electronics

Article

An Effective Hotel Recommendation System through Processing Heterogeneous Data †

Md. Shafiul Alam Forhad [1], Mohammad Shamsul Arefin [1,*], A. S. M. Kayes [2,*], Khandakar Ahmed [3], Mohammad Jabed Morshed Chowdhury [2] and Indika Kumara [4]

1 Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh; forhad0904063@cuet.ac.bd
2 Department of Computer Science and Information Technology, La Trobe University, Plenty Road, Bundoora, VIC 3086, Australia; m.chowdhury@latrobe.edu.au
3 College of Engineering and Science, Victoria University, 70/104 Ballarat Road, Footscray, VIC 3011, Australia; khandakar.ahmed@vu.edu.au
4 Jheronimus Academy of Data Science, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands; i.p.k.weerasingha.dewage@tue.nl
* Correspondence: sarefin@cuet.ac.bd (M.S.A.); a.kayes@latrobe.edu.au (A.S.M.K.)
† This Paper Is an Extended Version of Our Paper Published in the 5th International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2020.

Abstract: Recommendation systems have recently gained a lot of popularity in various industries such as entertainment and tourism. They can act as filters of information by providing relevant suggestions to the users through processing heterogeneous data from different networks. Many travelers and tourists routinely rely on textual reviews, numerical ratings, and points of interest to select hotels in cities worldwide. To attract more customers, online hotel booking systems typically rank their hotels based on the recommendations from their customers. In this paper, we present a framework that can rank hotels by analyzing hotels' customer reviews and nearby amenities. In addition, a framework is presented that combines the scores generated from user reviews and surrounding facilities. We perform experiments using datasets from online hotel booking platforms such as TripAdvisor and Booking to evaluate the effectiveness and applicability of the proposed framework. We first store the keywords extracted from reviews and assign weights to each considered unigram and bigram keywords and, then, we give a numerical score to each considered keyword. Finally, our proposed system aggregates the scores generated from the reviews and surrounding environments from different categories of the facilities. Experimental results confirm the effectiveness of the proposed recommendation framework.

Keywords: automated recommendation; hotel booking system; heterogeneous network data; data processing; points of interest; review analysis; score generation

Citation: Forhad, M.S.A.; Arefin, M.S.; Kayes, A.S.M.; Ahmed, K.; Chowdhury, M.J.M.; Kumara, I. An Effective Hotel Recommendation System through Processing Heterogeneous Data. *Electronics* **2021**, *10*, 1920. https://doi.org/10.3390/electronics10161920

Academic Editor: Rashid Mehmood

Received: 16 July 2021
Accepted: 6 August 2021
Published: 10 August 2021

1. Introduction

Recommendation systems play a vital role in making suggestions for items. They are used to filter information from different networks and predict the output based on the user's preferences. These systems have become extremely popular, and a relevant application of recommender systems is the travel industry. A large number of travel industries are benefiting from the recommendation systems in improving customer satisfaction and experience. In this way, they are making massive chunks of revenue, which is why most of them are turning to recommendation systems. In this paper, one of the main goals of our proposed approach is to provide a platform considering the analysis of the reviews of the customers and the surrounding facilities of the nearby areas of the hotels. Extraction of features from reviews is necessary for providing better recommendations.

Hotel reputation these days is strongly affected by the ratings provided by the guest [1]. Actually, guests are highly appreciated to rate hotels and comment on different aspects of

the hotels. Online reviews provided by the customers have a significant impact on hotel revenues [2]. Customers' trust has become a crucial factor when making decisions for online hotel booking. There has been an increasing effort in the current state-of-the-art literature [1–7] to analyze hotel reviews and ratings in the last decade. In this paper, we build a framework to generate scores from hotel reviews and ratings. We also consider the impact of nearby amenities of the hotels. Hotel selection heavily depends on the different types of P.O.I. (Points of interest), such as public transport, food, and shops. Figure 1 shows a comparative analysis of the overall ratings of a specific hotel for three different hotel booking websites. Ratings vary from website to website. One hotel which is considered average in terms of ratings in one of the hotel booking websites can be found better in other hotel booking websites.

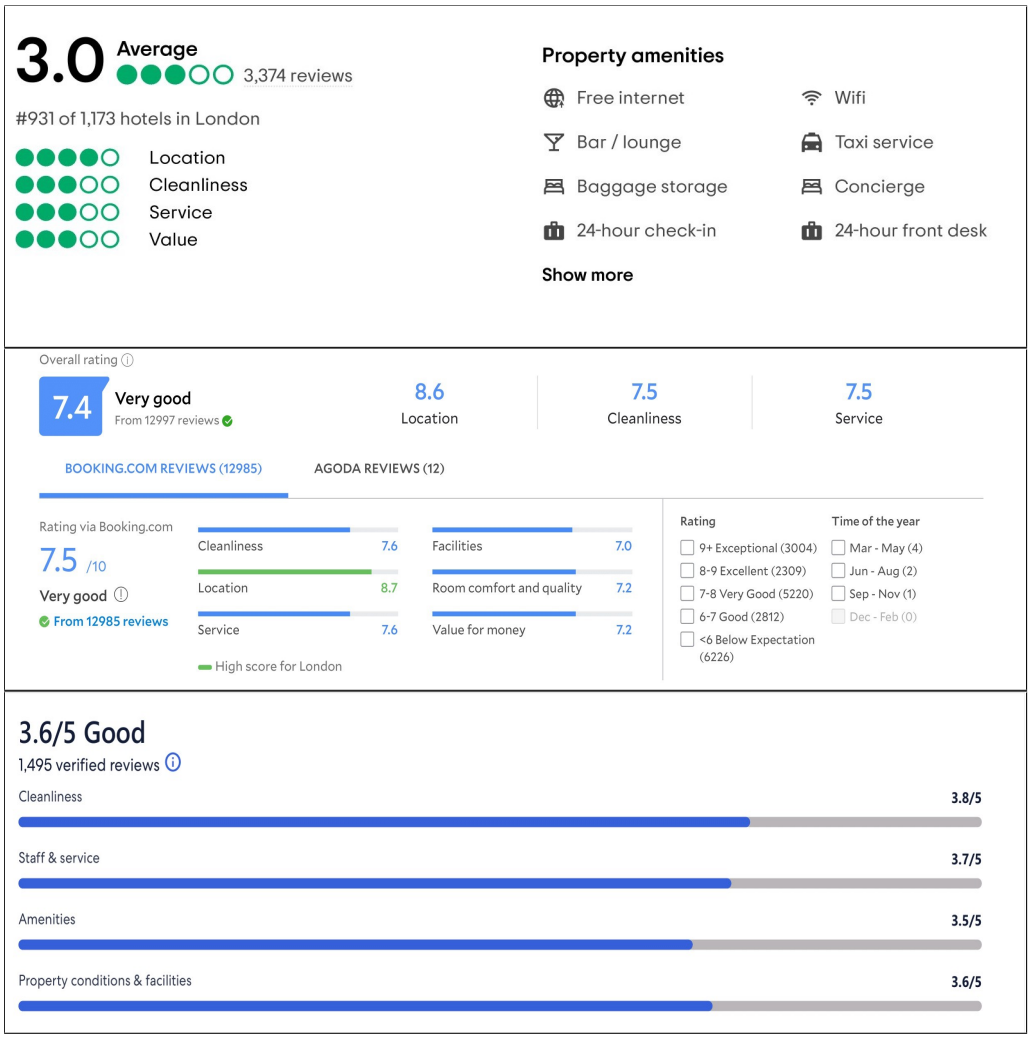

Figure 1. A comparative analysis of the overall ratings of a specific hotel for three different hotel booking websites, i.e., Tripadvisor.com (accessed on 29 June 2021), Agoda.com (accessed on 29 June 2021), and Expedia.com(accessed on 29 June 2021) (Image Source: [8–10]).

To relate the opinions of the guests with the hotel ratings and correlating with P.O.I. descriptions is difficult due to some reasons mentioned below:

- Reviews provided by the guests frequently miss an explicit description of the related context;
- Geo-location information is often missing in the hotel review dataset;
- Preparation and processing time of P.O.I. is time consuming as P.O.I. descriptions are often unstructured.

For this reason, understanding which point of interests are influencing the hotel reviews is difficult from the descriptions of the text. So, the recommendation generations by analyzing texts are not sufficient enough. In our proposed system, we considered the nearby P.O.I.s of the hotels by using Google Place API. Our system can rank hotels in four different ways considering (1) reviews and comments, (2) surrounding environments of the hotels, (3) numerical ratings, and (4) our proposed aggregated scores. Heterogeneous data are an unstructured data type which means a massive amount of data in diverse formats or nature. These unstructured data include text, numbers, images, demographic data, etc. Hotel booking websites contain this type of data. The analysis of the scores generated from the hotel reviews and surrounding P.O.I.s is necessary. We consider data from two famous hotel booking websites. The experimental outcomes give valuable insights into the viewpoints of the guests of the hotels. Figure 2 shows the surrounding facilities for a specific hotel for two widely used hotel booking platforms.

A comparative analysis of some reviewers' comments for two different hotel booking websites, i.e., TripAdvisor and Booking are shown in Table 1. The textual reviews can provide opinions, contextual information for recommender systems. For example, based on the reviews of the customers who stayed at the hotels, a recommender system can recommend a hotel which the previous customers liked.

The key contributions of this paper are as follows:

- We propose a hotel recommendation framework which is implemented by analyzing the
  (1) reviews generated by the customers of the hotels, and
  (2) nearby amenities of the hotels;
- The proposed framework computes scores from the customers' reviews and the nearby amenities of the hotels;
- The proposed method can be helpful for decision-makers, managers of the hotel industry to consider P.O.I.s, review scores for ameliorating the hotel recommendation except for the specific rating score;
- We consider data from multiple sources such as Tripadvisor and Booking.

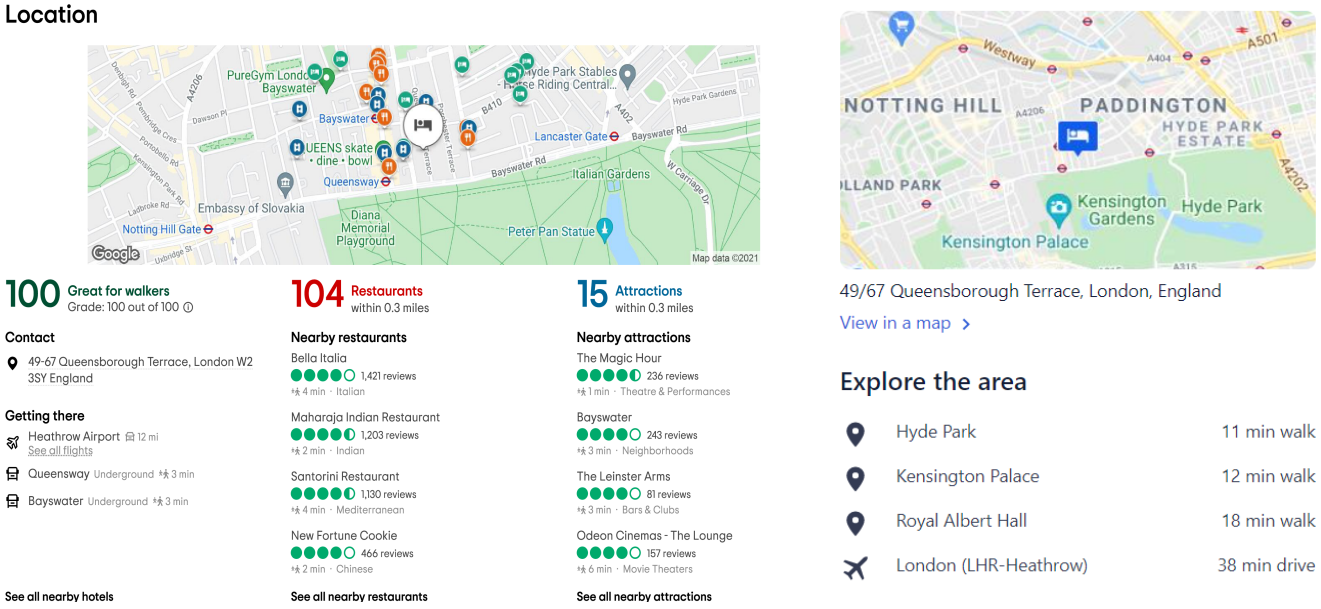

Figure 2. The surrounding facilities of the nearby areas of a specific hotel for two different hotel booking websites, i.e., TripAdvisor.com (accessed on 29 June 2021) and Expedia.com (accessed on 29 June 2021), respectively (Image Source: [8,10]).

Table 1. A comparative analysis of some reviews of two specific hotels for two different hotel booking websites , i.e., TripAdvisor [8] and Booking [11].

HID	Rating	Website	Review
H1	3.0/5	TripAdvisor	(1) Our stay was great. Rooms are clean, great bar with food available. Convenient to transport and sites. Half a block from the park. Rooms are a little small, but fine for touring. Would stay again. Breakfast available too. (2) It's a nice clean hotel and the staff are lovely and helpful. Location is amazing, on central line so easy to get everywhere. Only down side is there Is no air con so the rooms are unbearable to sleep in.
H1	7.4/10	Booking	(1) It was very clean and the location is amazing. The staff was very rude and not talkative at all, they did not even say hello. Also the thin walls in the room made it hard to sleep as the neighbour screamed all morning. (2) Great experience, clean room and staff were friendly and helpful.
H2	5.0/5	TripAdvisor	(1) We really loved this hotel. Our bed was super comfortable, room was clean and hotel staff was very helpful and friendly. I can recommend this hotel for everyone. You really made our holiday. Thank you! (2) Staff were amazingly helpful, very efficient and polite. Everything about the hotel was top class.
H2	9.7/ 10	Booking	(1) Excellent location and quiet. The service at the hotel is outstanding, very personal and they really make you feel special. Breakfast was delicious. The drinks are quite expensive but no different to what you would expect at a 5 star hotel in London. (2) Great staff - lovely atmosphere - delicious breakfast - perfect location. Would recommend. Our bathroom was a little dated - the shower in particular was not good.

The paper is organized as follows. Section 2 overviews the related literature review. In Section 3, we present the architecture of the proposed hotel recommendation system. The experimental outcomes are presented and discussed in Section 4. Finally, Section 5 concludes our work and discusses the future research directions.

2. Related Work

Data over the internet is growing so fast as the people's option to express their views about products or services is increasing rapidly. Due to the growing diversity of data generated from hotels worldwide, a turn of attention has been observed in recent studies in adopting numerous ways of managing these valuable data. In [3], they used a big data solution involving Hadoop to deal with the variety of numeric data as well as textual data in the heterogeneous form. Sharma et al. [12] used NLP (Natural Language Processing) in their work to determine the rating of the hotel used by the previous customers. The authors of [13] proposed the use of a unified deep NLP model, which analyzes sentences in reviews. They make use of BERT embedding to transform the raw text data into a unified review-POI latent space. It is necessary to extrapolate useful and essential information due to the potential effect that customer's opinions can have on businesses. Most of these data are textual data accompanied by a specific numerical rating. To increase customer satisfaction, researchers are building systems that can extract and leverage the knowledge from such reviews to offer guidance on the selection of hotels. The reviews typically contain the customers' opinions on the hotels and ratings, which indicate the sentiment towards the accommodation and fully characterizes the experience itself.

In [14], an approach to recommend hotels to the users by considering nearby facilities of the hotel was presented. The approach utilized the P.O.I. (Points of Interest) database to obtain the nearby amenities of the hotels. It measured the accommodation preferences of the users by using the reviews provided by the users and calculated the similarity score between the hotels and user preferences by using a similarity measure technique. The top-k hotels are suggested and recommended to the user. The experiments used a dataset collected from TripAdvisor. In [13], the authors proposed the use of a unified deep natural language processing (NLP) model which analyzes sentences

in reviews and uses public TripAdvisor hotel-review datasets to validate the approach experimentally. They addressed the challenge of investigating the similarities and dissimilarities between cities by considering the textual reviews and numerical ratings of the hotels and their correlation with the nearby P.O.I.s. They performed their experiment on public TripAdvisor hotel-review datasets and the results provided valuable insights into the viewpoints of hotel guests and suggested further investigation in this direction. Yang et al. [15] presented their effort at constructing a location-aware recommendation system that can model user preferences mainly based on the reviews of the users. They used datasets provided by Yelp. However, they have only included the textual reviews to grasp the nature of people's preference. Yang et al. [16] classified three different categories by considering all location-related factors. The three categories are accessibility to P.O.I., transport convenience, and the surrounding environment. The results confirmed that the presence of airports, public transport, attractions, universities, etc., are significant determinants. Chen et al. [17] combined the conventional recommendation technology with location-based services to provide recommendations. They considered price, service, the location of the tourist, etc., to provide recommendations. The results provided by their system can be nearest to the tourists' needs. The use of location-based social services has enabled opportunities for providing better services through P.O.I. P.O.I. recommendation is personalized, location-aware, and context-depended, unlike traditional recommendation tasks. Recently, many attempts have been dedicated to capturing user preference data from textual reviews for rating prediction purposes [18]. The critical challenge is to understand the key factors that contribute to customer dissatisfaction or satisfaction employing data-driven approaches [19]. Brett et al. [2] showed that the positive rating on customer actions is more influential than advertising strategies. So, review analysis and extraction of the hidden knowledge from reviews is particularly appealing.

Ramzan et al. [3] proposed a recommendation system that helps users find hotels by considering heterogeneous data. The experiments used two different hotel booking datasets that contain reviews, ratings, and ranks to represent data heterogeneity. Their proposed system generates polarity scores from the reviews by using NLP techniques and calculates the aggregated polarity score for each feature based on the reviews from selected websites. By aggregating numerical scores provided by ratings and polarity scores, it generates recommendations to the users. Final recommendations are generated by applying the fuzzy logic approach. Both qualitative and quantitative features of likeness can be achieved by using not only ratings but also reviews of the texts.

In [20], a text to score generation algorithm is proposed, which considers some keywords and their corresponding scores to generate scores from the reviews. They only used unigram keywords, and thus, pairwise combinations of words are neglected. Compared with their work, we consider the combination of unigram and bigram keywords. In our system, both single words and a pairwise combination of words are considered. Sharma et al. [12] examined a recommendation system by using a multi-criteria review-based approach. The approach is based on the user's reviews and preferences. They use various NLP approaches to find out the rating of a hotel from previous users. Instead of simple star ratings, their system also deals with the process to suggest hotels based on multi-criteria ratings. These ratings are derived from textual reviews. They only consider the TripAdvisor dataset for their experimental purpose.

In [21], a new feature and opinion extraction method based on the characteristics of online reviews was proposed to extract the user opinions from the user reviews effectively. They crawled a real online restaurant review dataset and collected 54,208 reviews. They selected 4000 reviews randomly and features and opinions extraction from these reviews are done manually. However, these systems process only homogeneous data, whereas most of the data on the web are heterogeneous. Chuhan et al. [6] have tried to express user preferences comprehensively by jointly analyzing hotel ratings and customer reviews. Zhang and Mao [22] suggested that appropriate recommender systems should be developed to achieve true and relevant recommendations according to the choice and preferences of the

customers. The deviation of various approaches, objective and advantages of the various recommendation systems are shown in Tables 2 and 3, respectively.

Table 2. Deviation of various approaches.

References	Ratings	P.O.I.	Reviews	Review/Polarity Score	Multi Data Source
[14]	No	Yes	Yes	No	No
[12]	Yes	No	Yes	No	No
[21]	Yes	No	Yes	No	No
[3]	Yes	No	Yes	Yes	Yes
[22]	Yes	No	Yes	No	Yes
Proposed approach	Yes	Yes	Yes	Yes	Yes

Table 3. The objective and advantages of the various RS (Recommendation Systems).

References	Objective	Advantages
[14]	Recommend hotels by using surrounding environments of locations	To provide recommendations by considering the nearby amenities of the hotels
[21]	To propose a new feature and opinion extraction method from online reviews	Their proposed method combine user preference and opinion for recommendation
[3]	To provide true recommendations considering multiple types of data	Recommend hotels by using heterogeneous data (Ranks, ratings and reviews)
[15]	Present a user preference based RS	A P.O.I. based recommendation system that models the preferences of the users' by considering user reviews
[12]	To select the best suited hotel in a city according to user reviews and preferences	To generate hotel recommendations based on multi criteria settings

3. Proposed System Architecture of Hotel Recommendation System

In this section, we will elaborate on the architecture of our hotel recommendation system. Our system contains the following modules: data pre-processing, storage, surrounding environment's evaluation, review analysis, and recommendation generation. Figure 3 shows our system architecture. In the review analysis module, scores are generated from pre-processed textual reviews. Score generation procedures from the nearby amenities of the hotels are performed in the surrounding environment's valuation module.

3.1. Dataset Description

We used two different hotel booking datasets for our experimental purposes. The datasets we used in our work are publicly available. We used the framework where the pre-processing stage is performed to the raw sentences, making it more understandable. The first dataset we used in our experiment was collected from Kaggle. This dataset contains about 515 K customer reviews and scoring of 1493 luxury hotels across Europe. For further analysis, geographical locations of hotels are also included here [23]. Table 4 shows the description of the dataset attributes. The file contained 17 attributes.

Another dataset is used for the reviews of hotels collected from TripAdvisor (259,000 reviews). This dataset was initially used for opinion-based entity ranking. We collected this dataset from [24]. We considered 875 hotels of London from these large datasets. We created a CSV file where we manually assigned a unique hotel ID for each hotel for our experimental purpose. The CSV file contains five fields which are shown in Table 5. By using Google API, we collected all considered facilities in the nearby area of the hotels. Our system categorizes the nearby amenities of each hotel by using different categories of the category tree shown in Figure 4.

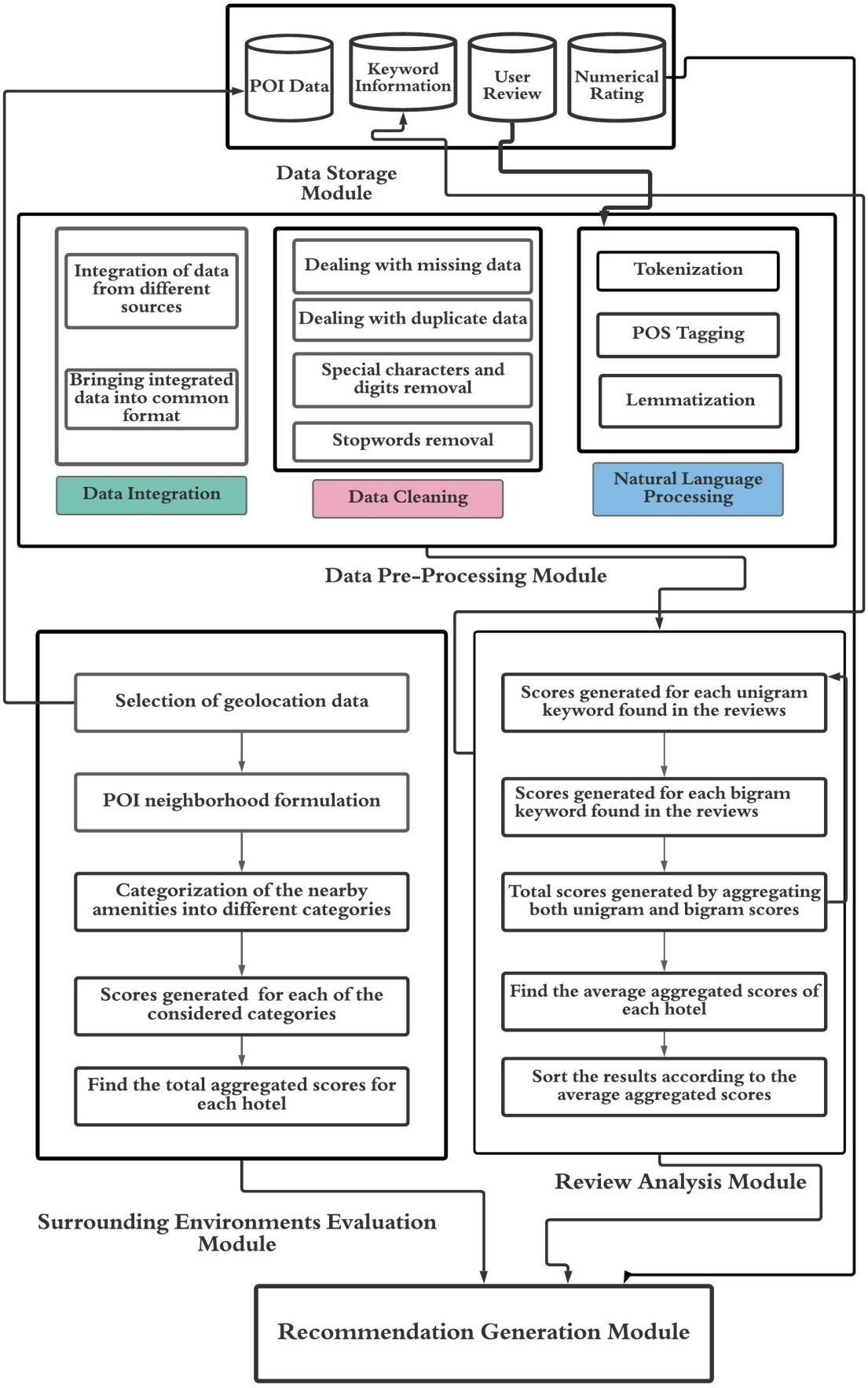

Figure 3. System architecture for generating hotel recommendation by analyzing heterogeneous data.

Table 4. Description of the attributes which we considered from the Booking Dataset.

Attributes	Description
Hotel_Address	Address of the hotel
Avg_Score	The average numerical scores of the corresponding hotels
HName	Name of the corresponding hotel
Negative_Review	Review (Negative) given by the reviewers
Positive_Review	Review (Positive) given by the reviewers
Total_Number_of_Reviews	Total number of valid reviews is represented by it
lat	It represents the latitude information of the hotel
lng	It represents the longitude information of the hotel

Table 5. Description of OpinkRank Hotel Dataset.

Attributes	Description
HID	It represents a unique Id for each hotel
Hotel_Name	It represents the hotel name
Review_Title	It represents the title of a review
Full_Review	It represents the full reviews given by each individual users

3.2. Storage Module

The storage module preserves the necessary information's for the purpose of generating recommendations. The four storages we used in our system and their functions are given below:

- User review database is used to store the textual reviews of the customers;
- Keyword database is used to store the extracted keywords for the purpose of score generation;
- P.O.I. database is used to store geolocation data about the nearby amenities of the hotels;
- A numerical rating database is used to contain the numerical ratings from hotel booking websites.

While some information may be put to use immediately, much of it will serve a purpose later on. When data are properly stored, the data can be quickly and easily accessed in the time of need. We use SQL (Structured Query Language) to store the data.

3.3. Data Pre-Processing

Data pre-processing is the process of removing incomplete and noisy data to clean data and put them in a formatted way while doing any operation with them. The kind of data we used in our work contains symbols and unusual text that need to be cleaned. Datasets may be of different formats for different purposes. We usually put the data into a CSV file.

Algorithm 1 shows our data pre-processing algorithm. Our algorithm is implemented in Python which is a high-level programming language and has a great number of data-oriented feature packages. These packages can speed up and simplify data processing, thus making it time-saving. In addition, it also has many excellent libraries for data analysis. Python can handle large datasets; it can more easily implement automated analysis. The pre-processing includes the steps of data integration, removal of missing values, removal of stop words, conversion to Lowercase, Tokenization, removal of special characters and digits, parts of speech tagging, lemmatization, etc.

Algorithm 1: Data Pre-Processing Algorithm.

```
Input: Review Text Data
Result: Pre-Processed Text Data
Data: User Review Data
begin;
if Dataset contains positive and negative reviews then
    Combined the reviews to get the overall reviews
end
for i = 1, ..., Number of hotels do
    for j = 1, ..., Number of reviews considered for Hotel_i do
        Drop rows with any empty cells
    end
end
function PROCESSING(text)
    Lower the text
    Tokenize the text
    Remove special characters
    Remove words which contain numbers
    Remove stop words
    POS Tag text
    Lemmatize text
    Join all
return text
for i = 1, ..., Number of hotels do
    for j = 1, ..., Number of reviews considered for Hotel_i do
        function PROCESSING(Review_ij)
    end
end
```

3.4. Review Analysis Module

The textual data need to be processed in order to retrieve more specific opinions. The keywords we consider in our system are categorized into ten different categories. The scores are calculated from the reviews of the customers. Table 6 lists some examples of keywords of different categories. The review-to-score generation procedure is shown in Algorithm 2.

The scores are calculated for a single review of a hotel by using the following Equations (1) and (2):

$$Review\ Score\ Unigram = \sum_{i=1}^{n} Occurrence(K_i) * Weight(K_i) \tag{1}$$

$$Review\ Score\ Bigram = \sum_{i=1}^{l} Occurrence(K_i) * Weight(K_i) \tag{2}$$

For each unigram/bigram keyword found in the review, multiply the keywords score ($weight(k_i)$) with the number of occurrences of the keyword present in the review. Then, total scores are generated by aggregating the scores considering the effect of n number of unigram/bigram keywords present in the review. The review score is computed by the following Equation (3):

$$Review\ Score = Review\ Score\ Unigram + Review\ Score\ Bigram \tag{3}$$

The total score generated by considering all of the k reviews of a particular hotel is computed by using Equation (4) given below:

$$Total\ Review\ Score = \sum_{i=1}^{k} Review\ Score_i \tag{4}$$

The total review score is computed by aggregating all k review scores.

The average review score generated for a single hotel is computed by using Equation (5) given below:

$$Scores\ Generated\ for\ a\ Single\ Hotel = \frac{Total\ Review\ Score}{k} \tag{5}$$

An average score is calculated for a single hotel by dividing the total review scores generated from all k reviews to the value of k.

Algorithm 2: RSG (Review to Score Generation) Algorithm.

```
Input: DB of pre-processed words and DB of keywords with corresponding score
Result: Generated score for each feedback
Data: Pre-Processed Review Data
begin
score = 0, count = 0, Overall_score=0, no_of_reviews=0, Average_score=0, scores
 _for_review=0, Scoreuni = 0, Scorebi =0, Total Score=0
for Each hotel in the dataset do
    for Each review of the hotel do
        for Each considered unigram keyword do
            if minimum one match found in the review then
                Scoreuni + = number of occurrences of unigram keyword * weight
                  of the corresponding unigram keyword;
                count += number of occurrences of unigram keyword
            end
        end
        for Each considered bigram keyword do
            if minimum one match found in the review then
                Scorebi + = number of occurrences of bigram keyword * weight of
                  the corresponding bigram keyword;
                count += number of occurrences of bigram keyword
            end
        end
        Total Score = Scoreuni + Scorebi;
    end
    Overall_score = Total Score / count
    scores_for_review=scores_for_review + Overall_score
    no_of_reviews + =1
end
Average_score=scores_for_review/no_of_reviews
```

3.5. Evaluation of Surrounding Environments

The P.O.I.s (Points of Interest) database is used in our system to evaluate the surrounding environments of the hotels. Using Google Place API, our system collected all considered facilities within five hundred meters of each hotel. We choose five hundred meters for our experimental purpose. By using a Category Tree (CT) shown in Figure 4, we classified different facilities into eight different categories. The internal nodes represent the types of facilities. The leaf nodes denote the objects of the facilities. Our system generates scores from the surrounding contexts of the hotels based on the information of the CT. The procedure of the surrounding environment's evaluation is shown in Figure 5. Our considered eight categories are shown in Table 7. Total scores are generated by aggregating the scores generated by all of the categories. Now, assume that there are two airports, four restaurants, one university, one movie theater, one bus station, and one night market

within five hundred meters from a specific hotel. Looking at the CT of Figure 4, we can see that two airports and one bus station are within the category "Travel and Transport", four restaurants are inside the category "Food", one university inside the category "College and University", one movie theater within the category "Arts and Entertainment" and one night market within the category "Nightlife Spot". In Figure 6, for different categories of surrounding facilities, the number of facilities is shown for a specific hotel H_1. Here, the number of facilities of H_1 for C_1 is 1, C_2 is 1, C_3 is 3, C_4 is 0, C_5 is 4, C_6 is 1, C_7 is 0, and C_8 is 0.

Table 6. Example of Keywords of different Categories.

Category No.	Category	Keywords	Score
1	Outstanding	Outstanding, Wonderful, Super friendly, Extremely clean	10
2	Excellent	Excellent, Comfortable,Bright, Delightful, Gorgeous, Good quality	9
3	Very Good	Very good, Very friendly, Beautiful, Peaceful	8
4	Good	Good, Friendly, Helpful, Charm, Kind, Convenient	7
5	Above Average	Above Average, Happy, Cool, Nice, Fine	6
6	Average	Average, Ordinary, Mean	5
7	Below Average	Below average, Dislike, Unhappy, Unclean	4
8	Poor	Poor, Bad, Sad, Difficult, Inadequate	3
9	Very Poor	Very poor, Rude, Very bad, Very expensive, Very unfriendly	2
10	Terrible	Dreadful, Damage, Spoil, Abnormal, Terrible, Dangerous, Insult	1

Table 7. Location categories in [25].

Category	Category Name	Sub-Categories	Example
C_1	Arts and Entertainment	36	Circus, Theater, Museum, Stadium, Aquarium, Zoo, ...
C_2	College and University	23	University, College, ...
C_3	Travel and Transport	34	Airport, Metro, Bus Station, Train Station, Cable Car, ...
C_4	Event	12	Street Fair, Conference, Festival, ...
C_5	Food	121	Chinese Restaurant, American Restaurant, ...
C_6	Nightlife Spot	7	Lounge, Night Market, Night Club, Bar, ...
C_7	Outdoors and Recreation	80	Beach, River, Botanical Garden, Mountain, Pool, Park, ...
C_8	Shop and Service	145	ATM, Chocolate Shop, Spa, Market, ...

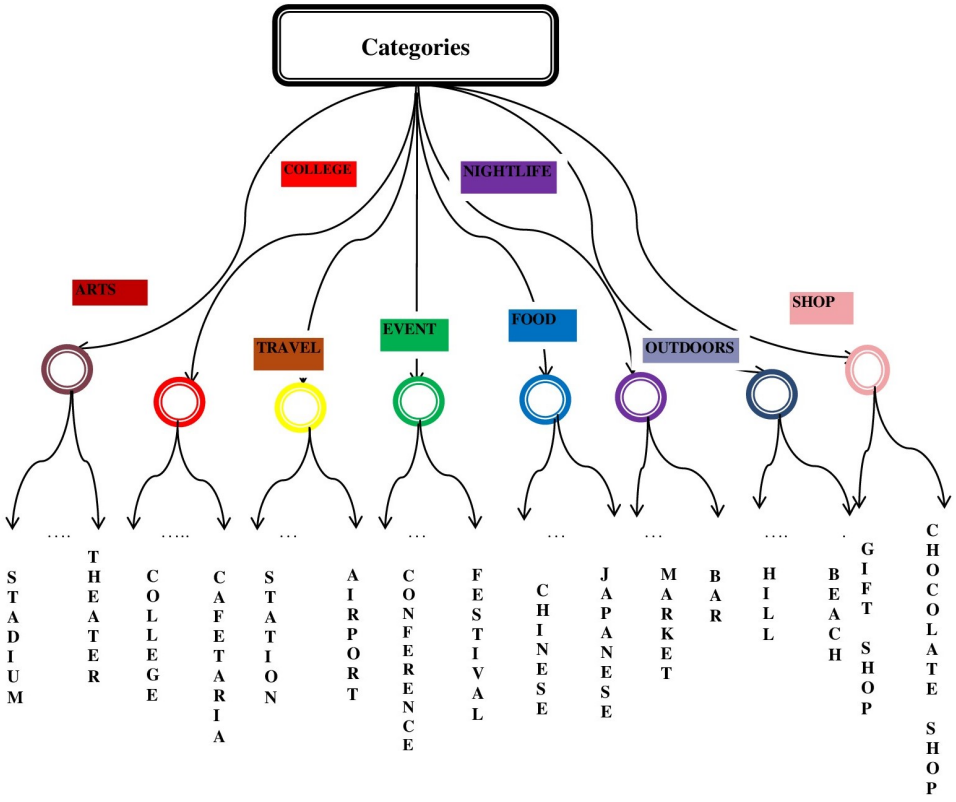

Figure 4. Category Tree.

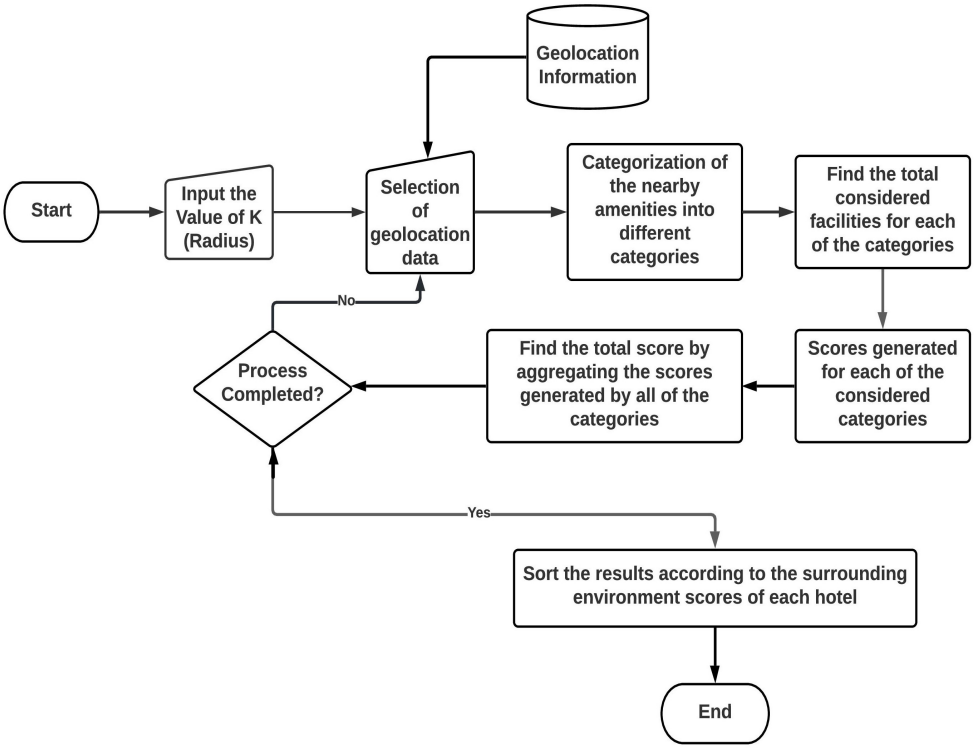

Figure 5. Score Generation Procedure from Surrounding Environments.

The scores are calculated for a single category are measured by using Equation (6):

$$S.E.S.(C_i) = \sum_{j=1}^{n} \sum_{k=1}^{l} F_{jk} * W_{jk} \quad (6)$$

Here, n denotes the total number of sub-categories for a specific category. F_{jk} represents the total number of facilities of type k for sub-category j. In our proposed method, we consider two types of weights, so the value of l is 2.

W_k represents the weight of the facility type;
C_i represents the ith category;
and $S.E.S.$ represents the surrounding environments score.

The scores are calculated for all of the categories are measured by using Equation (7):

$$Total\ Score\ for\ a\ Hotel\ = \sum_{i=1}^{8} S.E.S.(C_i) \tag{7}$$

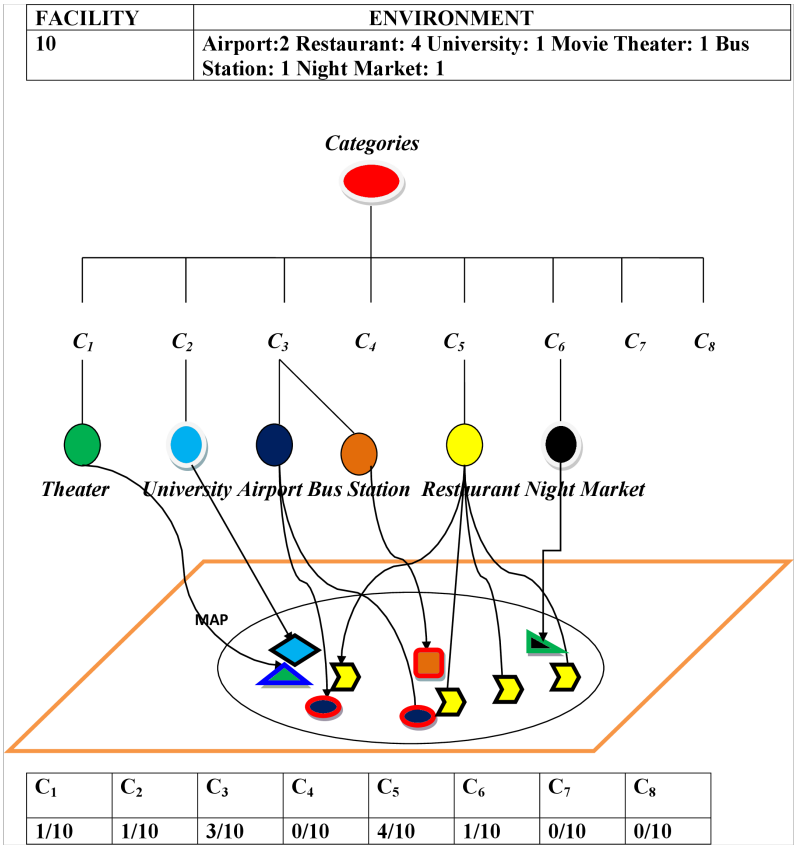

Figure 6. An Example of Surrounding Environment's Evaluation.

We give +1 score for the most important facilities and 0.5 for the other facilities. After determining the surrounding facilities of a hotel, the scores are generated by using Equations (6) and (7). Each of the considered categories are divided into some or many sub-categories. The overall surrounding environment score of a hotel is determined by aggregating the scores generated from all of the categories for that hotel. The scores are generated for each of the considered sub-category. Let us assume that there is a hotel which has 10 facilities in its surrounding areas within a specific range. Among them five facilities are under the category "Arts and Entertainment" and another five are in the category "Food". Then, the scores are calculated by adding the results obtained from the surrounding environment scores of all considered categories. There can be two or more sub-categories for each of the categories. For each sub-category, there are two types of weights we consider for the facility. The most important facilities are considered as type-1 facility and other facilities are considered as type-2 facility. For a specific category, scores are generated by adding the surrounding environment scores of all of the sub-categories of

the considered category. The surrounding environment score of a specific hotel is calculated by using Equation (7).

3.6. Recommendation Generation Module

The recommendation generation module generates recommendation by aggregating the scores generated from reviews and nearby amenities of the hotels. The aggregated score is the summation of S.G.R. (Score Generated from Review) and S.E.S. (Surrounding Environments Score). The scores are calculated by our system for a specific hotel that contains n number of reviews is computed by using Equation (8) given below:

$$Aggregated\ Score = \frac{(\sum_{i=1}^{n} S.G.R.(R_i)}{n} + \sum_{i=1}^{8} S.E.S.(C_i) \quad (8)$$

4. Experimental Results and Analysis

The top-10 recommendations based on different settings and using average numerical ratings of hotel bookings are discussed here. The dataset we considered here is collected from [23]. This dataset contains information on 1493 hotels. From Table 8, we can see that "Ritz Paris" is the topmost hotel by using average numerical ratings of Booking. The numerical rating score obtained for this hotel is 9.8. We can also see the top-10 recommended hotels by analyzing the reviews of the reviewers in Table 9.

Table 8. Top-10 recommended hotels for the 1493 hotels of booking.com based on the average numerical ratings of booking.com.

S. No.	Ratings	Hotel Name
1.	9.8	Ritz Paris
2.	9.6	Haymarket Hotel
3.	9.6	Hotel de La Tamise Esprit de France
4.	9.6	Hotel Casa Camper
5.	9.6	Hotel The Serras
6.	9.6	41
7.	9.6	H10 Casa Mimosa 4 Sup
8	9.5	Palais Coburg Residenz
9.	9.5	Waldorf Astoria Amsterdam
10.	9.5	Hotel Sacher Wien

By considering nearby amenities of the hotels, the top-10 recommended hotels for the 1493 hotels of booking are shown in Table 10. Finally, the top 10 recommendation generation based on our system is shown in Table 11. By using our developed RSG algorithm, our system generates scores from the reviews. The highest score obtained from the average review scores of each hotel is 6.91. The name of the hotel is "South Place Hotel". Next, our system analyzed the nearby amenities of the hotels. From Table 10, we can see that "Hotel Kaiserin Elisabeth" is the highest-ranked hotel. Finally, our system computes the aggregated scores of each considered hotel.

From Table 11, we can see that "Hotel Kaiserin Elisabeth" has the highest ranked hotel and the score generated for this hotel is 28.11. The "Hotel Casa Camper" is ranked as fourth by ratings of Booking but it is ranked as ninth by analyzing reviews. From Tables 12–14, the top-10 recommendation generation based on the different settings are shown. Top-10 recommendation generation uses the following parameters: review scores generated by using our developed RSG algorithm, scores generated from nearby amenities of the hotels and scores generated by our system. The TripAdvisor dataset we considered here is collected from [24].

Table 9. Top-10 recommended hotels for the 1493 hotels of booking.com by analyzing reviews.

S. No.	Scores Generated from Reviews	Hotel Name
1.	6.91	South Place Hotel
2.	6.79	Aparthotel Arai 4 Superior
3.	6.79	Hollmann Beletage Design Boutique
4.	6.77	Hotel Fabric
5.	6.76	Hotel Saint Paul Rive Gauche
6.	6.75	Petit Palais Hotel De Charme
7.	6.73	Boutique Hotel Konfidentiel
8	6.71	Hotel Daniel Paris
9.	6.70	Hotel Casa Camper
10.	6.68	Canal House

Table 10. Top-10 recommended hotels for the 1493 hotels of Booking based on surrounding environments.

S. No.	Surrounding Environments Score	Hotel Name
1.	22	Hotel Kaiserin Elisabeth
2.	22	Eurostars Ramblas
3.	21	Appartement Hotel an der Riemergasse
4.	20	Hotel K nig von Ungarn
5.	20	Hotel Das Tigra
6.	20	Sofitel London St James
7.	20	Austria Trend Hotel Astoria Wien
8	19	Catalonia Magdalenes
9.	19	Hotel Trianon Rive Gauche
10.	19	Monhotel Lounge SPA

When selecting a hotel for staying purposes, hotel attractions are very important as most customers of the hotels are tourists. Hotel review analysis is also very essential for the customers as well as the surrounding environments of the hotel. If two hotels have the same ratings, then from review scores, surrounding environments scores, a better decision can be taken by the customers. The rankings of the hotels by the surrounding environments can be important for someone who is only interested in the surrounding facilities of the hotels. Someone who is influenced by only the reviews of the previous customers, then, the review scores can be important to him/her. Scores generated from reviews reflect the opinions of the customers of the hotels and the scores generated from surrounding environments reflect the surrounding facilities of the nearby areas of the hotels. The integrated scores generated by our system are a different way of providing recommendations to the customers. The integrated score is the reflection of both review and surrounding environment scores.

Table 11. Top-10 recommended hotels for the 1493 hotels of Booking by using our system.

S. No.	Scores Generated by Our System	Hotel Name
1.	28.11	Hotel Kaiserin Elisabeth
2.	27.55	Eurostars Ramblas
3.	27.18	Appartement Hotel an der Riemergasse
4.	26.41	Hotel K nig von Ungarn
5.	26.15	Hotel Das Tigra
6.	26.09	Sofitel London St James
7.	25.57	Catalonia Magdalenes
8	25.48	Austria Trend Hotel Astoria Wien
9.	25.29	Monhotel Lounge SPA
10.	25.09	Hotel Trianon Rive Gauche

Table 12. Top-10 recommended hotels for the 875 Hotels of TripAdvisor by analyzing reviews.

S. No.	Scores Generated by Analyzing Reviews	Hotel Name
1.	7.44	No ten manchester street
2.	7.27	Home house
3.	7.23	Hanger hill hotel
4.	7.19	Odessa wharf
5.	7.11	London tower bridge apartments
6.	7.11	51 kensington court limited
7.	7.07	The chesterfield mayfair hotel
8.	7.05	The levin
9.	7.03	Royal over seas league
10.	7.03	Fraser place canary wharf

The rankings are different because it may be possible that a hotel that has a higher rank by considering ratings has reviews that are not overall good compared to a hotel that ranked as average by considering ratings. This is also possible if a hotel with high surrounding facilities has low ratings. So for these reasons, hotel rankings are varied. From Table 8, we can see that "Hotel Casa Camper" is ranked as 4th by average numerical ratings of Booking. It is ranked 9th by considering review scores. As the choice or taste of the customers can vary, so the different ways of providing hotel rankings can also be important.

From Table 12, we can also see that "No Ten Manchester Street" is the highest-ranked hotel among 875 considered hotels of London by analyzing the reviews of the hotels. "Hilton London Tower Bridge" is the highest-ranked hotel by both surrounding environments and scores generated by our system. In Table 14, the top-10 recommended hotels by using our system are shown.

Table 13. Top-10 recommended hotels for the 875 Hotels of TripAdvisor based on surrounding environments.

S. No.	Surrounding Environments Score	Hotel Name
1.	19	Hilton London tower bridge
2.	19	Sheraton park tower
3.	19	Norfolk plaza hotel
4.	18	London south kensington
5.	18	The rembrandt
6.	18	London americana hotel
7.	18	Westpoint hotel
8.	18	Apollo hotel bayswater
9.	17	St giles hotel london
10.	16	Haymarket hotel

Table 14. Top-10 Hotel Recommendation Generated by Our System for the 875 Hotels of TripAdvisor.com.

S. No.	Scores Generated by Our System	Hotel Name
1.	25.06	Hilton London towe bridge
2.	24.74	Sheraton park tower
3.	23.80	Citadines London south kensington
4.	23.75	Norfolk plaza hotel
5.	23.71	The rembrandt
6.	23.42	London americana hotel
7.	23.09	St giles hotel london
8.	23.03	Haymarket hotel
9.	23.00	Westpoint hotel
10.	22.41	City of London yha

There are 214 hotels that are common in the dataset of both of the hotel booking websites. Top-10 recommendation generation based on average numerical ratings of Booking is shown in Figure 7. Considering the two datasets of the common hotels, the top-10 hotels recommended by our system are shown in Figures 8 and 9, respectively. From Figure 7, we can see that "Haymarket Hotel" is ranked as 2nd by average numerical ratings of Booking. It is ranked as 3rd by considering the dataset of TripAdvisor and it is ranked as 5th by considering the dataset of Booking. From Figures 8 and 9, we can also see that "Hilton London Tower Bridge", "London Marriott Hotel County Hall", and "Cavendish Hotel" are also included in the top-10 recommended hotels by considering the dataset of both hotel booking websites.

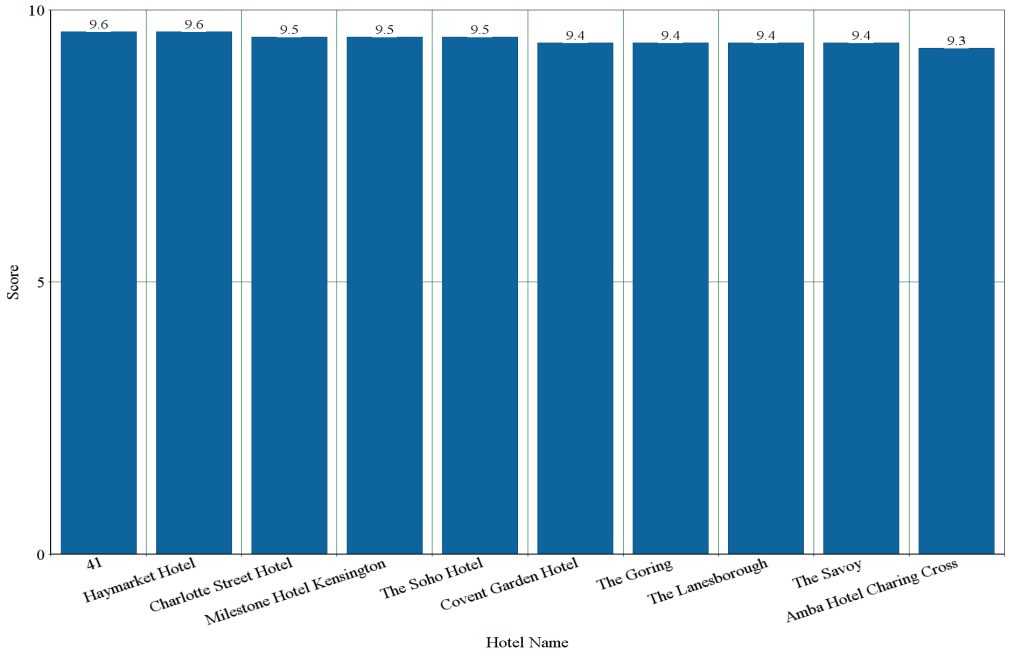

Figure 7. Top-10 Hotel Recommendation Generation for the common hotels Based on Ratings of Booking.

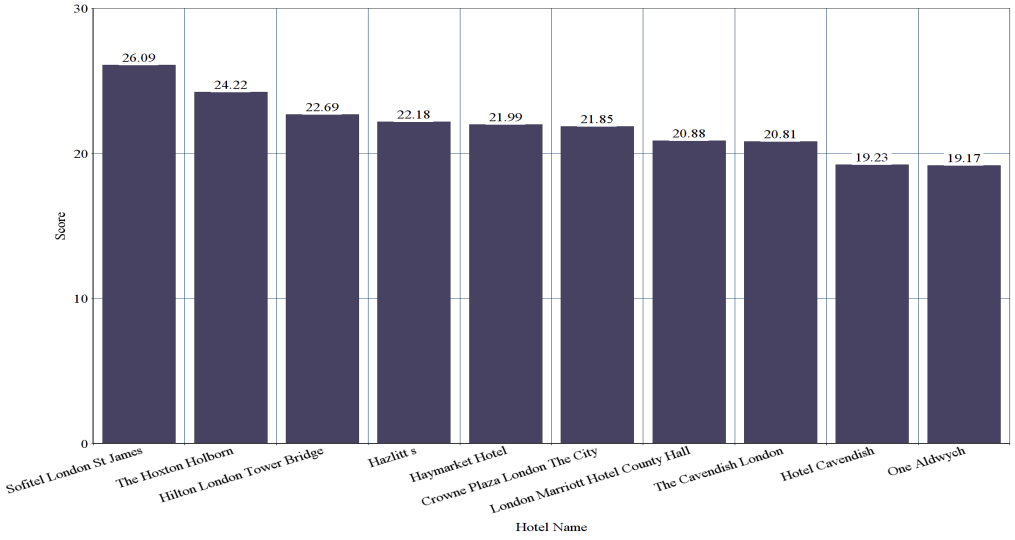

Figure 8. Top-10 Hotel Recommendation Generation by Our System for the 214 Common Hotels by considering the dataset of Booking.com.

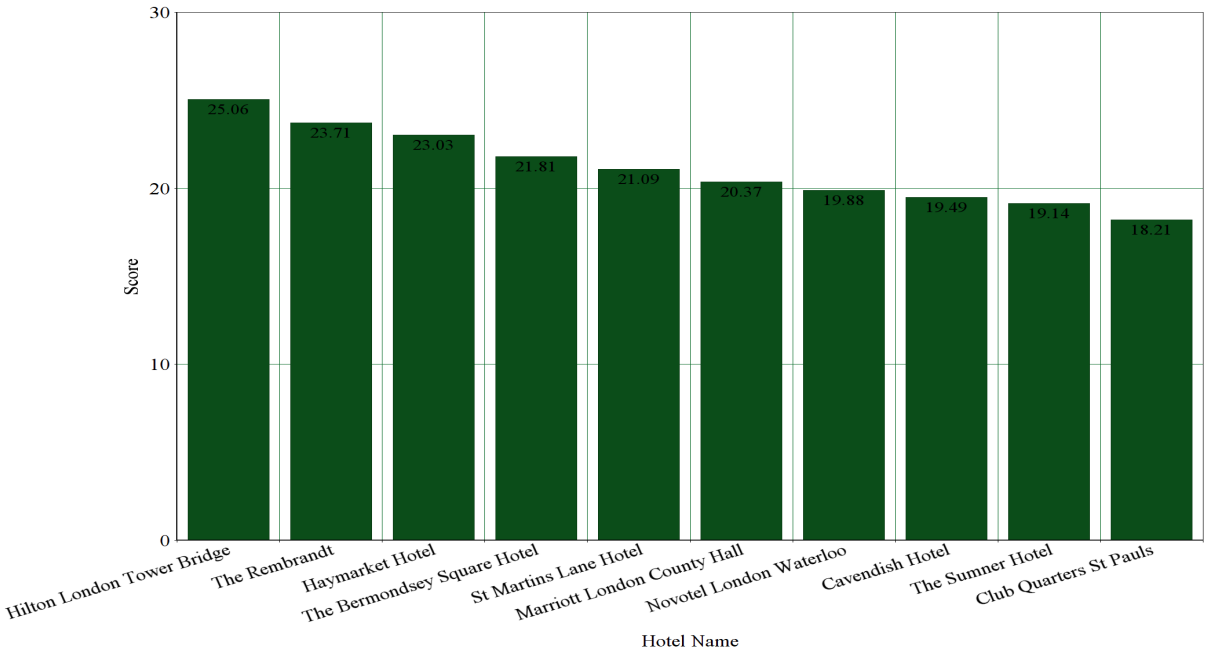

Figure 9. Top-10 Hotel Recommendation Generation by Our System for the 214 Common Hotels by considering the dataset of TripAdvisor.

There are 214 hotels which are common in both of the hotel booking datasets. The recommendation time of both of the hotel booking datasets for the selected 214 common hotels is given below:

We have compared the execution time of our proposed method with that of Liu et al. [21]. The execution time of our proposed method for the 214 common hotels by considering the data of both hotel booking websites is shown in Table 15. The runtime comparison of our proposed method with [21] is shown in Figure 10. The total execution time found in the method of [21] was about 27 s, whereas that of our method was about 6.55 s and 12.46 s for the considered two datasets, respectively. The reason for this difference is that they proposed a method for opinion-feature extraction from online reviews. They randomly selected 4000 reviews and manually extracted features and opinions from these reviews. The execution time of our method is less than that reported in [21]. The reason is that our system generates scores by considering the impacts of different important keywords present in the review and uses the RSG algorithm. As opinions may vary a lot in the reviews from different domains, the extraction is challenging and time-consuming. Experimental results show the effectiveness of the proposed recommendation method.

Table 15. Runtime comparisons of our proposed method for the 214 common Hotels of Booking and TripAdvisor.

S. No.	Dataset	Recommendation Time in Seconds
1	Tripadvisor	6.55 s
2	Booking	12.46 s

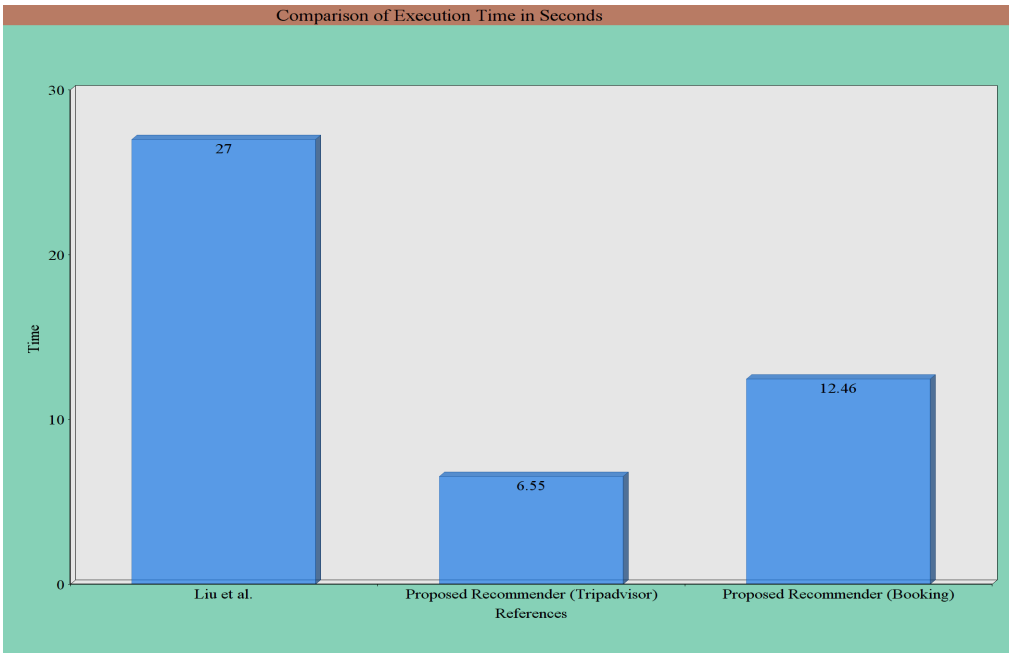

Figure 10. Runtime comparisons of our proposed method with [21].

5. Discussions

In this paper, we proposed a hotel recommendation system that considers the reviews of the reviewers collected from two famous hotel booking websites. Our proposed framework consists of a data storage module, review analysis module, surrounding environments evaluation module, data processing module, and recommendation generation module. To generate scores from the reviews of the hotels, we developed an RSG algorithm, which takes input as review text and generates scores by considering the impact of both single keywords and a pairwise combination of keywords as outputs. Then a method is used to generate scores by considering the nearby amenities of the hotels. By using Google Place API, the nearby amenities of the hotels are collected. The nearby amenities of hotels are categorized into eight different categories. The scores generated for each of the categories of hotels are aggregated. Then, by using our developed RSG algorithm, scores are generated from the reviews. Some hotel booking systems are available in the state-of-the-art for providing recommendations to the users. Our proposed framework considers the hotels' nearby amenities and analyzes reviews to generate better user recommendations. The data we used in our work were collected from two famous hotel booking websites, i.e., TripAdvisor and Booking, respectively.

6. Conclusions and Future Research Directions

With the increase of applications using the Internet, the sources of data are getting richer in heterogeneity. Therefore, the various factors in the new data bring new challenges. However, it is also a chance to create novel methods to achieve better recommendation results. So, for this reason, in this paper, we consider heterogeneous data to generate hotel recommendations for the users.

We proposed a hotel recommendation framework to predict top-rated hotels based on the scores generated from reviews and nearby amenities of the hotels through experimental analysis. We have used two reliable data repositories, TripAdvisor and Booking, containing a significant number of numerical ratings, textual reviews, geolocation information, to represent the heterogeneity of data. After data pre-processing, our system generates scores from the reviews of the selected hotel booking datasets. Review scores are aggregated with the surrounding environment scores of the hotels. These heterogeneous data sources, such as ratings, textual reviews, and P.O.I.s are used in our proposed approach, and final aggregated scores are obtained as shown in the experimental results section. The rank of

the topmost hotels by using the final aggregated scores are shown for different datasets in the experimental results section. We compared the results of our proposed system with the top-10 results produced by the baseline hotel booking website. In most of the existing recommendation systems, hotel ratings and rankings are typically calculated based on the reviews of previous users only, without considering the hotels surrounding environments.

When selecting a hotel for staying purpose, hotel attractions, such as tourist areas, shopping services, nightlife spots, restaurants, transportation, etc., are very important. More specifically, as most customers of the hotels are tourists, there is a need to consider the location of the hotels. Hotel review analysis is also very essential for the customers as well as the nearby amenities of the hotel. Hotel reviews shed light on the behaviors that had been perceived as pleasing or unpleasing by hotel customers. The proposed system can be helpful to the decision-makers, managers, etc., of the hotel industry to analyze online reviews on a regular basis for ensuring users' satisfaction. The proposed recommender system suggests the decision-makers of the hotels to consider the reviews, P.O.I.s, ratings, and the integration of P.O.I.s, review scores to improve the hotel recommendation systems. Our system can also help customers select the best-matched hotels when there are several hotels of the same category based on some features such as rank.

In the future, we will study methods and techniques which will improve our recommendation systems, and we will try to design the recommender system in a way that will consider dynamically updated data containing the reviews to provide better recommendations to the users. So, for example, the hotels which have improved their facilities after receiving low reviews will be considered. Another direction for future research might be using more data from different sources with different formats. Although a large-scale dataset was used in this paper for generating recommendations, more data with different parameters from other sources can be definitely helpful.

Author Contributions: Conceptualization, M.S.A.F. and M.S.A.; investigation, M.S.A.F., M.S.A., A.S.M.K., K.A., M.J.M.C. and I.K.; methodology, M.S.A.F., M.S.A. and A.S.M.K.; supervised the research; experiment, implementation and evaluation, M.S.A.F., M.S.A. and A.S.M.K.; writing—original draft preparation, M.S.A.F.; writing—review and editing, M.S.A., A.S.M.K., K.A., M.J.M.C. and I.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gellerstedt, M.; Arvemo, T. The impact of word of mouth when booking a hotel: Could a good friend's opinion outweigh the online majority? *Inf. Technol. Tour.* **2019**, *21*, 289–311. [CrossRef]
2. Hollenbeck, B.; Moorthy, S.; Proserpio, D. Advertising strategy in the presence of reviews: An empirical analysis. *Mark. Sci.* **2019**, *38*, 793–811. [CrossRef]
3. Ramzan, B.; Bajwa, I.S.; Jamil, N.; Amin, R.U.; Ramzan, S.; Mirza, F.; Sarwar, N. An intelligent data analysis for recommendation systems using machine learning. *Sci. Program.* **2019**. [CrossRef]
4. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
5. Hsieh, M.Y.; Chou, W.K.; Li, K.C. Building a mobile movie recommendation service by user rating and APP usage with linked data on Hadoop. *Multimed. Tools Appl.* **2017**, *76*, 3383–3401. [CrossRef]
6. Wu, C.; Wu, F.; Liu, J.; Huang, Y.; Xie, X. Arp: Aspect-aware neural review rating prediction. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; ACM: New York, NY, USA, 2019; pp. 2169–2172.
7. Xiang, Z.; Schwartz, Z.; Gerdes, J.H., Jr.; Uysal, M. What can big data and text analytics tell us about hotel guest experience and satisfaction? *Int. J. Hosp. Manag.* **2015**, *44*, 120–130. [CrossRef]
8. Tripadvisor.com. Available online: https://www.tripadvisor.com/ (accessed on 29 June 2021).
9. Agoda.com. Available online: https://www.agoda.com/ (accessed on 29 June 2021).
10. Expedia.com. Available online: https://www.expedia.com/ (accessed on 29 June 2021).
11. Booking.com. Available online: https://www.booking.com/ (accessed on 29 June 2021).
12. Sharma, Y.; Bhatt, J.; Magon, R. A multi-criteria review-based hotel recommendation system. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology, Ubiquitous Computing and Communications, Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, Liverpool, UK, 26–28 October 2015; pp. 687–691.

13. Cagliero, L.; La Quatra, M.; Apiletti, D. From Hotel Reviews to City Similarities: A Unified Latent-Space Model. *Electronics* **2020**, *9*, 197. [CrossRef]
14. Arefin, M.S.; Chang, Z.; Morimoto, Y. Recommending Hotels by Social Conditions of Locations. In *Tourism Informatics*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 91–106.
15. Yang, X.; Zimba, B.; Qiao, T.; Gao, K.; Chen, X. Exploring IoT location information to perform point of interest recommendation engine: Traveling to a new geographical region. *Sensors* **2019**, *19*, 992. [CrossRef] [PubMed]
16. Yang, Y.; Mao, Z.; Tang, J. Understanding guest satisfaction with urban hotel location. *J. Travel Res.* **2018**, *57*, 243–259. [CrossRef]
17. Chen, C.L.; Wang, C.S.; Chiang, D.J. Location-Based Hotel Recommendation System. In *International Wireless Internet Conference*; Springer: Cham, Switzerland, 2018; pp. 225–234.
18. Chen, L.; Chen, G.; Wang, F. Recommender systems based on user reviews: The state of the art. *User Model. User-Adapt. Interact.* **2015**, *25*, 99–154. [CrossRef]
19. Nicholas, C.K.W.; Lee, A.S.H. Voice of customers: Text analysis of hotel customer reviews (cleanliness, overall environment & value for money). In Proceedings of the 2017 International Conference on Big Data Research, Osaka, Japan, 22–24 October 2017; ACM: New York, NY, USA, 2017; pp. 104–111.
20. Mukta, R.B.M.; Arefin, M.S. An Agent Based Parallel and Secure Framework to Collect Feedbacks. *JCP* **2019**, *14*, 404–425. [CrossRef]
21. Liu, H.; He, J.; Wang, T.; Song, W.; Du, X. Combining user preferences and user opinions for accurate recommendation. *Electron. Commer. Res. Appl.* **2013**, *12*, 14–23. [CrossRef]
22. Zhang, J.J.; Mao, Z. Image of all hotel scales on travel blogs: Its impact on customer loyalty. *J. Hosp. Mark. Manag.* **2012**, *21*, 113–131. [CrossRef]
23. Kaggle.com. Available online: https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe (accessed on 29 June 2021).
24. Ganesan, K.; Zhai, C. Opinion-based entity ranking. *Inf. Retr.* **2012**, *15*, 116–150. [CrossRef]
25. Foursquare.com. Available online: https://www.foursquare.com/ (accessed on 29 June 2021).