MEF UNIVERSITY

AIRBNB HOST RECOMMENDATION ENGINE

Capstone Project

Batuhan Arslan

İSTANBUL, 2021

MEF UNIVERSITY

AIRBNB HOST RECOMMENDATION ENGINE

Capstone Project

Batuhan Arslan

Advisor: Prof. Dr. Özgür ÖZLÜK

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: AIRBNB HOST RECOMMENDATION ENGINE
Name/Last Name of the Student: Batuhan Arslan
Date of Thesis Defense: 25/01/2021

I hereby state that the graduation project prepared by Batuhan Arslan has been completed under my supervision. I accept this work as a "Graduation Project".

25/01/2021
Prof. Özgür ÖZLÜK

I hereby state that I have examined this graduation project by Batuhan Arslan which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

25/01/2021

Director
of
Big Data Analytics
Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Özgür ÖZLÜK	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Batuhan Arslan	25.01.2021	

EXECUTIVE SUMMARY

Airbnb Host Recommendation Engine

Batuhan Arslan

Advisor: Prof. Dr. Özgür ÖZLÜK

JANUARY, 2021, 21 pages

In this project, a fifth rule is proposed to reveal guests ' comments about hosts using the recommendation system and sentiment analysis for the super hosts' selection for Airbnb. This project is aimed to contribute to Airbnb's selection of Super hosts. In this study, sentiment analysis and comment data are examined, and polarity scores are created for use in suggestion systems. A collaborative filtering method is used for the recommendation system. The FunkSVD algorithm received the best RMSE score. Polarity scores are estimated for each latent user by looking at the host and listing id. The recommendation system developed ranked the polarity scores of hosts for each user.

Key Words: Recommendation Systems, Collaborative Filtering, Sentiment Analysis, Random Forest Classifier, Funk SVD

ÖZET

Airbnb Ev Sahibi Öneri Sistemi

Batuhan Arslan

Proje Danışmanı: Prof. Dr. Özgür ÖZLÜK

OCAK, 2021, 21 sayfa

Bu proje, Airbnb için süper ev sahipleri seçiminde öneri sistemi ve duygu analizi kullanılarak misafirlerin ev sahipleri hakkındaki yorumları ile beşinci bir kural ortaya çıkarılması amaçlamaktadır. Bu projenin Airbnb'nin Super ev sahibi seçimine katkı sağlaması amaçlanmaktadır. Bu çalışmada, duygu analizi ile yorum verisi incelenmiş ve öneri sistemlerinde kullanılmak için polarite skorları oluşturulmuştur. Öneri sistemi için işbirlikçi filtreleme yöntemi kullanılmıştır. Funk SVD algoritması en iyi RMSE skorunu almıştır. Ev sahibi ve listeleme numarasına bakılarak her bir gizli kullanıcı için polarite skorları tahmin edilmiştir. Geliştirilen öneri sistemi ile birlikte her bir kullanıcı için ev sahiplerinin polarite skorları sıralanmıştır.

Anahtar Kelimeler: Öneri Sistemleri, İşbirlikçi Öneri Sistemi, Duygu Analizi, Rastgele Orman Sınıflandırıcı, Funk SVD

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Through Airbnb data, the COVID-19 has been observed to affect negatively travel activity in many ways. As a result of the review Hu and Lee (2020), three main breaking points stand out: First quarantine in Wuhan, detection of local first cases, local quarantines. In the first case, it is observed that the travel activity of the Earth at different locations is inversely proportional to the distance between the destination and Wuhan. An 8.8% reduction in travel activity was observed in the first stage of the review, taking into account the number of Airbnb comments.

In the second phase, the emergence of local cases led to a second decline in travel activity. A 15% increase in Airbnb booking cancellations after the detection of local first cases also support this result. The third breaking point observes as a result of travel bans imposed through governments. Along with travel restrictions, a 57% decrease in booking activity and a four-and-a-half-fold increase in booking cancellation demand observed.

Nowadays the demand for short and long-term temporary accommodation is increasing thanks to easing travel conditions. This demand positively affects the number of online platforms that allow you to make reservations before traveling. Airbnb is one such platform, which allows travelers to make accommodation reservations based on the fact that the host leases all or part of his or her home to the traveler.

In the following, we discuss what conditions Airbnb users pay attention to when choosing accommodation and whether host selection is an important requirement for users.

Deshmukh (2019), Airbnb categorizes the people that open their homes to share, i.e., hosts, into two categories: host and Superhost. The categorization system is based on four different rules. Airbnb makes the rules it applies when choosing a super host every three months. Hosts are required to score 4.8 and above in the last year. Moreover, there must be more than ten stays in a year or more than 100 nights at least three times. At the same time, hosts must have a cancellation rate of less than 1 percent, except in emergencies. Finally, the return rate for messages within the last 24 hours should be at least 90%. Hosts that meet these conditions are candidates to become Superhost in the next quarter. Being Superhost airbnb.com there are three main benefits that it provides. Along with low service fees, there is an increase in their earnings. They're ahead in the

recommendation system. More promotions will be provided by Airbnb. Finally, there are special awards. Airbnb provides the NEST (home automation electronics) products with a 20% extra discount exclusive to super hosts and provides a $100 travel coupon every year to maintain super host status.

This research aims to increase the accuracy of Superhost selection and to develop a Superhost recommendation system by adding a fifth rule that recommends analyzing user comments to the runs of being Superhost.

1.1. Recommendation System Literature Survey

Panigrahi and Asha. T. (2018) published a paper about aspect level sentiment analysis for rating the hotels. Therefore, they also used the RHALSA (Ranking Hotels Using Aspect Level Sentiment Analysis) algorithm. The algorithm works with a dataset ranking user reviews of hotels, which is collected from TripAdvisor. They ranked in descending order of their Average Sentiment Score. The main ideas are the cleanliness and service aspect. The sentiment levels are “very negative”, “negative”, “neutral”, positive”, “very positive”. The result of algorithms could handle negative comments but also could be extended to handle discourse relations. When evaluating the approach of this article, it was observed how sentiment analysis affects the reviews dataset. Based on the methods used, the scores that will be generated by sentiment analysis on the Airbnb reviews data can result in a high accuracy result in the Airbnb Superhost proposal. At the same time, ranking the extracted scores reveals a method for the Superhost recommendation system.

Bhujade and Chandak (2018) have published a hotel recommendation system, where the main idea is to find the most appropriate recommendation list for the users. They used language modeling as a method. The algorithm is looking at probability distribution over the sequence of words. In the recommended system, they made a dictionary of hotel-related words. Therefore, it works when customers write preference. If their preference matches with extracted keywords in the dataset, the system will recommend more appropriate services. The article will also be able to produce better results because the data set used is performed according to the language. It will be used in this project. It is also a comparison source for reviews data, as it is done through sequence of words. In this article, very high-quality algorithms were used on the recommendation system and high accuracy results were obtained. This information can

be a source of comparative analysis for the recommendation system to be created in the source.

In (Thomas et.al, 2019), the cascade generalization aims to reduce the test errors on unseen data, which consists of combining the decisions of multiple classifiers. (Thomas et.al, 2019) also uses the PNR (Passenger Name Record) dataset. The main idea for this project is to increase session conversion, which is a session that leads to conversion when a customer starts to book some of the hotel offerings during the sessions. In (Thomas et.al, 2019), the authors have created a hotel recommendation system using hotel bookings and flight details. Among the different models implemented in the paper, LIME looks at the most important features, mean and maximum hotel conversion probabilities. It can help the Airbnb recommendation system to increase the model accuracy according to feature importance.

Ramzan et. al (2019) implemented heterogeneous and large-sized data using machine learning algorithms to make important recommendations for the expected customers. In this paper, Collaborative Filtering (CF) is implemented, which needs to handle the data in a big data Hadoop environment with Cassandra database for the high response time to make a recommendation system. The hotel dataset was also collected from many websites for the hotel recommendation system. At the preprocessing part, they used many types of methods for the textual reviews and implemented with NLTK (Natural language toolkit) which is a Python library. They also used Collaborative Filtering (CF) to generate recommendations. Therefore, some of the features are predicted by different filtering methods. The most useful method for the Airbnb Superhost recommendation system is CF. Therefore, the methods used in this article can be associated with CF and NLTK for the Superhost proposal.

The implementation of multiple methods and techniques to measure user behavior has been published. A different algorithm under the name FunkSVD was published by Simon (2006) in the recommendation system implementation section. The FunkSVD algorithm is a matrix factorization. In the Netflix competition, Simon (2006) presented a significantly successful practical solution along with the Netflix movie-by-user ratings matrix. This wonderful solution was implemented by Netflix. (2012) The algorithm factored two low-dimensional matrix multiplications over the user-item rating matrix, with the first row specifying each user and the second column specifying each item. Those associated with these specified are called latent factors. Using FunkSVD for recommendation systems over comments to be analyzed in the formation of the fifth

rule in Superhost selection can greatly contribute to suggesting polarity scores from user comments through the host and predicting some missing scores.

In this article, Cheng and Jin (2019) sentiment analysis and text mining were applied based on the experience of Airbnb users, and the online review comments collected during these analyses were investigated in a big data framework. As a result of analysis with text mining on the collected data, the most important ones are location amenities and host. At the same time, the most similar results were observed in the word host with a high percentage of sentiment analysis. Based on this, in the selection of Superhost, sentiment analysis can be analyzed to draw conclusions that may be of interest. The conclusion drawn from the article is that big data plays an active role in tourism and hospitality studies.

One of the best ways to understand Airbnb Boston's customers is to analyze reviews. Lawani et al. (2019) this article was analyzed using the AFINN dictionary (a lexicon based on unigrams) to understand customers ' feelings on comments. The methods used in the research can be used in Airbnb Istanbul data and compared with other sentiment analysis methods to achieve the best result. In this way, Airbnb Superhost selection can result in high accuracy results.

Jannach et al. (2013) on this article analyzed product ranges using multiple recommendation system algorithms using the Movielens dataset. The FUNK-SVD algorithm came first when the top 10 RMSE scores were calculated, which proved that the product's sales diversity decreased. Using FunkSVD over the methods used may affect the accuracy of the fifth rule in the Superhost selection. These algorithm methods can also be used as a comparison method.

García-Cumbreras et al. (2013) In this study, a new application of sentiment analysis was made in recommendation systems. The study aims to develop recommender systems based on sentiment analysis. This study analyzed the IMDB data set. Methods were analyzed using comments and ratings on the data set. KNN-80, Biased matrix factoring and factor-wise matrix factoring algorithms used for Recommendation systems, and RMSE scores considered. It proves that sentiment analysis has received successful results for recommender systems. It can be instrumental in the creation of polarity scores using comment features on the Airbnb dataset, as well as in the formation of a recommender system associated with sentiment analysis.

2. PROJECT DEFINITION

This project aims to develop the Superhost recommendation system. In addition to the existing Superhost rules in the Airbnb system, adding a new requirement to target the selection of Superhost. As a result of data analysis, it seems that some super hosts do not have great comments.

It aims to evaluate hosts as positive or negative by applying Sentiment analysis via the Istanbul reviews data set. When the accuracy of the results is measured to apply multiple sentiment analysis, it will continue with the model that gets the best score. Moreover, creating a new dataset by combining the Sentiment analysis results with the listing dataset. Accordingly, a super host recommendation system generates over the polarity scores that have been verified.

3. ABOUT THE DATA

The dataset describes the comments on the properties of homestay in Istanbul, Turkey since 2010. It includes listing id, date, review id, review name, and comments.

Airbnb presents the dataset to divide into countries and cities. There are three different datasets in these reports, which are listings, calendars, and reviews datasets. The reports are shared monthly. Airbnb listing and reviews data are the most common hotels open data sources not only in Turkey but also all over the world. The review dataset includes all properties renting online with the information about property listing id, review date, review id, review name, and reviewer comments. On the other hand, the listings dataset summarizes all details about the customer reservations.

The dataset is gathered from Airbnb Inside platform published by Airbnb (2020). The data sources are used together: "Istanbul Airbnb reviews dataset" as all listing hotel data Airbnb (2020), "Airbnb reviews" as a hotel review data Airbnb (2020).

3.1. Features

The features of Reviews dataset

- Listing_id: (numeric) | Unique identifier for each listing places
- Id: (numeric) | Unique identifier for each listing
- Date: (date) | comment date on this listing id
- Reviewer_id: (numeric) | Unique identifier for each reviewer
- Reviewer_name: (text) | Name of homestays
- Comments (text): | Reviews for each listing id Listing dataset of the features
- Id: (numeric) | Unique identifier for each listing
- Property_type: (categorical) | (e.g., Apartment, house, condo)
- Room_type: (categorical) | (e.g., Entire home/apt, private room)
- Amenities: (text) | Unstructured list separated by commas (e.g., tv, kitchen).
- Accommodates: (numeric) | Number of people the rental fits
- Bathrooms: (numeric) | Number of full and/or half baths
- First_review: (date) | How long ago the first review was left
- Host_response_rate: (numeric) | How often the host replies to inquiries (%)
- Last_review: (date) | Date of most recent hosting
- Latitude: (numeric)
- Longitude: (numeric)

- Name: (text) | Name of rental property.
- Neighborhood: (categorical) | Informal description of neighborhood (e.g., Brooklyn Heights, Downtown)
- Number_of_reviews: (numeric) | Total number of reviews given by guests
- Review_scores_rating: (numeric) | Mean rating of reviews given by guests
- Thumbnail_url: (numeric; we'll come back to this) | Link to primary photo of rental property.
- Bedrooms: (numeric) | Number of bedrooms in rental
- Beds: (numeric) | Number of beds in rental
- Host_verifications: (text) | communication to host

3.2. Exploratory Data Analysis

When examining the Listing data set, an attempt is made to understand what the data contains for the host_is_superhost column. NaN values found in the dataset were removed because of more meaningful conclusions. The required columns were selected from the Listing and Reviews data and then merged to analyze and understand data sets.

NaN value was first detected in a comment on the reviews dataset when the data examined in detail. The percentage of removing NaN values from the comment column is %1. In Figure 1, the scatter plot shows the relationship host_is_superhost. The Z score method is used to get rid of the outliers that appeared in Figure 1. Accordingly, data outside the range of + 3 to -3 removed. When this operation applied to all columns that were the outlier, 7% data removal occurred. After this process occurs, the chart on the right side shows the state without an outlier. When a scatter plot is drawn by categorizing review_scores_rating with the minumum_nights column, and these values as host and super host, the super host and the low review score rating appear in the chart. The Host_is_superhost column converted to numeric values.

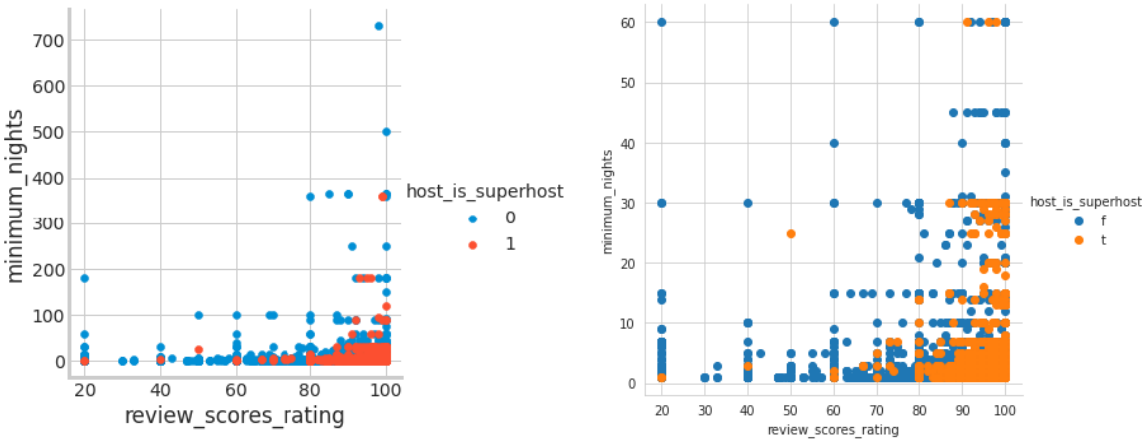

Figure 1: Two scatter plot between the reviews score rating and minimum nights values and into host or superhost categorical values

Figure 2a shows a graph of the number of comments Airbnb guests write about hosts by year. This chart shows the average number of comments made between 2010 and 2020. Most comments on the graph are known to be in 2016 and 2020, while the least observations get between 2010 and 2015 and 2018. The anomaly in these sections may be caused by social, cultural, or other reasons. Although there were very high average comments in the first period of 2020, there was a decrease in the number of comments due to Covid19.

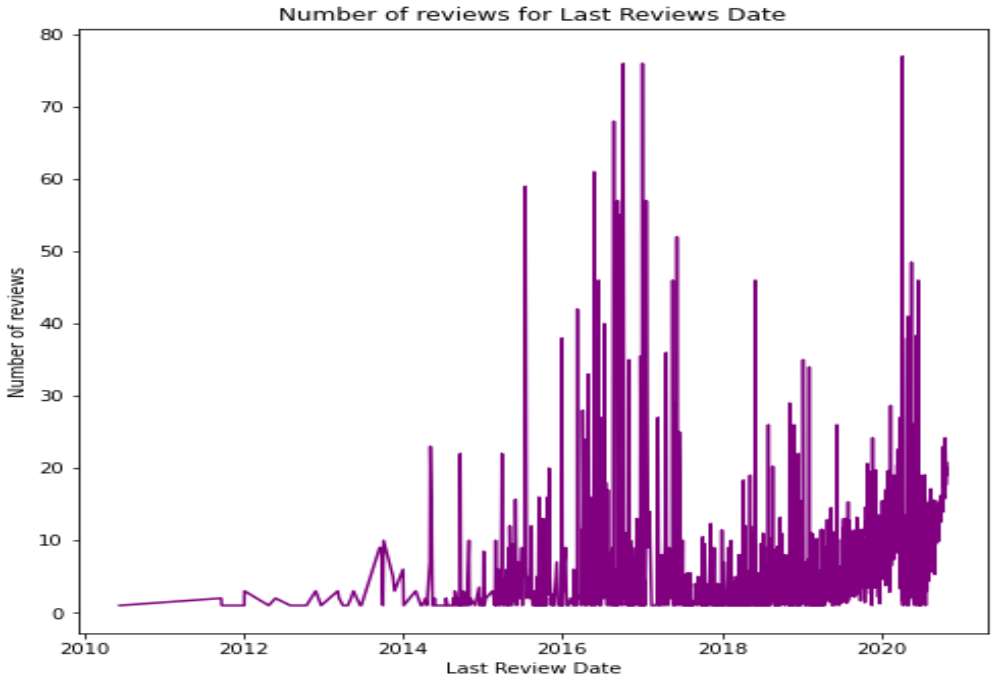

Figure 2a: Time series graph between average number of reviews and last reviews date

Figure 2b shows the average number of comments according to the months shown in this chart. April and June in the chart show a significant increase between 2016 and 2019. We can say that there is a seasonality in this part. February, May, March, July, and July of all years, the number of comments appears to have decreased. The number of comments in February, March, May, and July appears to have decreased. Moreover, there is an increase in other years except in December 2017. As a result, when we look at the first seven months, it is clear that there is seasonality.

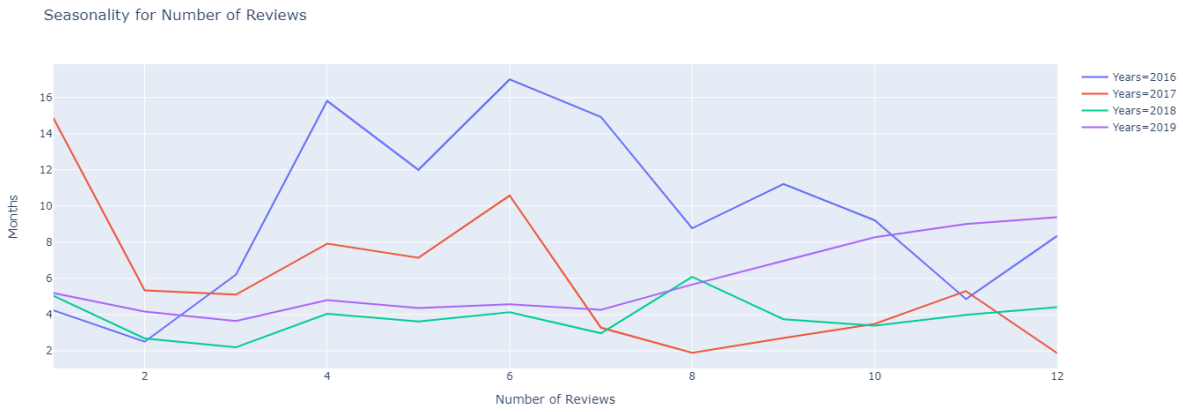

Figure 2b: Time series graph between average number of reviews and months for last reviews

The high number of comments in Figure 2c in 2016 appears to have fallen sharply in 2017. This is because of some terrorist attacks in Turkey, there has been a decline in the number of tourists coming to Istanbul. The decline in Airbnb comments is due to these reasons. Then in 2019, this number is increasing even more.

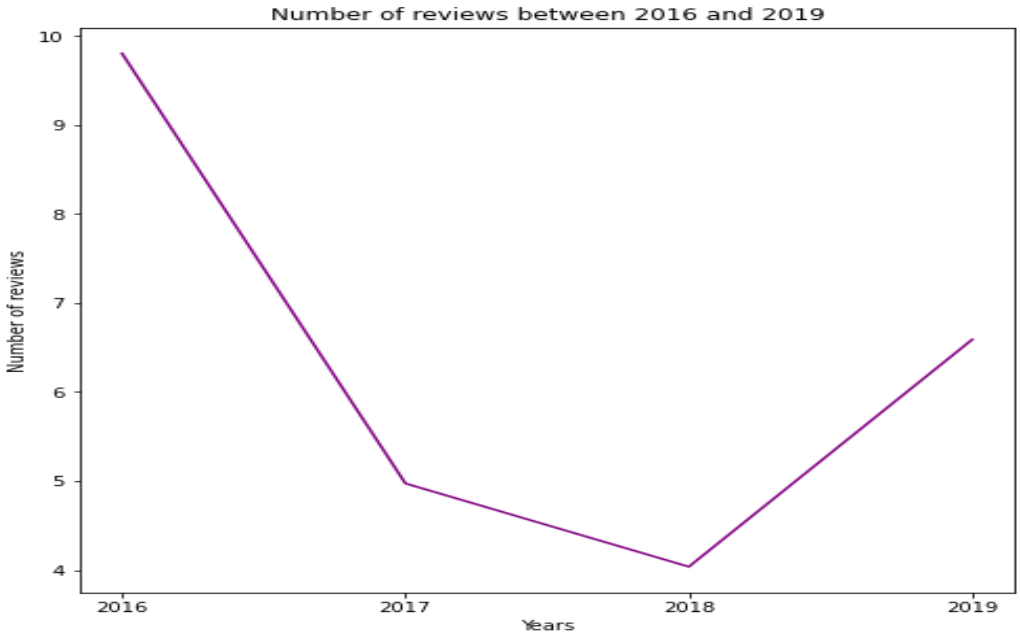

Figure 2c: Time series graph between average number of reviews and last reviews date between 2016 and 2019 years

Text preprocessing is a primary step for natural language processing (NLP). There are some steps for the text preprocessing step. Firstly, in the text data, 1938 sentences with 'the host canceled this reservation' removed. The reason for this is to make the data more meaningful. After that, emojis were found anywhere in the post detect. A cleaner data set created by removing 25% of the emojis in the entire text data.

In Figure 3, the 'en' language identified with langdetect, and the most used language was 'tr' by 28%. Although this rate was high, 126999 rows and eight columns remained when the data with the 'tr' language was completely removed.

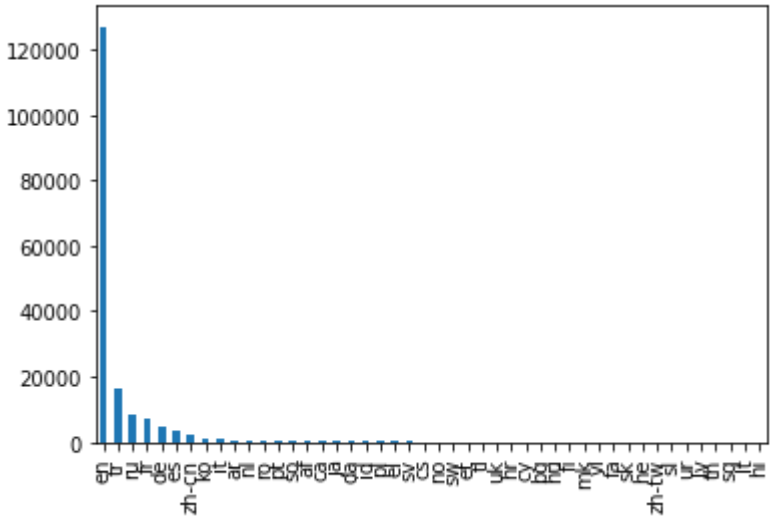

Figure 3: Distribution of after language detection analysis

A polarity prediction scores estimate with the remaining data set. In this part, comment data analyzed as neg, neu, pos, and compound. It split into columns. In Figure 4, the positive polarity score appears very intensely. When the similarity of the results compares, we can continue to recognize the data by explaining the relationship between polarity and review ratings with the correspondence analysis. The Compound score was determined as polarity, and the results showed on the scatter chart as in Figure 4.

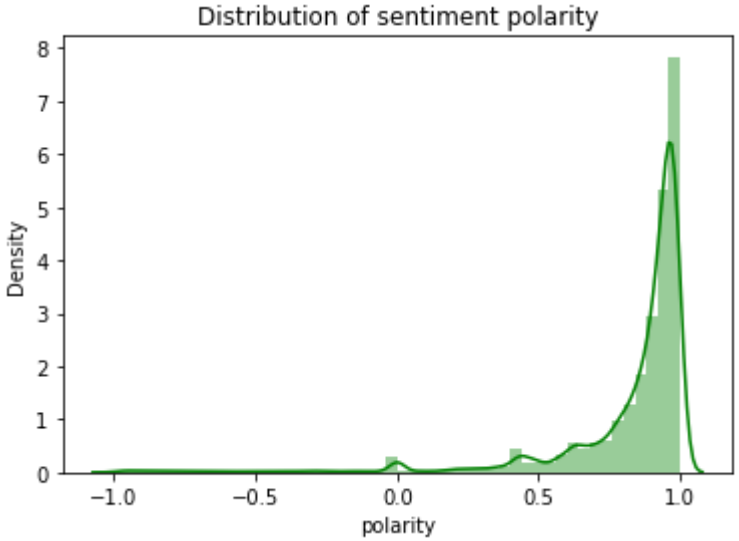

Figure 4: Distribution of positively polarity plot

In Figure 5, a scatter plot generated via a data set that finished the preprocessing step with Polarity scores. Based on Figure 5 benchmark with Figure 1, it shows clearly

from the positioning of the super hosts that there is a similarity between the polarity scores and the reviews score ratings.

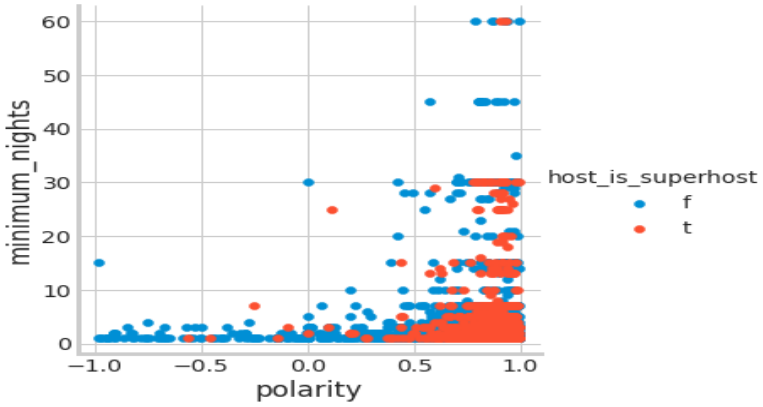

Figure 5: Scatter plot between polarity scores and minimum nights into host is super host values.

Neighborhoods with a polarity score above 0.5 examined, created by the data set cleared in Figure 6. First, the Beyoğlu neighborhood has the highest polarity score. The Beyoğlu neighborhood has more than 1200 Superhost in total, and it can also be associated with the accuracy of sentiment analysis, as it has a very high polarity score.

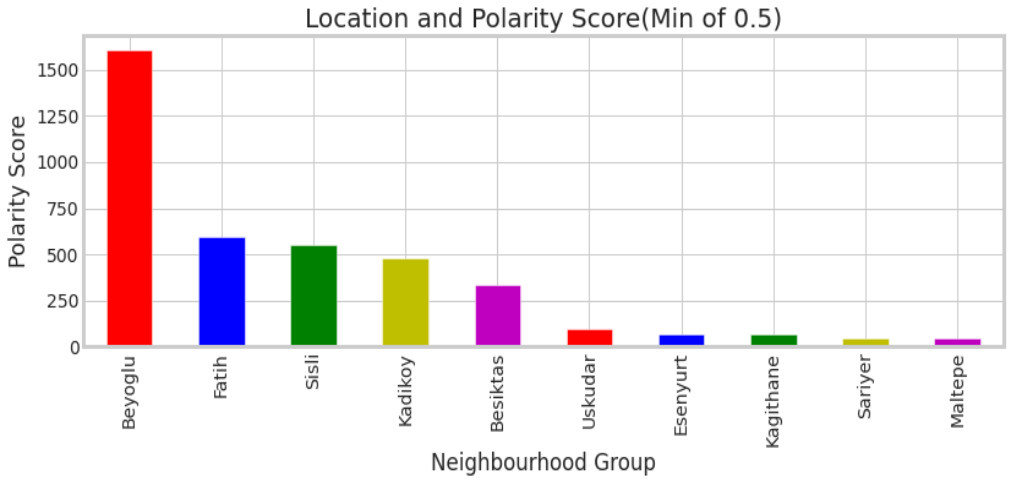

Figure 6: Distribution of Neighborhood with higher than 0.5 polarity scores

In Figure 7, Review score rating and polarity scores compare their similarities according to room types. In this chart, neighborhood and room types with polarity scores are very similar to each other. The main idea of creating the graph is to see the similarities between the scores and make sense of the newly created polarity feature.

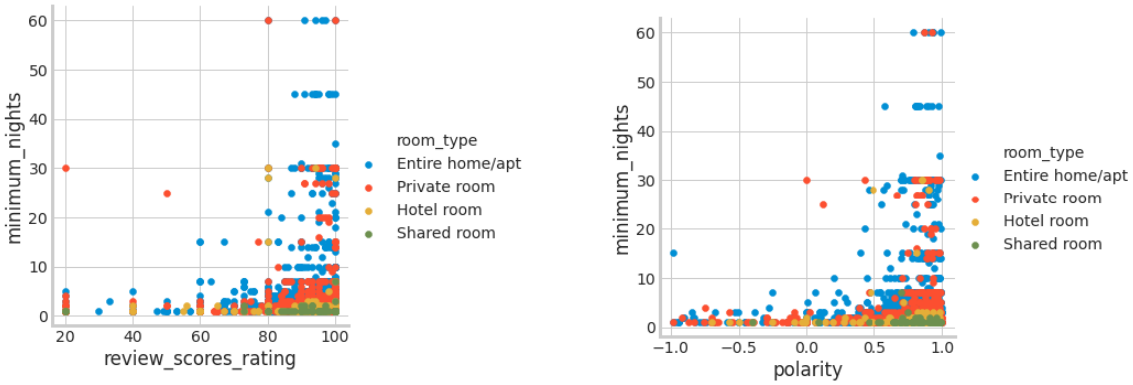

Figure 7: Two scatter plots between scores and minimum nights according to room types.

4. METHODOLOGY

In this project, machine learning methods were used in two different stages. The methods used are CountVectorizer, feature extraction, cosine similarity, spacy, SVD and RandomForest.

Below are the methods of applying the methods used in the stages.

4.1. Sentiment Analysis

The use of this analysis is to create expressions that make sense of the content of a text. These can be dissociated into positive, negative, neutral, or compound.

Sentiment analysis is used with methods such as natural language processing, text analysis. This analysis classifies the polarity of a given text in the document or sentences.

There are 3 different methods. Rule-based, feature-based, embedding-based methods. The VADER method is the most useful method for social media text sentiment analysis. It is designed with a focus on social media texts.

The results of sentiment analysis can be accessed using the RandomForestClassifier forecasting model for forecasting and model evaluation.

Feature extraction is the first machine learning method to form text sentiment analysis. The methods used in the preliminary part are important for data cleaning, as well as suitable for modeling classification to adapt to conditions. These methods include; remove numbers, stemming, part of speech tagging, remove puncture, lowercase, remove stop words.

When using this analysis, these posts should be cleared because the texts written on today's social network are a little confusing. Some methods are used for this.

In this project, the NLTK VADER rule-based method was used for sentiment analysis. In the text clearing section, the project was continued with only the 'most' using langdetect.

Canceled reservation comments removed, alphanumeric characters removed. Signs at the end of sentences, emojis have been removed. Tfidf ratios were then looked at using a vectorizer.

Polarity scores were measured along with vader_lexilon. This polarity measurement adapts to pos, neu, neg, and compound. By simply taking the compound from these values, a new feature was created on the data.

By creating a super host recommendation system together with polarity resulting from Sentiment analysis, you will be able to confirm the selection of super host according to the comments.

4.2. Recommendation System

In general, recommendation systems consist of algorithms that can present similar elements to users. Recommended application, articles, videos, etc. it's about the user. It analyzes the user's previous habits and makes recommendations. Each item shown to the user has a ranking. This sequence is based on the recommended system and is created by examining the user's historical data. This system consists of two separate categories. Content-based and Collaborative Filtering (CF) systems. The CF method consists of historical data between the user and the targeted elements. It is also based on the interactions of similar users.

It is a method that can offer items that users may like. In general, the steps of CF are as follows; finding similar users and items, estimating the order of items that have not yet been rated. RMSE and MAE are usually used to measure the accuracy of the results. There are 3 collaborative filtering methods. Model-based; models offer machine learning methods that allow users to keep unrated items in rating estimates in this approach. A few of them; Bayes Networks, clustering, Markov decision process, etc. This method is often applied to increase the accuracy of size reduction and model consolidation. One of the others is Hybrid; this approach is a combination of memory-based and model-based CF algorithms. As an example, the Google news recommender system. Finally, Deep-Learning is known as a CF method. The matrix factorization algorithm generalizes with a nonlinear neural architecture and takes advantage of new types of models known as Autoencoder. Deep learning methods can be applied in different scenarios; context-aware, sequence-aware, social tagging, etc. Another suggestion system method that has real effectiveness in a simple collaborative recommendation system is Content-Based, unlike CB, which uses user information in addition to making more content-based estimates. For example, it makes recommendations using demographic information such as the user's age, gender, and

behavior. For this reason, most websites inform users that some of their information will be processed upon request when entering the site, and they want to confirm it. In this method, if the similarity between the products is to be controlled, the Euclidean distance is calculated. But if textual elements are to be calculated, cosine is used and calculated with Jaccard for categorical data. The method used in this project is user-based CF. Pearson correlation and cosine similarity can be used primarily to determine the similarity between users. This method aims to use the polarity generated in the thought sentiment analysis to estimate the comments and then to measure the accuracy of whether the host is really super host. In practice, a suggestion system has been created using matrix factoring, spacy, Nearest Neighbors, cosine similarity methods.

5. RESULTS

As a result, Istanbul Airbnb data was examined in detail and the necessary preprocessing steps were made. Polarity scores were generated from the Airbnb reviews dataset using Sentiment analysis. Random Forest Classifier and Grid search models were used to measure the accuracy of predicted sentiment polarity scores. The best accuracy score was about 0.813 in the Random Forest Classifier. In Figure 8, ROC tells us how good the model is for distinguishing the given classes, in terms of the predicted probability. Moreover, the Sentiment analysis created a new dataset with polarity scores. After that, the recommendation system to be created for Superhosts began to be built.

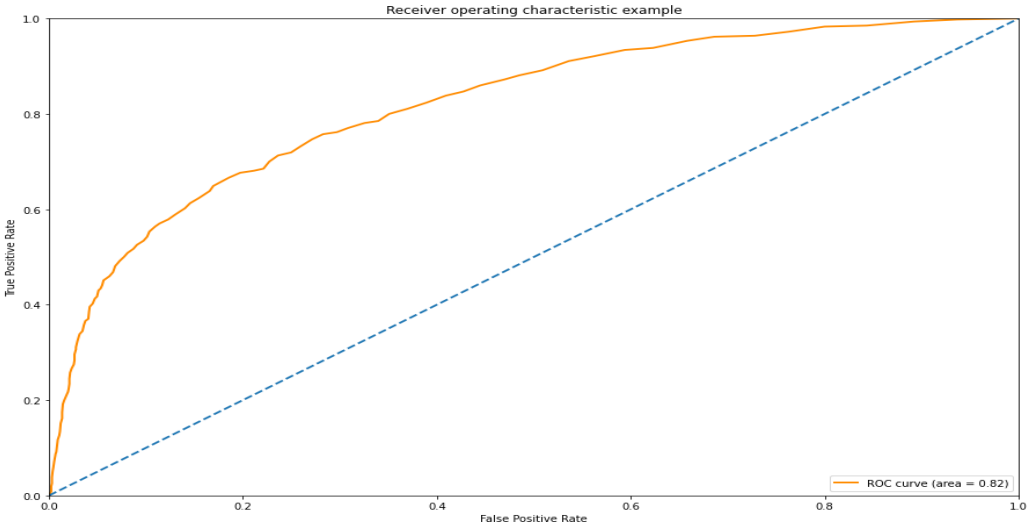

Figure 8: Sentiment Analysis results with ROC Curve

A matrix was generated from selecting the most important features along with the data set extracted from the sentiment analysis.

Two different recommendation models are created to predict the polarity scores of super hosts. The first method used was the surprise library. An RMSE score was obtained by selecting the best parameter with Grid search and estimating polarity scores with SVD. The RMSE score was used here as the accuracy metric, and the result was 0.26. The parameters used for the best score were "{'n_epochs': 10, 'lr_all': 0.002, 'reg_all': 0.4, 'n_factors': 100}". Then, three list suggestions print out for each host with the best parameters extracted.

Secondly, it was aimed to get a better result using matrix factoring. Matrix factorization can give the best results. Moreover, it gives us how much a user aligned with a set of k latent features or underlying tastes. SVD uses the recommendation

system. Before using SVD, a triple matrix was created along with host id, listing id, and polarity score. Since it was translated as a Pivot table, NaN values appeared, and these were removed.

A 0.249 rmse score revealed using SciPy's svds function. Later, Funk (2006) implemented the FunkSVD matrix factorization. Figure 9, the model with 100 iterations and a learning rate of 0.0001, along with a 73-minute running model, revealed an RMSE score of 0.001.

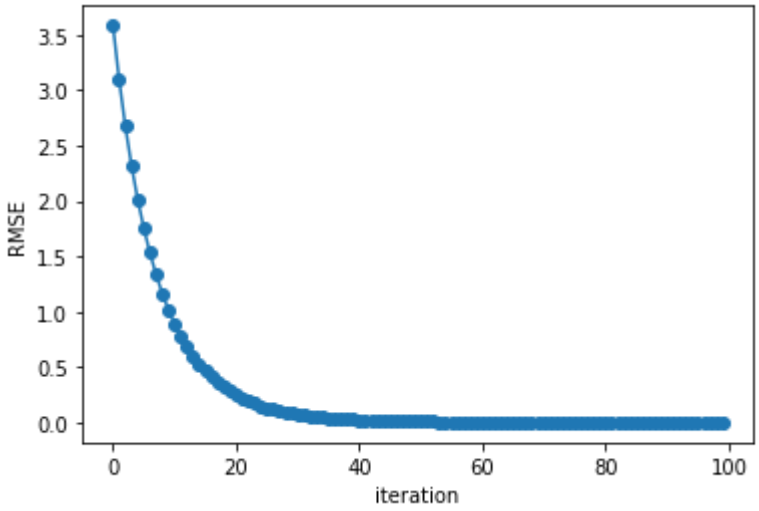

Figure 9: Recommendation engine RMSE results with FunkSVD

6. CONCLUSION

This project aims to create a fifth rule from guests' comments about hosts using the recommendation system and sentiment analysis in the selection of superhost for Airbnb. Using more than one model with Sentiment analysis, polarity scores were created to be used in the recommendation system. Moreover was created using algorithms such as FunkSVD, GridSearch, and SVD to achieve the best result in the recommendation system. The best result came out in FunkSVD and created a sample Recommendation system. Polarity scores were estimated for each latent user by looking at the Host-listing id. Each user then sorted the polarity scores of the recommended hosts. Along with this information, Airbnb can examine the polarity score for the host id it wants and decide whether it will be a host or superhost. For subsequent studies, model accuracy can be increased by feature engineering. It can also try for other accommodation companies with this method.

REFERENCES

[1] Panigrahi, N. & T, A. (2018). RHALSA: Ranking Hotels using Aspect Level Sentiment Analysis. *Journal of Computer Science, 14*(11), 1512-1520. https://doi.org/10.3844/jcssp.2018.1512.1520

[2] Bhujade, S. S., Chandak, M.B. (2018). A Hotel Recommendation System for Big Data Applications using a Keyword Aware Approach. *Journal of Engineering and Applied Sciences 13*(2): 523-528. https://www.researchgate.net/publication/323836792_A_hotel_recommendation_system_for_big_data_applications_using_a_keyword_aware_approach

[3] Thomas, E., Ferrer, A.G., Lardeux, B., Boudia, M., Haas-Frangii, C., & Agost, R.A. (2019). Cascaded Machine Learning Model for Efficient Hotel Recommendations from Air Travel Bookings. *RecTour*, 9 – 16. https://pdfs.semanticscholar.org/744e/a7451b9aee2c0e4e72fbf0528027dfb07ade.pdf?_ga=2.224273939.902502817.1610739261-760110415.1610739261

[4] Ramzan, B., Bajwa, I.S., Jamil, N., & Mirza, F. (2019). An Intelligent Data Analysis for Hotel Recommendation Systems using Machine Learning. https://arxiv.org/ftp/arxiv/papers/1910/1910.06669.pdf

[5] Deshmukh, A.S. (2019). The Effect of Superhost Status on Airbnb In Berlin Using Occupancy Rate and Revenue Per Available Listing. https://www.researchgate.net/profile/Ameya_Deshmukh3/publication/341482248_The_Effect_Of_Superhost_Status_On_Airbnb_In_Berlin_Using_Occupancy_Rate_And_Revenue_Per_Available_Listing/links/5ec3cd2da6fdcc90d682b7b2/The-Effect-Of-Superhost-Status-On-Airbnb-In-Berlin-Using-Occupancy-Rate-And-Revenue-Per-Available-Listing.pdf

[6] Najafi, S., & Salam, Z. (2019). Evaluating Prediction Accuracy for Collaborative Filtering Algorithms in Recommender Systems. https://www.diva-portal.org/smash/get/diva2:927356/FULLTEXT01.pdf

[7] Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management, (76)*, 58-70. https://www.journals.elsevier.com/international-journal-of-hospitality-management

[8] Lawani, A., Reed, M.M.R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in

Boston, *Regional Science and Urban Economics, (75),* https://doi.org/10.1016/j.regsciurbeco.2018.11.003.

[9] Jannach, D., Lerche L., Gedikli, F., & Bonnin, G. (2013). What recommenders recommend – An analysis of accuracy, popularity and sales diversity effects. *User Modeling, Adaptation, and Personalization, (7899),* 25–37. https://link.springer.com/chapter/10.1007/978-3-642-38844-6_3

[10] Hu, M.R., Lee, A.D. (2020). Airbnb, COVID-19 Risk and Lockdowns: Local and Global Evidence. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3589141

[11] García-Cumbreras, M.Á., Montejo-Ráez, A., & Díaz-Galiano, M.C. (2013). Pessimists and optimists: improving collaborative filtering through sentiment analysis, *Expert Systems with Applications, (40),* 6758–6765. https://doi.org/10.1016/j.eswa.2013.06.049

[12] Airbnb Insider. (2020). *Istanbul Airbnb* [Dataset] Available: http://insideairbnb.com

[13] Gastone, A. (2020, April 14). How to build a Recommender System for Airbnb in Python. *Medium*.

https://medium.com/@alexandra.gg150/how-to-build-a-recommender-system-for-airbnb-in-python-3a92ad500fa5.

[14] Santamicone, M. (2018, October 10). Seattle Confidential: unpacking Airbnb reviews with sentiment. *Medium.* https://medium.com/@mauriziosantamicone/seattle-confidential-unpacking-airbnb-reviews-with-sentiment-d421c15d8b8f

[15] Omar, A. (2019, September 10). Alternative Ways to Recommend Airbnb Listings Using Natural Language Processing. *towards data science*. https://towardsdatascience.com/alternative-ways-to-recommend-airbnb-listings-using-natural-language-processing-40fce2f1b

[16] Funk, S. (2006, December 11). Netflix Update: Try This at Home. https://sifter.org/~simon/journal/20061211.html