National College of Ireland

Reinforcing Hotel Recommendation Systems through the implementation of Hybrid Modeling

MSc Research Project
MSc in AI for Business

Ana Sofia Lara
Student ID: x23105879

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ana Sofia Lara
Student ID:	X23105879
Programme:	MSc in AI for Business **Year:** 2023/2024
Module:	Research Project
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	August 12th 2024
Project Title:	Reinforcing Hotel Recommendation Systems through the implementation of Hybrid Modeling
Word Count:	20
Page Count:	8,833

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ana Sofia Lara
Date:	12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Reinforcing Hotel Recommendation Systems through the implementation of Hybrid Modeling

Ana Sofia Lara
X23105879

Abstract

So far in 2024, 326 new hotels have been or are planned to be build across Europe. Nowadays choosing which Hotel to travel to is an overwhelming experience due to the multiple options available online, it is time consuming and tiring. Current studies in the implementation of emerging technologies specially machine learning algorithms have targeted the usefulness of recommendation systems for Travel and Tourism services such as Hotel Recommendations. Diverse papers have discussed this subject focusing on the development of hybrid recommendations, leaning to implement not only user-based recommendations but also content and context based ones. This research will focus on the integration of both collaborative filtering (CF) and content-based filtering (CBF) to produce hybrid recommendation that will not only take similar users into account but the item attributes to formulate patterns of suggestion. The findings positively concluded that hybrid recommendations are more likely to be more accurate than only CF or CBF models individually as well as maintaining a high identification of false positive outcomes.

1. Introduction

The number of new hotels forecasted to open in 2024 across Europe is 326, according to the analysts of LE (Lodging Econometrics), specialists in the Travel and Tourism(T&T) sector (Hospitality Design, 2024). The T&T industry plays a crucial role in many countries' economies due to the amount of international investment it brings. According to the European Commission, the T&T industry generates valuable economic flow continent-wise, increasing the cultural exchange and local employment(European Commission, 2024). The Commission's Tourism Manifesto states that this industry represents 9.5% of Europe's GDP since it both employs approximately 22.6 million people and receives almost 579 million international travelers across the continent (European Commission, "Tourism Manifesto 2024") (See Figure 1).

After the pandemic, the T&T industry suffered losses due to the travel and health restrictions. However, according to the 2024 Travel and Tourism Development Index of the World Economic Forum, the sector's trajectory has positively increased due to the new technologies available that have supported efficient ways of travel and promote remote destinations(T&T Development Index, 2024).

In business terms, the T&T industry expands its service across the globe, specially in Europe. Companies focusing on offering leisure and accommodation services have been implementing all the available resources such as customer satisfaction practices and emerging technologies to increase travel flow and boost service quality and tailored travel options. Newest developments are extremely relevant for supporting customers in making decisions related to travel destinations, accommodation preferences, and activity selection.

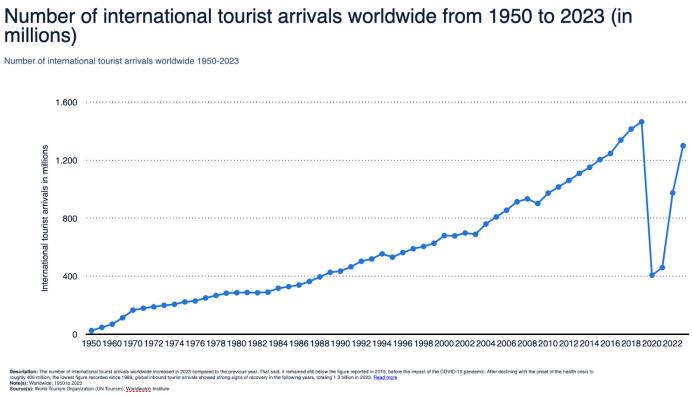


FIG 1 *- NUMBER OF TOURIST WORLDWIDE DURING 1950- 2023*[1]

The current conditions related to the digitalization of travel services such as online booking and travel recommendations has appeared as one of the most searched features nowadays due to the relief it offers to its users when browsing for options online who are overwhelmed by the number of options available. This ongoing issue has driven customers to to cancel travel plans, not for lack of information but an overflowing amount of it.

For this reason, one possible solution to attend to this issue, and support current conditions and user experience could be the implementation of AI-based models such as recommendation systems to impulse competitive environments by utilizing digital resources to identify customer preferences that will support the tailoring experiences, products, and services since Personalization, Prediction, and Forecasting of customers' needs and desires are some of the current requirements for customer-oriented industries such as Tourism (Bulchand- Gidumal, 2022).
The development of big data-based applications, AI-driven algorithms, and the deployment of specialized Machine Learning tools such as recommendation systems, chatbots, and other conversational systems converted into smart travel assistants focused on identifying, measuring, predicting, and forecasting Travelers behaviors and interactions, aiming to offer personalized treatments (Gavilan et al.,2018). Considering the latter, this research aims to understand customers' behavior by implementing a recommendation system to predict customer's travel choices such as Hotel selection. This project is looking to answer an initial research question: How *does implementing recommendation systems facilitate customer decisions on what Hotel to travel to next?*

To answer the research question and achieve the investigation objectives of generating a hybrid recommendation system integrating two models (Collaborative Filtering and Content-based Filtering) to recommend Hotels to users based on user data as well as to evaluate its performance by using metrics such as RMSE, Precision, and F1 Score; This paper will adhere to a detailed set of steps that will methodologically allow for the investigation to use, transform, and later on enforce the data into a model by first evaluating an EDA to later on preprocessed the data and finalized the final version of the dataset ready to be embedded in the two models that will be integrated to produce the hybrid recommendations. The paper

[1] Travel & Tourism Development Index 2024 M A Y 2 0 2 4. (n.d.). Available at: https://www3.weforum.org/docs/WEF_Travel_and_Tourism_Development_Index_2024.pdf.

will then be followed by an analysis of the results and a discussion of its implications, limitations, and future work in this area.

2. Related Work

Organizing travel plans can be quite the stressful and energy consuming task since it involves not only price comparison but also extensive decision making processes in terms of choosing plane tickets, itineraries, activities and of course hotel/accommodations. Recent technological developments have supported the quest of finding more effective ways to offer tailored services to its customers and lighten the load of having to choose between a too big band of options, reducing overwhelming sentiments related to traveling (Kiseleva et al, 2015).

2.1. Hotel Recommendation Systems

As emerging technologies increase in development and efficiency, the attention it receives has escalated as well, reaching more a more professionals in the different fields to analyze the benefits and challenges Machine Learning algorithms brings with them. Due to this, studies in the area of ML development have noticed that Recommendation systems (RS) started to appear in more recent years as one of the best models to enhance the predictions made on historic customer behavior and the best way to solidify the patterns discovered into offerings to particular clients (Kiseleva et al, 2015). Research propose RS as a very strong supporter for the identification of patterns, in the way it helps to ease the overload of data.

As the benefits of the recommendation systems collected more and more attention from organizations, the travel and tourism industry started to implement prediction techniques to better understand their customers. Hotels were not left behind, in the past there has been numerous studies about the different developments of recommender systems to support current operations. Some studies such as Tan et al (2021), covers the extent of the predictions based on the usage of recommendation systems in smart environments or rather industries such as e-tourism. Customers have grown more intelligent with time, not only tech wise but actually are expecting even more from the services and products offered to them. This particular subject has been treated by Linyouan L, et al (2012) and Goldenberg and Levi (2021), discussing on the user experience while using smart hotels and services; Due to new available technologies, predicting travel tendencies has arise the interest of specialist in the industry. Research has already explored and stated that accurate tendencies prediction over patterns and decisions is valuable and one of the most important foundations of new e-tourism practices (Isinkave et al. 2015).

Moreover, it has been mentioned that the past few years have experienced an incredible growth in the data that can be found online, both from companies and users, significantly rising the bar to the personalization needed to attract and successfully comply with the user's needs (Fararni et al, 2021). Studies have tackle this subject by introducing models based on predictive algorithms to use the information available online such as reviews, text comments, likes and any other user preferences to create tailored services, such as recommendation systems such as Hotels and Travel and Tourism products(Takuma et al, 2016).

2.2. Collaborative Filtering

As discussed, some of the models mostly seen during research about recommendation systems and prediction of customer preferences/behavior is collaborative filtering. According to some researchers like Kbajer, collaborative filtering has been the top priority when it comes to this kind of predictions due to its user-oriented nature(Kbaier et al, 2017). Collaborative filtering, focuses on the similarity between users or user-like items, such as ratings; its aim is to linked similar users together to recommend predictively liked items to the user in question(Chen et al, 2013). In the past, it has been the most used and studied model due to its flexibility and adaptable features to generate or rather determinate future user preferences based on past preferences and choices(Chen et al, 2013).

Recommender systems have been supported by collaborative filtering models due to the ability of using ratings as base of its performance. The past couple of years, with the development and popularity of platforms such as Netflix, Booking.com and any other streaming, social and booking services, an increase in the deployment of collaborative filtering based algorithms have been observed, since this platforms use recommendation systems as a tool for personalization of services which is the new business approach to targeting growing customer engagement(Bodhankar et al, 2019). Matrix factorization techniques such as SVD and ALS (Alternating Least Squares) are popular and the go to choice to implement said models, as well as other machine learning algorithms such as neural networks, K-means, K- , which target the grouping of similar patterns to generate recommendations(Kbaier et al, 2017).

All together, a review on past research on the subject of collaborative filtering algorithms provides a clear understanding of the approaches that have been analyzed in the past (Chen et al, 2021). Said analysis have stablished a consecutive review of the findings, highlighting the potential impact of the algorithm on the diverse industries must of all e-commerce platforms and customer-oriented processes.

2.2.1 Singular Value Decomposition

One of the methods encountered during the analysis of past studies on the subject of collaborative filtering based recommendation systems is Single Value Decomposition. Researchers have used this method to emphasize the matrix factorization of vectors to ensure the detailed evaluation of unseen patterns across the data (Sheng et al, 2005) . Studies such as (Sheng et al, 2005) and (Qilong et al, 2013) have focused on the use of SVD as main technique to develop due to its high quality recommendations and the accuracy it has proven to provide in the outcomes. The algorithm has been praised by past researches due to its adaptability towards different dataset sizes, since it allows to modify to dimensionality of the matrix ergo the dataset, supporting the correct functionality of the model when its computational size overpowers the tools of the research(Qilong et al, 2013).

2.3 Content-based Filtering

On the other hand, content-based filtering is not as old as collaborative filtering. This model was introduced later on due to new findings (Tang et al, 2014) signaling its useful ability for recommendation systems.

Later on, studies presented the benefits content-based filtering has for ML models such as RS due to its nature of using item characteristics to find linked interconnections and form patterns, including user preferences in its suggestions. As tailored services were the new features of customer oriented business, papers on the use of CBF highlight the great impact

this model made in the industry, now not only were recommendations based on similar users but it included user past preferences and particular likes of item attributes(Tang et al, 2014). Research assumes that content-based filtering in comparison to other models, allows and at the same time needs the data to be not so much complete but categorized into diverse and numerous attributes that will allow the model to perform recommendations based on different features, permitting personalized suggestions and more detailed analysis of the recommendations(Guo et al,2020). Moreover, studies on this subject, vary on the techniques used to perform CBF, some detailed the use of Euclidean Distance while other present the implementation of methods such as Pearson Correlation or more complex ones such as neural networks, TF-IDF and Word Embedding (Word2Vec), and the most use of all, Cosine Similarity(Biswarup et al, 2021)(Reddy et al, 2023).

2.3.1. Word2Vec and Cosine Similarity

Word Embedding is one of the techniques linked to content-based filtering or rather text vectorization and it functions as a word vector, giving similar words the same score to further model processing(Ozsoy, M. 2024). For models using text as one of their inputs, word vectorixation can support the numerical representation and in dense dimensionalities, reduces its computational complexity(Musto, C. Et al, 2016). It has been observed that along cosine similarity, Word2Vec is an incredible option to develop recommender systems specially to simplify the process of identifying similar words within reviews or large paragraphs, create patterns and control the understanding the model has of the data(Reddy et al, 2023).

2.4 Hybrid Recommendations

Although, models such as Collaborative filtering and Content-based filtering are quite effective when concerning with the prediction of patterns for recommendations, researches such as (Cui, et al. 2022) and later on (Zhang et al, 2024) argue that for a complete and accurate parameter of predicted connections, studies must not only take into account historical data but rather try to find supporting data that could sufficiently uphold the most precise recommendations such as user preferences, interests and behavior.Studies promoted by (Jalan et al, 2017) and (Zhang et al, 2024) highlight the benefits of implementing hybrid model to generate the recommendations, specially when CF and CBF are included since the combination of both their functions result in an increase in performance and overall accuracy. Since Collaborative filtering focuses on finding similar users or items by the identification of patterns, recommendations based on its results can suggest similarities between users in a same group whereas content-based filtering undertakes all the historical data from the user in question and creates suggestions based on similar items as the ones liked or purchased by the user.

Multiple studies have tackled this subject through the implementation of different techniques, seeking to encounter the best possible methodology to proceed with the development of the models and later on their deployment. Some researchers such as (Yassine et al, 2021) have discussed the implementation of hybrid recommendation through neural networks, while some have directed their research into the use of similarity functions between 2D arrays of user-item pairs to reduce dimensionality (Basilico et al, 2004). Collaborative filtering and content based complementations literature have decided to dive into the utilization of both models to explore the benefits it can bring the industry such as the development of movie

recommendations (Geetha et al, 2018), travel destinations and tourism planning (Goldenberg and Levi, 2021), as well as many other industries and operations. As on-going projects focus on the use of R.S though hybrid modeling, the need of accurate evaluation metrics increases not only to enforce precise methods to analyze performance both of the data and the two conjoint algorithms but rather due to the need to provide high quality recommendations to customers based on their historical data and preferences which is why metrics such as the Root Mean Square Error (Yan-Martin et al, 2021), Precision (Abuzir et al, 2021), recall and others like F-score (Zangerle and Bauer, 2022) are being used to sized the performance and most of all evaluate the accuracy of the recommendation by targeting a low difference between the predictions and the actual results(Beel et al, 2015).

3. Research Methodology

To be able to proceed with the research and satisfy the curiosity of finding out if it's possible to predict accurate recommendations about hotels based on customer reviews and ratings, this paper will follow a precise framework based on the CRISP-DM approach for data mining and statistical analysis, as well as the guidelines established by past research on Travel predictions specially the work of Dimitri Goldenberg and Pavel Levin in their paper "Booking.com multi-destination trip dataset" (Goldenberg and Levi, 2021)[2]. The methodology will be divided into 6 main steps as follows:

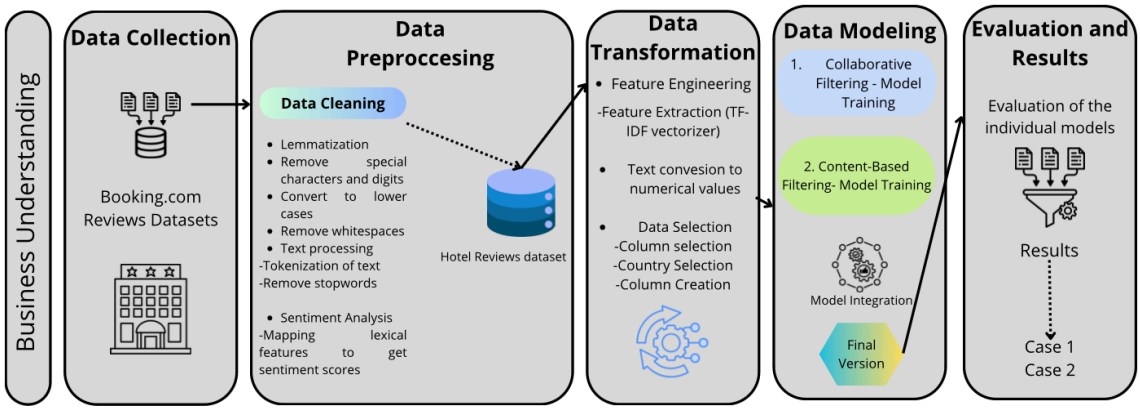


***Figure** 2: Methodology Framework for the research*

3.1 BUSINESS/ INDUSTRY UNDERSTANDING

As discussed above, the business in this case refers to the Travel and Tourism industry, most especially the Hotel sector. The situation trying to analyze and solve is the vast number of options overwhelming the user when deciding which hotel to travel to next and the need for tailored services willing to offer support in tough decisions such as this one.

3.2 DATA UNDERSTANDING/ DATA COLLECTION

This research will be based on the dataset "Hotel_Reviews.csv" originally called "Booking.com reviews dataset" collected from the official website of Booking.com through Crawl Feeds and downloaded from Kaggle. The dataset is available for everyone, being public domain and imported to the Kaggle platform 3 years ago in 2021 by @opensnippets. It

- 2 https://github.com/bookingcom/ml-dataset-mdt/blob/main/Dataset____Multi_Destination_Trips.pdf.

is a quite large dataset, sizing 47.02 MB. Regarding its content, the dataset counts 15 columns as can be seen in Figure 3, and 26,386 rows. Each row is a review of a hotel gathered as mentioned from booking.com.

review_title	Title of the review
reviewed_at	Date of the review
reviewed_by	User who wrote the review
Images	Images attached to the review
crawled_at	The date the review was tracked
URL	URL of the review
hotel_name	The name of the hotel
hotel_url	URL of the hotel in booking.com
avg_rating	Average rating
Nationality	Nationality of the user
# rating	Rating made by the user
review_text	Text of the review
raw_review_text	Raw text of the review
Tags	Tags of the review
Meta	Raw data of the review

FIG 3 - HOTEL_REVIEWS DATASET COLUMNS DEFINITION

During the understanding of the data, an initial Exploratory Data Analysis was done to gather insights about the data and its content. It was discovered that the data counted with only two countries as mentioned, Belgium and Belarus, being Brussels the city with the most hotels (2160 Hotels). The hotel with the most reviews is the Marivaux Hotel in Brussels with 449 reviews. Moreover, only 209 reviews had images attached to the review submitted and the dataset contains reviews from July 2018 to July 2021.For further understanding, the hotels are rated from 0 to 10, the reviewers come from 123 different nationalities, in its majority from the UK.

3.3 DATA PREPROCESSING

3.3. 1 Data Cleaning

The next stage in the methodology of this research is data cleaning to ensure that the data is ready for implementation and avoid or rather mitigate any potential malfunction or error in the following stages. Steps such as lemmatization of the text to reduced to root words, removal of special characters, whitespaces and conversion of text to lower cases is key to normalize the data. In addition, text preprocessing is needed to fully proceed with the preparation of the dataset, steps such as tokenization of data (split text into individual words “tokens”) and removal of stopwords, resulting in a cleaned/processed data-frame.

3.3. 2 Data Preparation

The original raw version of the dataset, as mentioned counted 15 columns, one of them was the hotel name and hotel URL which revealed the location of the hotel that in this case was needed to fully understand the different destinations within the dataset and the range of coverage of the research. Due to this, the first step of the data preprocessing was to add two new columns to the original dataset, “City” and “Country”. This was conquered by filtering the replicated hotels through Excel and filling out all the same hotel names with their city and country until the dataset was completed and the data was cleaned.

Moreover, as the dataset was compiled with both text and numerical data, it is important to review and clean the text by implementing lemmatization techniques as well as normalization

methods to correct inconsistent formatting and increase the uniformity of the data such as converting lower cases, removing digits and punctuation as well as removing whitespaces within the text. Further, it is necessary to remove stopwords within the review and tokenize the text to later on lemmatize it. In addition, categorical values need to be encoded so the machine learning algorithm can fully read them and have data consistency within the dataset. To do this, Sentiment Analysis through the NLTK, Vader library is used to generate the sentiment score reviews and have numerical values per review. In overview, it was needed to perform tokenization, stop word elimination, and lemmatization of the text for further processing.

3.3. 3 Data Transformation

Feature Engineering

As mentioned and following the preprocessing of the data, feature engineering is required to accommodate the data into its most useful version by organizing the information in a way that significantly supports the aim of the research. By using the sklearn library and methods such as TF-IDF, it was able to vectorize the text data and turn it into numerical values for the algorithm to be able to read. The vectors were then put into a tf-idf matrix to statistically measure the importance of each word (column) in a document (row) and the corpus(all the rows).

After transforming the text into numerical components, the dataset goes through data preparation efforts to add the needed columns, in this case, the creation of a new column "user_id" was required to identify each user.

Data Selection

Furthermore, Data selection was implemented to achieve the wanted organization of the data, instead of the initial 15 columns plus the 4 new columns generated during the preprocessing stage (cleaned_text, processed_text, sentiment_score_reviews, encoded tags), the final version of the dataset counts with 13 of the 19 focusing only in the columns valuable for the prediction. Along with this, during the data selection, specifically in the column "Country", the selection of only include Belgium was made as leaving Belarus could potentially decrease the accuracy of the model in terms of prediction logic and context. Predicting different hotels in only one country leads to more precise recommendations. The final dataset can be seen in Figure 4.

```
df_final = df[['hotel_name', 'City', 'Country', 'hotel_url', 'avg_rating', 'nationality', 'rating', 'tags',  'cleaned_text', 'processed_text', 'sentiment_
df_final.head()
```

	hotel_name	City	Country	hotel_url	avg_rating	nationality	rating	tags	cleaned_text	processed_text	sentiment_scc
0	Villa Pura Vida	Kortrijk	Belgium	https://www.booking.com/hotel/be/villa-pura-vi...	9.7	Poland	10.0	Business trip~Solo traveller~Junior Suite~Stay...	everything was perfect quite cozy place to relax	everything perfect quite cozy place relax	
1	Villa Pura Vida	Kortrijk	Belgium	https://www.booking.com/hotel/be/villa-pura-vi...	9.7	Belgium	9.0	Leisure trip~Couple~Deluxe Suite~Stayed 1 nigh...	very friendly host and perfect breakfast	friendly host perfect breakfast	
2	Hydro Palace Apartment	Ostend	Belgium	https://www.booking.com/hotel/be/hydro-palace....	9.2	United Kingdom	10.0	Leisure trip~Couple~Apartment with Sea View~St...	it was just what we wanted for a week by the b...	wanted week beach winter location fab apartmen...	
3	Villa Pura Vida	Kortrijk	Belgium	https://www.booking.com/hotel/be/villa-pura-vi...	9.7	Netherlands	10.0	Business trip~Solo traveller~Junior Suite~Stay...	my stay in the house was a experiencing bliss ...	stay house experiencing bliss luxury house she...	
4	Hydro Palace Apartment	Ostend	Belgium	https://www.booking.com/hotel/be/hydro-palace....	9.2	South Africa	9.2	Leisure trip~People with friends~Apartment wit...	the building itself has a very musty smell in ...	building musty smell hallway despite built apa...	

***FIG** 4 - VISUALIZATION OF THE FINAL DATASET*

3.4 *MODELING:*

3.4.1. Collaborative Filtering:

The First filtering model is collaborative filtering since it will allow supporting the recommendations with personalized features due to prior rating systems and similar characteristics (Cheng et al,2021). In this research, this model allows to use of the hotel name list and its ratings to create a ratings system, this system later on implemented in the model, will grant tailored suggestions to the user based on their history and past preferences. The method used to achieve the user-based recommendations is SVD (Single Value Decomposition) by developing matrix factorization to the columns "hotel_name", "user_id" and "rating".

3. 4.2. Content-based filtering:

Regarding the second model, content-based filtering will be focusing on the grouping of similar hotels by analyzing its attributes and generate recommendations based on similar Hotels to the ones liked by the particular user. Content-based filtering was developed by the implementation of Word Embedding, using the Word2Vec model. Matrix factorization will enable the model to use vectors in the text reviews to be able to later on apply cosine similarity and found the proper recommendations.

In more detail, the gensim library allows the Word2Vec to utilize the tokenized reviews already gathered and with the addition to the generated aggregated tokenized reviews, allowing to cluster unique hotel reviews in a group will later on be the foundation for the matrix factorization needed to calculate similarities of hotels based on reviews by the resulting similarity scores that delivers the cosine similarity.

3. 4.3. Hybrid Modeling - Integration of the Models

Finally, the models will be integrated to confirm the effectiveness of both sides of the spectrum. As collaborative filtering focuses on similar users based on multiple and content-based filtering in the user characteristics and preferences, the integration of both models can result in a better understanding of the user and their future decision-making processes.

3.5 EVALUATION

To evaluate the performance and accuracy, three distinct metrics were implemented to compare their results and analyze the model execution Root Mean Square Error, Precision, and F1 Score. The three metrics were calculated by the comparison of both actual and predicted ratings, to oversee how close was the model results to the prediction.

4. Design Specification

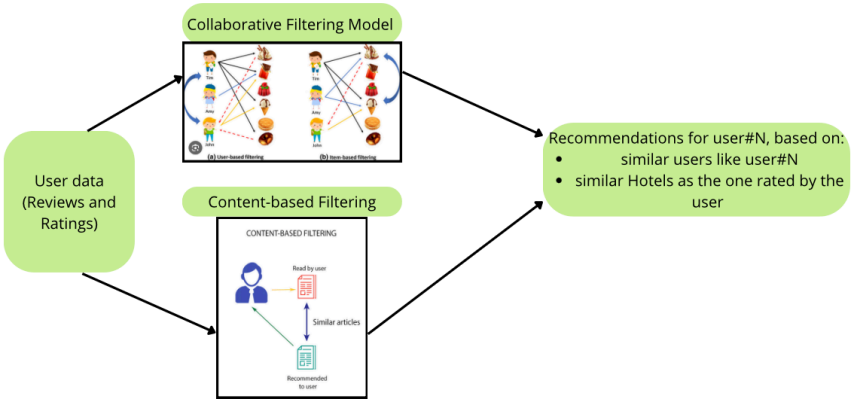


FIG 5 *- Design Specifications of the Hybrid Recommendation - The Model Architecture*

4.1 USER DATA

As the figure 5, shows, the design will be based on the user data being embedded in the model. For the side of collaborative filtering, the ratings made by the diverse users will be needed whereas for the content-based filtering which focuses on item to item similarities, the reviews will be equipped into processing-able data to be the core of the functionality of the CBF model recommendations.

Further, the user data represents the final version of the data mining, showing the result of the earlier processes carry out involving data collection, data preprocessing and data transformation which were developed by taking into account guidelines such as the ones proposed by Deepanshi (2024), Mbaabu (2023), Zhao (2022) and Python Programming official guidelines such as the one presented by Medium (2023). Regarding special steps within the mentioned processes, like sentiment analysis were generated by the frameworks developed by past researchers on this field, like GitHub in the paper "TextSentimentAnalysis" and Suvrat Arora (2024) through Analytics Vidhya. As well, feature engineering sequences and guidelines were introduced to have further understanding of the data, such as studies from Raymond Cheng published in Towards Data Science (2023) and Priya Muthu guidelines in Kaggle (2021).

4.2 COLLABORATIVE FILTERING

Later on, during the Modeling stage, two different models were selected as has been mentioned since past literature on this subject suggests the better quality and accuracy of recommendations the integration of both models generates instead of focusing the entire implementation in only one of them (Eticha,A. 2020). Although, collaborative filtering is the model most used for research in the past to produce tailored recommendations based on one user or item similarity, in the limitations of said work was found that additional user information can be a game changer. Collaborative filtering (CF) itself finds the similarity between users or items using matrix factorization methods to be able to understand underlined patterns within the data. By identifying connections between the users, the model is able to create predictions of what the user would like based on its own preferences which are alike other users. As can be observed in the Figure 5, CF can actually use multiple users to stablished interconnections and group them by vectors to categorized the data and later on make suggestions(Peng et al, 2024).

To achieve CF, the method of SVD was implemented due to its high functionality with larger datasets as well as its attributes of finding similar users that like the same item, clustering users by their preferences. Studies have used this method to develop their algorithms reasoning that is the method most suitable to generate model-based collaborative filtering (Sheng et al, 2005) (Qilong et al, 2013) The main components needed to stablished the base line to develop said model are user_id, hotel_name and ratings.

4.3 CONTENT-BASED FILTERING

Regarding the second model, CBF was implemented by the cooperation of two methods: Word Embedding and Cosine Similarity. Past research varies on the methods used to achieve recommendations based on similar item targets, ensuring the user has an array of possible suggestions. In this case, investigations close in the methodology discussed diverse methods

such as TF-IDF, Latent Semantic Analysis (LSA), Word2Vec, Doc2Vec, and Cosine Similarity, the choice resides in the context of the problem, objectives, and data collection. After careful analysis, this report goes ahead with the implementation of the method Word2Vec which with its neural network-based nature can support Word Embeddings-based models by generating vector representation of words and text values (Tang et al, 2014). Although other methods also show promising results, research has found that when comparing, Word2Vec shows the best performance and accuracy (Musto, C. Et al, 2016) (Ozsoy, M. 2024).
On the other hand, cosine similarity was inserted in the model as well to increase its end performance, studies have suggested the collaboration of both methods to achieve founded and precise recommendations in this case, in the recommendation of Hotels (Reddy et al, 2023). Cosine Similarity supports the model by grouping hotels aka vectors by their similar components/attributes, helping to cluster Hotels by their similar reviews.

4.4 Evaluation

Finally, the evaluation metrics chosen to gauge the performance of the two diverse models and later on their integration are as mentioned RMSE, Precision, and F1 Score. The reason behind the selection was the suitability of the metrics themselves, studies have presented evidence that for recommendation systems these are the most optimal for detailed analysis of the model performance; Precision measures the competence of the model (Abuzir et al, 2021), according to literature, is one of the most common and globally understood measure when it comes to recommendation systems and algorithms (Yan-Martin et al, 2021). However, additional metrics have been encountered as useful to evaluate model performance and accuracy such as Root Mean Square Error (specially for the evaluation of predicted and actual ratings), Recall, and F-scores (Beel et al, 2015). As the second and third measure, the paper will implement RMSE and F1-score due to their natures, while RMSE is key to analyzing rating predictions, the F1 score is valuable to evaluate classification results which in this case is useful when classifying pertinent hotels based on user ratings(Zangerle and Bauer, 2022).

5. Implementation

How the methodology was developed is key to understanding the results of the hybrid recommendation and the overall performance of the research.

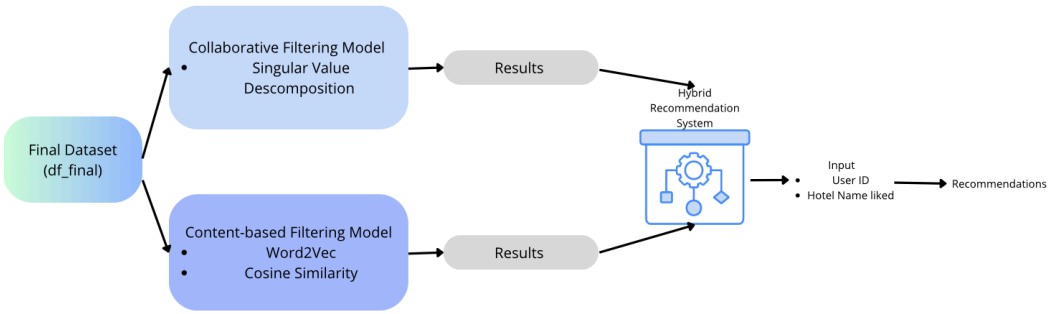


FIG 6- IMPLEMENTATION FRAMEWORK

1. **Collaborative Filtering**

The following steps reflect how the methodology was implemented to develop both the individual models and the hybrid recommendation as the last phase as seen in Figure 6. Regarding the first part of the implementation, Collaborative Filtering was developed by the installation of the surprise library since it was not available within the Colab dictionary. From the library, diverse components were imported such as the class Dataset Reader, two matrix factorization algorithms: Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF), cross-validation, train/test split, Gridsearch CV to find the best parameters for evaluation and finally accuracy.

The model development follows a set of steps after the library importation, first, it identifies the reader within a rating scale of 1-10 since that is the actual rating scale observed in the dataset. Then, by embedding the data into the model, the three required components were stablished : User id, Hotel name and Ratings. Further, the SVD algorithm was developed and later on the results of the training set analyzed through cross validation using 3 folds, the results generated test_rmse = 0,224619 which suggested that the average measure of difference between the actual rating and the prediction was close to 0.22. Regarding the test_mae= 0,121091, it can be said that the model predictions in general are distinct from the actual ratings by 0,12. As per the fit_time and test_time which are 0,66 and 0,008 correspondently, it was observed that the time the model took to both train and test the model for each fold is quite good and suggests a good performance. The evaluation of the model by using a test set, will be discussed later on the Evaluation section. The guidelines taken into account to produced this model were found Klaudia Nazarko “Model Based Collaborative Filtering Recommender” research (Nazarko, 2020) and Medium blogger Amy in her 2020 article “Recommendation System: User based Collaborative Filtering” (GrabNGoInfo, 2022) and video tutorials.

2. Content-based Filtering

On the other hand, the second model was designed by importing a particular branch of the Sklearn Library, metrics.pairwise to develop cosine similarity between pairs of vectors. Initially, the first step needed to proceed with the implementation was to make sure that all the reviews were properly tokenized as this second model will be based on the text reviews themselves. As a result, the data Fram “tokenized reviews” serve as a sentence aka a list of lists of words needed to produced the first technique the model will be founded on: Word Embedding. To do this, the Word2Vec model was generated through the importation of the gensim library and then using the tokenized reviews to train the model itself. The parameters set to train the model were initially larger than the final version, since it was needed to readjust due to lack of computational capacity. The vector size is 100, which set the dimensionality of the vectors within enough information to process the recommendation, as well as the window being 5, which is a exponential decision based on past research and the use of content-based-Word2Vec models (Musto et al, 2016). As per the minimum count, 1 is the perfect number to avoid all vectors lower than 1 token, supporting the Continuous bag of Words model, which was chosen with sg=0, the latest parameter was imposed due the lack of existence of

Moreover, since the dataset contains multiple reviews of the same hotels, only being 795 unique hotels for all of the reviews, it set a limitation that needed to be handle in order to

continue with the development of the content-based model, resulting in the generation of unique vectors for each repeated hotel that later on would be linked together to create clusters of reviews (group_reviews); the new dataset, was constructed of only unique hotels and the aggregated tokenized texts.
Further, the fourth step within the framework of implementing content based filtering was to calculate average word vectors for each group of reviews, meaning to generate list of tokenized words. By producing a list of word vectors for each word in a review, it ensured that only the words present in the Word2Vec model were included to avoid run disruptions and being able to calculate the mean of the word vectors later needed when the second model was going to be introduced: Cosine Similarity. By using the already imported library, the function was aiming to calculate the cosine similarity between two pairs of word vectors, to do this, it was needed to first transform the review vectors into a sequence to later stack using np.stack into a 2D infrastructure where each row was a review vector (group of vectors of a unique hotel). Next, the matrix factorization was imposed and a cosine similarity matrix generated between all pairs in the 2D array, the final step was to convert the matrix into a working data frame for later testing.
The prior was the result of the analysis of diverse guidelines and past studies such as Deva Lindey with her work in "Recommender Systems using Word Embeddings" (Lindey, 2024), Luong Vuong in the paper "Content-Based Collaborative Filtering using Word Embedding: A Case Study on Movie Recommendation" (Guo et al,2020) and the video tutorials of Krish Nail on how to build a Content-based Recommendation System.

3. Hybrid Recommendation

Finally, the hybrid recommendation was the result of the integration of the two prior models. Notebook Implementation Part III shows the development of the final stage. As first instance, it was required to upload the needed csv files of the two prior models as this was a new notebook, however, first it was compelled to first import the libraries involved to upload them such as pandas and gensim, this last one to decipher the Word2Vec model. As mentioned, the csv files were uploaded (Collaborative_filtering_model.csv and the following for the content-based filtering: Word2Vec_model.model, aggregated_reviews.csv for the group reviews vectors and cosine_similarity.csv for the similarity scores).

Moreover, the integration was initiated by first ensuring that the content-based recommendations were in order by running the function over again def get_content_based_recommendations. After successfully ensuring its functionality, it was time to generate the hybrid recommendations by embedding the function df get_hybrid_recommendations and directing it on the selection of user id and hotel name. It must be clear that the hybrid recommendations are the integration of the content based recommendations based on text reviews and the collaborative filtering predictions based on user ratings, which is why the next step was to merge both models by the hotels names, combining the rating scores with similarity scores through the enforcement of weighted average, in addition to ensuring that if there was a rating or score missing this will be fill with 0, securing the running time of the model.
To ensure the right balance between the scores, the alpha parameter was set up in 0.5, meaning that both scores were equally important. The alpha score reflected the CBF similarity scores at 0,5 and (1-alpha) represented the CF prediction ratings. This was done to

delivery offer balanced recommendations based equally on both models. The development of the final stage was done based on past research: Basilico et al (2004), Geetha et al(2018), Haidar (2024) and video tutorials such as the one from Spencer Pao “Building a Recommendation System in Python”.

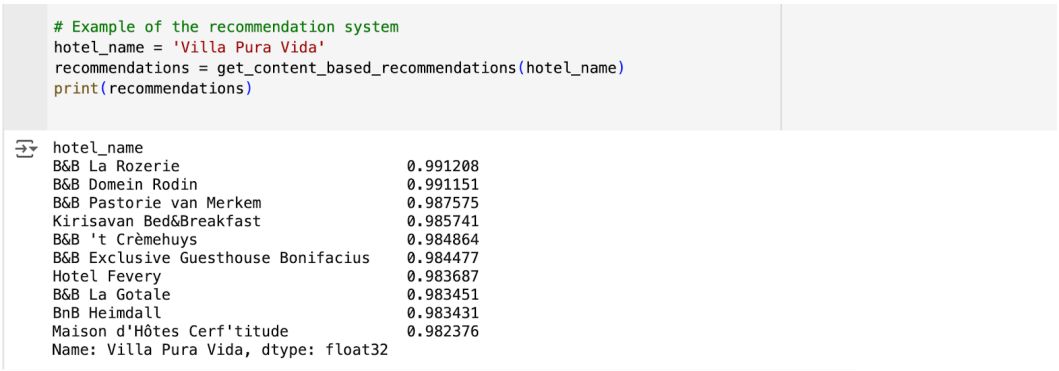
6. Evaluation

Regarding the Evaluation phase, as seen already, this research more than to only answer the research question, seeks to reach a tight level of accuracy to prove not only successful implementation of the model but rather precise recommendations and proper prediction of customer travel decisions in terms of hotel selection. After the implementation was completed, the model casted some insightful results, valuable to conquer final analytical thoughts on the performance of the hybrid recommendation.

6.1 Content-Based Filtering Results

On the other hand, the content based filtering model, which was developed by taking into account the text reviews as base of the model since whereas collaborative filtering targeted similar users to based its recommendations, content based focused on similarity scores of the reviews of the unique Hotels by clustering them and suggest similar hotels from the one the user already reviewed (Tang et al, 2014).

The Figure # 8 shows the initial testing of the model, by interpreting the text reviews using Word Embedding and group similar reviews by implementing cosine similarity.



```
# Example of the recommendation system
hotel_name = 'Villa Pura Vida'
recommendations = get_content_based_recommendations(hotel_name)
print(recommendations)
```

```
hotel_name
B&B La Rozerie                          0.991208
B&B Domein Rodin                        0.991151
B&B Pastorie van Merkem                 0.987575
Kirisavan Bed&Breakfast                 0.985741
B&B 't Crèmehuys                        0.984864
B&B Exclusive Guesthouse Bonifacius     0.984477
Hotel Fevery                            0.983687
B&B La Gotale                           0.983451
BnB Heimdall                            0.983431
Maison d'Hôtes Cerf'titude              0.982376
Name: Villa Pura Vida, dtype: float32
```

FIGURE 8 - CONTENT-BASED RECOMMENDATION EXAMPLE

6.2 Collaborative Filtering Results

First instance, the collaborative filtering model throw an idea of the accuracy of the recommendation based on user based items such as the ratings in this case. As mentioned above, this particular model used based characteristics to cluster similar users in groups and predict future suggestions based on similar likings (ratings). The Hotels given as recommendation are the most similar rating wise as the ratings gave by this particular user.

As can be observed in the Figure # 9 , which shows the RMSE (Root Mean Square Error) of the collaborative filtering during the testing phase, it can be deduced that the performance of the first model is quite positive, mainly due to the RMSE value which is the one this report will be focusing on, that is 0.1835 leading to a small difference between the predicted ratings

and the actual ones, resulting in not so far off recommendations as can be seen in the following analysis (Beel et al, 2015).

6.3 Hybrid Recommendation Systems Results

Finally, the hybrid recommendation is based on the integration of both models, preparing to deliver recommendations based on not only user like items like similar ratings but also taking into account the reviews themselves by suggesting similar hotels. As has been discussed, the collaboration of both approaches will allow to offer more complete recommendations, which is why evaluating its performance its critical to debate its functionality. For this purpose, three different evaluation metrics had been incorporated to analyze the accuracy and overall performance of the hybrid recommendation model. Before inserting the evaluation metrics functions, the data was split into two groups, training and testing sets as well as the development of predictions to as its name suggest predict the future recommendations and test how far off are those predictions from the actual numbers (RMSE, AskPython, 2020).

The three metrics are as follows:

PERFORMANCE METRICS COMPARISON			
	Root Mean Square Error	Precision	F1-Score
Collaborative Filtering	0.1835	N/A	N/A
Hybrid Modeling	0.1483	1	0.8

FIGURE 9 -PERFORMANCE METRICS COMPARISON

As per the RMSE, it is one of the most common used evaluation metrics when recommendation systems or rather prediction models are involved. In this case, as it targets the difference between predicted ratings and actual ones, it is the most valuable metric to measure collaborative systems models. As can be observed in the Figure 9, the RMSE of the model is approximately 0.15, meaning that there is little difference between the predictive ratings and the actual ones, giving the recommendation a better accuracy.(RMSE, AskPython. 2020). According to past studies on this subject, the lower the RMSE value, the better the model(Ahoudi et al, 2021);in comparison to other papers who have developed hybrid recommendations based on both collaborative and content based filtering, the results of the RMSE are promising and insightful enough to suggest that the predictions of the methods used to implement CF and CBF are a positive match, beneficial for the overall outcome.
Further, it is insightful to compare both the collaborative filtering model RMSE results (0.1835) and the hybrid recommendation system results (0.1483), leading to deduct that the hybrid recommendation which is the collaborative filtering model with the support of the content-based filtering model actually does seem improved by the integration and the difference between predicted ratings and actual ones is less.

In second instance, this performance review metric will allow to measure the precision (as its name suggests) of the recommendation system. The metric excels in identifying positive predictions and increasing the accuracy of the model itself. As the precision metric goes from 0 to 1, the Figure 9 allows to observe that the precision is 1, meaning that the predictions of

the model, the positive ones are indeed correct and there are no false positives in the model recommendations. Past studies such as the ones made in the subject of hybrid recommendation (Ahoudi et al, 2021), suggest that a good precision metric identifies the possibility of errors or off recommendations inside the implementation, it is quite the good signal that the measurement in this case represents the verity of the accuracy of the final suggestions (Scikit-learn, 2024).

Regarding the last evaluation metric, F1 score allows us to target a more complete overview of the integration of both precision and recall. This metric as it concentrates itself also in the identification of false positives and negative predictions, permits the analyst to evaluate the accuracy of the final outcomes(Grossman ,M. 2023). The reason to include it in the final evaluation of the model is that F1 score permits to analyze the balance between precision and recall, as well as contextually give insights on the existence of false negative or false positives which in this context is quite important(Grossman, M. 2023). This model pretends to focus on a high customer-oriented industry, false positives or negatives in the final recommendation suggest lack of accuracy and the potential negative consequences it could have in the overall customer engagement. Referring to the metric itself, Figure 9 allows to observe a F1 score of 0,8 signaling to a high and well maintain balance between precision and recall and the model successfully or rather positively identifies false negatives or positives in its outcomes(Grossman, M. 2023).

As an additional metric of evaluation, there was an weighted average comparison between different percentages to analyze how the balance between the models in the hybrid recommendation affected its results. After using different configurations: (0,5-0,5) (0,7-0,3) and (0,2-0,8) it was discovered that there were no significant changes in the overall evaluation metrics and that the recommendation in its essence did not change its results.

ADDITIONAL PERFORMANCE ANALYSIS - PERFORMANCE IN THE CONTEXT OF RUNNING TIME

TIME/PERFORMANCE COMPARISON BETWEEN THE 3 MODELS (in seconds)		
Collaborative Filtering	Content-Based Filtering	Hybrid Model
68.90	29.86	398.8

FIGURE 10 - RUNNING TIME PERFORMANCE (IN SECONDS)

As shown by Figure 10, the collaborative filtering model developed has a running of 68.90 seconds. This particular model which was created by using Singular Value Decomposition (SVD), generates recommendations based on user/item preferences such as ratings in this case to recommend hotels liked by users who have similar preferences/attributes. SVD supports the creation of vectors for identifying said similarities, although, the model's efficiency is significantly dependent to the size of the data not only in its running time but in the final accuracy of the predictions.

On the other hand, the Content-based filtering model was developed by using Word2Vec and Cosine Similarity to as mentioned, generate recommendations based on the user's hotel

review history and travel purchase interactions. This second model as seen by Figure 10, had a running time of 29.86 seconds which might be the result of lighter methods applied during the development like Word2Vec instead of TF-IDF or lower parameters of training and testing to reduce computational expenditure.

Finally, the third model, the hybrid recommendation as result of the combination of the two first models had a running time of 398.79 seconds approximately, equivalent to 6 minutes. The efficiency of the model depends mainly on the right implementation of the two models being used as pillars for the final recommendation as well as more general factors such as data size and the overall model parameters which is weighted average in this case that for example is 0.5, creating a balance between collaborative filtering and content based filtering in the final result as can be seen in both Figure 10 and Figure 11.

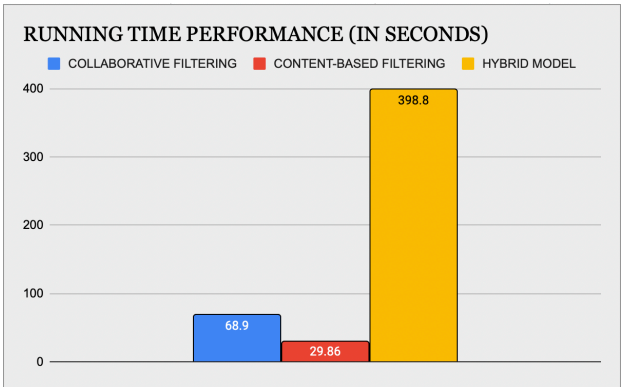


FIGURE 11 - RUNNING TIME PERFORMANCE COMPARISON (IN SECONDS)

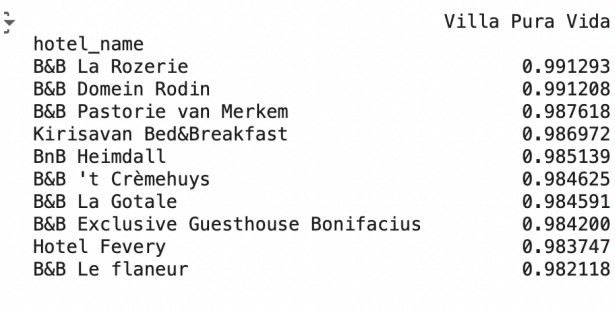
6.4 Experiments

To fully evaluate the model, the following experiments were done by randomly selected two users. The first case, has User=1who liked the hotel Novotel Gent Centrum, after running the model, the following hotels were recommended based on user similarities and hotel attributes.

hotel_name	Novotel Gent Centrum
ibis Gent Centrum St. Baafs Kathedraal	0.997302
nhow Brussels Bloom	0.996685
Theater Hotel	0.996657
Thon Hotel Brussels City Centre	0.995709
Marivaux Hotel	0.995557
Aris Grand Place Hotel	0.995085
pentahotel Leuven	0.995060
Hotel 't Putje	0.994968
Sputnik Hotel	0.994422
Hotel Britannia	0.994391

***FIGURE** 12 - EXPERIMENT 1*

The second case has User=12 who attended and later reviewed the hotel Villa Pura Vida. After running the model, these were the top hotels recommended for this user to travel to.



hotel_name	Villa Pura Vida
B&B La Rozerie	0.991293
B&B Domein Rodin	0.991208
B&B Pastorie van Merkem	0.987618
Kirisavan Bed&Breakfast	0.986972
BnB Heimdall	0.985139
B&B 't Crèmehuys	0.984625
B&B La Gotale	0.984591
B&B Exclusive Guesthouse Bonifacius	0.984200
Hotel Fevery	0.983747
B&B Le flaneur	0.982118

***FIGURE** 13 - EXPERIMENT 2*

6.5 Discussion

As past literature has already discussed, the implementation of recommendation systems has shown promising results in terms of customer engagement. Understanding the user is key to tailor their experience and improve overall performance. When it comes to the Travel and Tourism Industry which is highly customer oriented, the user experience is a crucial component to ensure the best quality of service. Hotels, more now than ever once they have learned to adapt to the new technological environment and regulations past COVID pandemic, have found significant improvement in their business analysis when ML algorithms as well as other emerging technologies are applied to comprehending user behavior (Tan et al, 2021). The use of recommender systems as well as other AI bots, have proven to be revolutionary for this particular industry since it offers around-the-clock service and more than that tailored assistance to customers anywhere, improving efficiency (Das M, 2023).

This particular research focuses on recommending customers' future hotel choices based on user historical experiences and preferences as well as predicting the best possible Hotel for each user. By the end of the development and after the embedding of the data into the hybrid recommendation model, insightful results were observed. The hybrid model results suggest the hotels with that similar users used and hotels with similar reviews from the ones made by the particular user, attempting to predict the more likely hotels that will successfully fill all the customer needs and preferences.

Therefore, what do the results entail in the context of the travel industry?
Developing a successful recommendation system as has been seen in the review of past literature is useful and effective for developing supported systems of assessment of user decision-making processes. As technological advancement surged in the different industries, it has affected the T&T industry in great depth (Reddy et al, 2023). Overall recommendation systems have appeared to have improved the quality of its services, providing tailored options to customers as well as guiding them through the numerous offerings, reducing the time-consuming activity of looking for suitable Hotels (Abuzir et al, 2021).

In particular, the development of hybrid recommendation systems based on the integration of two different models especially when it comes to Collaborative filtering and Content-based filtering helps to support more accurate recommendations, not only basing the results on user-based items (such as ratings) but also complementing those outcomes with content-based factors such as reviews giving recommendations based on both preferences of similar users

and similar items based on user history (Guo et al, 2020). Past literature as disclosed earlier in the report has found this integration to be resourceful in generating veritable recommendations to users, especially in the industry of Travel and Tourism where user experience and tailored services have been found crucial to increase profitable operations (Guo et al, 2020)(Tang et al,2014) in addition to fighting correct issues such as the Cold Start Problem (Collaborative filtering) and the lack of data versatility (CBF). The implementation of said algorithms and ML models can enhance the performance of digital e-tourism platforms such as booking.com and serve as a complement to ongoing initiatives. This research will prove to be valuable in measuring the usefulness of implementing the collaboration of different approaches in one model, ensuring the production of hybrid recommendations and increasing the personalization of the outcome which although has been already studied by a couple of past researchers, is still in the process of being refined by the analysis of the mix and match of different methods to develop said ML models. As mentioned before, there is a variety of models used in past studies regarding the integration of CF and CBF, and still, the search for the best match of methods is open which is why this research collaborates on finding the best possible infrastructure and algorithms to work efficiently and accurately to develop high-quality recommendations.

7. Conclusion and Future Work

In this project, the previously stated research question lays the ground to develop an enhanced analysis of available solutions to attack or rather improve the actions to satisfy the current demand in the Tourism and Hospitality Industries. The benefits a successful outcome could have in the industry, reach out to be valuable for companies in this sector to take into account, since the addition of technological and digital tools in the market, allows for innovative solutions to be found and ergo excel as a result in a pool of never-ending options.
Both Collaborative filtering and content-based filtering are useful models to generate tailored recommendations based on users. Past studies in this area have used one of the two models to illustrate the convenience of utilizing machine learning algorithms to improve user experience and precision in personalized services and products, however, only recently, researchers have gone the extra mile and integrated both models into one, not only offering one side of the recommendation meaning based on similar user characteristics but rather combine to sides of the coin and offer customized recommendations based both on similar customers and similar items to be purchased/used so the recommendations integrate the user preferences, historical data and the item similarities.

As per the research question, it can be concluded that implementing recommendation systems can indeed facilitate customer decisions on what Hotel to travel to next since the hybrid model generates suggestions based on that particular user preferences by clustering similar users into groups and recommending the Hotels that the other members have liked in this case by the ratings they have disclosed, in addition to this, the content-based filtering supports the recommendation by finding similar Hotels within the reviews that can be then recommended to the user in question.

7.2 Limitations

Moreover, throughout the research some limitations were found within the framework established. First, by using Google Colab as the main platform of development, there were some limitations regarding the RAM and storage available (12.67 GB) for the algorithms to run. The production of the three main parts of the hybrid model had to be done in three individual notebooks as one would not have been able to hold all the code snippets and run the cells. Later on, when developing the second model CBF, some limitations were found in terms of algorithm weight, at first the model was meant to be implemented using TF-IDF and cosine similarity, however, the first method was too heavy for Colab to run completely, due to this and to avoid additional processing time, the final method implemented was Word Embedding which was able to give similar outputs with lesser storage space requirements. On the other hand, another limitation found was the dataset itself. Although this data served as the proper foundation for the development of the model, the initial dataset was not easily comprehensible, and further preparation was needed due to this.

7.3 Future Work

As has been presented throughout the literature review, there have been diverse attempts to find the best match of methods to perform both Collaborative Filtering and Content-based Filtering specially since it was found that the integration of both models could result in more accurate recommendations as well as precise personalization of the service, although there was a very detailed and extensive search on different investigations, future work can be guided through a more complex analysis of all the different techniques to dutifully encounter the most effective methods. Further, the use of more extensive datasets could give an additional value to the overall methodology not only depending on one dataset to implement the model into but extra data to base the recommendations on complementary analysis such as larger datasets, datasets from other sources, and additional overall information about the user preferences as well as the Hotel characteristics.

Regarding the research as a whole, it can be concluded that future work can be done in better and more effective ways to use emerging technologies such as ML algorithms based on user history and purchase behavior. This type of development could increase even more the user experience throughout the whole operation and ensure that the customers do not feel the overwhelming feeling of the paradox of choice but rather use technological tools to reduce the options and support users to opt for the tailored recommendation since it’s based on their needs and likes.

Moreover, the addition of extra metrics for recommendation such as time, social media posts, language, deeper reviews and comments, can give even more personalized recommendations that could actually identify user attributes in a way that could guide not only suggest your next Hotel selection or where to travel to.

References

- **Basilico, J. and Hofmann, T., 2004.** Unifying collaborative and content-based filtering. *Twenty-first international conference on Machine learning - ICML '04*. Available at: https://doi.org/10.1145/1015330.1015394.
- **Bulchand-Gidumal, J., 2022.** Impact of Artificial Intelligence in Travel, Tourism, and Hospitality. In: Xiang, Z., Fuchs, M., Gretzel, U., and Höpken, W. (eds.) *Handbook of e-Tourism*. Cham: Springer. Available at: https://doi.org/10.1007/978-3-030-48652-5_110.

- **Gavilan, D., Avello, M. and Martinez-Navarro, G., 2018.** The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, 66, pp. 53–61.

- **Geetha Mohan, Safa Iqubal, Fancy Chelladurai and Saranya, D., 2018.** A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System. *[online]* Available at: https://www.researchgate.net/publication/324763207_A_Hybrid_Approach_using_Collaborative_filtering_and_Content_based_Filtering_for_Recommender_System.

- **Zhang, Y., Tan, H., Jiao, Q., Lin, Z., Fan, Z., Xu, D., Xiang, Z., Law, R. and Zheng, T., 2024.** A Predictive Model Based on TripAdvisor Textual Reviews: Early Destination Recommendations for Travel Planning. *SAGE Open*, 14(2). Available at: https://doi.org/10.1177/21582440241246434.

- **Goldenberg, D. and Levin, P., 2021.** Booking.com Multi-Destination Trips Dataset. *International ACM SIGIR Conference on Research and Development in Information Retrieval*. Available at: https://doi.org/10.1145/3404835.3463240.

- **Hospitality Design, n.d.** Lodging Econometrics Releases European Hotel Pipeline Data. *[online]* Available at: https://hospitalitydesign.com/news/business-people/lodging-econometrics-european-hotel-pipeline-q1-2024/.

- **Data.world, 2018.** data.world. *[online]* Available at: https://data.world/opensnippets/bookingcom-reviews-dataset/workspace/project-summary?agentid=opensnippets&datasetid=bookingcom-reviews-dataset.

- **Travel & Tourism Development Index 2024 M A Y 2 0 2 4, n.d.** Available at: https://www3.weforum.org/docs/WEF_Travel_and_Tourism_Development_Index_2024.pdf.

- **Singgalen, Y., 2024.** Implementation of CRISP-DM for Social Network Analysis (SNA) of Tourism and Travel Vlog Content Reviews. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 8, pp. 572-583. Available at: https://doi.org/10.30865/mib.v8i1.7323.

- **Musto, C., Semeraro, G., de Gemmis, M. and Lops, P., 2016.** Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems. In: Ferro, N., et al. (eds.) *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science*, vol. 9626. Cham: Springer. Available at: https://doi.org/10.1007/978-3-319-30671-1_60.

- **Ozsoy, M., n.d.** From Word Embeddings to Item Recommendation. *[online]* Available at: https://arxiv.org/pdf/1601.01356.

- **Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B., 2014.** Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. *[online]* Association for Computational Linguistics, pp. 1555–1565. Available at: https://aclanthology.org/P14-1146.pdf.

- **Reddy, M.S., Kumar, P.T.R., Siddarth, L.M. and Mothukuri, R., 2023.** Designing Recommendation System for Hotels Using Cosine Similarity Function. In: Ranganathan, G., EL Allioui, Y. and Piramuthu, S. (eds.) *Soft Computing for Security Applications. ICSCS 2023. Advances in Intelligent Systems and Computing*, vol. 1449. Singapore: Springer. Available at: https://doi.org/10.1007/978-981-99-3608-3_1.

- **Abuzir, Y. and Dwieb, M., 2021.** Hotel Recommender System based on Knowledge Graph and Collaborative Approach. *International Journal of Computing*, pp. 63–71. Available at: https://doi.org/10.47839/ijc.20.1.2093.

- **Yan-Martin Tamm, R., Damdinov, R. and Vasilev, A., 2021.** Quality Metrics in Recommender Systems: Do We Calculate Metrics Consistently? In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21)*. New York, NY, USA: Association for Computing Machinery, pp. 708–713. Available at: https://doi.org/10.1145/3460231.3478848.

- **Beel, J. and Langer, S., 2015.** A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems. In: Kapidakis, S., Mazurek, C. and Werla, M. (eds.) *Research and Advanced Technology for Digital Libraries. TPDL 2015. Lecture Notes in Computer Science*, vol. 9316. Cham: Springer. Available at: https://doi.org/10.1007/978-3-319-24592-8_12.

- **Zangerle, E. and Bauer, C., 2022.** Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys*, 55(8), pp. 1–38. Available at: https://doi.org/10.1145/3556536.

- **Tan, R., Chen, H., Jing, X., Jin, Z. and Deng, S., 2021.** Customer Experience of Smart Hotel Based on Network Evaluation Information. In: Ahram, T.Z. and Falcão, C.S. (eds.) *Advances in Usability, User Experience, Wearable and Assistive Technology. AHFE 2021. Lecture Notes in Networks and Systems*, vol. 275. Cham: Springer. Available at: https://doi.org/10.1007/978-3-030-80091-8_61.

- **Guo, W., Nguyen, T.-H. and Jung, J.J., 2020.** Content-Based Collaborative Filtering using Word Embedding. *[online]*Available at: https://doi.org/10.1145/3400286.3418253.

- **Chen, Y.C., Hui, L. and Thaipisutikul, T., 2021.** A collaborative filtering recommendation system with dynamic time decay. *Journal of Supercomputing*, 77, pp. 244–262. Available at: https://doi.org/10.1007/s11227-020-03266-2.

- **Zhang, S., Wang, W., Ford, J., Makedon, F. and Pearlman, J., 2005.** Using singular value decomposition approximation for collaborative filtering. In *Seventh IEEE International Conference on E-Commerce Technology (CEC'05)*, Munich, Germany, pp. 257-264. Available at: https://doi.org/10.1109/ICECT.2005.102.

- **Ba, Q., Li, X. and Bai, Z., 2013.** Clustering collaborative filtering recommendation system based on SVD algorithm. In *2013 IEEE 4th International Conference on Software Engineering and Service Science*, Beijing, pp. 963-967. Available at: https://doi.org/10.1109/ICSESS.2013.6615466.

- **Cui, C., Wei, M., Che, L., Wu, S. and Wang, E., 2022.** Hotel recommendation algorithms based on online reviews and probabilistic linguistic term sets. *Expert Systems with Applications*, 210, p. 118503. Available at: https://doi.org/10.1016/j.eswa.2022.118503.

- **Sundermann, C., Antunes, J., Domingues, M. and Rezende, S., 2018.** Exploration of Word Embedding Model to Improve Context-Aware Recommender Systems. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Santiago, Chile, pp. 383-388. Available at: https://doi.org/10.1109/WI.2018.00-64.

- **Kiseleva, J., Bernardi, M., Davis, C., Kovacek, I., Einarsen, M.S., Kamps, J., Tuzhilin, A. and Hiemstra, D., 2015.** Where to Go on Your Next Trip? *arXiv*. Available at: https://doi.org/10.1145/2766462.2776777.

- **Isinkaye, F., Folajimi, Y.O. and Ojokoh, B.A., 2015.** Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3), pp. 261–273. Available at: https://doi.org/10.1016/j.eij.2015.06.005.

- **Takuma, K., Yamamoto, J., Kamei, S. and Fujita, S., 2016.** A Hotel Recommendation System Based on Reviews: What Do You Attach Importance To? In *2016 Fourth International Symposium on Computing and Networking (CANDAR)*, Hiroshima, Japan, pp. 710-712. Available at: https://doi.org/10.1109/CANDAR.2016.0129.

- **Ray, B., Garain, A. and Sarkar, R., 2021.** An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98, p. 106935. Available at: https://doi.org/10.1016/j.asoc.2020.106935.

- **Kbaier, M.E.B.H., Masri, H. and Krichen, S., 2017.** A Personalized Hybrid Tourism Recommender System. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia, pp. 244-250. Available at: https://doi.org/10.1109/AICCSA.2017.12.

- **Chen, J.-H., Chao, K.-M. and Shah, N., 2013.** Hybrid Recommendation System for Tourism. In *2013 IEEE 10th International Conference on e-Business Engineering*, Coventry, UK, pp. 156-161. Available at: https://doi.org/10.1109/ICEBE.2013.24.

- **Jalan, K. and Gawande, K., 2017.** Context-aware hotel recommendation system based on hybrid approach to mitigate cold-start problem. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, pp. 2364-2370. Available at: https://doi.org/10.1109/ICECDS.2017.8389875.

- **Bodhankar, P.A., Nasare, R.K. and Yenurkar, G., 2019.** Designing a Sales Prediction Model in Tourism Industry and Hotel Recommendation Based on Hybrid Recommendation. In *2019 3rd International Conference on Computing Methodologies and*

Communication (ICCMC), Erode, India, pp. 1224-1228. Available at: https://doi.org/10.1109/ICCMC.2019.8819792.

- **Fararni, K.A., Nafis, F., Aghoutane, B., Yahyaouy, A., Riffi, J. and Sabri, A., 2021.** Hybrid recommender system for tourism based on big data and AI: A conceptual framework. *Big Data Mining and Analytics*, 4(1), pp. 47-55. Available at: https://doi.org/10.26599/BDMA.2020.9020015