism industry: A literature review," *Cogent Business & Management*, vol. 11, no. 1, 2024, doi: https://doi.org/10.1080/23311975.2024.2367088.

[18] C. Troussas, A. Krouska, P. Tselenti, D. K. Kardaras, and S. Barbounaki, "Enhancing personalized educational content recommendation through cosine similarity-based knowledge graphs and contextual signals," *Information*, vol. 14, no. 9, Art. no. 505, 2023, doi: https://doi.org/10.3390/info14090505.

[19] H. Simanjuntak, D. Tarigan, I. Sibarani, C. J. Hutapea, R. Lumbantoruan, and M. Sigiro, "Weighted hybrid recommendation system for Toba tourism based on Google Review data," in *Proc. IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM)*, Medan, Indonesia, Oct. 2022, pp. 1–8, doi: https://doi.org/10.1109/ICOSNIKOM56551.2022.10034911.

[20] I. Y. Choi, J. K. Kim, and Y. U. Ryu, "A two-tiered recommender system for tourism product recommendations," in *Proc. Annual Hawaii International Conference on System Sciences (HICSS)*, Kauai, HI, USA, 2015, pp. 3354–3363, doi: https://doi.org/10.1109/HICSS.2015.405.

[21] N. Wijaya and A. Furqan, "Coastal tourism and climate-related disasters in an archipelago country of Indonesia: Tourists' perspective," *Procedia Engineering*, vol. 212, pp. 535–542, 2018, doi: https://doi.org/10.1016/j.proeng.2018.01.069.

[22] A. Umanets, A. Ferreira, and N. Leite, "GuideMe – A tourist guide with a recommender system and social interaction," *Procedia Technology*, vol. 17, pp. 407–414, 2014, doi: https://doi.org/10.1016/j.protcy.2014.10.248.

[23] T. Izumi, T. Kitamura, and Y. Nakatani, "A suggestive recommendation method to make tourists 'feel like going,'" *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 573–578, 2016, doi: https://doi.org/10.1016/j.ifacol.2016.10.606.

[24] M. E. B. H. Kbaier, H. Masri, and S. Krichen, "A personalized hybrid tourism recommender system," in *Proc. IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia, 2017, pp. 244–250, doi: https://doi.org/10.1109/AICCSA.2017.12.

[25] J. D. C. L. Soares, Suyoto, and A. J. Santoso, "M-guide: Hybrid recommender system tourism in East Timor," in *Proc. International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)*, Bali, Indonesia, 2017, pp. 303–309, doi: https://doi.org/10.1109/ICSIIT.2017.16.

[26] C. S. Wang, C. Y. Li, and C. C. Yeh, "Intelligence traveling schedule recommender based on commonsense reasoning algorithm," in *Proc. International Conference on Computer and Communication Engineering (ICCCE)*, Kuala Lumpur, Malaysia, 2010, pp. 1–5, doi: https://doi.org/10.1109/ICCCE.2010.5556814.

[27] G. Hirakawa, G. Satoh, K. Hisazumi, and Y. Shibata, "Data gathering system for recommender system in tourism," in *Proc. 18th International Conference on Network-Based Information Systems (NBiS)*, Taipei, Taiwan, 2015, pp. 521–525, doi: https://doi.org/10.1109/NBiS.2015.78.

[28] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Intelligent travel recommendation system by mining attributes from community contributed photos," *Procedia Computer Science*, vol. 50, pp. 447–455, 2015, doi: https://doi.org/10.1016/j.procs.2015.04.014.

[29] Q. Liu, E. Chen, H. Xiong, Y. Ge, Z. Li, and X. Wu, "A cocktail approach for travel package recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 278–293, 2014, doi: https://doi.org/10.1109/TKDE.2012.233.

[30] L. T. Yong, "A collaborative awareness framework for mobile tourist recommender system," in *Proc. 3rd International Conference on Computer Research and Development (ICCRD)*, Shanghai, China, 2011, pp. 329–333, doi: https://doi.org/10.1109/ICCRD.2011.5764030.

[31] E. Ashley-Dejo, S. M. Ngwira, and T. Zuva, "A context-aware proactive recommender system for tourist," in *Proc. 3rd International Conference on Advances in Computing, Communication and Engineering (ICACCE)*, Durban, South Africa, 2016, pp. 271–275, doi: https://doi.org/10.1109/ICACCE.2016.8073760.

[32] A. Kumar, S. Gupta, S. K. Singh, and K. K. Shukla, "Comparison of various metrics used in collaborative filtering for recommendation system," in *Proc. 8th International Conference on Contemporary Computing (IC3)*, Noida, India, 2015, pp. 150–154, doi: https://doi.org/10.1109/IC3.2015.7346670.

[33] E. Cavenaghi, L. Camaione, P. Minasi, G. Sottocornola, F. Stella, and M. Zanker, "A re-rank algorithm for online hotel search," in *Springer Proceedings in Business and Economics*, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 53–64, doi: https://doi.org/10.1007/978-3-031-25752-0_5.

[34] H. Hu and X. Zhou, "Recommendation of tourist attractions based on Slope One algorithm," in *Proc. 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 2017, vol. 1, pp. 418–421, doi: https://doi.org/10.1109/IHMSC.2017.102.

[35] J. Ghahramani, M. Sadeghzadeh, and A. Broumandnia, "A hybrid location-based recommender system for trip planning and mobility using geo-tagged media," *Journal of Intelligent Information Systems*, 2025, doi: https://doi.org/10.1007/s10844-025-00992-y.

[36] F. Majeed, A. Siddique, M. Zafar, and K. Shaukat, "A fusion of recommendation techniques to deliver personalized tourism experience," *IET Software*, vol. 20, no. 1, 2026, doi: https://doi.org/10.1049/sfw2.9218059.

[37] D. M. G. Muñoz, J. D. C. Bernal, and J. I. R. Molano, "Machine learning applications in tourism: Case study from Colombia and its post-conflict areas," *Journal of Tourism & Development*, no. 40, pp. 93–102, 2023. **(SD: el DOI https://doi.org/10.23919/CISTI58278.2023.10211549 corresponde a una ponencia CISTI y no a este artículo de revista).**

[38] N. Luh Putu Melya Wati, I. Made Dwi Putra Asana, N. Wayan Suardiati Putri, K. Jaya Atmaja, and I. Gede Iwan Sudipa, "Comparison of automation testing on card printer project using Playwright and Selenium tools," *Computing, Networking, Architecture, and High Performance Computing*, vol. 6, no. 3, 2024, doi: https://doi.org/10.47709/cnahpc.v6i3.4362.

[39] A. Jadon and A. Patil, "A comprehensive survey of evaluation techniques for recommendation systems," in *Artificial Intelligence and Machine Learning for Recommendation Systems*, Cham, Switzerland: Springer Nature Switzerland, 2025, pp. 281–304, doi: https://doi.org/10.1007/978-3-031-71484-9_25.

[40] Z. Shen, "Outlier detect using vector cosine similarity by adding a dimension," in *Proc. International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Osaka, Japan, Feb. 2024, pp. 255–259, doi: https://doi.org/10.1109/ICAIIC60209.2024.10463442.

[41] J. M. Luna, "Introduction to data mining," in *Periodic Pattern Mining*, vol. 2, Singapore: Springer Singapore, 2021, pp. 1–22, doi: https://doi.org/10.1007/978-981-16-3964-7_1.

[42] Z. Arifin, M. R. Ibrahim, and H. R. Hatta, "Nearest tourism site searching using Haversine method," in *Proc. 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2016, pp. 293–296, doi: https://doi.org/10.1109/ICITACEE.2016.7892458.

[43] A. Vulevic, S. Roy, R. Alexandre Castanho, M. Franco, and G. Couto, "Exploring the geo-tourism potential and its accessibility in Danube Region, Serbia: A geo-statistical approach," in *Urban Agglomeration – Extracting Lessons for Sustainable Development*. London, U.K.: IntechOpen, 2024, doi: https://doi.org/10.5772/intechopen.1004744.

[44] M. Garramone and M. Scaioni, "Adapting the '15-minute city' concept to medium-sized urban contexts: A case study in Desio, Italy," *Regional Science Policy & Practice*, vol. 18, no. 3, Art. no. 100283, 2026, doi: https://doi.org/10.1016/j.rspp.2026.100283.

[45] S. V. Madangarli, G. Suganeshwari, S. P. Syed Ibrahim, M. S. Kavitha, and N. Setozaki, "Semantic and temporal-aware hybrid embedding for transformer-based sequential recommendation," *Artificial Intelligence Review*, vol. 59, no. 4, Art. no. 110, 2026, doi: https://doi.org/10.1007/s10462-025-11483-5.

[46] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. New York, NY, USA: Springer, 2011, doi: https://doi.org/10.1007/978-0-387-85820-3.

[47] R. Burke, "Evaluating the dynamic properties of recommendation algorithms," in *Proc. 4th ACM Conference on Recommender Systems (RecSys)*, Barcelona, Spain, 2010, pp. 225–228, doi: https://doi.org/10.1145/1864708.1864753.

[48] R. Cañamares, P. Castells, and A. Moffat, "Offline evaluation options for recommender systems," *Information Retrieval Journal*, vol. 23, no. 4, pp. 387–410, 2020, doi: https://doi.org/10.1007/s10791-020-09371-3.

[49] M. Jesse, C. Bauer, and D. Jannach, "Intra-list similarity and human diversity perceptions of recommendations: The details matter," *User Modeling and User-Adapted Interaction*, vol. 33, no. 4, pp. 769–802, 2023, doi: https://doi.org/10.1007/s11257-022-09351-w.

[50] S. Marukatat, "Tutorial on PCA and approximate PCA and approximate kernel PCA," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5445–5477, 2023, doi: https://doi.org/10.1007/s10462-022-10297-z.

[51] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, no. 3, pp. 491–504, 2014, doi: https://doi.org/10.13053/CyS-18-3-2043.

[52] M. Rafiq, Y. S. Chauhan, and S. Sahay, "Efficient implementation of Mahalanobis distance on ferroelectric FinFET crossbar for outlier detection," *IEEE Journal of the Electron Devices Society*, vol. 12, pp. 516–524, 2024, doi: https://doi.org/10.1109/JEDS.2024.3416441.