# Appendix A

## Analysis of the first received ratings

Figure A1 shows the mean values of the first 6 received ratings by sequence. On the Airbnb platform, an aggregated score is visible after a host has received three ratings.

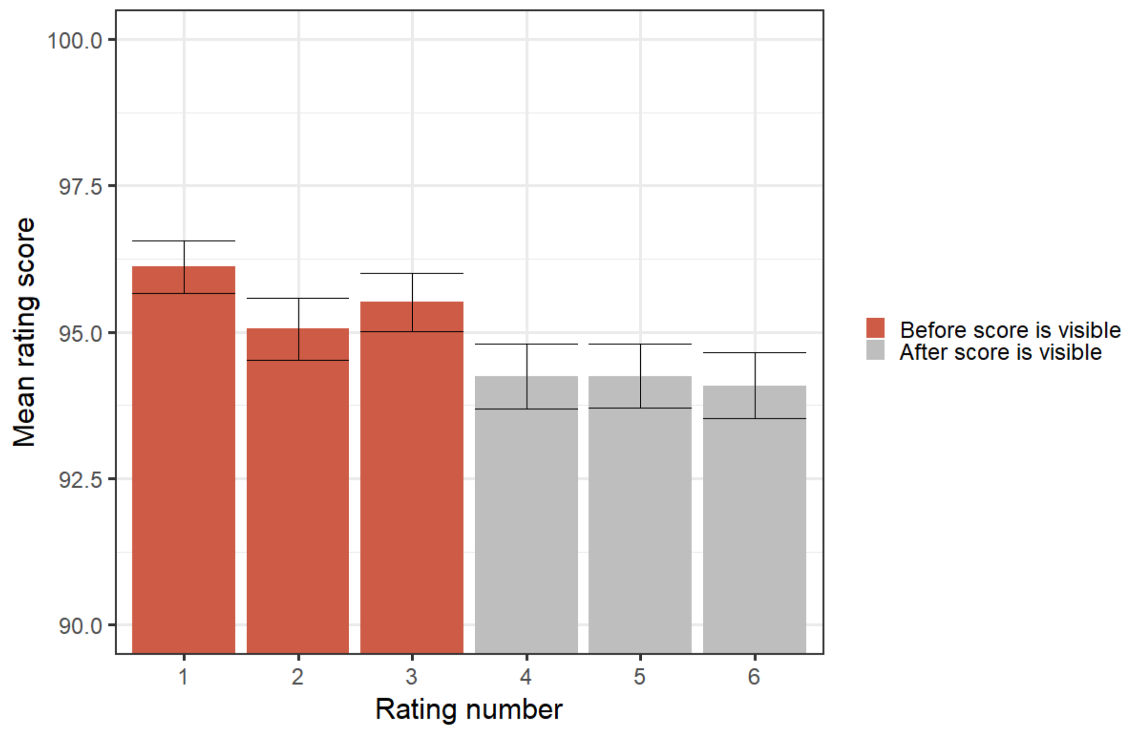


**Figure A1: Average received rating score by sequence (error bars indicate standard errors)**

There occurs a substantial difference in received rating scores once the aggregated score is visible. In particular, the difference between the 3rd and 4th review is even larger than the difference between first and second review. This yields a potential proportion of fake reviews of around 24% in the third review. In light of this, our approach to identify fake reviews in the first rating seems rather conservative, yet already suggesting a considerable proportion of fake reviews.

# Appendix B

### Analysis of guest attributes on five-star ratings

In Table B1, we consider the effect of reviewer location on the likelihood of a listing receiving a five-star rating using logistical regressions. Specifically, we investigate the role of reviewers that indicate to be located in the same city as the listing (i.e., the destination of their trip). We observe a significant effect of reviewer location (same city vs. different city) on the likelihood of five-star ratings overall (Model 1). Additionally, we see that five-star ratings are overall more likely to occur in first reviews than in second reviews (Model 2), these relationships also remain significant when analyzed jointly (Model 3). Importantly, we find a significant positive interaction term between the “same city” and “first review” variables (Model 4), showing that guests from the same city are more likely to leave a five-star rating in the first review. These results point to the possibility of hosts recruiting friends and family (who will likely live close by) to book their listing and provide an initial first five-star review.

**Table B1: Logit regression results – reviews from the same city**

| **Independent Variable** | **Dependent variable: Received 5-star rating (0=no, 1=yes)** | | | |
|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | **Model 4** |
| Reviewer from same city (no=0, yes=1) | 0.114*** (0.027) | | 0.112*** (0.027) | **0.052 (0.038)** |
| Is first review (no=0, yes=1) | | 0.123*** (0.015) | 0.106*** (0.016) | **0.093*** (0.017)** |
| SameCity × FirstReview | | | | **0.121* (0.054)** |
| **# Observations** | 99,269 | 116,536 | 99,269 | 99,269 |
| Standard errors in parentheses,****p*<0.001; ***p*<0.01; **p*<0.05 | | | | |

Conducting an analogous analysis for reviewers who are also hosts on the platform, we see an interaction as well (significant on the 10% level), pointing towards a higher proportion of five-star ratings left by other hosts in the first review than in the second (Table B2).

**Table B2: Logit regression results – reviews from other hosts**

| **Independent Variable** | **Dependent variable: Received 5-star rating (0=no, 1=yes)** | | | |
|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | **Model 4** |
| Reviewer is host (no=0, yes=1) | 0.240*** (0.058) | | 0.235*** (0.058) | **0.138 (0.081)** |
| Is first review (no=0, yes=1) | | 0.123*** (0.015) | 0.123*** (0.015) | **0.119*** (0.015)** |
| SameCity × IsHost | | | | **0.194 (0.116)** |
| **# Observations** | 116,536 | 116,536 | 116,536 | 116,536 |
| Standard errors in parentheses,****p*<0.001; ***p*<0.01; **p*<0.05 | | | | |

Note that in our initial analysis, we also see substantially more hosts in the first review overall than in the second (1,270 vs. 1,089). As the host attribute is known across our entire dataset, we can also see that hosts are significantly more likely to book listings without reviews (i.e., leave the first review; see Figure B1).

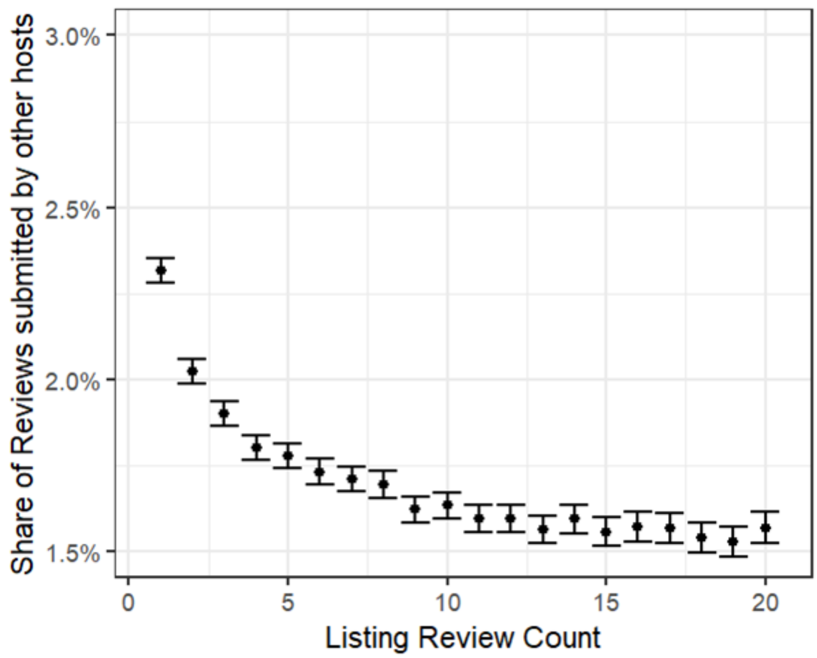


**Figure B1: Proportion of reviews from other hosts (error bars indicate 99% confidence intervals)**

# Appendix C

## Analysis of Yelp and Google reviews

In an additional analysis of **Yelp reviews**, we use the publicly available dataset provided by Yelp.com [17]. The dataset contains 150,346 businesses with around 7 million reviews from 14 US states. For this comparison, we consider the first and second rating score distributions. As the data contains a wide range of businesses, we first filter the dataset for hotels. This yields a total of 5,858 businesses. Second, filtering for states with at least 300 businesses (for a sufficiently sized sample of businesses per state) yields 4,926 businesses from 8 US states (retaining 84% of the businesses). Applying the KL-divergence approach outlined in this paper yields a rather low overall share of potential fake reviews of around 2.7% with high regional variations of up to 15.2% (see Figure C1). Please note that there are substantial rating differences between Yelp' (open) and Airbnb's (closed) reputation systems. For instance, the mean score of Yelp reviews across all domains is 3.82 stars for the first, and 3.74 stars for the second rating (on the 1 – 5 stars scale). For hotels, these values are 3.51 and 3.43 stars, respectively. For reference, on Airbnb, these values are 4.73 and 4.69 stars.

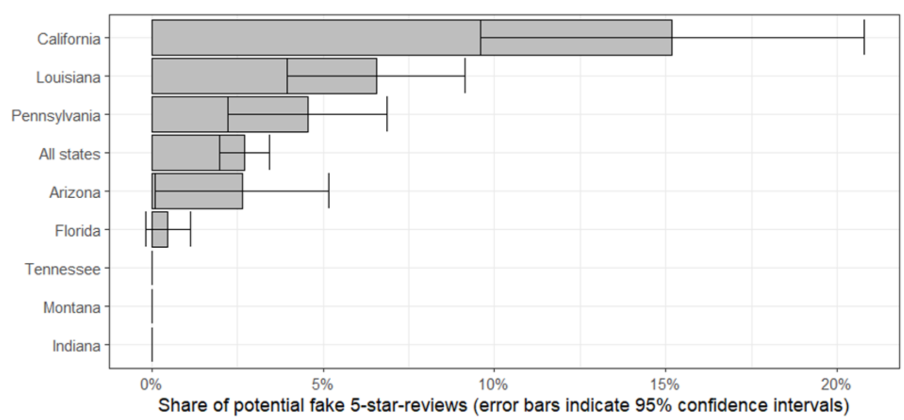


**Figure C1: Share of potential Yelp fake reviews by state**

For our analysis of **Google reviews**, we use data published by the University of San Diego, comprising more than 70 million Google Maps reviews from the state of California belonging to more than 500,000 businesses [8,18]. Again, we see significant differences to the Airbnb reputation system with a mean score of 4.19 stars for the first and 4.20 stars for the second rating (and more 5-star ratings in the first review than in the second). Filtering for hotels yields a sample of n=5,463 and new mean rating scores of 3.79 for the first and 3.82 stars for the second rating. Given that there is no positive difference between the first and second average rating score, this does not suggest any fake reviews on this dataset.

The data and results illustrate the fundamental differences between open and closed reputation systems. Aside from different dynamics of reputation on Yelp and Google, we also observe more negative feedback on both platforms as well as a higher degree of dispersion of rating scores – for example, the standard deviation of the mean score on Yelp is markedly higher than on Airbnb (1.43 and 1.52 vs. 0.68 and 0.73 for the first and second rating, respectively).

# Appendix D

## Listing attributes and share of fake reviews

Table D1 provides an overview of different listing attributes and the respective shares of estimated fake reviews. We see that fake reviews are particularly likely in listings that are priced above the median, as well as listings that are hosted by “superhosts” (i.e., a seal awarded by Airbnb to distinguish hosts “who [go] above and beyond to provide excellent hospitality”). Additionally, we see a difference of potential fake reviews between listing types with more fake reviews when entire homes are offered than private rooms.

**Table D1: Potential shares of fake review for different listing and guest attributes**

| Listing Attribute | | Share of potential fake reviews |
|---|---|---|
| Price | Above median | 15.2% |
| | Below median | 8.5% |
| Hosted by “superhost” | Yes | 20.6% |
| | No | 10.2% |
| Listing type[1)] | Entire home | 12.4% |
| | Private room | 8.4% |

1) For the types ‘shared room’ and ‘hotel room’, we identified less than 300 listings and consider the sample size too small to calculate the share of fake reviews

# Appendix E

## Comparison of datasets with all available listing data

To identify possible biases in our dataset resulting from the necessity to rely on listings for which we have observations both when they have one review *and* two reviews (later on), we compare our dataset to all available listing data. To assess the degree of potential bias, we first need to ensure comparability. Most importantly, our dataset of listings represents listings at their early stages on the platform**.** As research has shown that listings with more reviews tend to increase prices over time, we compare our data with all available data we have on early-stage listings (for more research on the dynamics of prices on Airbnb, we refer to Ikkala and Lampinen [6] and Teubner et al. [15]). Thus, we compare the selected listings to the overall set of early-stage listings on the platform (i.e., the data available when the listing has exactly one review). Note that further data (such as price etc.) is not available for all listings.

For our user dataset (especially the geographic analysis), we further filter our dataset for listings for which the respective user profiles were known, and whose profiles contained valid location data. We apply the same comparison to this subset of our listing dataset as well.

Table E1 provides a comparison of a variety of listing features between our final dataset and all available (early-stage) listings.

| **Table E1: comparison – all available listings vs. final dataset** | | | |
|---|---|---|---|
| **Attribute** | **All listings** | **Listing dataset** | **User dataset** |
| Number of listings | *256,423* | *42,420* | *21,443* |
| Mean price (USD) | 186.95 | 186.62 | 178.73 |
| Median price (USD) | 115 | 118 | 115 |
| Mean size (guests) | 3.75 | 3.88 | 3.94 |
| Median number of host's listings | 2 | 2 | 2 |
| Share of superhosts (%) | 13.4% | 18.6% | 18.9% |
| Share of apartment types<br>• Shared room<br>• Private room<br>• Entire home<br>• Hotel room | <br>1.2%<br>26.6%<br>71.2%<br>1.0% | <br>0.8%<br>20.5%<br>78.0%<br>0.7% | <br>0.5%<br>17.7%<br>80.9%<br>0.8% |

We observe only marginal differences between our dataset and all available listings in terms of price, size, as well as the number of listings a host has on the platform. Our final dataset contains a slightly higher proportion of superhosts (however, given that the superhost-status requires the completion more than 10 stays, the respective hosts must have gained the status with a different listing; [1]). Additionally, our final dataset contains a higher share of listings of the type "entire home" and less "private rooms" than the entirety of available listings. The same holds true for our dataset limited to users with valid location data. Overall, we conclude that the subset of listings is sufficiently similar to the overall dataset.

# Appendix F

## Comparison of guest analysis results

Rerunning the guest analysis on the subset of 5-star ratings in the first and second reviews, respectively, we attain similar results as with the overall dataset. Please note that focusing on five-star ratings yields two datasets of different lengths (as we have more five-star ratings in the first than in the second review). Table F1 shows the results of the analysis. Please note that this analysis does not lead to any substantial changes in the significance of results compared to the analysis of the overall dataset.

**Table F1 Overview of differences between first and second reviews/ reviewers**

| Feature | 1st Review | | 2nd Review | |
|---|---|---|---|---|
| | Full dataset | Five-star ratings | Full dataset | Five-star ratings |
| Share of deleted profiles | 1.4% | 1.4% | 1.0% | 1.0% |
| Share of users with no ID verification | 12.3% | 12.0% | 11.2% | 11.0% |
| Share of users with no validated phone number | 0.09% | 0.09% | 0.04% | 0.04% |
| Share of users with no validated email address | 15.5% | 15.6% | 16.5% | 16.6% |
| Share of users with 1 or 0 reviews received | 19.6% | 19.2% | 20.0% | 19.5% |
| Mean number of reviews | 12.2 | 12.7 | 11.1 | 11.4 |
| Share of guests from close vicinity (<50km) | 17.5% | 17.8% | 16.4% | 16.7% |
| Share of guests from the same location | 12.1% | 12.3% | 11.2% | 11.4% |
| Share of guests who are also hosts | 2.2% | 2.3% | 1.8% | 1.9% |

# Appendix G

## Analysis of time between reviews

Assessing the impact of the time between reviews on the likelihood of a listing receiving a five-star rating in the second review, we do not find a significant association. Figure G1 summarizes the relationship between those variables. Notably, we see that listings with a five-star rating in the first review take an average approximately 25 days less to receive a second review than listings without a five-star rating. Additionally, we see that listings that received a five-star rating in the first review are significantly more likely to receive a five-star rating in the second review as well.

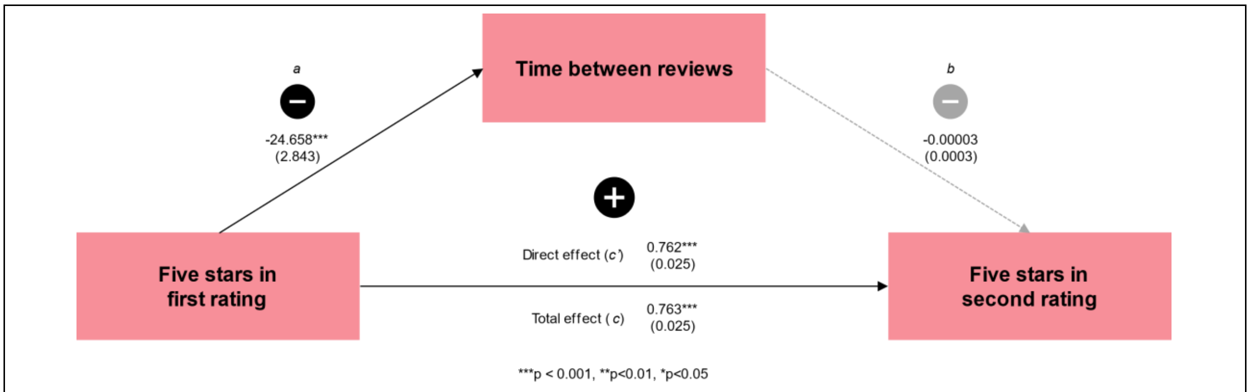


**Figure G1: Overview – regressions on five-star ratings and time between reviews with coefficients (standard errors in parentheses)**

# Appendix H

## Short survey among Airbnb hosts

In order to further substantiate our analysis, we conducted an explorative ad-hoc survey among 21 Airbnb hosts by reaching out on in the Airbnb community center forum [2]. Additionally, the survey was advertised on Facebook and Reddit within groups that specifically cater to Airbnb hosts. Table H1 lists the questions asked in the survey.

**Key results:**

- 14.3% of hosts admitted to having used fake reviews themselves to boost their listings (e.g., asked friends and family to book their listing without actually staying there).
- 23.8% of hosts stated to be aware of fellow hosts having procured fake reviews.
- 42.9% of hosts stated that if – hypothetically – they were to procure fake reviews, they would do (or would have done) so for their first review.

The survey serves as anecdotal evidence to supplement our analysis of early-stage fake reviews. While the results complement our finding of a prominent role of early-stage fake reviews on Airbnb, several limitations exist. First, the survey has a small sample size. Second, hosts might not answer honestly. This *social desirability bias* is a common phenomenon and constraint in research where respondents tend to answer questions in a manner they believe to be viewed favorably by others. This can lead to over-reporting of socially desirable behaviors or attitudes and under-reporting of undesirable ones, thereby compromising the accuracy and reliability of the data collected [7]. In the context of our research, the bias would likely lead to an under-reporting of any involvement with fake reviews. The 14.3% share of hosts who report to have procured fake reviews might thus be considered a conservative estimate.

**Table H1: List of questions in the explorative survey**

| Section | Question |
|---|---|
| Demographic and platform control questions | Are you a host on the Airbnb platform? |
| | How old are you? |
| | What is your gender? |
| | In which country do you offer your room(s)/house(s)? |
| | In which city do you offer your room(s)/house(s)? |
| | How many listings do you have? |
| | Do you use other platforms than Airbnb to rent your space? |
| | Which other platforms do you use? |
| | Do you offer private or shared space? |
| | How long have you been a host on Airbnb? |
| | Is your Airbnb listing your main source of income? |
| Familiarity with fake reviews | Have you ever heard about fake reviews on the internet? |
| | Have you ever heard about untruthful reviews being present on Airbnb? |
| | Have you ever heard about the possibility to buy reviews for Airbnb listings? |
| | Have you ever heard about fake reviews being published on Airbnb? |

| | |
|---|---|
| | Are you aware of any Airbnb hosts that have either bought reviews or asked their friends/family to write them? |
| Own involvement in fake reviews | Have you ever bought a review? |
| | What was the reason for buying a review? |
| | How many times have you bought a review? |
| | Have you ever asked someone you personally know to write a review for your listing without them actually staying there? |
| | What was the reason for asking someone to write you an untruthful review? |
| | How many times have you asked someone to write a review for you? |
| Hypotheticals – fake reviews | If you were to buy a review or ask someone to write it, when would you do it? |
| | Could you explain why you chose the specific time frame in the question 1.? |
| | Which rating would you choose for yourself if you were to buy a review or ask someone to write it for your listing? |
| | What do you think, how common are untruthful reviews on the Airbnb platform? |
| | What do you think, how many hosts buy reviews or ask someone to write them for their listings? |
| | Could you explain why you chose the specific answer in the previous question? |
| | What is your opinion on the topic of fake reviews on Airbnb? |

# Appendix I

## Supervised classifiers based on training data from open reputation systems

To test the suitability of supervised classifiers for the detection of fake reviews in our dataset, we trained three different fake review classifiers with available training data from an open reputation system. Specifically, we used the dataset curated and provided by Ott et al. [11] consisting of authentic and fake hotel reviews (from TripAdvisor, Expedia, and more – all open review systems). In total, the dataset contains 1,600 reviews:

- 400 authentic positive reviews,
- 400 authentic negative reviews,
- 400 fake positive reviews, and
- 400 fake negative reviews.

The dataset has been used in numerous studies [4,10,12]. Note that negative fake reviews represent an issue more or less unique to open reputation systems and cannot reasonably assumed to be found in closed reputation systems. We thus deliberately train our classifier without the negative fake reviews, focusing only on authentic reviews and positive fake reviews (n=1,200).

We preprocessed the dataset using a variety of steps proposed by Alsubari [3], including the removal of non-English and too-short reviews, punctuation, stopwords, capitalization, and whitespaces. Additionally, the dataset was lemmatized to reduce word variants. Next, we extracted features using uni-, bi-, and trigrams, which is a common feature set used in fake review detection [3,5,13].

We trained three different commonly-used supervised algorithms: Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and Random Forest (RF). For more details on the respective approaches and their use in fake review detection, we kindly refer to Li et al. [9] and Singhal and Kashef [14].

Training the different classifiers using n-grams yields the following performance metrics (using a random 80/20 split of our dataset for training and testing) summarized in Table I1.

**Table I1: Performance metrics for supervised classifiers (fake review share: 33.3%)**

| Classifier / Performance | Accuracy | Precision | Recall | F1- Score | ROC AUC |
|---|---|---|---|---|---|
| *XGBoost* | 0.87 | 0.89 | 0.92 | 0.91 | 0.82 |
| *SVM* | 0.85 | 0.85 | 0.96 | 0.90 | 0.89 |
| *RF* | 0.82 | 0.85 | 0.91 | 0.88 | 0.76 |

Note that the accuracy of our approach varies by the specific training / test split and reaches values of >0.90. Overall, we find values for accuracy, precision, and recall that are well in line with what is reported by other studies in the literature. For a comprehensive overview of literature on fake review detection, we kindly refer to Wu et al. [16]. There, numerous studies on supervised fake review detection report accuracies of 0.80 to 0.95.

To apply our classier to our data (i.e., the Airbnb reviews), we preprocessed these reviews along the same steps as we did with the training/test data. Table I2 summarizes the reported shares of fake reviews from our data.

| Table I2: Identified shares of fake reviews on Airbnb data by different classifiers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Fake review Share** | **XGboost** | | **SVM** | | **RF** | | ***Number of reviews*** | |
| | First rating | Second rating | First rating | Second rating | First rating | Second rating | *First rating* | *Second rating* |
| **All reviews** | 21.1% | 19.2% | 6.4% | 6.2% | 11.4% | 10.7% | *20,557* | *20,554* |
| **only 5-star reviews** | 24.1% | 22.8% | 7.1% | 7.3% | 13.0% | 12.6% | *17,167* | *16,567* |

Note that even though the classifiers have very similar metrics on the training and test data, we see substantial differences on the Airbnb dataset: The share of identified fake reviews varies substantially – critically hinging on the classifier used. Two out of three classifiers report slightly higher shares of fake reviews in the first rating distribution (as compared to the second rating distribution).

These results underscore the problem of transferring supervised machine-learning-based approaches from one domain of interest to another. While hotel reviews might be similar to Airbnb reviews, we see that the remaining difference renders a reliable identification of fake reviews infeasible. We, therefore, argue that our approach of leveraging score distribution provides a valuable addition, if not a more resilient approach, to estimating the share of fake reviews in the absence of labeled training data.

# References


1. Airbnb. About Superhosts. 2024. https://www.airbnb.com/help/article/828?_set_bev_on_new_domain=1704733314_Yzg2MjVhZmZjMGFm&locale=en.
2. Airbnb. Community Center - Airbnb Community. 2024. https://community.withairbnb.com/t5/Community-Center/ct-p/community-center.
3. Alsubari, S.N. Data Analytics for the Identification of Fake Reviews Using Supervised Learning. *Computers, Materials & Continua*, *70*, 2 (2022), 3189–3204.
4. Dewang, R.K. and Singh, A.K. Identification of Fake Reviews Using New Set of Lexical and Syntactic Features. In *CTMCDC 2015 Proceedings*. pp. 115–119.
5. Elmogy, A.M., Tariq, U., Mohammed, A., and Ibrahim, A. Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications*, *12*, 1 (2021).
6. Ikkala, T. and Lampinen, A. Monetizing Network Hospitality. In *CSCW 2015 Proceedings*. 2015, pp. 1033–1044.
7. Larson, R.B. Controlling social desirability bias. *International Journal of Market Research*, *61*, 5 (2019), 534–547.
8. Li, J., Shang, J., and McAuley, J. UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. In *ACL Proceedings 2022*. 2022, pp. 6159–6169.
9. Li, S., Zhong, G., Jin, Y., Wu, X., Zhu, P., and Wang, Z. A Deceptive Reviews Detection Method Based on Multidimensional Feature Construction and Ensemble Feature Selection. *IEEE Transactions on Computational Social Systems*, *10*, 1 (2023), 153–165.
10. Lu, J., Zhan, X., Liu, G., Zhan, X., and Deng, X. BSTC: A Fake Review Detection Model Based on a Pre-Trained Language Model and Convolutional Neural Network. *Electronics*, *12*, 10 (2023), 2165.
11. Ott, M., Choi, Y., Cardie, C., and Hancock, J.T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *ACL 2011: HLT Proceedings*. 2011, pp. 309–319.
12. Shunxiang, Z., Aoqiang, Z., Guangli, Z., Zhongliang, W., and KuanChing, L. Building Fake Review Detection Model Based on Sentiment Intensity and PU Learning. *IEEE transactions on neural networks and learning systems*, *34*, 10 (2023), 6926–6939.
13. Siagian, A.H.A.M. and Aritsugi, M. Robustness of Word and Character N-gram Combinations in Detecting Deceptive and Truthful Opinions. *Journal of Data and Information Quality*, *12*, 1 (2020), 1–24.
14. Singhal, R. and Kashef, R. A Weighted Stacking Ensemble Model With Sampling for Fake Reviews Detection. *IEEE Transactions on Computational Social Systems*, *11*, 2 (2024), 2578–2594.
15. Teubner, T., Hawlitschek, F., and Dann, D. Price Determinants on Airbnb: How Reputation pays off in the Sharing Economy. *Journal of Self-Governance and Management Economics*, *5*, 4 (2017), 53.
16. Wu, Y., Ngai, E.W.T., Wu, P., and Wu, C. Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, *132*, (2020), 113280.
17. Yelp. Yelp dataset. 2023. https://www.yelp.com/dataset.

18. Zhankui He, A.Y., Li, J., Zhang, T., and McAuley, J. Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. In *SIGIR Proceedings 2023*. 2023, pp. 2251–2255.