An Analysis of Customer Reviews of the Airbnb online platform

Marcell P. Granát & Zója M. Szabó

April 15, 2021

Contents

Abstract

Public perception of shared goods and/or services has changed significantly in the last few years. Shared accommodations have gained so great popularity, that house and flat sharing platforms like Airbnb now rival some of the world's largest businesses in hospitality. Sharing of personal properties provides an opportunity for owners to lower the transaction costs of operating short-term rentals and online rental marketplaces connect people who want to rent out their dwellings with the ones who are looking for accommodations. This study is aimed at determining the perceived behavior of individuals choosing Airbnb and exploring the factors that influence user ratings and consumer adoption of Airbnb while assuming that customer feedbacks contribute significantly to consumer choice. We also analyze the market trends of the Hungarian Airbnb accommodations as primary examples of sharing or collaborative economy. Weekly data was collected for the Hungarian accommodation establishments all over the country. We aimed to build a complete dataset of the active suppliers by using automated "web scraping" techniques during a certain window of time. Our database contained customer ratings, reviews and pieces of public information concerning the rooms. We performed a TF-IDF analysis and a Lasso-based feature selection on the aforementioned variables. Our key findings were that four attributes form the vast majority of online review comments. These are 'amenities', 'host', 'location' and 'cleanliness'. Contrary to our expectations, 'price' was not identified as a key determinant of customer satisfaction. A positivity bias can be detected in Airbnb users' comments (this means an overwhelmingly large number of positive comments), and a higher degree of intimacy between users and hosts than in the case of traditional hotels. Negative feedback is usually related to 'location' (safety issues), 'noise' and bad quality of 'amenities'.

Keywords— Airbnb, Customer reviews, Topic modelling

List of Tables

List of Figures

```
Error in readChar(con, 5L, useBytes = TRUE): cannot open the connection
```

Introduction

The phenomenon of collaborative consumption including peer-to-peer (P2P) accommodations has gained great popularity over the past decade. To understand what P2P accommodation means, we first define sharing or collaborative economy (the term 'collaborative economy' is interchangeably used with the term 'sharing economy'). The term 'collaborative economy' – as the Commission refers to it – "is business models where activities are facilitated by online platforms that create an open market place for the temporary user of goods or services often provided by private individuals" (COM, 2016, 356 final, page 3). Peer-to-peer accommodation occurs when two individuals (service providers and the users of these services) transact business without an intermediary third party.

Several factors drive the use of shared accommodations. Sharing of personal properties provides an opportunity for owners to (1) *lower* the *transaction costs* of operating short-term rentals (and this allows travelers to consider destinations and tourism activities that are otherwise of a prohibitive cost (Tussyadiah & Pesonen, 2015)). Online rental marketplaces (2) establish a *direct connection between people* who want to rent out their dwellings with the ones looking for accommodations. The (3) *availability of traditional experiences* – allowing travelers to "live like a local" – also has an impact on consumer decisions (Bridges & Vásquez, 2016). Another significant factor is (4) *social acceptance*: "sharing economies, ones such as accommodations, are a more sustainable alternative to traditional travel lodging through the consumption of less energy and resources, the production of less waste, and the overall theme of sustainability that is portrayed through many hosts and users of the service" (Midgett et al., 2018). Further key drivers explaining the growth in the use of Airbnb are "positive word of mouth", "digital infrastructure and literacy of the population" and "vocal support from the government" (PwC, 2016).

As a result of the factors mentioned above, house and flat sharing platforms like Airbnb now rival some of the world's largest businesses in hospitality. Since the founding of Airbnb in 2008, more than 800 million[1] guests have checked in at Airbnb listings around the world – and the number of guest arrivals is still growing rapidly.

Figure 1 presents the proportion of online bookings of the main types of tourist accommodation by country and partner type. Approximately 25 percent of partners – domestic and outbound together – book accommodation online in Hungary. This supports our expectation that there is a significant number of Airbnb users in the country. "A related study by PwC shows that while in 2013 the sharing economy companies [...] where the new business model is the most prevalent earned sales revenue of 15 billion dollars, by 2025 this will have risen to 335 billion dollars, so half of the revenues in these markets will go to companies with a sharing-based model" (PwC, 2015).

The main objective of this study is *to determine the perceived behavior of individuals choosing Airbnb accommodations through the analysis of online reviews and investigate the attributes that influence user ratings and reviews on the aforementioned collaborative platform.* In this paper, we examine the entire territory of Hungary.

Accordingly, our research question is the following: *What are the key factors that influence customer satisfaction concerning the Airbnb flats in Hungary?* To give a detailed answer, we analyzed online review comments gained from the Hungarian Airbnb website by using web scraping techniques. Big data analytics with over 75,000 observations (customer feedbacks) enabled us to better understand customer preferences, behavior and sentiment. Machine learning analysis was performed to extract all the above-mentioned information. We used two main text analysis tools: TF-IDF (term frequency-inverse document frequency) analysis – which was also supported by the closely related literature (Barbosa, 2019; Cheng, 2019; Zhang & Fu, 2020) – and the lasso regression model. By using these tools, *Cheng* (2019) identified that customers' experience with Airbnb and traditional hotels have some points in common, and as a result, users often evaluate their experience with Airbnb stays based on their experience with traditional hotels. The author detected three factors ('location', 'amenities' and 'host') that contribute significantly to customer satisfaction.

[1] https://news.airbnb.com/about-us/

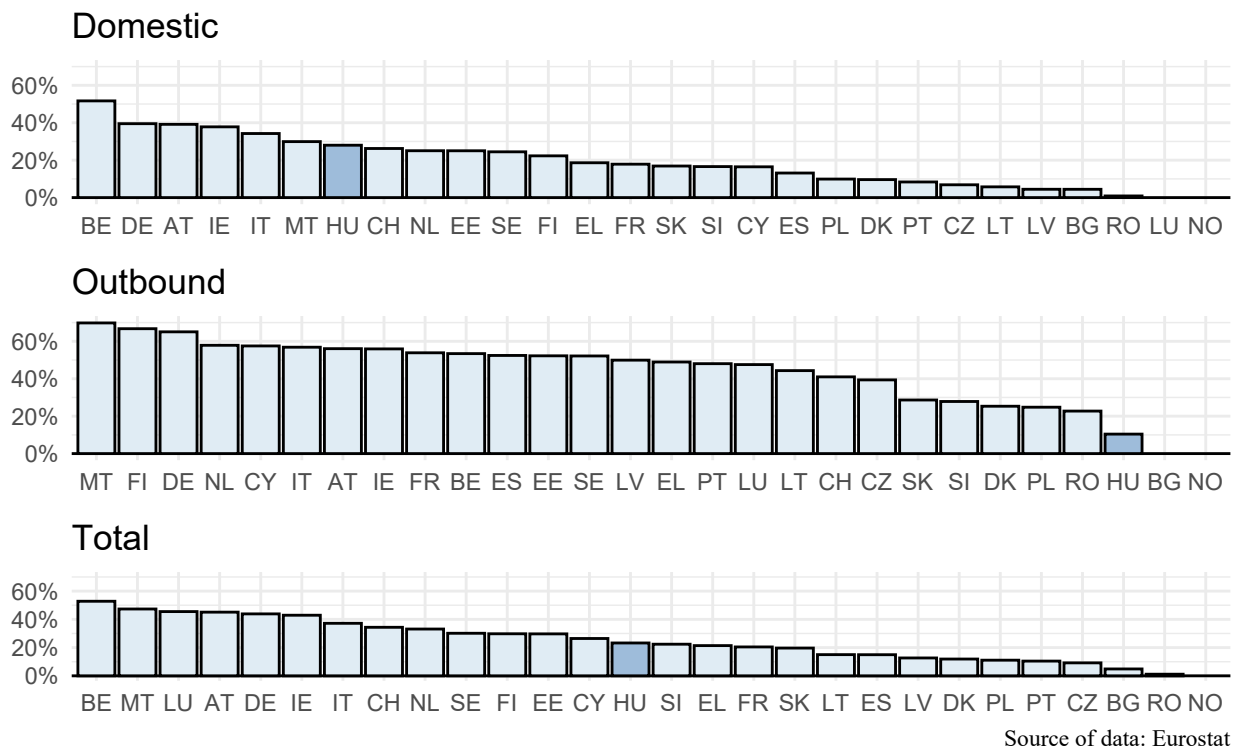


Figure 1: Proportion of internet bookings of the main means of accommodation by countries and the partner type

Contrary to expectations, 'price' was not statistically significant when investigating its impact on user ratings, which confirms our findings. *Cheng* also found that Airbnb users tend to use the names of the hosts in their reviews, which led the author to the conclusion that a higher degree of intimacy between hosts and costumers can be observed – we came to the same conclusion. According to *Zervas* (2015, p. 12), there is empirical evidence that user ratings on the Airbnb platform are „dramatically high". Our results also support this statement: we found that 60 percent of the observations (properties) have a user rating of 4.75 or above.

Since the phenomenon of Airbnb is quite new, only a small amount of research has been done about the preferences and motivation of its users. This study contributes to a broader understanding of the experiences and preferences of Airbnb users in Hungary, while extends the scientific discourse in many directions by exploring research gaps. Methodologically, it contributes to the existing literature by providing an accurate and up-to-date database (created by using web scraping techniques in R software), which contains data about Airbnb users all over Hungary. The database is available in a public GitHub repository[2].

Data

Data collection

Our research is based on the customer reviews available on the Hungarian Airbnb website. We chose to examine the high season[3] – the time interval between May 31 and August 29 – which enabled us to investigate the entire population of Airbnb users in Hungary during the summer. Airbnb's search engine has three characteristics that are important to mention: (1) it lists properties within a given radius, (2) there is a necessity of using an iterative price filter to find every apartment in a settlement, and (3) it provides a limited

[2]https://github.com/MarcellGranat/airbnb-research
[3]KSH (2019): Helyzetkép a turizmus, vendéglátás ágazatról, figure 7

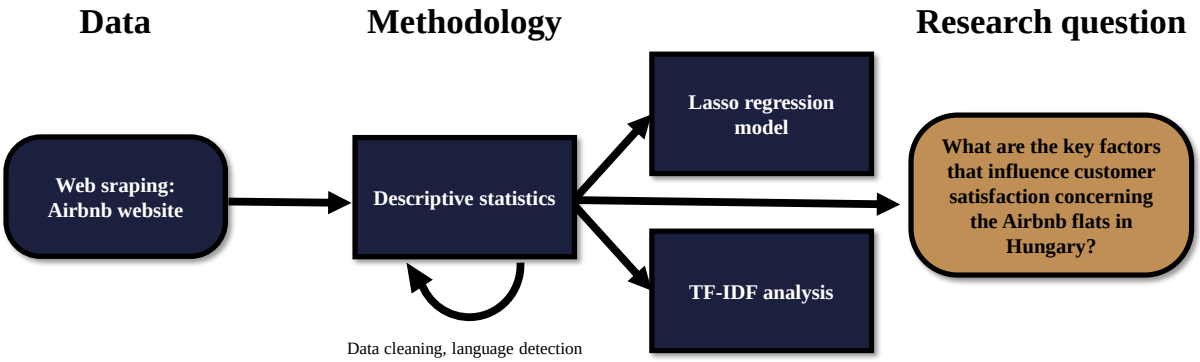


Figure 2: Main steps of the study

availability of comments (about a maximum of 40 per property). These will be discussed later in this study.

We collected information about over 8,700 Airbnb properties listed in Hungary (downloaded on March 5, 2021). For each property, the database we generated contained the property's unique id, its URL address, the type of place (entire place, private room, shared room), the name of the host, an approximate location (Airbnb does not provide an accurate location for listings before booking), price, the number of reviews, star ratings (for the overall experience and for specific categories including cleanliness, accuracy, communication, location, check-in, and value-for-money), a description and customer reviews. We aimed to collect data about the whole territory of Hungary, which enabled us to detect territorial typologies concerning travelers' accommodation preferences and identify as many patterns in the data as possible.

The first step in the data collection process was creating a list of the potential settlements in Hungary. The list included the 50 largest settlements in Hungary by population[4] supplemented by the most visited settlements in Hungary[5]. The reason why we combined the aforementioned datasets is that we aimed to cover all regions of Hungary, which – in some cases – meant that we added settlements manually to the list in order to fill the gaps on the map. As a result, we got a list of 115 settlements (Figure 3). As previously mentioned, there was a necessity to use an iterative price filter in order to find every Airbnb flat in a settlement. We faced this difficulty in the case of locations with more than 300 listings, because Airbnb only displays 300 listings at a time. If a price filter had not been used, the rest of the listings would have been omitted from the results. To avoid this, we introduced a price filter. The search results now presented only the listings that were within the chosen range of price, which reduced search results.

As a second step, we collected information from the Airbnb website by searching for properties in every settlement we listed previously. Airbnb's search algorithm provides listings within a certain radius from the center of the search. Although we had to remove duplicates and the properties that are not located in Hungary, we could easily detect and add a large number of new cities and villages to the list, which now consisted of 575 settlements all over the country, 8700 apartments (rooms) and over 78,000 review comments.

[4]Hungarian Central Statistical Office (2020): 50 largest settlements in Hungary, available at: https://www.ksh.hu/stadat_files/fol/en/fol0014.html

[5]KSH (2019): Kereskedelmi szálláshelyek vendégforgalma – Magyaroszág kereskedelmi szálláshelyei, available at: https://www.ksh.hu/turizmus-vendeglatas

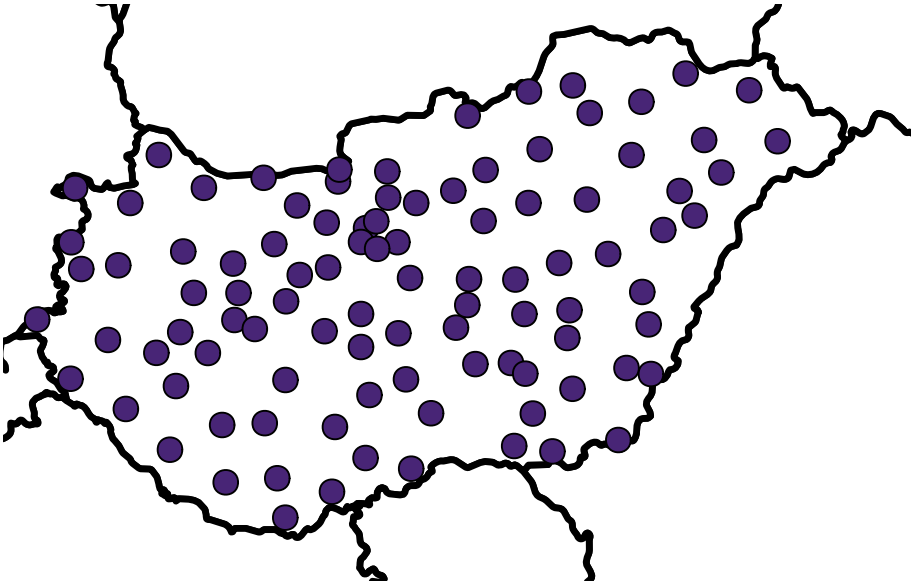


Figure 3: Starting points of our scraping algorithm

Table 1: Number of avaiable accomodations by counties on the Airbnb website

County	Number of settlements
Bács-Kiskun	84
Baranya	74
Békés	69
Borsod-Abaúj-Zemplén	210
Budapest	5104
Csongrád	222
Fejér	108
Győr-Moson-Sopron	150
Hajdú-Bihar	239
Heves	314
Jász-Nagykun-Szolnok	59
Komárom-Esztergom	38
Nógrád	46
Pest	239
Somogy	537
Szabolcs-Szatmár-Bereg	41
Tolna	33
Vas	149
Veszprém	525
Zala	455

Table 1 displays the distribution of Airbnb flats (rooms) in Hungary by county. It can be seen, that more than 5000 properties are located in Budapest. Although most of the English comments belong to the properties located in the capital, the large proportion of these accommodations in the population greatly determines the overall distribution of the language of comments (Figure 4). A higher ratio of Hungarian comments in other counties indicates that domestic tourism is more significant in these regions – according to data published by the *Hungarian Central Statistical Office* (2019) the number of domestic tourism nights spent in commercial accommodation is higher in the case of properties in the most popular rural destinations such as Hajdúszoboszló and settlements near Lake Balaton.

To identify the language in which a review was written, we used text categorization based on character n-gram frequencies. This function is included in an R extension package (*textcat*) which is used "for natural language processing, in particular by providing the infrastructure for general statistical analyses of frequency distributions of (character or byte) n-grams" (Mair, 2013). A drawback of the algorithm is that since it

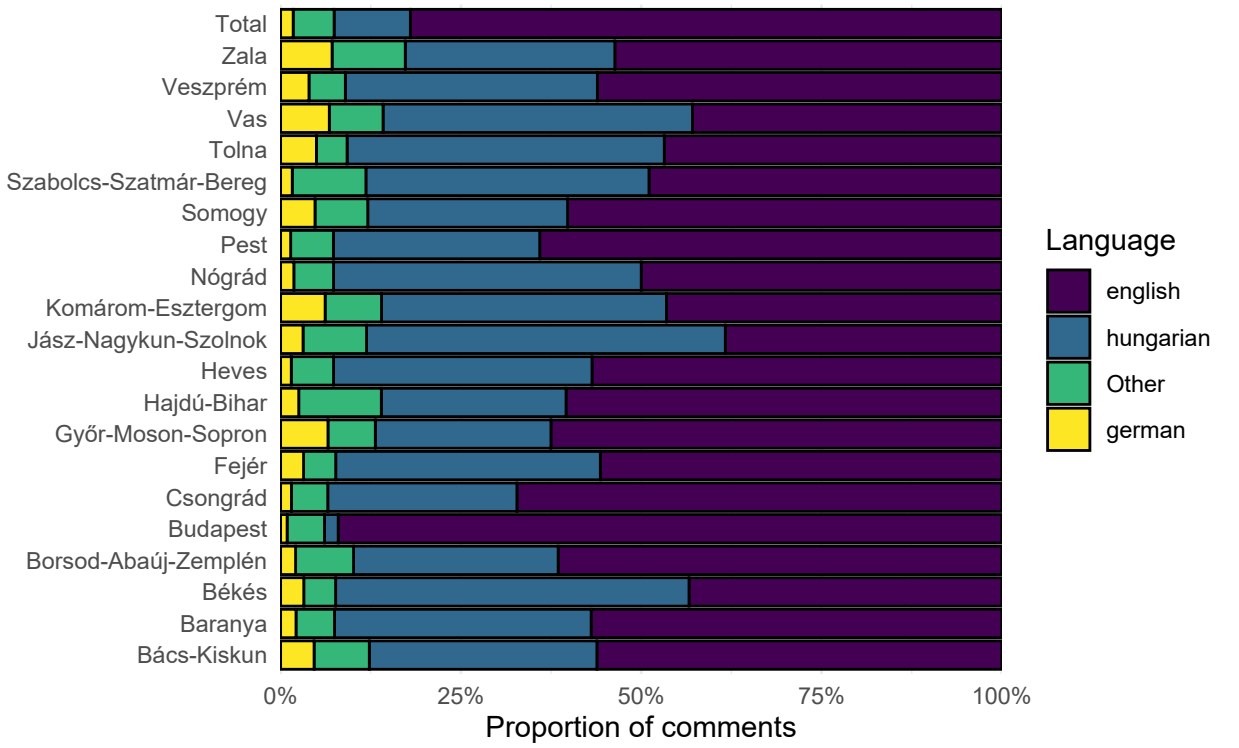


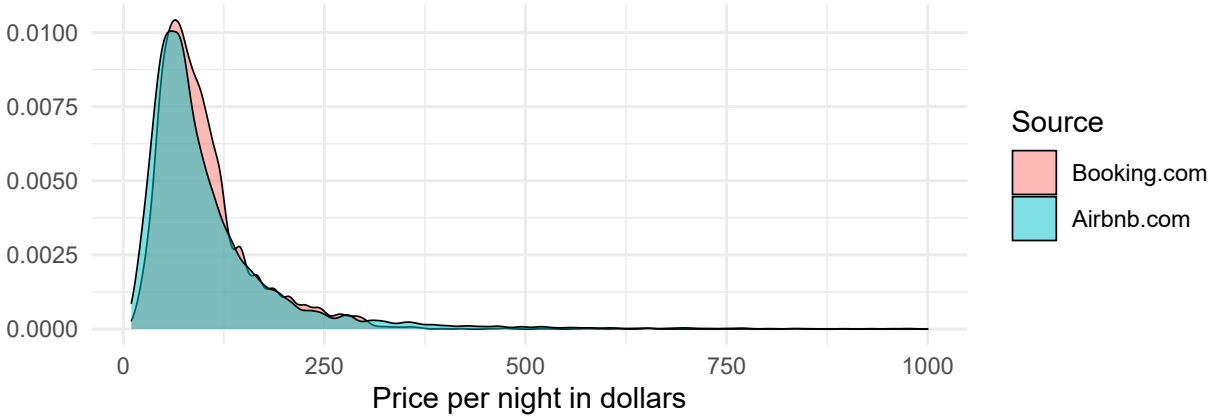
Figure 4: Most common languages found in the comments by counties

hotels (shown in Figure 5). For this, we collected information about over 16,000 hotels from Booking.com by using web scraping techniques. The distribution of prices for Airbnb properties and traditional hotels has a large kurtosis, which means that there are high probabilities of extremely high and extremely low prices in both cases. The empirical density function indicates a median of 79 dollars for Airbnb and 85 dollars for hotels, and a mean of 130 dollars for Airbnb, and 126 dollars for hotels. In conclusion, even though the top motivation of tourists to choose Airbnb is its comparatively low cost (Guttentag, 2017), we found that the distribution of prices is very similar in the case of the two types of accommodation.

Figure 6 shows the number of Airbnb accommodations per capita by territorial units. It can be seen, that the relative number of those properties located in the immediate vicinity of Lake Balaton is very high. According to data published by the *Hungarian Central Statistical Office* (2021), among the first 100 most-visited settlements concerning domestic tourism nights in 2020, there were 25 located near Lake Balaton.

Figure 7 illustrates the relationship between price and user ratings. A slope of zero means that the value of user rating is constant no matter the value of price. Knowing the price of the apartment does not reduce the uncertainty associated with user ratings. *Cheng* (2019) also found, that "price was not treated as important as other attributes when evaluating Airbnb experiences" (Cheng, 2019 p. 61). This was predictable, since – in contrast to other factors determining customer satisfaction such as cleanliness (as discussed later), – the exact value of the price is already known to the customer before using the service.

The empirical cumulative distribution function (ECDF) provides an alternative visualization of distribution (Figure 8). In the graph, the x-axis is assessment (or user rating) and the y-axis is cumulative density corresponding to the assessment. We found that there is a "positivity bias" (Zervas et al., 2015) towards the hosts in the comment reviews: over 60 percent of the observations (properties) have a user rating of 4.75 or above. The existing literature also supports this: "several empirical papers have analyzed the rating distributions that arise on major review platforms, most arriving at a similar conclusion: ratings tend to be overwhelmingly positive, occasionally mixed with a small but noticeable number of highly negative reviews" (Zervas et al., 2015). According to *Hu et al.* (2009), "this implies two biases: (1) purchasing bias – only



Distribution of prices on Booking.com is drawn based on 16059 observations.

Source: https://www.booking.com

Download at 2020-05-02

Figure 5: Comparison of prices per night: Airbnb and Booking.com

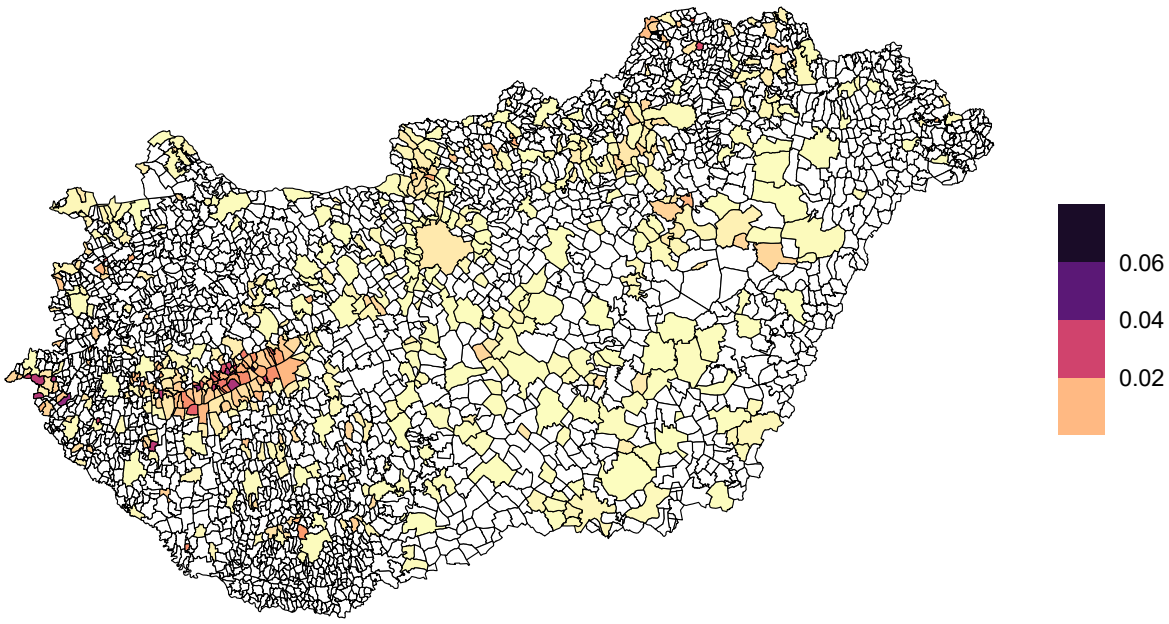


Figure 6: Number of Airbnb accomodations per capita by territorial units

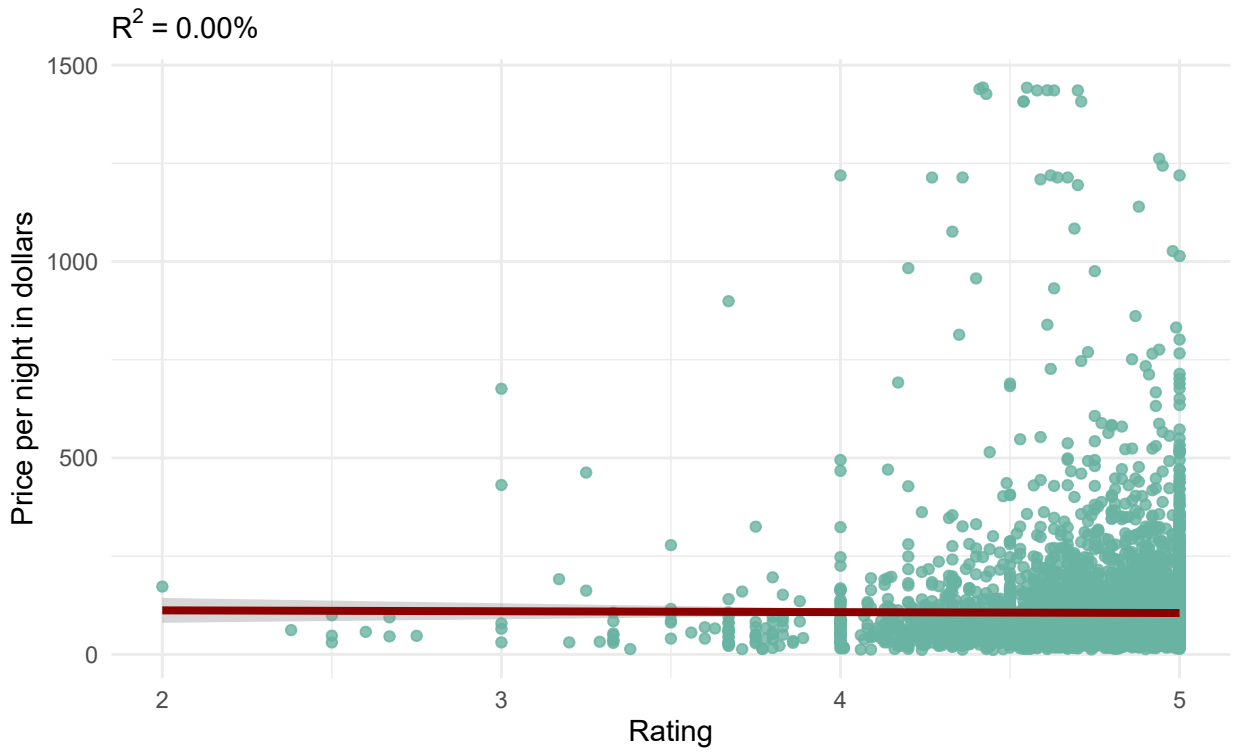


Figure 7: Scatter plot of overall ratings and prices

consumers with a favorable disposition towards a service purchase the service, and have the opportunity to write a review, and (2) under-reporting bias – consumers with polarized (either positive or negative) reviews are more likely to report their reviews than consumers with moderate reviews". The distribution of the online review comments is also referred to as a J-shaped distribution (shown in Figure 8). Figure 8 also illustrates the empirical cumulative distribution of reviews.

It is plain to see, that where the value of user rating is higher than 4.9, the number of comments is lower. As mentioned previously, only approximately 40 review comments per property (room) is visible on the Airbnb website, and as a result, we only had limited information to process when assessing the relationship between the number of reviews and user ratings – in the case of properties with a large number of reviews, it led to statistical noise. However, below 40 comments, it is visible that the probability that the number of reviews takes on a value less or equal to a number is higher, if user ratings are higher than 4.9. This led to the conclusion that there is a considerable amount of fake reviews. According to *Valant* (2015), "tools for increasing consumer awareness and raising their trust in the market should not, however, mislead consumers with fake reviews, which, according to different estimates, represent between 1% and 16% of all 'consumer' reviews."

Since the focus of our study is customer preferences, we presented the key factors that determine overall ratings in Figure 9. Our theoretical consideration is that the overall rating is the explained and the others are explanatory variables. To filter out multicollinearity (the situation where explanatory variables are highly related) we illustrated the partial linear correlation coefficients. 'Accuracy'[6] was not identified as a key influencer of user ratings, which was contrary to our expectations, considering that 'accuracy' is mainly viewed as the 'host trustworthiness', and the currency of sharing economy is trust. Other researchers also argue that "the level of hosts' trustworthiness, mainly as inferred from their photos, affects listings' prices and probability of being chosen, even when all listing information is controlled for" (Ert et al., 2016). According

[6]This includes the timing of the experience, the name of the host, the location, a list of what is provided, and what guests will do. (Airbnb, 2018, available at: https://blog.atairbnb.com/accuracy/)

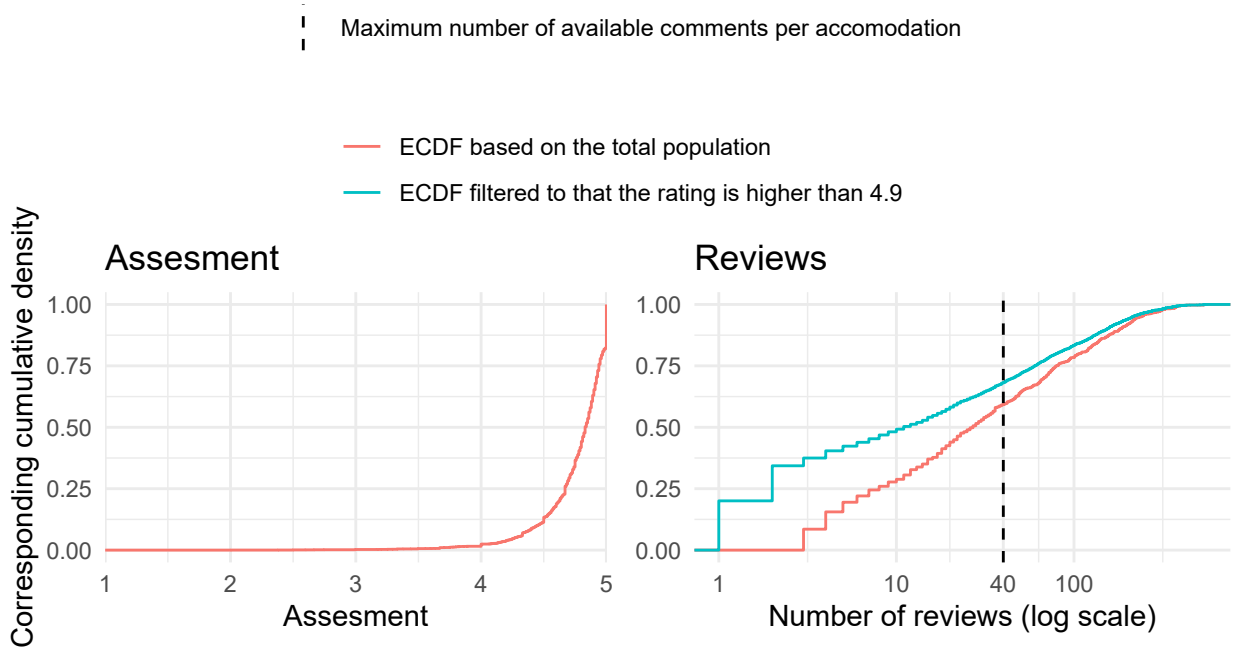


Figure 8: Empirical cumulative distribution functions of overall rating scores and the number of reviews

to Figure 11, 'location' and overall user ratings are not highly related. Cheng (2019) also found that 'location' was not statistically significant in influencing user satisfaction. 'Cleanliness', however, plays a significant role in people's satisfaction (Bridges & Vazquez, 2018). It is important to note that the factors indicated are the ones that appear on the website of Airbnb, not the ones that we flagged as key determinants of customer reviews and satisfaction later in the study. However, common elements were found by using text mining techniques where we also gave explanations for the relationships.

Term frequency-inverse document frequency (TF-IDF) analysis

Text mining (also referred to as text analysis or text data mining) is the process of deriving information from unstructured text by transforming it into structured data to identify patterns. In order to quantify what a review comment is about, we used the term frequency-inverse document frequency analysis (hereinafter TF-IDF). TF-IDF is a technique that evaluates how relevant a word is to a document in a collection of documents (here: categories by user rating). In this study, we aimed to define the words (gained from user comment reviews) from which we can infer a conclusion, that whether an Airbnb property or room can be categorized into the upper decile group, the bottom decile group or the intermediate category (the dataset of Airbnb properties was divided into three sub-datasets: the upper and the bottom decile group of properties and a group of the rest of the data by overall ratings, which enabled us to perform text mining). The value of the upper and bottom decile is 4.99 and 4.5. To calculate a term's TF-IDF, two metrics are multiplied: *term frequency* (how frequently a word appears in a document) and the term's *inverse document frequency*, which is the frequency of the term (local frequency) adjusted for how rarely it appears in the collection of documents (global frequency). This decreases the weight of the commonly used words, and increases the weight of the words that are not frequent in a set of documents (Silge & Robinson, 2021. The inverse document frequency of a term is defined as:

$$idf(\text{ term }) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right) \tag{1}$$

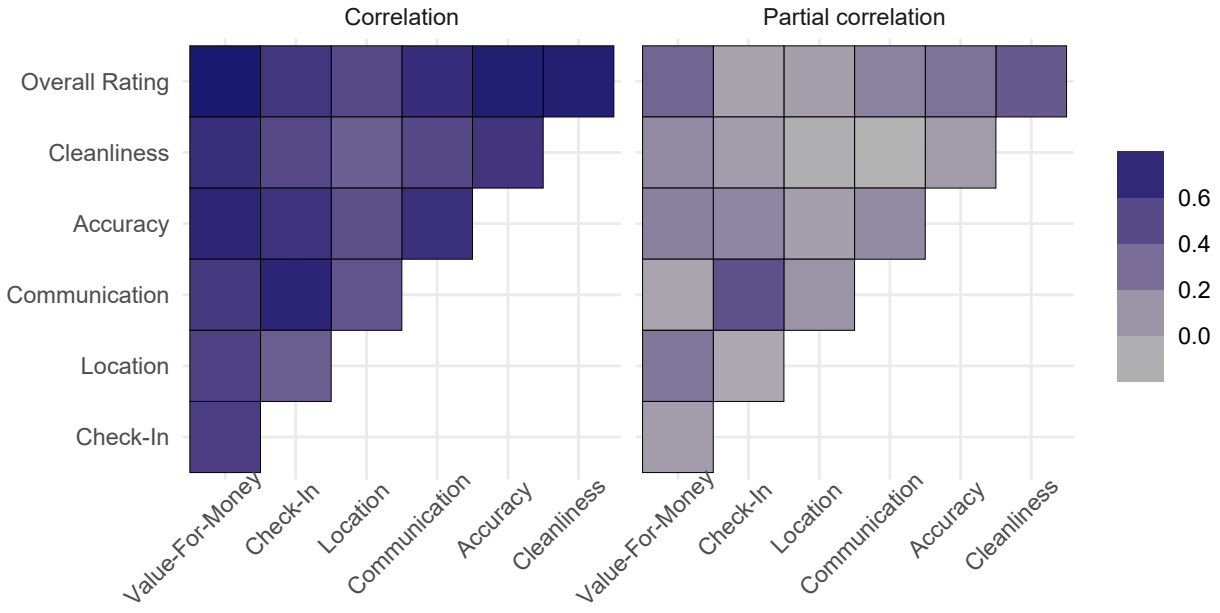


Figure 9: Correlation and partial correlation coefficients among customer ratings for different categories

If the word is very common and occurs in (almost) every document, this number will approach 0. Otherwise, it will approach 1. For instance, the word 'refund' is very frequent in the bottom decile, but nowhere else, so the number of its inverse document frequency approaches 1. This means, that this term has a higher relative frequency among Airbnb flats with lower user ratings. Thus, we marked it as a 'negative' word (Figure 10). The term 'refund' appeared to be the most related word to accommodations with relatively low ratings. More specifically, this indicates that service users were not completely satisfied with the flexibility of the host. The high relative frequency of the words 'cancel' and 'compensation' also suggests this. We found, that robberies were more frequent in the case of these properties. Terms 'dust', 'mold', 'stains', 'leaking [roof, tap, boiler, etc.]' and 'bugs' are more common where a lower rating was given, and suggest that – in some cases – the services did not suit the convenience of customers, and the cleanliness of the apartments was insufficient. Location-related safety issues were also reported by *Barbosa* (2019). It can be seen, that the relative frequency of names is extremely high among properties with high user ratings. This confirms our assumptions that there is a higher degree of intimacy between the hosts and service users concerning peer-to-peer accommodation. "The name carries a personal touch in this space" (Cheng, 2019). This is a factor that determines the confidence of the users and influences guest satisfaction to a great extent. Two examples for the appearance of the above-mentioned terms ('robbery' and 'flexibility of the host'):

> "We were robbed in this Airbnb! Robbed of at least $2,500.00 worth of valuables; laptop, jewelry, camera, clothing all GONE. After this, we were made to sleep in the same apartment that was robbed that very day. [...] The place itself is nice, clean and tidy. But the way that we were treated after going through something as awful as a robbery is despicable. Do not stay here." (Airbnb ID: 18167202, average rating: 4.15)

> "They accepted by booking late which meant I had then booked another Airbnb. I explained this and asked to cancel.. or just give me part refund.. neither of which happened.. Shame.." (Airbnb ID: 23819249, average rating: 4.15)

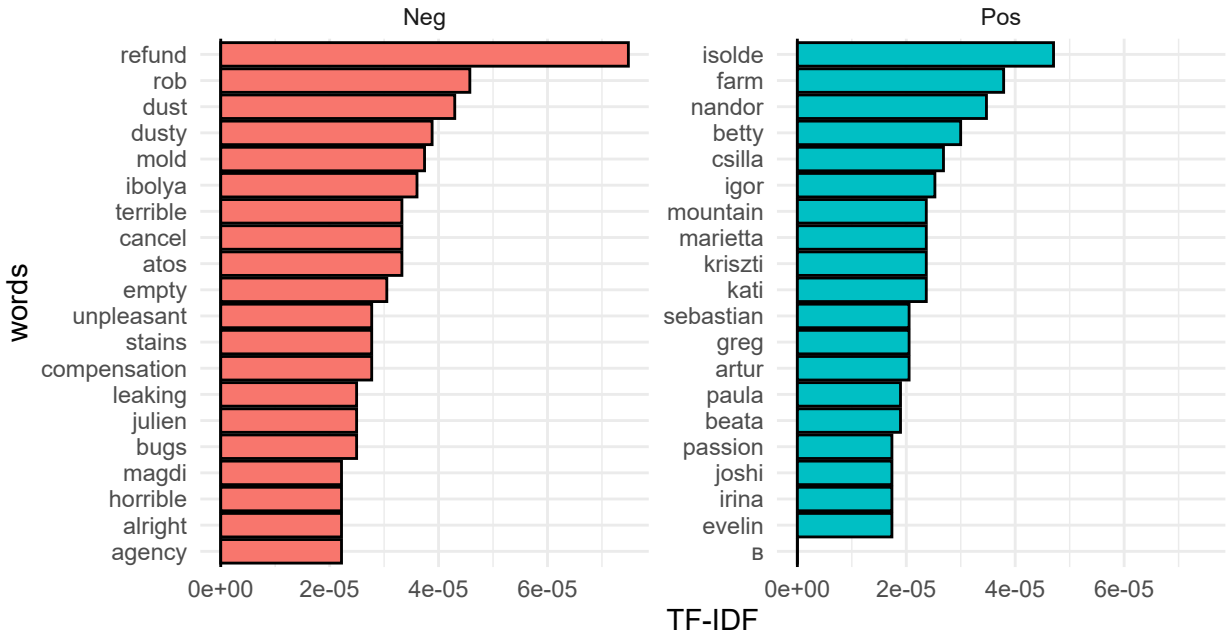


Figure 10: TF-IDF

Lasso-based feature selection

The TF-IDF analysis is a great tool to detect whether a word is more frequent among comments related to higher or lower-rated accommodations, but there are some drawbacks as well. The algorithm fails to leave out words that have a high TF-IDF, even though the global frequency of the word is extremely low. For instance, the word 'robbery' only appeared in the comments of low-rated Airbnb flats (so its relative frequency is high), but because of its absolute frequency, it could not be used for predicting user ratings. Another issue raised with the use of the TF-IDF analysis when the words related to lower-rated flats correlated with each other. For example, 'horrible' and 'cancel' appeared in the same comments:

> "I made a reservation. The host didn't write me, he didn't respond. After I have called them, they told me they didn't see the reservation. That the system is broken, and they couldn't see my reservation. I had to cancel it before arriving and had to look for another accommodation at 9 pm! Horrible" (Airbnb ID: 37383207, average rating: 3.5).

To handle correlation between words, and find the truly useful ones for prediction, we performed a lasso regression-based classification of the words.

The method of categorizing comments as "positive" and "negative" is equivalent to the one presented in the previous section. If a flat has a rating equal to or above the upper decile, it is positive, and negative if it is equal to or below the bottom decile. The difference compared to the previous model is that the intermediatory category has been omitted. The reason for that is the classifier is a logistic regression, which only predicts binary outputs.

For this, the key idea is to generate random samples from the population of the reviews, and divide them into training and testing sets. Logistic lasso regression determines the coefficient of each word, and categorizes the comment reviews as "negative" or "positive". Its performance (so that if the frequency of a term provides a good prediction of a review's positive or negative nature) is validated on the test set. After committing this multiple times, we measure the contribution of a term to the reduction of prediction error (variable importance). The most important terms are reported in Figure 11. We created 25 repetitions, and the samples contained about 9000 reviews as a training set and 3000 as a test set.

This model framework outperforms simple TF-IDF-based categorization. (1) If the relative frequency of a term differs in the case of high-rated and low-rated properties' comments, but its total frequency is low, the word will not show significant contribution to the categorization of a text (here review comments). (2) If a combination of two words frequently occurs together, the logistic lasso regression model will omit the variable which does not contain new information.

This attribute is one of the main disadvantages of the model, since the more meaningful word may be excluded. For example, the word 'bit' by itself does not provide much information, yet appears because it represents the information that the Airbnb flat is located in a "bit noisy" surroundings.

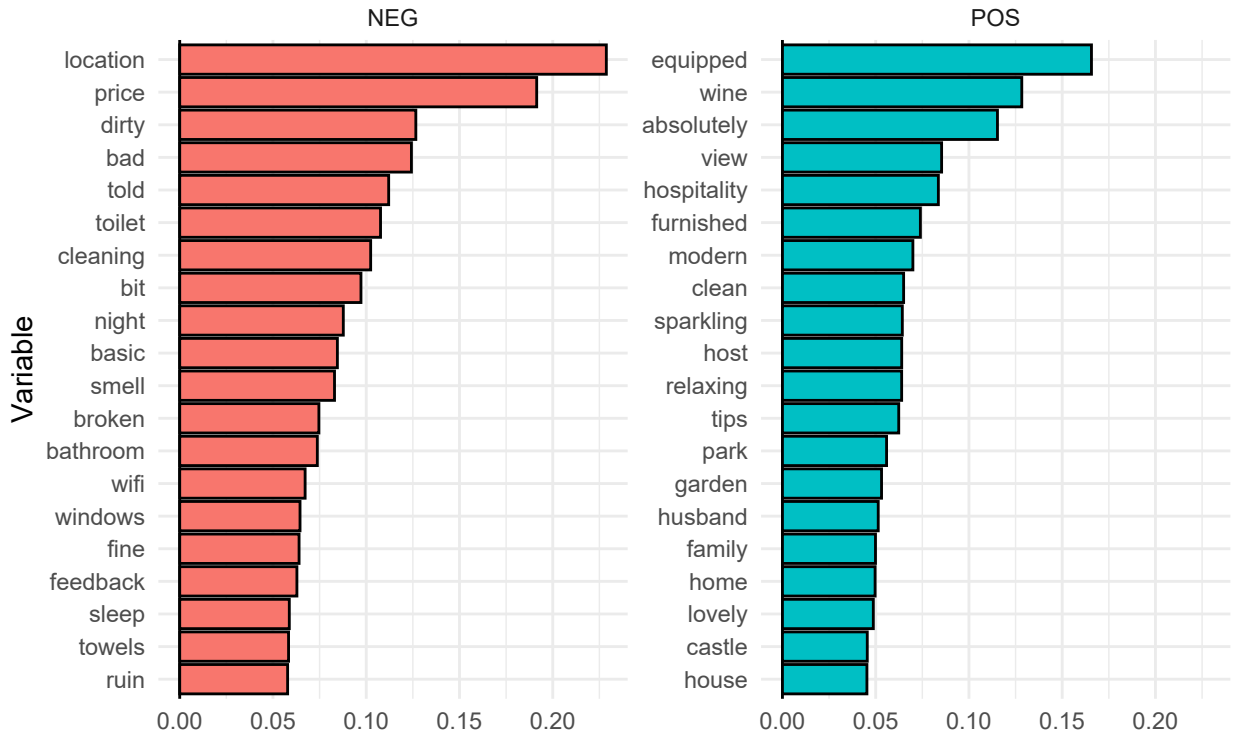


Figure 11: Most important positive and negative words based on lasso selection

As Figure 11 shows, the flat's location and amenities are the attributes, that most stand out among review comments, contributing to guest satisfaction in an unfavorable way. The literature studied also supports these findings (Barbosa, 2019). Words related to amenities such as 'toilet', 'broken [equipment]', 'bathroom', 'windows', 'towels' and 'wifi' mostly appeared in reviews that belonged to properties with ratings of 4.9 stars (upper decile) or above. According to *Lee et al.* (2019), a bad Wi-Fi connection in an Airbnb flat can ruin the entire customer experience. We, as humans tend to value things only after we have lost is – this is also the case with internet connection. Earlier in this paper, 'price' has not been defined as a key determinant of overall user ratings, but in this model, it is ranked very high. The reason for this is that the word 'price' in customer reviews refers to 'value-for-money' numerous times. We also examined which factors contribute favorably to user satisfaction. 'Equipped' and 'furnished [apartments]' appeared as people's top priority. 'Location' was also a significant factor. We removed some of the adverbs and adjectives like "beautifully" and "amazing" associated with the location, which enabled us to gain concrete information. However, the presence of positive opinions concerning location in the comments gave evidence that it highly determines users' positive experience. *Barbosa* (2019) said that several people participate in peer-to-peer accommodation with their families (high frequency of words 'family' and 'husband'). They rent whole houses with many amenities ('equipped', 'furnished' apartments), so they have an opportunity to 'relax', while enjoying the 'view' and sipping 'wine' in the 'garden'. Others were provided with destination 'tips' by the locals, and were more interested in obtaining traditional experience. 'Hosts' also play an important role in evaluating

customer experiences.

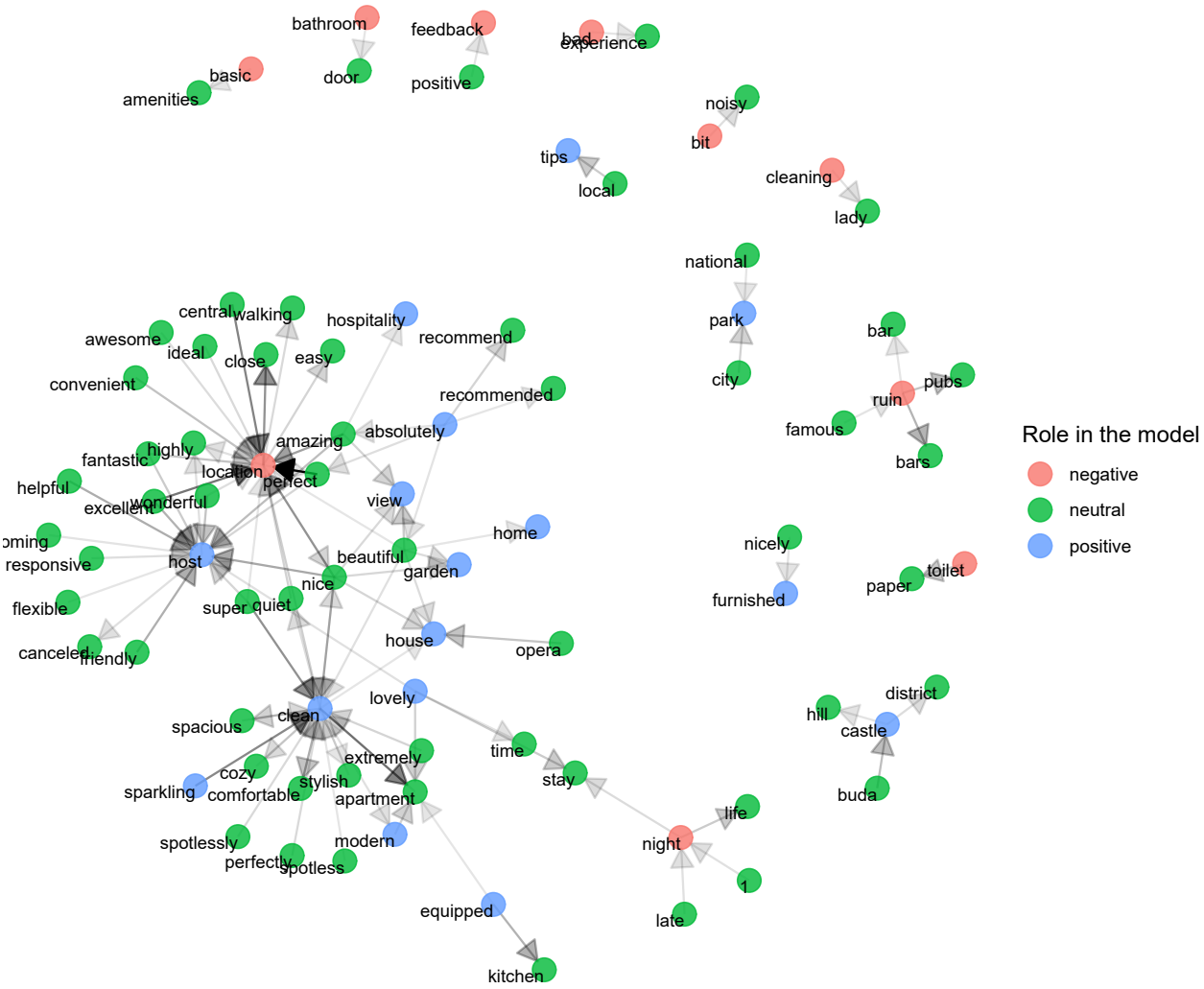


Figure 12: Graph of bigramms related to the model results

Figure 12 illustrates the graph of bigrams related to the lasso regression model results. The words presented were classified into three categories: positive words (belonging to the flats with a user rating equal to or above 4.9), negative words (belonging to the flats which were given 4.5 stars or below) and neutral words. For example, the word 'bit', which also appeared in Figure 11, now became interpretable: in some cases, Airbnb users found the environment a "bit noisy". It is confirmed by Cheng (2019), that negative sentiment is mostly caused by 'noise'. Another example is the word 'ruin', which is also coupled with low user ratings, makes no sense by itself. However, supplemented by the word 'bars' or 'pubs', it is easily understood. From this comment review below, we can see that apartments located near nightclubs are more likely to get negative feedback:

> "There's a nightclub on the first floor with an open sky, you can hear the music all night long as the windows aren't noised canceling... not something we were expecting at all... Otherwise, the location is great, the breakfast is a bit expensive for what they serve but it's unlimited... If you plan to party all night long it's a good place, otherwise, be prepared to have trouble sleeping."

Conclusion

In this paper, we aimed to understand what factors drive the use of shared accommodations and what are the key influencers of customer satisfaction concerning the tourist accommodation service provided by Airbnb all over Hungary. We obtained data from the Airbnb website by using web scraping techniques and examined these through text analysis. Two main statistical tools were used to quantify and visualize Airbnb users' online review comments: the TF-IDF (term frequency-inverse term frequency) analysis and Lasso-based feature selection (lasso regression) – both by using R software. To investigate customer satisfaction and its key influencers, we chose to analyze user ratings and reviews comments. Our study has provided empirical evidence for several factors, that play an important role in evaluating customer experiences. Most of them corresponded with the ones suggested by the extant literature. In some cases, different statistical models provided different – a couple of times contradictory – results. In the following, we report the results of descriptive statistics.

A positivity bias was identified towards the hosts in the comment reviews: the vast majority of the apartments had a user rating of 4.75 or above. The related literature also supports our finding that user ratings tend to be unambiguously positive. This phenomenon can be explained with two types of bias: (1) purchasing bias – only consumers with a favorable disposition towards a service purchase the service, and (2) under-reporting bias – consumers with polarized reviews are more likely to report their reviews (Hu et al., 2009). We found that there is a considerable number of fake reviews (written by the host themselves or friends), because the probability that the number of reviews takes on a value less or equal to a given number is higher, if user ratings are outstandingly high.

Surprisingly, 'accuracy' (the timing of the experience, the name of the host, the location, a list of what is provided, and what guests will do) was not identified as a key determinant of user ratings. According to these, we concluded that a hosts' 'trustworthiness' is not as important as collaborative consumption suggests. Although 'location' was considered a highly valued aspect by some of the other papers, we found that it is not highly related with overall user ratings. 'Cleanliness', however, seemed to play a significant role in customer satisfaction.

In the TF-IDF part, we derived information from customer reviews by transforming them into structured data. TF-IDF analysis is a technique that evaluates how relevant a word is to a document across a set of documents. We aimed to categorize words from review comments into the upper decile group, the bottom decile group or the intermediate category by overall user ratings.

We found that service users were *not completely satisfied* with the 'flexibility of the host' concerning low-rated flats by identifying keywords such as 'compensation' and 'cancel' in the reviews. 'Robberies' were also mentioned several times in the case of these properties. Terms related to amenities and cleanliness also appeared in a negative aspect.

The research revealed that the relative frequency of names is extremely high among properties with high user ratings. This confirmed our assumptions that the Airbnb platform establishes a direct and more intimate relationship between host and user. In conclusion, 'host' as an attribute *positively influences* guest satisfaction to a great extent.

In the Lasso-based feature selection part, we performed a lasso regression-based classification of the words. Words were classified into three categories: positive words (flats with a user rating equal to or above 4.9), negative words (flats which were given 4.5 stars or below) and neutral words. This method was very similar to the one used in the TF-IDF part.

Based on this model, 'location' and 'amenities' are the attributes, that stood out mostly among review comments, contributing to guest satisfaction in an *unfavorable way*. Although 'price' has not been defined as a key determinant of overall user ratings by previous tools, this model ranked it very high in determining (*negative*) customer experience. However, it is not directly 'price', but 'value-for-money' is the term commonly used.

Factors that contribute *favorably* to user satisfaction are 'equipped' and 'furnished [apartments]' (with many amenities), which appeared as people's top priority. 'Location' was also a significant factor. The high presence

of positive opinions about location in the comments gave evidence that it highly determines users' overall satisfaction. 'Hosts' also play an important role in evaluating customer experiences. Although 'price' has not been defined as a key determinant of overall user ratings by the previous tools, this model ranked it very high in determining (negative) customer experience. The reason for this is that the word 'price' in customer reviews refers to 'value-for-money' numerous times.

We also illustrated the graph of bigrams related to the lasso regression model results. The words displayed were classified into three categories based on the method presented in the previous model. This tool is useful when a term by itself does not provide meaningful information, yet appears because it represents the information of the other word connected. In conclusion, the answer to our research question that "*what are the key factors that influence customer satisfaction concerning the Airbnb flats in Hungary?*" is the following: the research identified four main attributes, that form the majority of online review comments on the Hungarian Airbnb website. These are 'amenities', 'host', 'location' and 'cleanliness'.

By investigating the nature of customer review comments, we acquire a deeper understanding of customer preferences. This includes exploring other segments of collaborative consumption and the future of short-term rentals such as Airbnb. The results obtained in this paper can serve as benchmark data in further research, while enabling us to look at time trends.

References

[A Europan Agenda for the collaborative economy, 2016] A Europan Agenda for the collaborative economy (2016). : A europan agenda for the collaborative economy.

[Barbosa, 2019] Barbosa, S. R. P. (2019). *Airbnb customer satisfaction through online reviews*. PhD thesis.

[Bridges and Vásquez, 2018] Bridges, J. and Vásquez, C. (2018). If nearly all airbnb reviews are positive, does that make them meaningless? *Current Issues in Tourism*, 21(18):2057–2075.

[Cheng and Jin, 2019] Cheng, M. and Jin, X. (2019). What do airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70.

[Ert et al., 2016] Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism management*, 55:62–73.

[Feinerer et al., 2013] Feinerer, I., Buchta, C., Geiger, W., Rauch, J., Mair, P., and Hornik, K. (2013). The textcat package for n-gram based text categorization in r. *Journal of statistical software*, 52(6):1–17.

[Guttentag and Smith, 2017] Guttentag, D. A. and Smith, S. L. (2017). Assessing airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management*, 64:1–10.

[Hu et al., 2009] Hu, N., Pavlou, P. A., and Zhang, J. J. (2009). Why do online product reviews have a j-shaped distribution? overcoming biases in online word-of-mouth communication. *Communications of the ACM*, 52(10):144–147.

[Lee et al., 2019] Lee, C. K. H., Tse, Y. K., Zhang, M., and Ma, J. (2019). Analysing online reviews to investigate customer behaviour in the sharing economy: The case of airbnb. *Information Technology & People*.

[Midgett et al., 2018] Midgett, C., Bendickson, J. S., Muldoon, J., and Solomon, S. J. (2018). The sharing economy and sustainability: A case for airbnb. *Small Business Institute Journal*, 13(2):51–71.

[PWC, 2016] PWC (2016). *Sharing or paring?*

[PwC, 2016] PwC, U. (2016). Assessing the size and presence of the collaborative economy in europe. *Report Delivered to EC*.

[Silge and Robinson, 2017] Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Sebastopol, CA.

[Tussyadiah and Pesonen, 2016] Tussyadiah, I. P. and Pesonen, J. (2016). Impacts of peer-to-peer accommodation use on travel patterns. *Journal of Travel Research*, 55(8):1022–1040.

[Valant, 2015] Valant, J. (2015). *Consumer protection in the EU. Policy overview*.

[Zervas et al., 2017] Zervas, G., Proserpio, D., and Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, 54(5):687–705.

[Zervas et al., 2020] Zervas, G., Proserpio, D., and Byers, J. W. (2020). A first look at online reputation on airbnb, where every stay is above average. *Marketing Letters*, pages 1–16.

[Zhang and Fu, 2020] Zhang, Z. and Fu, R. J. (2020). Accommodation experience in the sharing economy: A comparative study of airbnb online reviews. *Sustainability*, 12(24):10500.

Appendix: R codes

```r
# Packages ------------------------------------------------------------------------------

library(tidyverse)
library(knitr)
library(tidytext)
library(rvest)
library(parallel)
library(RSelenium)
library(rnaturalearth)
library(rnaturalearthdata)
library(topicmodels)
library(stm)
library(quanteda)
library(tidytext)
library(vip)
library(tidymodels)
library(textrecipes)

# Introduction --------------------------------------------------------------------------

f.plot_eurostat <- function(x){
  eurostat::get_eurostat('tour_dem_ttorg', time_format = 'num') %>%
    filter(trip_arr %in% c('ACC_WEB', 'TOTAL') & duration == 'N1-3' & time == 2017 &
             purpose == 'TOTAL') %>%
    pivot_wider(names_from = trip_arr, values_from = values) %>%
    mutate(
      value = ACC_WEB/TOTAL
    ) %>%
    filter(partner == x) %>%
    mutate(geo = fct_reorder(geo, -value)) %>%
    ggplot() +
    aes(geo, value, fill = geo == 'HU') +
    geom_hline(yintercept = 0) +
    geom_col(color = 'black') +
    scale_fill_brewer(palette = 3, guide = F) +
    scale_y_continuous(labels = scales::percent, limits = c(0, .7)) +
    labs(
      x = NULL, y = NULL, title = case_when(
        x == 'DOM' ~ 'Domestic',
        x == 'OUT' ~ 'Outbound',
        T ~ 'Total'
      )
    )
}

ggpubr::ggarrange(
  f.plot_eurostat('DOM'),
  f.plot_eurostat('OUT'),
  f.plot_eurostat('WORLD') +
    labs(caption = 'Source of data: Eurostat'),
  ncol = 1
)
```

```r


cities <- readxl::read_excel("cities.xlsx")
hun_cities <- read_csv('worldcities.csv') %>%
  filter(iso2 == 'HU')

world <- ne_countries(scale = "large", returnclass = "sf")

merge(cities, hun_cities, by = 'city') %>%
  tibble() %>%
  ggplot() +
  geom_sf(data = world, size = 1.2, fill = 'white', color = 'black') +
  coord_sf(xlim = c(16, 23.4), ylim = c(45.5, 48.7), expand = FALSE) +
  geom_point(aes(x = lng, y = lat), size = 4, alpha = 1, color = 'black',
             shape = 21, fill = viridis::viridis(1, begin = .1)) +
  theme_void()

dat_rbnb %>%
  merge(rename(hun_counties, geo = city)) %>%
  count(county) %>%
  knitr::kable(caption =
                 'Number of avaiable accomodations by counties on the Airbnb website',
               align = c('l', 'c'), col.names = c('County', 'Number of settlements'))

dat_comments %>%
  merge(dat_rbnb) %>%
  select(id, language, geo) %>%
  mutate(
    geo = ifelse(str_detect(geo, 'Budapest'), 'Budapest', geo)
  ) %>%
  merge(rename(hun_counties, geo = city)) %>%
  tibble() %>%
  mutate(
    language = fct_lump(language, n = 3) %>%
      fct_infreq()
  ) %>%
  {rbind(., mutate(., county = 'Total'))} %>%
  mutate(county = fct_reorder(county, county == 'Total')) %>%
  select(county, language) %>%
  na.omit() %>%
  ggplot() +
  aes(y = county, fill = language) +
  scale_x_continuous(labels = scales::percent, limits = c(0,1), expand = c(0,0)) +
  geom_bar(color = "black", position = position_fill()) +
  scale_fill_viridis_d() +
  labs(x = 'Proportion of comments', y = NULL, fill = "Language")

ggpubr::ggarrange(
  (booking_prices * 0.0033) %>%  # 1 HUF means 0.0033 US $ -> 2021-05-08 <-
    {tibble(type = 'Booking.com', price = .)} %>%
    rbind(tibble(type = 'Airbnb.com', price = dat_rbnb$price)) %>%
    mutate(type = fct_reorder(type, price, .desc = F)) %>%
    ggplot() +
    aes(price, fill = type) +
```

```
    geom_density(alpha = .5, color = 'black', size = .3) +
    scale_x_continuous(limits = c(10, 1000)) +
    labs(x = 'Price per night in dollars', y = NULL, fill = 'Source',
         caption =
           'Distribution of prices on Booking.com is drawn based on 16059 observations.
         \n Source: https://www.booking.com
         \n Download at 2020-05-02')
)
(booking_prices * 0.0033) %>%  # 1 HUF means 0.0033 US $ -> 2021-05-08 <-
  {tibble(type = 'Booking.com', price = .)} %>%
  rbind(tibble(type = 'Airbnb.com', price = dat_rbnb$price)) %>%
  pivot_wider(names_from = type, values_from = price) %>%
  skimr::skim()
dat_rbnb %>%
  transmute(NAME = ifelse(str_detect(geo, 'Budapest'), 'Budapest', geo)) %>%
  count(NAME) %>%
  merge(rename(hun_counties, NAME = city), all = T) %>%
  mutate(n = n/pop) %>%
  merge(
    sf::read_sf('kozighatarok/admin8.shp'), all.y = T
  ) %>%
  ggplot() +
  geom_sf(aes(fill = n, geometry = geometry), size  =.2, color = 'black') +
  scale_fill_viridis_c(direction = -1, option = 'magma', na.value = 'white',
                       guide =
                         guide_colorsteps()) +
  labs(fill = NULL) +
  theme_void()

dat_rbnb %>%
  filter(assesment != 0 & price < 1500) %>%
  ggplot() +
  aes(assesment, price) +
  geom_point(color="#69b3a2", alpha=0.8) +
  geom_smooth(method = 'lm') +
  labs(x = 'Rating', y = 'Price per night in dollars', subtitle = expression(R^2~'='~'0.00%'))

dat_rbnb %>%
  filter(assesment != 0 & price < 1500) %>%
  lm(formula = assesment ~ price) %>%
  {broom::tidy(.); broom::glance(.)}

max_comment <- max(pull(count(dat_comments, id), n))

ggpubr::ggarrange(
  dat_rbnb %>%
    ggplot() +
    stat_ecdf(aes(x = assesment, color = 'ECDF based on the total population')) +
    geom_blank(aes(color = 'ECDF filtered to that the rating is higher than 4.9')) +
    scale_x_continuous(limits = c(1, 5), expand = c(0,0)) +
    labs(x = 'Assesment', y = 'Corresponding cumulative density', title = 'Assesment',
         color = NULL, linetype = NULL) +
    scale_linetype_manual(values = c(2)) +
```

```
    theme(
      legend.position = 'bottom',
      legend.box = 'vertical'
    ) +
    geom_vline(aes(xintercept = as.numeric(NA),
                   linetype = 'Maximum number of available comments per accomodation')),
  dat_rbnb %>%
    ggplot() +
    stat_ecdf(data = filter(dat_rbnb, assesment > 4.9),
              mapping = aes(n_reviews,
                            color =
                              'ECDF filtered to that the rating is higher than 4.9')) +
    stat_ecdf(aes(x = n_reviews, color = 'ECDF in the total population')) +
    scale_x_log10(breaks = c(1, 10, max_comment, 100)) +
    geom_vline(aes(xintercept = max_comment,
                   linetype = 'Maximum number of available comments per accomodation')) +
    labs(x = 'Number of reviews (log scale)', y = NULL,
         title = 'Reviews', linetype = NULL) +
    scale_linetype_manual(values = c(2)) +
    theme(
      legend.position = 'bottom',
      legend.box = 'vertical'
    ), common.legend = T
)

dat_rbnb %>%
  select(assesment, starts_with('stars')) %>%
  na.omit() %>%
  {list(cor(.), ppcor::pcor(.)$estimate)} %>%
  lapply(function(x) {
    rownames_to_column(data.frame(x), var = 'x') %>%
      pivot_longer(-1, names_to = 'y') %>%
      mutate(r = row_number())
  }) %>%
  reduce(rbind) %>%
  mutate(
    r = cumsum(ifelse(r == 1, 1, 0))
  ) %>%
  mutate_at(1:2, .funs = function(x) {
    case_when(
      x == 'assesment' ~ 'Overall rating',
      str_remove_all(x, 'stars_') == 'value' ~ 'Value-for-money',
      str_remove_all(x, 'stars_') == 'checkin' ~ 'Check-in',
      T ~ str_remove_all(x, 'stars_')
    ) %>%
      str_to_title()
  }
  ) %>%
  mutate_at(.vars = 1:2, function(x) factor(x, ordered = T,
                                            levels =
                                              c(setdiff(unique(x), 'Overall rating'),
                                                'Overall rating'))) %>%
  mutate(value = ifelse(x < y, value, NA)) %>%
```

```r
  na.omit() %>%
  ggplot() +
  aes(y, x, fill = value) +
  geom_tile(color = 'black') +
  facet_wrap('r', labeller = as_labeller(c('1' = 'Correlation',
                                           '2'= 'Partial correlation'))) +
  scale_fill_gradient(low = 'grey70', high = 'midnightblue',
                      guide = guide_colorsteps()) +
  labs(x = NULL, y = NULL, fill = NULL) +
  theme_minimal() +
  scale_x_discrete(limits = rev) +
  scale_y_discrete(limits = rev) +
  theme(axis.text.x = element_text(angle = 45, hjust = .8))

merge(dat_comments, dat_rbnb) %>%
  filter(language == 'english' & !is.na(assesment)) %>%
  tibble() %>%
  select(assesment, text) %>%
  mutate(
    rating = as.numeric(Hmisc::cut2(assesment, g = 10, levels.mean = T)),
    rating = case_when(
      rating == min(rating) ~ 'Neg',
      rating == max(rating) ~ 'Pos',
      T ~ 'Middle'
    ),
  ) %>%
  select(rating, text) %>%
  unnest_tokens(words, text) %>%
  mutate(SnowballC::wordStem(words = words)) %>%
  group_by(rating, words) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  bind_tf_idf(term = words, document = rating, n = n) %>%
  arrange(desc(tf_idf)) %>%
  filter(rating != 'Middle') %>% # TODO remove?
  filter(n > 10 & !str_detect(words, '\\d') & words != 'pt') %>%
  anti_join(rename(stop_words, words = word)) %>%
  group_by(rating) %>%
  group_modify(~ head(.x, 20)) %>%
  ungroup() %>%
  mutate(words = fct_reorder(words, tf_idf)) %>%
  ggplot() +
  aes(tf_idf, words, fill = rating) +
  geom_vline(xintercept = 0) +
  geom_col(color = 'black', show.legend = F) +
  facet_wrap(~rating, scales = 'free_y') +
  labs(x = 'TF-IDF')

reviews_parsed <- merge(dat_comments, dat_rbnb) %>%
  filter(language == 'english' & !is.na(assesment)) %>%
  tibble() %>%
  select(assesment, text) %>%
  mutate(
```

```r
    rating = as.numeric(Hmisc::cut2(assesment, g = 10, levels.mean = T)),
    rating = case_when(
      rating == min(rating) ~ 'Neg',
      rating == max(rating) ~ 'Pos',
      T ~ 'Middle'
    ),
  ) %>%
  select(rating, text) %>%
  filter(rating != 'Middle')

set.seed(123)
review_split <- initial_split(reviews_parsed, strata = rating)
review_train <- training(review_split)
review_test <- testing(review_split)

review_rec <- recipe(rating ~ text, data = review_train) %>%
  step_tokenize(text) %>%
  step_stopwords(text, language = "eng") %>%
  step_tokenfilter(text, max_tokens = 500) %>%
  step_tfidf(text) %>%
  step_normalize(all_predictors())

review_prep <- prep(review_rec)

lasso_spec <- logistic_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet")

lasso_wf <- workflow() %>%
  add_recipe(review_rec) %>%
  add_model(lasso_spec)

lambda_grid <- grid_regular(penalty(), levels = 40)

set.seed(123)
review_folds <- bootstraps(review_train, strata = rating)

set.seed(2020)
lasso_grid <- tune_grid(
  lasso_wf,
  resamples = review_folds,
  grid = lambda_grid,
  metrics = metric_set(roc_auc, ppv, npv)
)

best_auc <- lasso_grid %>%
  select_best("roc_auc")

final_lasso <- finalize_workflow(lasso_wf, best_auc)

lasso_words <- final_lasso %>%
  fit(review_train) %>%
  pull_workflow_fit() %>%
  vi(lambda = best_auc$penalty) %>%
```

```
  filter(!str_remove(Variable, "tfidf_text_") %in% c(stop_words$word,
                                                     'highly',
                                                     'absolutly',
                                                     'beautiful',
                                                     'beautifully',
                                                     'wonderful',
                                                     'love',
                                                     'perfect',
                                                     'hosts',
                                                     'absolutly'
                                                     )) %>%
  group_by(Sign) %>%
  top_n(20, wt = abs(Importance)) %>%
  ungroup() %>%
  mutate(
    Importance = abs(Importance),
    Variable = str_remove(Variable, "tfidf_text_"),
    Variable = fct_reorder(Variable, Importance)
  )

lasso_words %>%
  ggplot(aes(Importance, Variable, fill = Sign)) +
  geom_vline(xintercept = 0) +
  geom_col(color = "black", show.legend = F) +
  facet_wrap(~ Sign, scales = 'free_y') +
  labs(x = NULL)

library(igraph)
library(ggraph)
set.seed(2021)

f_colorise <- function(x) {
  ifelse(x %in% filter(lasso_words, Sign == 'POS')$Variable, 'positive',
         ifelse(x %in% filter(lasso_words, Sign == 'NEG')$Variable, 'negative', 'neutral')
  )
}

merge(dat_comments, dat_rbnb) %>%
  filter(language == 'english' & !is.na(assesment)) %>%
  tibble() %>%
  select(assesment, text) %>%
  mutate(
    rating = as.numeric(Hmisc::cut2(assesment, g = 10, levels.mean = T)),
    rating = case_when(
      rating == min(rating) ~ 'Neg',
      rating == max(rating) ~ 'Pos',
      T ~ 'Middle'
    ),
  ) %>%
  select(rating, text) %>%
  filter(rating != 'Middle') %>%
  select(text) %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
```

```
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(word1 %in% lasso_words$Variable | word2 %in% lasso_words$Variable) %>%
  count(word1, word2, sort = TRUE) %>%
  head(100) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = grid::arrow(type = "closed", length = unit(.15, "inches")),
                 end_cap = circle(.07, 'inches')) +
  geom_node_point(aes(color = f_colorise(name)), size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1, size = 3) +
  theme_void() +
  labs(color = 'Role in the model')
```