A Comprehensive Machine Learning-Based Hybrid Approach for Personalized Tourism Recommendations in Ireland

MSc Research Project
MSc Data Analytics

Mohib Tariq
Student ID: 23259850

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….Mohib Tariq ………………………………..……………………………………………………

Student ID: ……x23259850……………………………………………………………………………………..……

Programme: ………MSc Data Analytics…………………………… **Year:** ………2024…………..

Module: ……Research Project……………………………………………………….………

Supervisor: ……Teerath Kumar Menghwar…………………………..…………………….………

Submission Due Date: ………12th December 2024………………………………………………………….………

Project Title: …A Comprehensive Machine Learning – Based Hybrid Approach For Personalized Tourism Recommendations in Ireland………………………

Word Count: …………8864……………………………… **Page Count**………28……………………….……..

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: …………………Mohib Tariq………………………………………………………………

Date: ………………12th December 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):	

A Comprehensive Machine Learning-Based Hybrid Approach for Personalized Tourism Recommendations in Ireland

Mohib Tariq
x23259850

Abstract

Recommendation systems have become very important in providing personalized recommendations to users across many domains, including tourism. The main aim of this research study is to develop and evaluate a recommendation system to promote tourism in Ireland. A hybrid approach is implemented in this study, which is a combination of content-based filtering, collaborative filtering, and cluster-based techniques. The study uses three distinct datasets that contain information about Ireland's tourist attractions, Ireland's accommodations, and cultural landmarks in Ireland. Throughout this study, a detailed methodology was used involving data pre-processing, feature engineering, and machine learning modeling techniques. The evaluation of the models was done using evaluation metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score, which ensured a strong performance comparison. The findings showed that the weighted hybrid approach performed better than the individual content-based and collaborative filtering methods and showed accurate and relevant recommendations. This research adds an important contribution to the field of tourism recommendation systems by offering a unique and innovative approach to improving user experiences.

Keywords: Recommendation System, Content-Based Filtering, Collaborative Filtering, Cluster-Based Techniques, Hybrid Approach

1 Introduction

Nowadays the demand for unique needs and recommendations has given more importance to recommendation systems. These include our daily interactions with online platforms and services like video streaming websites such as YouTube and e-commerce websites and so on as discussed in the study by Delimayanti et al. (2022). The usage of these recommender systems is dependent upon user data such as user preferences. In this way, these systems help to improve customer engagement rates and user satisfaction.

In recent times, the tourism sector has offered much popularity to recommendation systems which are used by many travelers and tourists to discover attractions and famous places that match with their interests which helps to enhance their experience. The growing tourism in Ireland, emphasizes the need for accurate recommendation systems that can provide relevant recommendations of destinations and places to tourists which can make their experience even more enjoyable.

There have been many studies carried out on recommendation systems which mainly had a focus on the traditional methods: content-based filtering and collaborative filtering. Both of these methods have some limitations such as the cold start problem also highlighted by Sarkar et al. (2023). There was a need for more advanced approaches which could help to overcome the problems faced by the traditional methods. Hybrid recommendation systems can further help in solving the limitations of the traditional methods as they provide more efficient solutions some of which are discussed by Yang et al. (2022).

1.1 MOTIVATION AND IMPORTANCE

This research was carried out after studying the limitations of traditional recommendation approaches like content-based filtering and collaborative filtering. Content-based filtering depends on the content i.e. item features. Hence, this approach suffers from the limitation that it is not able to make more recommendations which are based on the similarity of recommendations between different users as also pointed out in the research by Manikandan & Kavitha (2023). On the other hand, collaborative filtering requires a lot of user-item interaction data to make accurate recommendations. This causes the cold start problem which means that the technique is unable to provide recommendations for new users as discussed by Sarkar et al. (2023). This can have a bad effect on user satisfaction and can also affect the overall performance of the system.

In order to increase user satisfaction, it is important to provide solutions to these limitations. The problems can cause the system to struggle with personalization which can lead to a reduced engagement of users. This research aims to fill the gap by implementing and studying a hybrid recommendation model that combines content-based filtering, collaborative filtering, and cluster-based features to overcome the limitations that were caused by the traditional methods and provide efficient tourist recommendations. By using a combination of multiple techniques, the hybrid approach aims to improve the accuracy and diversity of recommendations. This helps to develop a more robust solution that also deals with user preferences effectively.

1.2 RESEARCH PROBLEM AND HYPOTHESES:

As discussed above, the main problem addressed in this research is the limitations of traditional recommendation systems, such as the cold start problem, limited personalization, and less accurate recommendations. To address these problems, this research proposes a hybrid recommendation approach that combines multiple machine-learning techniques to improve the quality of recommendations.
The research question that gives strength to this study is: “**How does a hybrid recommendation approach, combining content-based filtering, collaborative filtering, and cluster-based features, improve the accuracy and relevance of tourist recommendations of Ireland compared to traditional methods like content-based filtering and collaborative filtering**?”

The main objectives of this research are:

- To study the content-based, collaborative, and hybrid approaches in recommendation systems and evaluate their performance.
- To design and implement a hybrid approach that uses multiple features and metrics to provide improved recommendations.
- To evaluate the hybrid approach and compare the effectiveness with the traditional approaches using evaluation metrics like RMSE, MAE, precision, and recall.
- To use the hybrid approach to provide accurate recommendations and promote tourism in Ireland.

1.3 CONTRIBUTION OF THIS RESEARCH:

The study provides insights on how the hybrid approach can enhance recommendation accuracy and personalization. The research offers a valuable solution for industry practitioners, in the context of tourist attractions in Ireland by providing more accurate and personalized recommendations. This research can help to improve user satisfaction, increase customer engagement, and help tourists seek recommendations for popular destinations in Ireland.

The following structure is used for this research thesis: Section 2 presents a literature review of previous studies in recommendation systems. Section 3 outlines the research methodology, including feature selection, model training, and evaluation techniques used in the development of a hybrid recommendation model. Section 4 discusses and evaluates the results obtained from the different approaches that are used to analyze the performance of the recommendation approaches. Finally, the last section summarizes the findings of the research and suggests some of the future work that can be carried out to enhance the effectiveness of recommendation systems in the tourism and accommodation domains.

2 Related Work

2.1 Overview Of Recommendation Systems

Over the last few years, the need for recommendation systems has increased across many industries. To implement these recommendation systems, there are many algorithms which are used to predict the preferences of users based on past user behavior, content features, and other demographic features. In the domain of tourism, recommendation systems play a very important part as they help to provide personalized recommendations to tourists regarding destinations, accommodations and other activities. . The study by Manikandan & Kavitha (2023) supported this point by discussing the importance of recommendation systems. Furthermore, Sarkar et al. (2023) emphasized the importance of tourism recommendation systems and discussed that tourist recommendation systems must take additional factors like Environmental factors, geolocations, and user preferences for more high-quality recommendations.

Lu (2022) discussed that there are some complexities in the case of tourism recommendation systems. If we compare the user preferences of the tourism industry with that of e-commerce and social media, the tourism experiences are diverse and dynamic which requires the recommendation system to take into account external factors like weather, season, etc. Many studies discuss the importance and need for scalable, and region-specific recommendation systems which are capable of handling large datasets and adapting to the changing preferences of tourists.

2.2 Motivation And Scope Of The Literature Review

The rise of the digital era has transformed the travel industry. In Ireland, where tourism is one of the most important sources of the economy, the need for advanced recommendation systems arises. The study conducted by Chongwarin et al. (2024) discusses the challenges in the existing recommendation systems, which are data sparsity, scalability, and region-specific datasets. This literature review aims to explore methodologies and critically evaluate their strengths, and weaknesses with respect to the research question.

2.3 Content-Based Filtering:

Content-based filtering focuses on the characteristics or content of items like descriptions, tags, features, etc. to recommend similar items to users. In the tourism industry, content-based filtering has been used for analyzing unique attributes like destination features, types of accommodations, or descriptions of activities. Manikandan & Kavitha (2023) mention the use of TF-IDF Vectorization and cosine similarity which can recommend tourist attractions based on descriptions of user preferences. Many applications in the tourism industry have features like personalized planning, which suggests recommendations based on past history. Lu (2022) demonstrated how content-based filtering models can recommend scenic spots by utilizing content metadata. On the other hand, Vaishnavi & Kalpana (2023) highlighted the application of filtering tourist packages that were tailored to the user preferences. Moving on, Delimayanti et al. (2022) discussed the use of cosine similarity and KNN to recommend movies. This use can be followed in tourism recommendation systems too. The research done by Luostarinen & Kohonen (2013) applied Latent Dirichlet Allocation (LDA) for extracting features in a news recommender system which shows the potential for analyzing tourist reviews. The study by Sivanaiah et al. (2020) implements a content-boosted hybrid filtering system which solves the problem of providing recommendations to pessimistic users as well.

Content-based filtering offers transparency as the recommendations can be interpreted, as they are based on clear item features. Furthermore, they are also user-independent as it does not always rely on other user’s data for individual personalization. However, some studies have discussed the limitations encountered by content-based filtering methods. Manikandan & Kavitha (2023) discuss the feature engineering challenges. For more effective implementation, it is important to get more meaningful item descriptions which can be resource-intensive to generate.

Moving on, several solutions to these limitations have been proposed in many studies. Lu (2022) discussed the possibility of integrating semantic analysis with feature extraction techniques such as deep learning to solve the cold start problem. Furthermore, many hybrid approaches which are discussed later, can mitigate the problems as well.

2.4 Collaborative Filtering:

Collaborative Filtering focuses on user preferences to make recommendations. The patterns in user-item interactions are studied and analyzed in order to produce effective recommendations. There are two approaches commonly used in collaborative filtering methods which are User-Based Collaborative Filtering, and Item-Based Collaborative Filtering. In the Item-Based Collaborative Filtering approach, similar items are suggested to those previously rated or suggested by the same user. Moving on, the study conducted by Alabduljabbar (2023) used SVD and Non-Negative Matrix factorization to develop a restaurant recommendation system in Riyadh. This helps to analyze that these methods are also adaptable in large diverse datasets.

Adding to this, Chang et al. (2025) proposed a trust-aware collaborative filtering algorithm that used interval-valued matrix factorization to integrate user-trust relationships with ratings. This was done to achieve better accuracy in more sparse datasets. On the other hand, Xu et al.

(2023) carried out research that incorporated privacy-preserving techniques like differential privacy which added security concerns in recommendation systems.
Many studies have reported the scalability problems in Collaborative Filtering. Matrix factorization techniques like Alternating Least Squares and Singular Value Decomposition (SVD) are widely used to solve the scalability problems in collaborative filtering. Chongwarin et al. (2024) discuss the application of these approaches and discuss that these approaches are used to decompose the user-item interaction matrix to gain insights from latent features that support the predictions of unseen attractions. Moving on, the work done by Lu (2022) introduced a hybrid model utilizing user-item interactions and other factors to improve the quality of scenic spot recommendations. The study by Manikandan & Kavitha (2023) also explored collaborative filtering approaches for e-learning and provided insights that can be used in the context of tourism recommendation systems.
Many challenges like data sparsity, cold-start problems, and scalability are encountered by collaborative filtering methods. Lu (2022) discusses that external data sources like user-generated content can be integrated which increases the prediction accuracy by enriching the data matrix.

2.5 Hybrid Approach

Hybrid recommendation systems are able to combine content-based filtering collaborative filtering and other techniques to overcome the limitations faced by individual techniques. Xu et al. (2023) proposed a hybrid model that combined collaborative filtering with clustering and privacy-preserving techniques. This helped to achieve improved accuracy in tourism point-of-interest recommendations. There are various approaches to hybrid recommendation systems like Weighted Hybrid models, Switching Models, and Cascade Models which are used. Switch models alternate between content-based filtering and collaborative filtering depending on the scenario. The research by Vaishnavi & Kalpana (2023) discusses the method of cascade models which uses the output of one method as input for the other to improve the quality of recommendations. Furthermore, Lu (2022) also demonstrated the efficiency of combining user-item factorization with semantic content analysis to recommend tourist attractions, which helps to achieve more efficient results. The model proposed by Alabduljabbar (2023) used a hybrid recommendation system using collaborative filtering methods with contextual data, which helped to gain many improvements in user satisfaction rates. Furthermore, the research by L & B (2024) used a ridge regression model to implement a recommendation system which prevented over-fitting. Adding to this, Ruchika et al. (2023) developed a hybrid recommendation approach using deep learning including multiple layers and stages. Yao et al. (2021) perform a simulation of users to see the impact of the recommendation under different types of user behavior. The study by Yang et al. (2022) provides an overview of many intelligent recommendation algorithms including hybrid approaches which are most commonly used in industries.

The evaluation of hybrid models is performed to verify their ability to mitigate the cold-start problem by using content features when there is no user-item interaction data. Moreover, the combination of user behavior with item features allows for more enhanced and personalized recommendations. The study by Lu (2022) discusses the Context-Aware Hybrid Models which integrate contextual data into hybrid frameworks to enhance the recommendations dynamically. Further on, Sarkar et al. (2023) highlighted the ability of the hybrid recommendation systems to balance accuracy and diversity. Manikandan & Kavitha (2023) discuss that hybrid systems have a layered nature where it can be challenging to ensure interpretability.

2.6 Clustering Techniques In Tourism Recommendation Systems

Clustering groups the users or items based on similarities, which helps to increase the quality of personalized recommendations. The study conducted by Vaishnavi & Kalpana (2023) discusses that these techniques are mainly used in hybrid approaches where segmentation helps to increase the accuracy of predictions. There are many types of clustering techniques that are used, like K-Means clustering, Hierarchical clustering, and density-based clustering. In the study by Weiwei et al. (2022), users were segmented based on travel preferences, and those interested in adventure were grouped in a similar group. In the study by Chongwarin et al. (2024), destinations were grouped according to attributes like popularity, location, and user ratings for more accurate recommendations. Compared to this, the study by Appadoo et al. (2020) segments users based on different factors like salaries, experience e.t.c to propose a job recommendation system. In the study by Xu et al., (2023), clustering was performed along with collaborative filtering to identify private neighbor groups which enhanced the accuracy of point-of-interest recommendations. Further on, Xu et al. (2023) reported that satisfaction and precision were improved by using the clustering approach.

2.7 Evaluation Metrics For Recommendation Systems

Evaluation metrics are very important to measure the performance of recommendation systems. In the context of tourism recommendation scores, several metrics have been used. Root Mean Square Error and Mean Absolute Error (MAE) are used to evaluate the systems that measure the error in predicted user ratings compared to actual ratings. Vaishnavi & Kalpana (2023) used Precision, Recall, and F1 Score to evaluate the quality of top-k recommendations.

Xu et al. (2023) discussed that the hybrid recommendation system, which used collaborative filtering, helped to achieve a higher precision and recall score. Furthermore, Chang et al. (2025) showed that interval-values matrix factorization improves the RMSE when compared to standard collaborative filtering methods.

2.8 Gaps In Literature

Despite all of the advancements in recommendation systems, there are still some gaps, especially in the context of tourism. These gaps reflect that there is still a need for further exploration to enhance the efficiency and effectiveness of recommendation systems in the real world. According to Chongwarin et al. (2024) , there is a limited availability of datasets for every region or country like Ireland so tourist recommendation systems can be developed for each country.

There is a very limited integration of contextual factors like geo-political issues and safety concerns into tourism recommendation systems. Syahputra et al. (2022) discuss the importance of using contextual factors in the recommendation system. However, according to the study by Vaishnavi & Kalpana (2023) have yet to explore how negative events like political instability can affect user decisions in the context of tourism.

Moving on, another gap is regarding the unavailability of travel times in tourism recommendation systems. The study by Karurukan et al. (2018) provides regression-based models to estimate travel times. The studies by Zhang & Balog (2020) and Xu et al. (2023) discussed the influence of travel time on user experience in the context of tourism planning.

Moreover, using social network-based information in collaborative filtering models is also an underdeveloped area. There are still some studies like Kim et al. (2018) and Weiwei et al. (2022) where they use social network intimacy scores to make recommendations more personalized, but still, this approach is not widely adopted across the tourism industry.
In addition, there is a lack of research on the integration of clustering techniques in recommendation systems. As discussed in the studies by Zhang & Balog (2020) , and Weiwei et al. (2022), the use of clustering has not been explored fully as most studies still focus on simple segmentation approaches.
Finally, there are fewer evaluation metrics that measure the performance of tourism recommendation systems. The existing metrics that are used, like RMSE and MAE, focus more on accuracy than other factors like diversity and user satisfaction. While there are many studies, like that of Vaishnavi & Kalpana (2023) and Karurukan et al. (2018), that provide insights regarding the evaluation metrics, there is still a need for an evaluation framework that involves both predictive accuracy and user-centered metrics.
Concluding the above studies, the research conducted has laid a good foundation for recommendation systems in the field of tourism. However, for further advancement in the area, it is essential to further move into the field. Future research can contribute to more robust, dynamic, and user-focused recommender systems that provide personalized tourism experiences.

In Summary, there is a significant progression in developing personalized recommendation models using content-based filtering, collaborative filtering, and hybrid approaches. However, there are still many existing solutions do not take into external factors. Furthermore, there is a possibility of getting more personalized and diverse recommendations by integrating social networking data and advanced clustering techniques. Despite the advancements in the type of evaluation metrics, there are still many models that have a primary focus on accuracy, hence user satisfaction is neglected. These gaps describe the limitations of the current methodologies, which support the need for more dynamic and context-aware approaches to tourism recommendations. Therefore, this research study aims to address the limitations by developing a hybrid recommender system that includes contextual factors and using advanced clustering techniques that offer a more robust, and user-centric solution that is tailored to the dynamic needs of users in regard to tourism.

3 Research Methodology

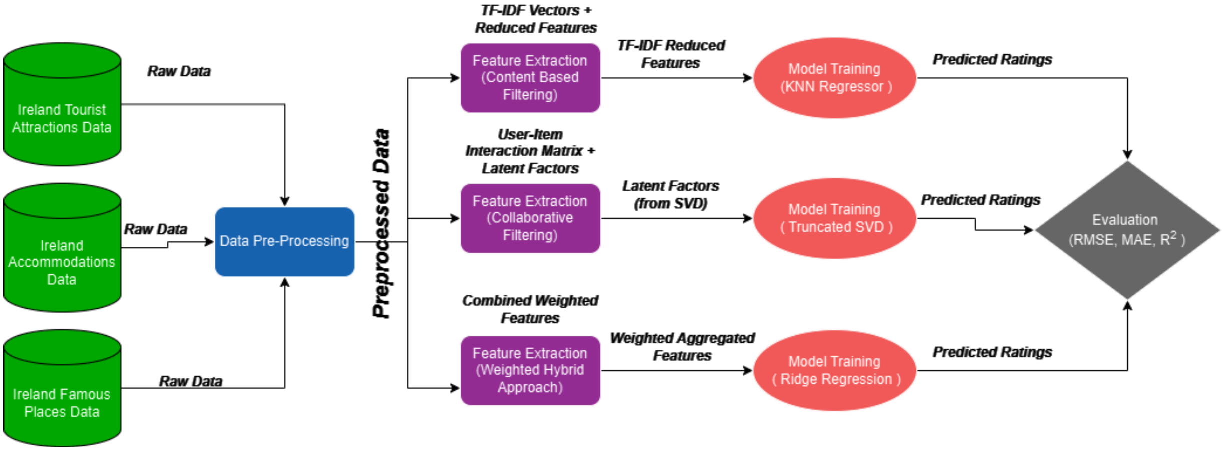

Ireland Tourist Attractions Data
Ireland Accommodations Data
Ireland Famous Places Data
Raw Data
Raw Data
Raw Data
Data Pre-Processing
Preprocessed Data
TF-IDF Vectors + Reduced Features
Feature Extraction (Content Based Filtering)
User-Item Interaction Matrix + Latent Factors
Feature Extraction (Collaborative Filtering)
Combined Weighted Features
Feature Extraction (Weighted Hybrid Approach)
TF-IDF Reduced Features
Latent Factors (from SVD)
Weighted Aggregated Features
Model Training (KNN Regressor)
Model Training (Truncated SVD)
Model Training (Ridge Regression)
Predicted Ratings
Predicted Ratings
Predicted Ratings
Evaluation (RMSE, MAE, R2)

Figure 1: Overall Methodology Diagram

3.1 Overview Of Methodology

An overview of the methodology can be seen in Figure 1. This study follows the CRISP-DM framework for the research methodology. The CRISP-DM framework consists of a set of six stages which are business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. We will be using this framework in our research for building, training, and evaluating the recommendation approaches and models that are implemented in this study. This study doesn't include the Deployment stage as the models are not deployed in a production environment. However, the outputs are studied and evaluated to understand the impact and effectiveness of promoting tourism in Ireland. Figure 2 gives an overview of the CRISP-DM methodology steps which are followed in this study.

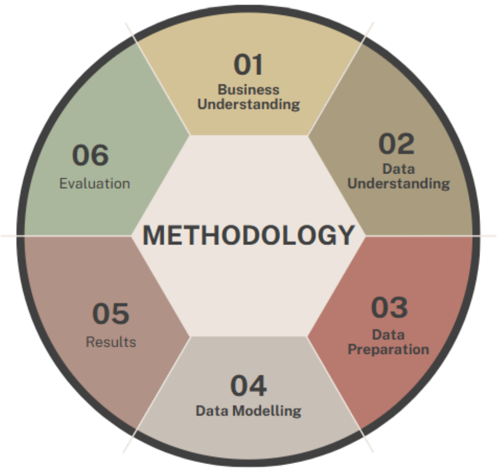


Figure 2: Methodology Stages (CRISP-DM)

3.2 Research Procedure

The main aim of this study is to develop an efficient hybrid recommendation model that recommends the different tourist attractions along with accommodations in the tourism industry. This recommendation model can prove to be very beneficial in order to promote tourism in Ireland. In this study the following outline was followed:

3.2.1 Problem Identification (Business and Data Understanding):

The main problem which gave rise to this study was the limitations of the traditional recommendation systems like content-based filtering, and collaborative filtering, in providing a personalized environment for recommendations. A hybrid recommendation model was implemented to solve the above problems. The hybrid approach can support multiple features and enhance the quality of the recommendations.

3.2.2 Data Gathering (Data Understanding):

To evaluate and test the performance of the recommendation models more effectively, 3 datasets were used.

Dataset 1 (Ireland's Tourist Attractions Dataset): This dataset was taken from the data.gov.ie website. The dataset consists of tourist attractions in Ireland. The dataset contains important features like Place ID, Name, Latitude, Longitude, Address, Tags, and Ratings of each tourist attraction. The tags are short words that describe each tourist attraction. The dataset contains 5954 rows and 8 columns. The Google Places API was used to fetch average ratings of each tourist attraction. The API fetched Ratings directly from Google Maps which ensures that our data is accurate and reliable.

Dataset 2 (Ireland's Accommodations Dataset): This dataset was taken from the Failte Ireland open data API. This dataset contains features like accommodation_id, Sector (type of accommodation eg. Hotel/guest house), Account Name (Name of Accommodation), Rating, Address, Latitude, Longitude, and Total Units (Number of rooms). The dataset contains 1904 rows and 12 columns.

Dataset 3 (Ireland Famous Places Dataset): This tourist attractions dataset was taken from Google Maps using the Google Places API. It consists of external features like latitude, longitude, tags, and other description features. The dataset contains 1086 rows and 6 columns. The dataset includes features like Name, Ratings, Latitude, Longitude, Address, and Tags (short words that describe each place).

3.3 Data Pre-Processing (Data Preparation)

Data pre-processing is a very important step where the raw data collected is prepared for further analysis and modeling. For this project, there were consistent pre-processing steps applied on all of the three datasets to ensure uniformity and effective comparison. The key steps that were involved during this process were handling missing values, analyzing some descriptive statistics, and conducting effective exploratory data analysis (EDA) to understand the structure and distribution of our datasets. The main aim of this step was to ensure that high-quality datasets are prepared which can be used as effective inputs for our machine learning models.

A number of steps were applied to each of the three datasets to ensure efficient data pre-processing. For each of the datasets, the null and missing values were addressed to ensure that the data was complete. For example, the numerical columns like 'Rating' in the 2nd and 3rd datasets were replaced with their mean or mode values, while any irrelevant columns like 'address' in the 2nd dataset were dropped. Further on, standardization and feature extraction techniques were applied. The textual attributes like 'Tags' were visualized using a Word Cloud to gather insights, and then geographical attributes like 'Latitude' and 'Longitude' were retained for geospatial analysis. Furthermore, any categorical values were completed by replacing the null or irregular values with the modes of the column. Moving on, categorical features were also label encoded to prepare the data for analysis. For example, 'Address County' in the 2nd dataset was transformed using a custom mapping technique, where a unique numeric value was assigned to each type of County e.g (Wicklow = 1, Galway = 2). These steps helped to ensure that the data could be utilized efficiently in the recommendation models.

In summary, the data pre-processing steps helped to prepare efficient data and ensured that the datasets are clean, consistent, and ready for further analysis. The steps involved like handling missing values, label encoding e.t.c provided a strong foundation to build robust recommendation models.

3.4 Feature Selection and Engineering (Data Preparation)

From all of the features, the relevant features were selected for each type of recommendation model. In content-based filtering, features like tags, ratings, and geographical coordinates were used. In the case of Collaborative Filtering, a user-item matrix was created which was used as a feature. The hybrid approach combined the features from the content-based filtering and collaborative filtering approaches along with cluster-based features.

3.5 Model Development (Data Modeling)

The main focus of the Modelling phase was on implementing the three recommendation approaches which are Content-based Filtering, Collaborative Filtering, and Weighted Hybrid Approach. Each of the recommendation approaches was applied on each of the three datasets: Ireland's Tourist Attractions, Ireland's Accommodations, and Ireland's Famous Places. Each approach used unique characteristics from the datasets to provide accurate and relevant recommendations.

1. **Content-Based Filtering:**
   Content-Based Filtering uses descriptive tags and other numerical features like rating, Latitude, longitude e.t.c to recommend similar items based on their content and user preferences. The textual features like tags were vectorized using TF-IDF, while the numerical attributes were normalized using MinMaxScaler. A K-Nearest Neighbors (KNN) regressor was applied and predictions were generated.

2. **Collaborative Filtering:**
   User-Item matrices were created to capture the latent patterns of user preferences. Truncated Singular Value Decomposition (SVD) was applied to reduce dimensionality and the ratings were reconstructed for prediction. The number of latent factors was optimized for each dataset to balance the accuracy and complexity.

3. **Weighted Hybrid Approach:**
   The hybrid approach combines Content-Based Filtering, Collaborative Filtering, and clustering techniques using a weighted approach. This helped to combine the strengths of each of the individual models. Ridge Regression was used to optimize the weights for better generalization. Furthermore, cluster-based features were also combined to get more diverse recommendations.

The above models were evaluated using evaluation metrics like RMSE, MAE, R^2 to measure the prediction accuracy and variance explanation. The results showed improvement in the implementation of the hybrid approach to address the limitations of the individual methods.

3.6 Evaluation

After applying the above steps, the performance of the recommendation models was evaluated based on certain factors. The factors used to evaluate were accuracy, predictive reliability, and computational efficiency. The Evaluation metrics that were used for the evaluation are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. Each of the evaluation metrics provides insights regarding the model performance from different angles.

1. **Mean Absolute Error (MAE):**

The MAE measures the average magnitude of prediction errors. It basically provides an idea of how far the predicted values are far from the original values. MAE is calculated from the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|$$

Where $\widehat{y_i}$ is the predicted value, y_i is the actual value, and *n* is the number of observations.

2. **Root Mean Squared Error (RMSE):**

The RMSE gives a weighted measure of error. It gives more weight to larger errors which helps to make it more sensitive to outliers. RMSE is calculated from the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}$$

3. **R^2 Score:**

The R^2 Score explains the variance in the data hence it tells how generalizable the data is. The R^2 is calculated from the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\bar{y}$ is the mean of the actual values.

Each of the models was trained and tested on three datasets and the results were compared to determine the effective recommendation approach. The data was split into the ratio of 80/20 where 80% was used for training and 20% was used for testing.
The results were further supported by visualizations of residuals, scatter plots, and performance metrics across each of the datasets. This approach helped to identify the strengths and limitations of each of the approaches used in this research.

4 Design Specification

The proposed tourist recommendation system is designed on three approaches: Content-Based Filtering, Collaborative Filtering, and a Weighted Hybrid Approach. The main aim of the system is to provide personalized recommendations by using a combination of textual, numerical, and user ratings from tourist attractions. Each of the filtering approaches uses different techniques TF-IDF is used for content-based filtering, user-item interaction matrices are used in collaborative filtering, and weighted feature aggregations are used in the hybrid approach. This is ensured by using evaluation metrics like RMSE, MAE, and R^2. Figure 3 shows the overall design of the hybrid approach which is a combination of the individual methods of Content-based Filtering, Collaborative Filtering, and Cluster-Based Filtering.

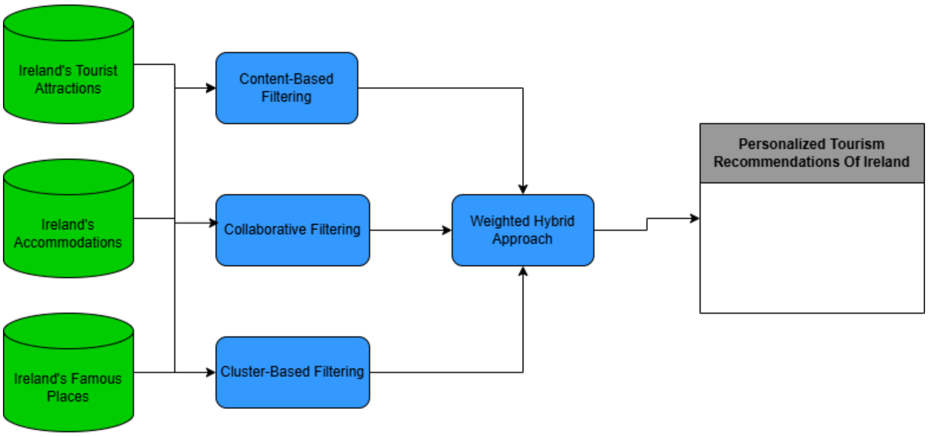


Figure 3: Overall Design (Hybrid Approach)

Figure 1 in the methodology section further describes the design architecture of the recommendation system which integrates the three filtering approaches into a unified architecture. The proposed design ensures scalability and robustness in the recommendations that are produced. Each of the steps produces useful outputs which are used as inputs for further steps. This helps to ensure a consistent pipeline that connects all of the implemented steps.

In the initial process of Data Pre-processing, the raw datasets are cleaned and structured individually. This step helps to produce pre-processed data which is further used for feature extraction. The Content-Based Filtering Module then transforms the textual features like tags into numerical vectors using TF-IDF Vectorization. Dimensionality Reduction using Truncated SVD was applied to the vectorized matrix, which helps to enhance computational efficiency. The final TF-IDF reduced features assist in understanding the significance of the tags, which are passed to the KNN regressor for similarity-based user-item predictions.

In parallel, The Collaborative Filtering module constructs a user-item matrix from the simulated user ratings, which is then reduced using Truncated SVD to extract the latent features that represent user preferences. This process understands the relationships between the users and the items and produces latent factors that are used by the collaborative model for personalized recommendations.

The Weighted Hybrid Approach aggregates the features from the content-based filtering and collaborative filtering with the cluster-based insights that were gathered from applying K Means clustering. A hybrid matrix is created using the weights of each of the filtering techniques. This hybrid matrix is passed to a Ridge Regression Model, which balances the contributions of different features for accurate predictions.

In the final step, the system is evaluated using metrics like RMSE, MAE, and R^2.

5 Implementation

In this section, the final implementation of the tourist recommendation systems on each of the three datasets is discussed. The implementation phase involved collecting data, transforming the raw data into structured formats, building effective recommendation models, and developing a hybrid approach to enhance the accuracy and effectiveness of recommendations. This section gives an overview of the data pre-processing steps, the machine learning models utilized, the tools and technologies used, and the final outputs.

5.1 Data Pre-Processing

5.1.1 Simulation Of User Ratings:

Since real user ratings were not available, the user ratings were simulated using a Gaussian Distribution with a mean of 3.5 and a standard deviation of 1.0 to make it realistic. The following values of the mean and standard deviation were chosen based on the assumption that most ratings are around the scale of 3 to 4. The Gaussian Distribution helps to capture the natural variability in the user ratings. Furthermore, to ensure that the ratings are in a certain realistic range, the np. clip() function was used which prevented any unrealistic ratings. The above steps ensured realistic ratings which helped to better evaluate the models. The same method was used to generate user ratings for each of the datasets which ensured consistency.

5.1.2 Data Cleaning:

Dataset 1 (Ireland's Tourist Attractions): The missing or null values in the dataset were handled properly. The Rating column in the dataset contained 346 null values which were replaced by imputing the mean of the values. The 'County' column consisted of 27 null values which were removed by replacing the values with the mode of the column. The 'Url' and 'Telephone' columns were dropped because they had less significance.

A) **Removing Duplicates:** The duplicates were identified and removed using the drop_duplicates() function to make sure that the dataset contains unique rows for each tourist attraction.

B) **Tokenization Of Tags:** Furthermore, the 'Tags' column was tokenized for further use. Only the common words were kept and the rest were filtered out which helped to optimize the dataset for similarity-based calculations.

Dataset 2 (Ireland's Accommodations): The missing values were detected in some columns. The rows containing null values of latitude, longitude, and Eircode were dropped. The column of 'Address' and 'Proprietor Description' was dropped. The duplicate records were removed using the drop_duplicates() function.

a) **Label Encoding:** The 'Sector' Column which represented the accommodation types like (Hotel, guest House), etc. was label encoded using a mapping dictionary. Each unique type of Sector was assigned a unique numerical value to ease the process of further analysis. Furthermore, in the same way, the 'Rating' column was also label encoded where each type was assigned a specific numerical value. The below figure shows the original values and the values after encoding the 'Rating' column. The 'County' column was also labeled and encoded the same way as the other 2 columns. Table 1 shows the actual values of the 'Rating' column and the values after performing label encoding.

Table 1: Original Values and Values After Encoding

Original Value	Value After Label Encoding
'Hotel – 5 Star'	5
'Hotel – 4 Star'	4
'Welcome Standard – Approved'	3
'Holiday Cottage Approved'	3

b) **Feature Extraction:** A unique Accommodation ID was assigned to each accommodation which was used as a unique identifier for each type of accommodation.

c) **Data Transformation:** The data was reordered and the column 'accommodation_id' was set as the index. The categorical and numerical columns were transformed appropriately.

Dataset 3 (Ireland's Famous Places): The missing values in the 'Rating' Column were replaced with the mean rating of the data which helped to preserve the critical attributes. Furthermore, the duplicate values in the dataset were dropped using the drop_duplicates() function.

a) **Tokenization of Tags:** The 'tags' column was tokenized to further process the tags and gain useful insights.

b) **Feature Engineering:** The 'Rating' Column was scaled to a range of 1 to 5 to standardize the ratings and ensure that all the values are consistent.

5.2 Exploratory Data Analysis

5.2.1 Ireland's Tourist Attractions Dataset (Dataset 1)

The bar chart in Figure 4 indicates that Dublin has the highest number of entries of attractions which is followed by County Cork and Kerry.

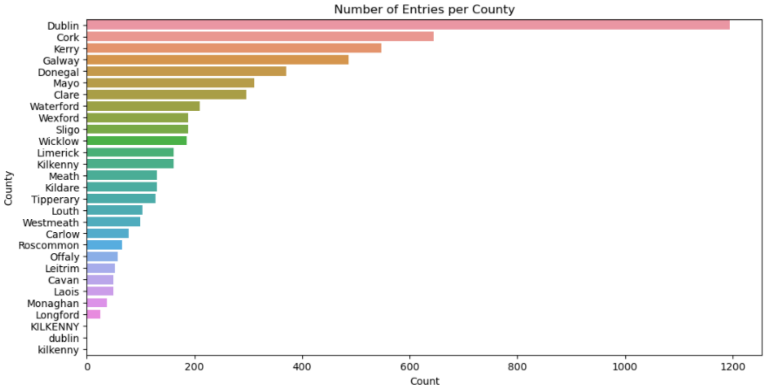


Figure 4: Dataset 1 - Number of Records for Each County

The histogram in figure 5 illustrates the frequency distribution of the average ratings of each of the tourist attractions. It can be seen from the figure that the majority of the tourist attractions have high ratings which indicates that most of the tourist attractions have helped to create positive user experiences.

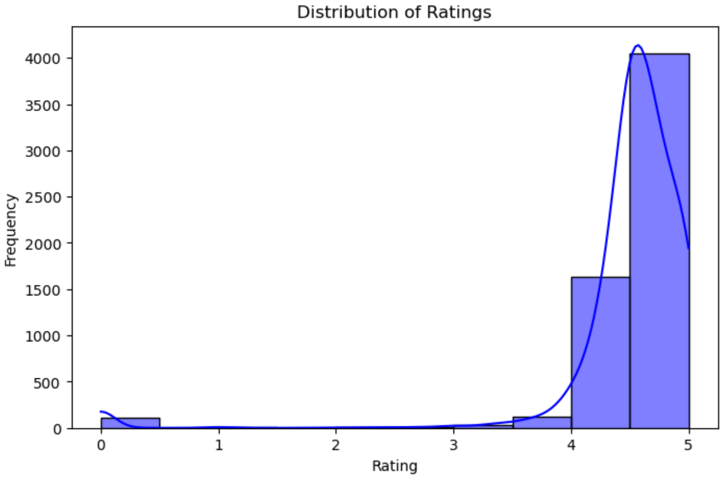


Figure 5: Distribution Of Ratings in Dataset 1

Furthermore, the word cloud in Figure 6 shows the most frequently occurring tags. The text in larger sizes indicates a higher frequency. It can be seen that the most prominent tags are "Activity", "Experience", and "Operator" which provide insights regarding the popularity of the type of tourist attractions.

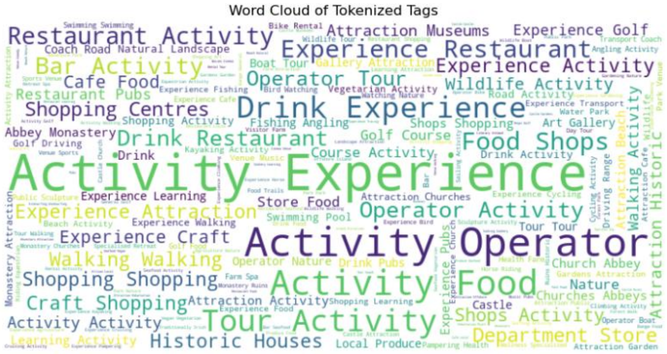


Figure 6: Most Commonly Used Tags in Dataset 1

5.2.2 Ireland's Accommodations Dataset (Dataset 2)

The correlation matrix in Figure 7 illustrates the relationships between the numerical values in the dataset. The 'Rating' variable shows a positive correlation of 0.19 with 'Total Units', which suggests that accommodations with more units may have higher ratings.

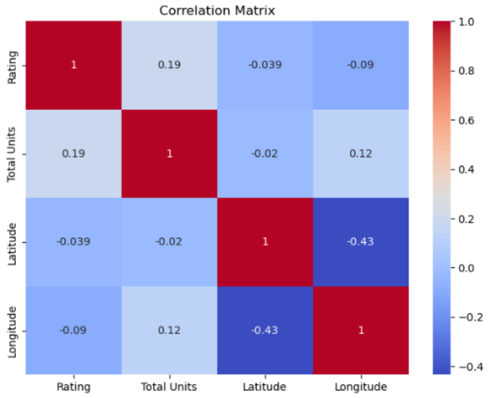


Figure 7: Dataset 2 (Ireland's Accommodations): Correlation Matrix Of Features

5.2.3 Ireland’s Famous Places Dataset (Dataset 3)

The correlation matrix in Figure 8 explores the relationships between the numerical variables of the dataset. The above matrix shows a positive relationship between the latitudes, longitudes, and the ratings.

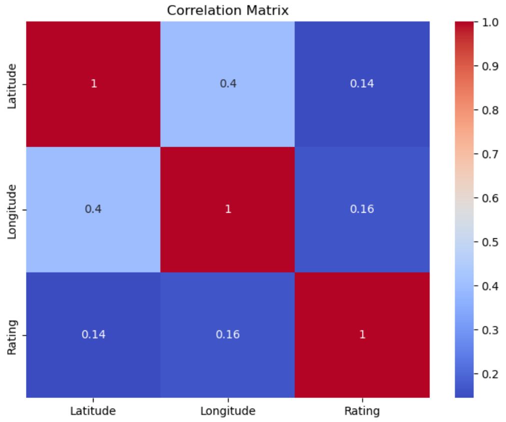


Figure 8: Dataset 3 - Correlation Matrix of Numerical Features

Furthermore, The rating distribution in Figure 10 shows that most of the places have high ratings, and a peak value of 4.6 tells that overall the places have positive reviews. Figure 9 shows the wordcloud, which tells us the description of the different types of common places.

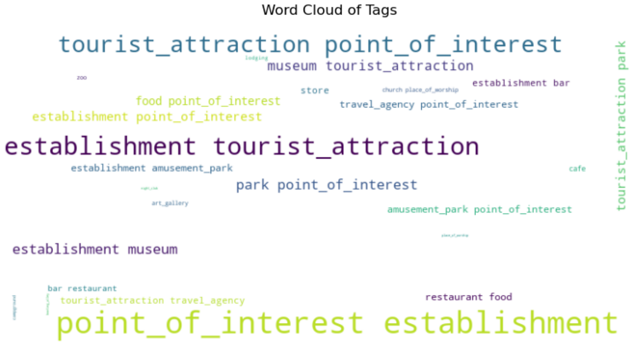


Figure 9: Dataset 3 – Wordcloud Of Tags

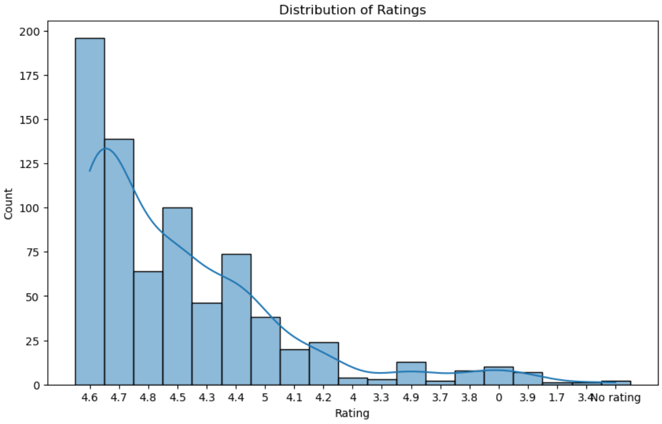


Figure 10: Dataset 2 - Distribution of Ratings

The plot of Geospatial distribution in Figure 11 highlights the clustering of attractions based on the latitudes and longitudes and hence it indicates the areas which have more tourist activities.

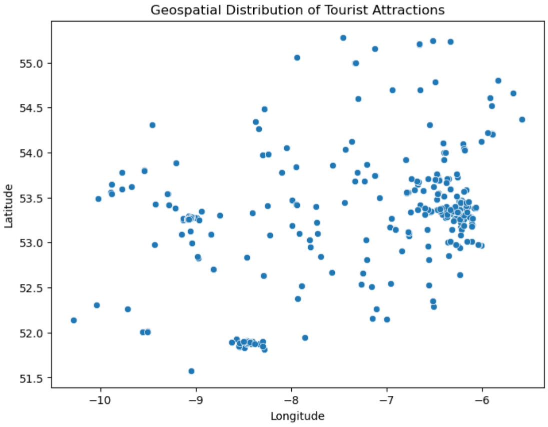


Figure 11: Dataset 3: Geospatial Distribution Of Tourist Places

5.3 Model Development:

5.3.1 Content-Based Filtering:

Content-based filtering is implemented to recommend tourist attractions based on the content and the similarity of the features. We can see the detailed steps of the implementation across all datasets to ensure consistency and efficiency in the recommendations. In the end, Table 2 summarizes the application of Content-Based filtering on each of the datasets. A summarized implementation can be seen in Figure 12.

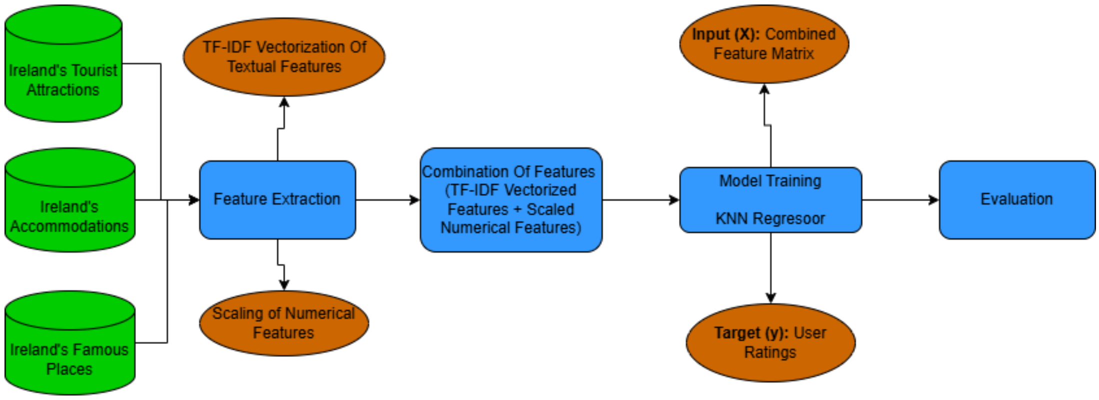

Ireland's Tourist Attractions
Ireland's Accommodations
Ireland's Famous Places
TF-IDF Vectorization Of Textual Features
Feature Extraction
Scaling of Numerical Features
Combination Of Features (TF-IDF Vectorized Features + Scaled Numerical Features)
Input (X): Combined Feature Matrix
Model Training
KNN Regresoor
Target (y): User Ratings
Evaluation

Figure 12: Implementation Of Content-Based Filtering

5.3.1.1 Feature Extraction:

A number of feature extraction methods were applied to the datasets to fully utilize each of the features and get more efficient and optimized results.

In the case of the textual features like '*Tags*' (Dataset 1 and 3), and '*Names*' (Dataset 2), the columns were transformed using the TF-IDF vectorizer. TF-IDF helps to identify the importance of each of the tags as it assigns weights to each term. The stop words were removed to ignore any irregular and invalid words. Furthermore, Dimensionality Reduction was applied to the TF-IDF reduced matrix using Singular Value Decomposition (SVD). The number of components (*n_components*) was adjusted for each dataset according to the complexity of the data.

The numerical features like 'Rating', 'Latitude', and 'Longitude' were normalized using MinMaxScaler which scaled all values in the specified range of 0 to 1.

5.3.1.2 Combination of Features

After the extraction of all features, both of the textual attributes (TF-IDF Reduced Matrix) and the Numerical features (Scaled Features) were concatenated in the form of a combined feature matrix. This matrix helped to identify each item and was used as input to each of the recommendation models.

5.3.1.3 Model Training:

The K-Nearest Neighbors (KNN) regressor was used as the prediction algorithm. It was chosen for its simplicity and suitability for similarity-based modeling.

a) Input and Target Variables:

- **Input (*X*):** The combined feature matrix.
- **Target (y):** User Ratings

b) Data Splitting

- The dataset was split into 80% training and 20% testing.

c) Selection of Parameters

- The number of neighbors (*k*) was set to 5, which helped to balance the model complexity and the prediction accuracy.

- Cosine Similarity was chosen as the distance metric as it helps to measure the angle between vectors which emphasizes more on the similarity rather than magnitude.

Table 2: Summary of Key Parameters across all 3 datasets (Content-Based Filtering)

Parameter/Figure	Dataset 1: Ireland's Tourist Attractions	Dataset 2: Ireland's Accommodations	Dataset 3: Ireland's Famous Places
Text Feature	Tags	Account Name	Tags
TF-IDF Components	100	20	100
Dimensionality Reduction	Truncated SVD	Truncated SVD	Truncated SVD
Numerical Features	Rating, Latitude, Longitude	Rating, Latitude, Longitude, Sector, Total Units	Rating, Latitude, Longitude
Scaling Method	MinMaxScaler	MinMaxScaler	MinMaxScaler
Combined Feature Matrix	TF-IDF + Scaled Numerical Values	TF-IDF + Scaled Numerical Features	TF-IDF + Scaled Numerical Values
Model	K-Nearest Neighbors Regressor	K-Nearest Neighbors Regressor	K-Nearest Neighbors Regressor
Neighbours (k)	5	5	5
Distance Metric	Cosine Similarity	Cosine Similarity	Cosine Similarity
Train-Test Split	80% Train / 20% Test	80% Train / 20% Test	80% Train / 20% Test

5.3.2 Collaborative Filtering:

Collaborative Filtering works by identifying patterns in the user-item interactions by using the user preferences and item similarities. In this research thesis, collaborative filtering was applied to each of the three datasets. A unified approach can be seen below in Figure 13.

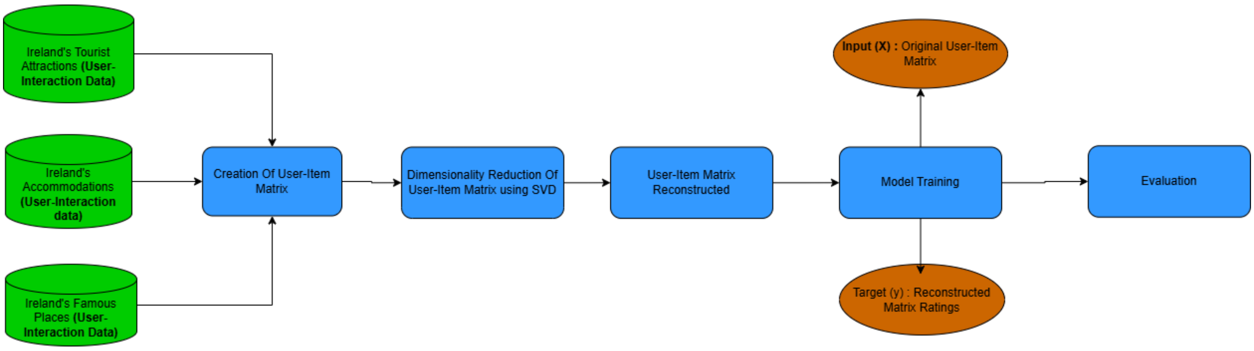

Ireland's Tourist Attractions (User-Interaction Data)
Ireland's Accommodations (User-Interaction data)
Ireland's Famous Places (User-Interaction Data)
Creation Of User-Item Matrix
Dimensionality Reduction Of User-Item Matrix using SVD
User-Item Matrix Reconstructed
Model Training
Input (X) : Original User-Item Matrix
Target (y) : Reconstructed Matrix Ratings
Evaluation

Figure 13: Implementation Of Collaborative Filtering

5.3.2.1 Data Preparation:

A user-item matrix was created where the rows represent the users and the columns represent the unique items. The cell values contain the user ratings for each item. The resulting user-item matrix was highly sparse.

5.3.2.2 Dimensionality Reduction:

Truncated Singular Value Decomposition (SVD) was applied to the datasets to see latent user and item factors. The application of SVD helped to decompose the matrix into latent factors that represent hidden patterns in user preferences. The number of latent features (*n_components*) was adjusted for each of the datasets based on the size and complexity of the data.

- **Dataset 1:** 25 components
- **Dataset 2:** 20 components
- **Dataset 3:** 25 components

5.3.2.3 Model Training

After applying dimensionality reduction on the user-item matrix, the matrix was reconstructed using the decomposed latent factors to predict the ratings.

a) Input and Target Variables

- **Input (*X*):** Original user-item matrix
- **Target (y):** Reconstructed matrix ratings

b) Data Splitting:

The data was split into 80% training and 20% testing.

Table 3: Overview of Collaborative Filtering across all 3 datasets

Dataset	Matrix Size	Latent Features	Training/Testing Split
Dataset 1 (Ireland's Tourist Attractions)	Users x Attractions	25	80% Train, 20% Test
Dataset 2 (Ireland's Accommodations)	Users x Accommodations	20	80% Train, 20% Test
Dataset 3 (Ireland's Famous Places)	Users x Places	25	80% Train, 20% Test

Table 3 provides a summary of the implementation of collaborative filtering.

5.3.3 Weighted Hybrid Approach

The weighted hybrid approach combines the individual recommendation techniques of Content-Based Filtering and Collaborative filtering, and cluster-Based features to enhance the quality of recommendations. A unified approach can be found below which was applied across all 3 datasets. Table 4 provides an overall summary of the weighted hybrid approach while Figure 14 provides a concise view of the implementation.

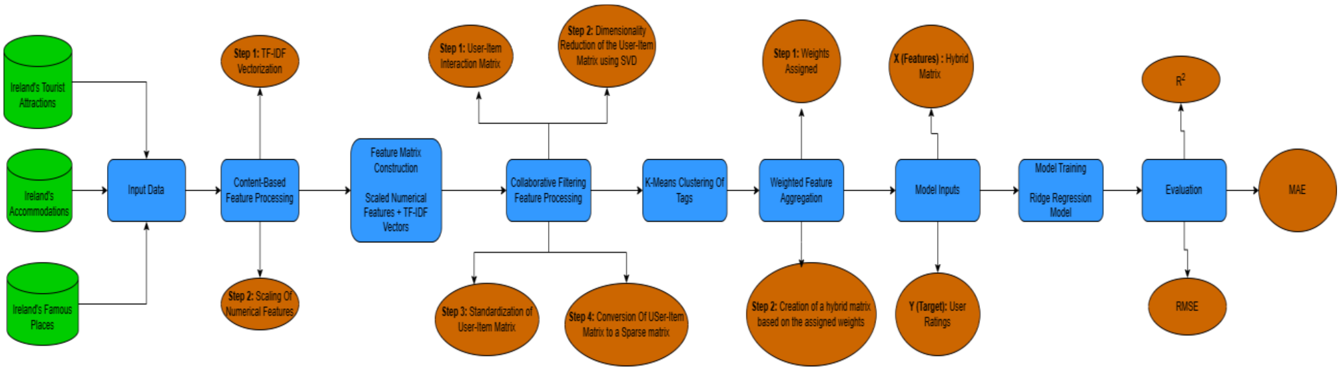

Ireland's Tourist Attractions
Ireland's Accommodations
Ireland's Famous Places
Input Data
Step 1: TF-IDF Vectorization
Content-Based Feature Processing
Step 2: Scaling Of Numerical Features
Feature Matrix Construction
Scaled Numerical Features + TF-IDF Vectors
Step 1: User-Item Interaction Matrix
Step 2: Dimensionality Reduction of the User-Item Matrix using SVD
Collaborative Filtering Feature Processing
Step 3: Standardization of User-Item Matrix
Step 4: Conversion Of USer-Item Matrix to a Sparse matrix
K-Means Clustering Of Tags
Step 1: Weights Assigned
Weighted Feature Aggregation
Step 2: Creation of a hybrid matrix based on the assigned weights
X (Features) : Hybrid Matrix
Model Inputs
Y (Target): User Ratings
Model Training
Ridge Regression Model
R^2
Evaluation
RMSE
MAE

Figure 14: Implementation Of a Weighted Hybrid Approach

5.3.3.1 Feature Extraction

a) **Content-Based Features:**
TF-IDF has applied to the textual features like 'Tags' again. A combined feature matrix was created by combining the TF-IDF reduced features with the scaled numerical features. The dimensionality of the matrix was reduced using Truncated SVD.

b) **Collaborative Filtering:**
A user-item matrix was created with users used as rows, items used as columns, and cell values containing the user ratings. Truncated SVD was applied again on the matrix to extract the latent features. These latent features represented the relationships between users and items.

c) **Cluster-Based Features:**
The items (Attractions, Accommodations, Places) were grouped on the basis of textual features like tags. This was done using K-Means Clustering. Each of the items was assigned a cluster.

5.3.3.2 Weighted Feature Integration:

A hybrid feature matrix is created by combining matrices from each of the above approaches. A unique weight was assigned to each of the methods according to the nature of the dataset and the importance of each type.

- **Content-Based Matrix :** Feature matrix created from content-based filtering
- **Collaborative Matrix:** Matrix containing latent features from collaborative filtering.
- **Cluster-Based Features:** Features obtained by assigning clusters according to similarities.

5.3.3.3 Model Training:

The combined hybrid feature matrix was passed as input to the prediction model. A Ridge regression Model was applied to predict the user ratings. Ridge Regression was chosen for its ability to handle the multi-collinearity in the hybrid matrix.
The Regularization parameter (α) was set to 1.0 to prevent overfitting and control the complexity of the model. The data was split into 80% training and 20% testing.

Table 4: Summary Of Weighted Hybrid Approach

Parameters/Feature	Dataset 1 (Ireland's Tourist Attractions)	Dataset 2 (Ireland's Accommodations)	Dataset 3 (Ireland's Famous Places)
Content-Based Matrix	TF-IDF + Scaled Numerical (60%)	TF-IDF + Scaled Numerical (50%)	TF-IDF + Scaled Numerical (50%)
Collaborative Matrix	Truncated SVD Of User-Item Matrix (30%)	Truncated SVD Of User-Item Matrix (30%)	Truncated SVD Of User-Item Matrix (30%)
Cluster-Based Matrix	K-Means (10 Clusters, 10%)	K-Means (10 Clusters, 20%)	K-Means (10 Clusters, 20%)
Regularization Parameter	Ridge Regression (α = 1.0)	Ridge Regression (α = 1.0)	Ridge Regression (α = 1.0)
Train-Test Split	80% Train / 20% Test	80% Train / 20% Test	80% Train / 20% Test

5.4 Tools And Technologies

There were many tools and technologies which used in the implementation of the recommendation systems to ensure efficient performance and flexibility.

- **Python:** Python was used as the primary language for data manipulation, model development, and implementation. There were many python libraries which were used like Pandas, Numpy, Scikit-Learn which were used in pre-processing data and implementing the recommendation models.
- **Jupyter Notebook:** Jupyter Notebook was used for the development and testing as it provides a good environment for running and executing code.
- **Matplotlib and Seaborn:** Both of these libraries were used for data visualizations.

The tools helped in carrying out complex data processing tasks , model training, and evaluation processes involved in the research.

6 Evaluation

This section aims to analyze the performance of the recommendation approaches, which are Content-Based Filtering, Collaborative Filtering, and the Weighted Hybrid Model, across the three datasets: Ireland Tourist Attractions, Ireland Accommodations, and Ireland Famous Places. The analysis helps to answer the research question by comparing the accuracy and performance of the hybrid approach against the traditional methods using evaluation metrics like RMSE, MAE, and R^2. Table 5 summarizes the evaluation metrics of each of the datasets.

Table 5: Evaluation Metrics Across All 3 Datasets

Dataset	**Method**	**RMSE**	**MAE**	**R^2**
Dataset 1 (Ireland's Tourist Attractions)	Content-Based Filtering	1.0303	0.8377	-0.1854
	Collaborative Filtering	0.5156	0.2281	-5.1043
	Weighted Hybrid Approach	0.9480	0.7760	-0.0043
Dataset 2 (Ireland's Accommodations)	Content-Based Filtering	0.51284	0.4141	-0.2079
	Collaborative Filtering	0.4981	0.4012	-0.1520
	Weighted Hybrid Approach	0.0069	0.0049	0.1819
Dataset 3 (Ireland's Famous Places)	Content-Based Filtering	0.8178	0.6179	0.2538
	Collaborative Filtering	0.5554	0.2837	-3.4067
	Weighted Hybrid Approach	0.9642	0.7937	-0.0117

6.1 Experiment 1: Performance Of Content-Based Filtering

The main aim of this experiment is to evaluate the performance and the effectiveness of the Content-Based Filtering Approach to generate accurate recommendations across the three datasets: Ireland's Tourist Attractions, Ireland's Accommodations, and Ireland's famous Places. This experiment helps to analyze how well were the textual features like tags and numerical attributes utilized to generate relevant recommendations.

The Nearest Neighbours (KNN) Regressor was applied with cosine similarity used as the distance metric. The number of neighbors (k) was set to 5 to balance the computational complexity and the prediction accuracy. The results of the Content-Based Filtering Approach can be seen in the table, which summarizes the results on all 3 of the datasets. Table 6 shows the evaluation metrics obtained from the implementation of Content-Based Filtering.

Table 6: Content-Based Filtering Evaluation Metrics

Dataset	**RMSE**	**MAE**	**R^2**
Ireland’s Tourist Attractions	1.0303	0.8377	-0.1854
Ireland’s Accommodations	0.5128	0.4141	-0.2079
Ireland’s Famous Places	0.8178	0.6179	0.2538

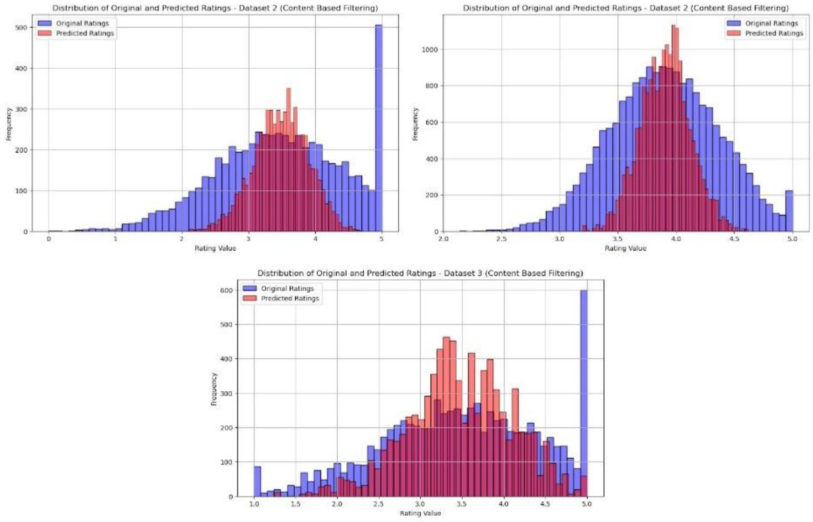


Figure 15: Distribution Of Actual V/s Predicted ratings across all 3 datasets(Content-Based Filtering)

In Figure 15, the overlap between the actual values in blue and the predicted values in red shows the accuracy of the predictions. It can be seen that the distribution between the actual and predicted values varies across the dataset where Dataset 2 shows the highest overlap the RMSE and MAE are minimal and Dataset 1 has fewer overlaps which tells that the data might not be able to provide as accurate recommendations as dataset.

6.2 Experiment 2: Performance Of Collaborative Filtering

This experiment is carried out to evaluate the performance of Collaborative Filtering in predicting the user ratings for each of the three datasets. The main focus is to understand how well the latent features can capture user preferences and improve the quality of recommendations.

A user-item interaction matrix was constructed for each dataset, and Truncated SVD was applied to reduce the matrix to 25 latent factors. The ratings were re-constructed from the SVD-applied matrix to predict the user preferences. The data was split into 80% training and 20% testing.

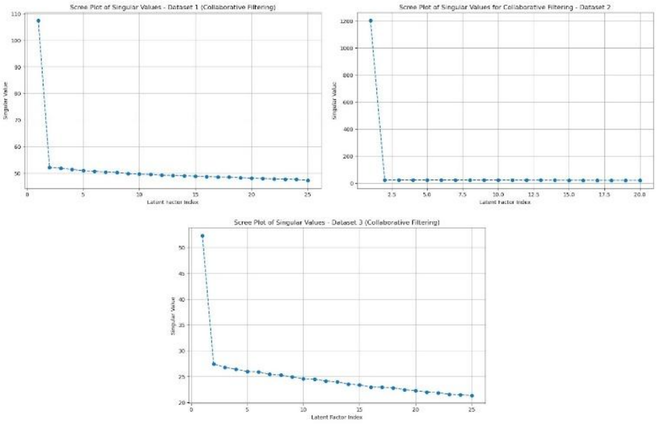


Figure 16: Scree Plots of Singular values for Collaborative Filtering Across all 3 datasets

Table 7: Collaborative Filtering Evaluation Metrics

Dataset	**RMSE**	**MAE**	**R^2**
Ireland’s Tourist Attractions	0.5156	0.2281	-5.1043
Ireland’s Accommodations	0.4981	0.4012	-0.1520
Ireland’s Famous Places	0.5554	0.2837	-3.4067

The scree plots of each of the datasets can be seen in Figure 16. There is a steep decline initially, which shows that a small number of latent factors capture most of the variance in the user-item matrices. After the sharp decline in the singular values, the plots flatten.

The low R^2 value in Datasets 1 and 3 may be due to the insufficient variance captured by the latent factors. Compared to datasets 1 and 3, dataset 2 shows a more stable performance due to more dense interactions. The need for a hybrid approach can be seen as collaborative filtering struggles to generalize in sparse datasets.

6.3 Experiment 3: Performance Of Weighted Hybrid Approach

This experiment evaluates the effectiveness of the Weighted Hybrid Approach which is a weighted combination of Content-Based Filtering, Collaborative Filtering, and Cluster-Based features. The main aim of this experiment is to address the research question by determining if the hybrid approach improves the quality of recommendations across the three datasets.

The Hybrid Model combined Content-Based Filtering, Collaborative Filtering, and a Cluster-Based approach where clusters were given to each tourist attraction dependent upon the tags. A weighted mechanism was used to aggregate the predictions from the individual methods where unique weights were assigned to each type.

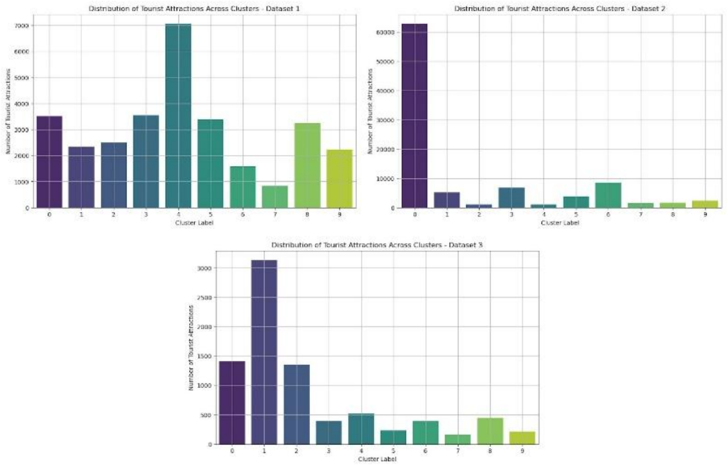


Figure 17: Distribution Of Ratings Across Clusters For All 3 Datasets

The clustering shown in Figure 17 is based on shared tags which helps to improve the diversity of recommendations in the hybrid approach.

A Ridge regression Model was applied to combine the predictions from content-based filtering and collaborative filtering. The data was split into 80% training and 20% testing. The evaluation metrics like RMSE, MAE, and R^2 were used to evaluate the hybrid approach.

It can be seen from table 8 that the hybrid approach provided improved overall results compared to the individual approaches. The weighted hybrid approach is able to make a balance between accuracy and generalization which helps to achieve an improved value of R^2.

Table 8: Weighted Hybrid Approach Evaluation Metrics

Dataset	**RMSE**	**MAE**	**R^2**
Ireland’s Tourist Attractions	0.9480	0.7760	-0.0043
Ireland’s Accommodations	0.4682	0.3781	-0.0069
Ireland’s Famous Places	0.9642	0.7937	-0.0117

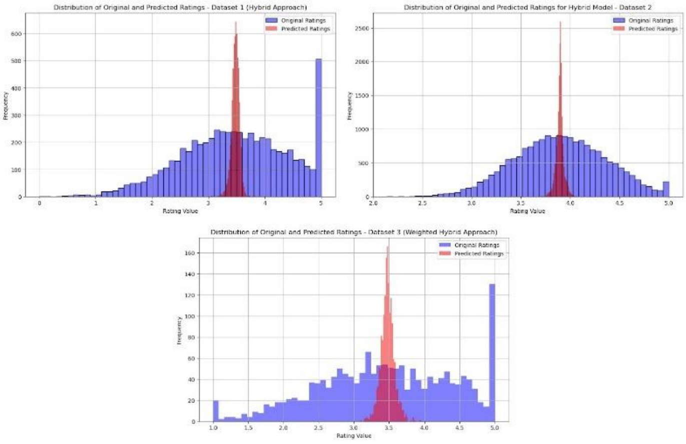
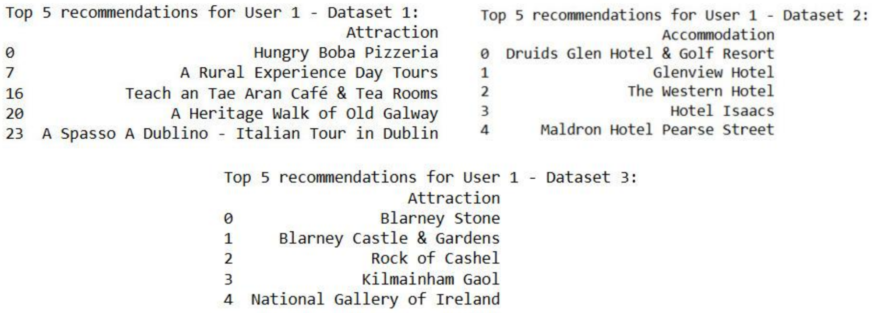


Figure 18: Distribution of Original and Predicted Ratings for the Hybrid Approach across all 3 datasets

Figure 18 shows the distribution of the actual and the predicted values and the overlap between them indicates the accuracy of the model. The above figure also helps to support the research question by showcasing the improved performance of the hybrid approach after integrating Content-Based and Collaborative Filtering. It can be seen that there is a high overlap between the predicted and the actual ratings in Dataset 2 which is also justified by the low RMSE and MAE values.



```
Top 5 recommendations for User 1 - Dataset 1:
                                   Attraction
0                       Hungry Boba Pizzeria
7                  A Rural Experience Day Tours
16            Teach an Tae Aran Café & Tea Rooms
20                 A Heritage Walk of Old Galway
23   A Spasso A Dublino - Italian Tour in Dublin
```

```
Top 5 recommendations for User 1 - Dataset 2:
                                Accommodation
0  Druids Glen Hotel & Golf Resort
1                   Glenview Hotel
2                The Western Hotel
3                     Hotel Isaacs
4       Maldron Hotel Pearse Street
```

```
Top 5 recommendations for User 1 - Dataset 3:
                   Attraction
0               Blarney Stone
1     Blarney Castle & Gardens
2              Rock of Cashel
3              Kilmainham Gaol
4  National Gallery of Ireland
```

Figure 19: Top 5 Tourist Recommendations generated by the Hybrid Approach for test Users across all 3 datasets

Moving on, Test Functions were generated to generate the top 5 recommendations for test users, which can be seen in Figure 19. The recommendations reflect a balance of user preferences and content-driven preferences to suggest high-quality recommendations.. This further helps to demonstrate the effectiveness of the hybrid approach.

7 Discussion

The hybrid recommendation approach for the tourist recommendation system showed great ability to provide an answer to the research question by improving the accuracy and relevance of the recommendations across the three datasets. The combination of Content-Based Filtering, Collaborative Filtering, and clustering techniques within the hybrid approach showed a great balance in the personalization of recommendations. The hybrid approach takes many factors into account which helps the approach to provide high-quality recommendations and this is the reason the hybrid approach stands out from the traditional approaches. However, there are certain limitations in data sparsity, and feature weighting showed areas for improvement. The discussion helps to critically analyze the findings, identify strengths and weaknesses, and suggest areas for improvement.

- **Collaborative Filtering:** The collaborative filtering approach performed great on diverse datasets like Dataset 2, however, there was a problem of sparsity which resulted in negative R^2 values for datasets that had limited interactions.
- **Cluster-Based features:** The cluster-based features were able to improve the recommendation diversity by grouping the items based on tags, but this was dependent upon the quality of tags which again varied across the different datasets.
- **Sparsity:** The Sparse User-Item Matrices in Datasets 1 and 3 were affected in poor generalization and negative values of R^2.
- The missing values were handled by replacing them with the mean or modes. However, this could have caused potential noise in the data which may have affected the overall accuracy of the predictions.

The above limitations can be improved by following the steps which can also help to enhance the performance of the tourist recommendation system:

- Dynamic Weighting Techniques can be implemented in the hybrid approach which will help to optimize the integration process according to each dataset.
- Advanced clustering algorithms like DBSCAN can be used for clustering techniques. Furthermore, additional contextual features can be added which can increase the quality of the information of clusters.
- Addition of more user-interaction data, like click-through rates can help to solve the data sparsity issues and can also improve model generalization.

The hybrid approach helped to solve the research question by solving problems like data sparsity and also enhanced the quality of the recommendations. However, some limitations have opportunities for further improvements. overall, the results show the high performance of the hybrid approach and also provide a foundation for further research to optimize the tourist recommendation systems for more diverse datasets and user preferences.

8 Conclusion and Future Work

The main aim of this research study was to address the research question: ‘How does a hybrid recommendation approach, combining Content-Based Filtering, Collaborative Filtering, and clustering techniques, improve the accuracy and the relevance of Ireland’s Tourist Attractions?’. The main objective was to evaluate the performance of the hybrid approach across the three datasets: Ireland’s tourist Attractions, Irelan’s Accommodations, and Ireland’s famous places - while handling problems like data sparsity.

Using evaluation metrics like RMSE, R^2, and MAE, this research was able to achieve its goals and objectives by showing the ability of the hybrid approach to improve the quality of recommendations by balancing accuracy, relevance, and diversity.

The following findings were achieved:

- The Content-Based Filtering Module used textual features like tags to offer personalization of recommendations.
- Collaborative Filtering performed good in dense datasets by capturing the latent patterns but struggled with data sparsity.

- The Hybrid model combined the strengths of both the models which caused the R^2 value to improve and helped to offer relevant recommendations.

This study explains the value of combining traditional methods to improve the quality and the relevance of recommendations. Therefore, this hybrid approach can also be applied in many real-world tourism platforms that will help to increase customer engagement rates by offering personalized recommendations.

While there were some limitations like data sparsity, future work can be carried out to address these limitations:

- Dynamic Weighting Approaches can be used in Hybrid Approaches.
- The hybrid approach can be extended to other real-world domains like e-commerce, healthcare, etc.
- More advanced clustering techniques can be carried out to improve performance.
- More diverse data sets can be used to optimize the hybrid approaches.

This hybrid approach can also be commercialized and integrated into many real-world tourist platforms and accommodation booking websites. This will help to increase customer retention rates by bringing more user engagement and participation.

References

Alabduljabbar, R. (2023). Matrix Factorization Collaborative-Based Recommender System for Riyadh Restaurants: Leveraging Machine Learning to Enhance Consumer Choice. *Applied Sciences*, *13*(17), 9574. https://doi.org/10.3390/app13179574

Appadoo, K., Soonnoo, M. B., & Mungloo-Dilmohamud, Z. (2020). Job Recommendation System, Machine Learning, Regression, Classification, Natural Language Processing. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6. https://doi.org/10.1109/CSDE50874.2020.9411584

Chang, J., Yu, F., Ouyang, C., Yang, H., He, Q., & Yu, L. (2025). An interval-valued matrix factorization based trust-aware collaborative filtering algorithm for recommendation systems. *Information Sciences*, *686*, 121355. https://doi.org/10.1016/j.ins.2024.121355

Chongwarin, J., Manorom, P., Chaichuay, V., Boongoen, T., Li, C., & Chansanam, W. (2024). Enhancing Book Recommendation Accuracy through User Rating Analysis and Collaborative Filtering Techniques: An Empirical Analysis. *Journal of*

Telecommunications and the Digital Economy, *12*(3), 51–72. https://doi.org/10.18080/jtde.v12n3.976

Delimayanti, M. K., Laya, M., Warsuta, B., Faydhurrahman, M. B., Khairuddin, M. A., Ghoyati, H., Mardiyono, A., & Naryanto, R. F. (2022). Web-Based Movie Recommendation System using Content-Based Filtering and KNN Algorithm. *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 314–318. https://doi.org/10.1109/ICITACEE55701.2022.9923974

Karurukan, V. I., Nugraha, I. G. B. B., & Kumiawan, N. B. (2018). The Estimation Model of Travel Time Between Point of Interest for Tourist Travel Recommendations with Regression Analysis. *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*, 324–329. https://doi.org/10.1109/ICITSI.2018.8696070

Kim, D., Song, J., Yoo, S., & Jeong, O. (2018). A Social Travel Recommendation System using Item-based collaborative filtering. *Journal of Internet Computing and Services*, *19*(3), 7–14. https://doi.org/10.7472/JKSII.2018.19.3.7

L, P., & B, C. (2024). Personalized Millet Recommendations through BMI Analysis and Ridge Regression-based Production Trends. *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 917–923. https://doi.org/10.1109/I-SMAC61858.2024.10714833

Lu, J. (2022). Personalized Recommendation Algorithm of Smart Tourism Based on Cross-Media Big Data and Neural Network. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/9566766

Luostarinen, T., & Kohonen, O. (2013). *Using Topic Models in Content-Based News Recommender Systems*.

Manikandan, N. K., & Kavitha, M. (2023). RETRACTED: A content recommendation system for e-learning using enhanced Harris Hawks Optimization, Cuckoo search and DSSM. *Journal of Intelligent & Fuzzy Systems*, *44*(5), 7305–7318. https://doi.org/10.3233/JIFS-213422

Ruchika, ., Sharma, M., & Hossain, S. A. (2023). Efficient Machine Learning Algorithms in Hybrid Filtering Based Recommendation System. *Journal of Information Technology Management*, *15*(3). https://doi.org/10.22059/jitm.2023.93631

Sarkar, J. L., Majumder, A., Panigrahi, C. R., Roy, S., & Pati, B. (2023). Tourism recommendation system: A survey and future research directions. *Multimedia Tools and Applications*, *82*(6), 8983–9027. https://doi.org/10.1007/s11042-022-12167-w

Sivanaiah, R., Sakaya Milton, R., & Mirnalinee, T. T. (2020). Content boosted hybrid filtering for solving pessimistic user problem in recommendation systems. *Intelligent Data Analysis*, *24*(6), 1477–1496. https://doi.org/10.3233/IDA-205244

Syahputra, M. E., Achmad, S., Fahrain, F., Mackenzie, A. J., Putra Panghurian, F., & Santoso Gunawan, A. A. (2022). Smart Tourism using Attractive and Safe Travel Recommendation Technology. *2022 IEEE Creative Communication and Innovative Technology (ICCIT)*, 1–5. https://doi.org/10.1109/ICCIT55355.2022.10118828

Vaishnavi, N., & Kalpana, B. (2023). Sentiment Based Product Recommendation System Using Machine Learning Techniques. *Journal of Engineering Science and Technology Review*, *16*(6), 170–176. https://doi.org/10.25103/jestr.166.21

Weiwei, Z., Lin, Z., Mengmeng, C., Yanan, S., Yunna, G., & Wei, L. (2022). Research on travel scheme recommendation model based on logistic regression algorithm. *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, 789–793. https://doi.org/10.1109/ICCEAI55464.2022.00165

Xu, Z., Hu, Z., Zheng, X., Zhang, H., & Luo, Y. (2023). A matrix factorization recommendation model for tourism points of interest based on interest shift and differential privacy. *Journal of Intelligent & Fuzzy Systems*, *44*(1), 713–727. https://doi.org/10.3233/JIFS-211542

Yang, H., Zhou, H., & Li, Y. (2022). A Review of Academic Recommendation Systems Based on Intelligent Recommendation Algorithms. *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, 958–962. https://doi.org/10.1109/ICIVC55077.2022.9886104

Yao, S., Halpern, Y., Thain, N., Wang, X., Lee, K., Prost, F., Chi, E. H., Chen, J., & Beutel, A. (2021). *Measuring Recommender System Effects with Simulated Users* (arXiv:2101.04526). arXiv. https://doi.org/10.48550/arXiv.2101.04526

Zhang, S., & Balog, K. (2020). Evaluating Conversational Recommender Systems via User Simulation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1512–1520. https://doi.org/10.1145/3394486.3403202