Sentiment Analysis on Airbnb Reviews

Team 12 : Nikita Agrawal, Rajani Goudar, Urvi Midha, Christine Olds, Prathyusha Pateel, Raniya Syeda

1 INTRODUCTION

Vacationers these days rely on various reviews from previous vacationers, and weigh positive and negative aspects to make informed decisions on where to stay. Though sites like Airbnb have reviews, manually reading through every review can be time-consuming, even though reviews can be a better quality signal than star-based ratings. There is a need for a tool that can provide a concise summary of the general sentiment within the reviews, making it easier for guests to assess their overall quality and experience of a property. This tool would also be helpful for existing hosts to maintain the quality of their listing by incorporating customer feedback, and for potential hosts to determine whether a location has the potential to be lucrative.

Using public datasets acquired from "Inside Airbnb: Get the Data" [1], we plan to create an interactive dashboard visualization tool that accurately and concisely summarizes the overall sentiment for each Airbnb listing in various US cities to help the guests make a decision. The tool will use sentiment analysis to analyze each review and assign it a score between -1, 0, or 1 based on a negative, neutral, or positive sentiment. The scores will be averaged for each listing and plotted on a map by the listing's respective latitude and longitude with a corresponding color on a spectrum from red to white to green.

Sentiment analysis scores will be validated by comparing the output to the star-based ratings collected by Airbnb. There is no risk with the project or short-term payoffs. The long-term payoff is creating a tool to help people vacation better. Since the datasets are very large, we will use Google Colaboratory for computation and store our datasets in Google Drive. There will be no cost to this project.

2 LITERATURE SURVEY

Usually, online scaled ratings, say {1-5}, are one of the first methods for measuring opinions on vacation spots. However, several studies of these sites have found a consistent positivity bias, which results in a J-shaped distribution of online ratings (usually 4-5 stars). A smaller amount of ratings is negative, with the least amount being the 2-3 star category[5]. Although this paper is not useful for our project, as it explored similarities between reviews left by guests and hosts, it suggests that there is no substantial connection between reviews and star ratings. This implies that the two metrics may represent distinct quality concepts[18]. However, it wil be interesting to see if there is a positivity bias in our analysis of reviews.

Thus, we now move towards textual-based reviews for our evaluations. One such dataset is affiliated with Airbnb customer reviews. This paper[3] evaluates if this dataset is real or not, and it concludes that it is in fact legitimate, which helped us in picking this dataset. However, it does not analyze cross-city comparison metrics. Therefore, we then looked into multiple sentiment analysis techniques conducted on this data to analyze their different aspects. In [13] and [15], many existing models, recently proposed enhancements, and various applications are presented briefly in this survey. This is useful for our project, as we can conclude by selecting the best model based on document-level sentiment analysis strategies for our use case.

Going one step ahead, we also explored aspect-based sentiment analysis. In [6], text-mining was conducted on Sydney city reviews to unveil important themes, including “location”, “amenities”, and “host”. Their observations are useful to foresee basic aspect-based learning rules such as positive listings are in the vicinity of shopping malls, beaches, and other public areas, and bad listings are near busy streets and high-noise areas. Similarly, in [11] and [9] they employ topic modeling to classify the reviews into specific topics, thereby facilitating a more detailed comprehension of the dominant feature aspects that are of greatest importance to guests. [4] uses novel Latent Dirichlet Allocation (LDA) method for the above topic modeling.

For training our machine learning model, various existing algorithms have achieved notable accuracy, among which there are three predominant approaches.

(1) Firstly, the VADER algorithm[8] is popular for determining the associated polarity of a review using lexicon and rule-based approaches at a sentence level. [17] and [3] use the above algorithm on Stanford CoreNLP and NLTK tool sets respectively for POS tagging and dependency parsing.

(2) Secondly, TextBlob is a library also built on NLTK and is common for natural language processing.
(3) Thirdly, Bag of Words is another approach using TF-IDF Vectorization. Another tool called Senti Strength[16] is useful, which is a human-labeled lexical dictionary enhanced by ML on two sharing economy and one traditional economy platforms. In another paper[10], AFINN dictionary was used for sentiment matching. However, this was not useful to us as the focus was more on how reviews affect price and not the overall sentiment.
(4) Lastly, deep learning models are useful for exploring higher accuracy. [14] is one of the few papers that experimented with GRUs, RNNs, and LSTMs.

Nevertheless, a shortcoming of all the above papers is that they have not transformed their results into real-world and deployable interactive tools. To create such an interface, we would need an understanding of the end-to-end software engineering aspects for the final UI tool available to the general users. [7] and [19] are very good references for the same. To come up with meaningful features of our tool, we looked into general characteristics that vacationers look for in listings using Graphical and Visual inferences. Through [12] and [2] papers, we concluded our literature survey on a note that location-based analysis and super host identifications are the most in-demand services needed.

3 PROPOSED METHOD

Sentiment Analysis Algorithms

The code utilizes both VADER and TextBlob for sentiment analysis, to measure the sentiment score of reviews for each listing in the 34 cities available on "Inside Airbnb: Get the Data". VADER is chosen for this analysis due to its ability to handle informal language, including slang, colloquial expressions, and emoticons, commonly encountered in online reviews. Moreover, its rule-based approach enables it to capture the context and intensity of sentiment, providing a nuanced understanding of customer experiences. This makes it particularly suitable for analyzing Airbnb reviews, where guests often express their opinions and emotions in diverse linguistic styles. Additionally, TextBlob is also utilized for comparison. TextBlob is built on top of NLTK, a powerful library for natural language processing.

The code initiates by importing the required libraries and then scraping two CSV's per city. The first CSV, called listings, contains the ID, name, latitude, longitude, and price for each Airbnb listing in the respective city, as well as several other columns such as the number of bedrooms and percentage of year that the Airbnb is normally booked that we dropped. The latitude and longitude were crucial columns to keep for our visualization. The second CSV, called reviews, contains the listing ID and comments, as well as a few other columns such as host name that we dropped. The extracted data is then transformed into a Pandas DataFrame. Due to the computationally intensive nature of NLP algorithms, iterating through each review in every city proved highly inefficient. The process took several hours to complete for one of the largest review datasets from Los Angeles. By utilizing Pandas DataFrames, we were able to apply the algorithm to the "comments" column of each city, greatly decreasing the computational intensity and time required. Next, the sentiment analysis algorithm is applied to the reviews DataFrame for each city, creating a new column for each listing containing the sentiment analysis score calculated by the algorithm. The algorithm assigns a score between -1 and 1, with -1 indicating a negative sentiment, 0 a neutral sentiment, and 1 a positive sentiment. Reviews in different languages and Emojis were excluded from consideration. Since there are multiple reviews for each listing, the DataFrame is then grouped by the unique listing ID and subsequently merged with the listings DataFrame. While grouping, the sentiment analysis scores for each review are averaged to calculate the overall average sentiment score of the listing. Additionally, to compare the scores calculated by TextBlob and VADER, we also calculated the overall average of each city. These results are shown on a bar graph in Figure 5 and will be discussed in the Experiments/Evaluation section.

User Interface

We utilized Tableau to create an interactive dashboard visualization, shown in Figure 1. The data was in CSV format and contained over 200,000 Airbnb listings in 34 cities across 20 states. For each row, the CSV file contained columns populated with each listing's unique ID, name, latitude, longitude, price, average rating, state, city, and calculated sentiment score. Listings are shown on a map of the United States, plotted by Tableau using each Airbnb's respective latitude and longitude. As illustrated by the sentiment score color gradient, dots are assigned a color between red and green, with red

Figure 1: TextBlob Visualization

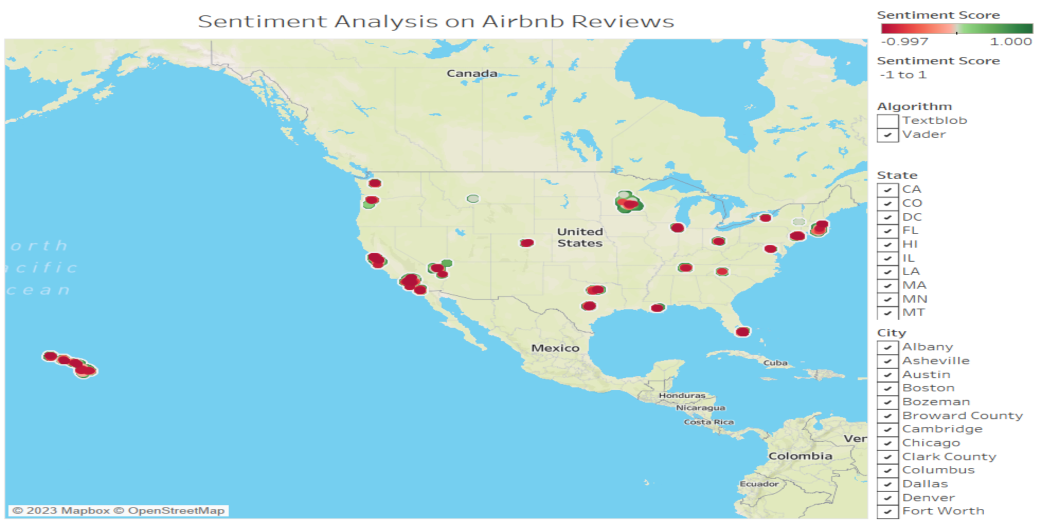

Figure 2: VADER Visualization

representing a negative score and green representing a positive score. Under the color gradient is an interactive slidebar that the user can adjust to filter out listings with too low, or maybe even too high, of a score. The Algorithm options allow a user to display the scores from TextBlob, Vader, or an average of the two. Finally, the city and state filters allow users to select which state and city they are looking to plan their next vacation. Hovering over a dot on the map will display the listing's name, nightly price, sentiment score, city, and state.

Scalability

In an effort to make this an end-to-end & real time product, we attempted to devise our entire pipeline utilizing Amazon Web Services, AWS. We created an account under their free tier and wrote a scraping script on AWS Lambda to extract the data. The Lambda function was scheduled to run every three months to ensure regular updates. Collected data was subsequently stored in organized S3 buckets. This s3 event triggered a Sagemaker script containing our sentiment analysis algorithms where the averaged sentiment score was

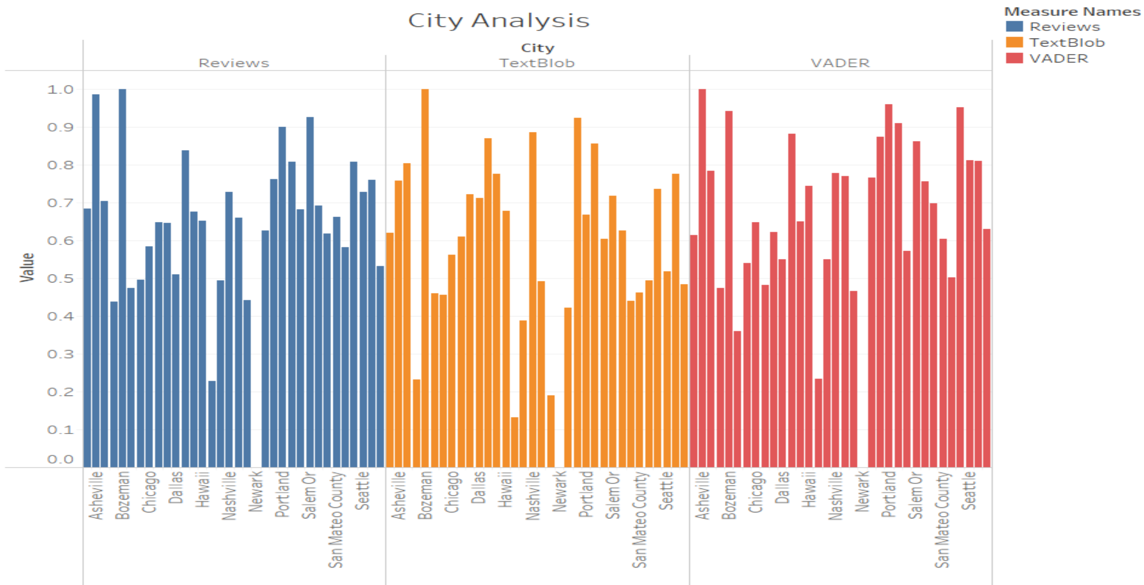

Figure 3: Algorithm Analysis

calculated and this updated data was written back to the S3 bucket. Finally, this updated data was uploaded to Tableau through S3 data source server, where the UI interface displayed the map visualization detailed above. An example flowchart is given in Figure 6, making the visualisation real time. We were able to make all the components run but we could not find any sentiment analysis libraries which had very fast inference time, so the entire pipeline became very slow. Moreover, due to the large amounts of data and required computational power, we exceeded the limits of the free tier after just one city (Los Angeles) by $214.09. We also realized that we had used a lot of pricing for our testing codes which resulted in us exceeding the price limit. As students with limited resources, we had to explore cost-effective alternatives to handle the data. For the development and execution of our code, we leveraged Google Colaboratory (Colab), a cloud-based Jupyter notebook environment that offers free access to GPU resources, integration with Google Drive, and real-time collaboration. Colab proved to be an invaluable resource as it was affordable and allowed us to collaborate on code simultaneously. The VM used for Colab had the following specifications: 13GB RAM, Intel Xeon Processor 2.20 GHz, 2 vCPU and 33GB HDD.

4 EXPERIMENTS/EVALUATION

Sentiment Analysis Algorithms

City analysis of the algorithms is illustrated in Figure 3, where it is evident that VADER and TextBlob produced a similar distribution of sentiment scores. The left-most bar graph, featuring blue lines, depicts the average five-star rating for each city. This information was included in the data extracted from each city's listings dataset. The second bar graph, displaying yellow lines, illustrates the average scores calculated by the TextBlob algorithm. Lastly, the third bar graph, indicated with red lines, showcases the average scores calculated by VADER for each city. Given that the star ratings initially ranged from 0 to 5, and the sentiment scores were on a scale from -1 to 1, the output was scaled to allow for a side-by-side comparison. Upon manually reviewing the algorithm-assigned scores, we observed that VADER tended to assign more extreme scores, while TextBlob tended to score more conservatively around 0. For example, in analyzing the following comment "I had a pleasant stay here. The bed was indeed very comfortable I wish I could have one like that at home. Efrat and her husband went above and beyond my expectations. Very grateful," VADER assigned a score of 0.9399 while TextBlob assigned a score of 0.3633. In cases of neutral statements, such as, "The reservation was canceled days before arrival. This is an automated

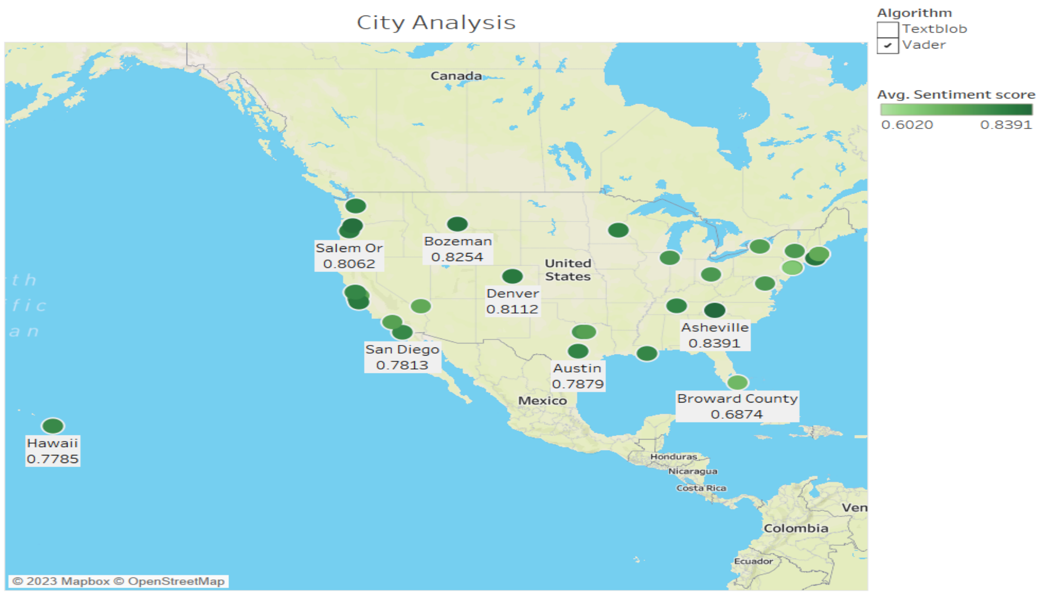

Figure 4: Algorithm Analysis

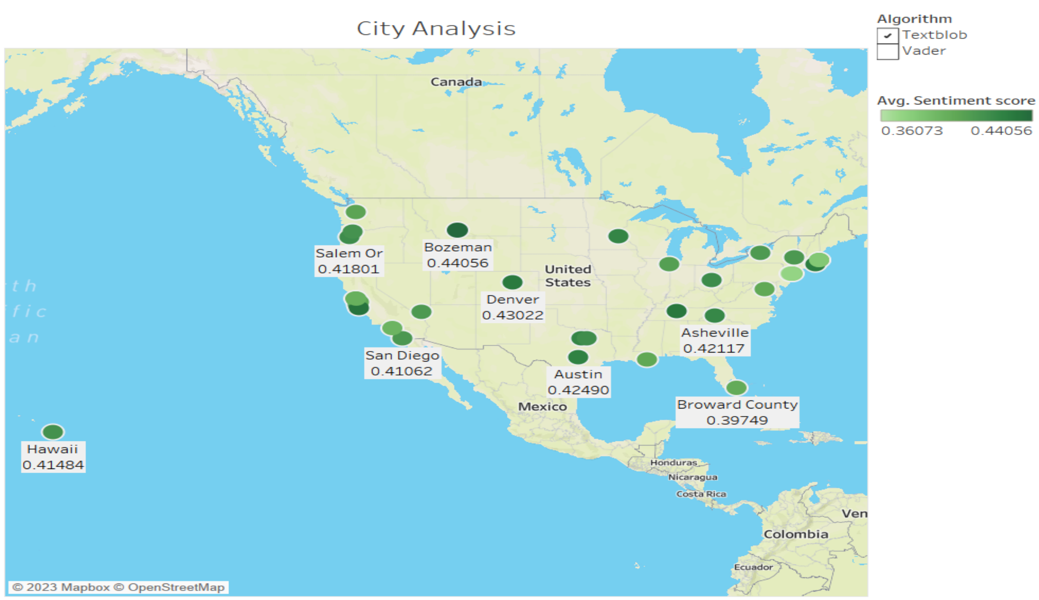

Figure 5: Algorithm Analysis

posting," both VADER and TextBlob assigned a neutral score of 0. For a negative review such as, "Worst Airbnb we have set foot in We did not stay Airbnb support was great . . . We could see how dirty the carpets were. Hard surfaces seemed sticky under the shoes," VADER assigned a score of -0.9198 while TextBlob assigned a score of -0.1299. This example is particularly noteworthy because the user mentions that Airbnb support was "great," yet VADER correctly classifies it as an extremely negative comment, whereas TextBlob classifies it as only mildly negative. In the analysis of Airbnb reviews across the United States, see Figures 4 and 5, for

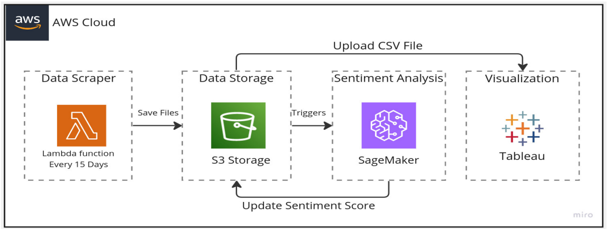

Figure 6: Attempted AWS Cloud Pipeline

Figure 7: Current Colab Pipeline

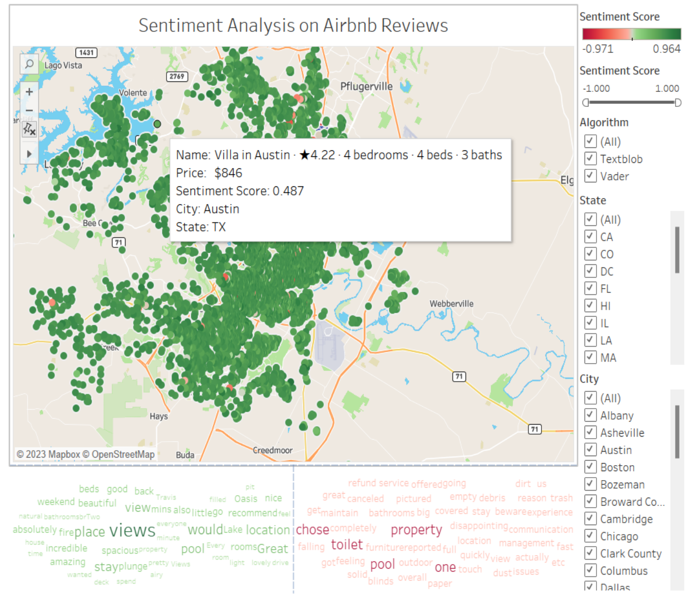

Figure 8: Word Cloud

VADER and TextBlob outputs, respectively. The number displayed here is the average of the scores for each city. All averages are positive and displayed in green, aligning with research from our literature survey in which we found there is often a positivity bias when conducting sentiment analysis. These figures illustrate the same information as Figure 3 in a different format, allowing the user to analyze how sentiment scores vary from city to city. The trends illustrated here also align with our observations that VADER tends to assign more extreme scores while TextBlob tends to assign more conservatively.

Word Cloud Generation

We created two word-clouds for each Airbnb listing, showcasing the top 50 frequently used words in user reviews. We segregated the reviews into positive and negative comments based on their sentiment score, to give an understanding of the corresponding factors impacting the overall listing score. We implemented a dynamic hover feature, allowing users to explore both positive and negative word clouds when interacting with a specific listing. This approach aimed to provide immediate insights into the sentiments associated with each property. However, we encountered a challenge due to the substantial size of the dataset and the resource-intensive nature of sentiment analysis and word cloud generation. This complexity led to performance issues within Tableau, compromising the intended user experience. In response, we opted to exclude this specific feature from the final dashboard to prioritize overall dashboard performance.

5 CONCLUSIONS & DISCUSSIONS

With all the information available to plan a vacation, it can be difficult, even impossible, to sift through all the available reviews and ratings to make an informed decision. Our tool provides a practical implementation of natural language processing algorithms to facilitate the vacation-planning process. Using two different sentiment analysis algorithms, the interactive dashboard allows users to filter the algorithm and compare two different outputs. Additionally, the dashboard allows filtering options for state and city to allow maximum personalization of the Airbnb search. This tool can also be helpful for hosts to better understand the overall sentiment around their listing and make data-driven decisions to enhance overall customer satisfaction and service quality. Overall, the dashboard is unique & interactive, successfully facilitates the vacation-planning process. It is published and available on Tableau Public for use.

All team members contributed equally.

REFERENCES

[1] AirBnB. 2023. Inside AirBnB, Get the Data. http://insideairbnb.com/get-the-data/

[2] Aljoharah Almjawel, Sahar Bayoumi, Dalal Alshehri, Soroor Alzahrani, and Munirah Alotaibi. 2019. Sentiment Analysis and Visualization of Amazon Books' Reviews. In *2019 2nd International Conference on Computer Applications Information Security (ICCAIS)*. 1–6. https://doi.org/10.1109/CAIS.2019.8769589

[3] Abdulkareem Alsudais and Timm Teubner. 2019. Large-Scale Sentiment Analysis on Airbnb Reviews from 15 Cities. In *Twenty-fifth Americas Conference on Information Systems at Cancun*.

[4] H.M.A.P. Bandara, J. Charles, and L. S. Lekamge. 2022. Using Sentiment Analysis to Explore the Accommodation Experience in the Sharing Economy through Topic Modeling. In *2022 4th International Conference on Advancements in Computing (ICAC)*. 156–161. https://doi.org/10.1109/ICAC57685.2022.10025122

[5] Judith Bridges and Camilla Vásquez. 2018. If nearly all Airbnb reviews are positive, does that make them meaningless? *Current Issues in Tourism* 21, 18 (2018), 2057–2075. https://doi.org/10.1080/13683500.2016.1267113 arXiv:https://doi.org/10.1080/13683500.2016.1267113

[6] Mingming Cheng and Xin Jin. 2019. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management* 76 (2019), 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004

[7] Marvin Hagge, Moritz von Hoffen, Jan H. Betzing, and Jörg Becker. 2017. Design and Implementation of a Toolkit for the Aspect-Based Sentiment Analysis of Tweets. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, Vol. 01. 379–387. https://doi.org/10.1109/CBI.2017.68

[8] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[9] George Joseph and Vinu Varghese. 2019. *Analyzing Airbnb Customer Experience Feedback Using Text Mining*. Springer Singapore, Singapore, 147–162. https://doi.org/10.1007/978-981-13-6339-9_10

[10] Abdelaziz Lawani, Michael R. Reed, Tyler Mark, and Yuqing Zheng. 2019. Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. *Regional Science and Urban Economics* 75 (2019), 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003

[11] Yi Luo and Rebecca (Liang) Tang. 2019. Understanding hidden dimensions in textual reviews on Airbnb: An application of modified latent aspect rating analysis (LARA). *International Journal of Hospitality Management* 80 (2019), 144–154. https://doi.org/10.1016/j.ijhm.2019.02.008

[12] Heyam Madhi and Muna Alhammah. 2021. What Drives Airbnb Customers' Satisfaction in Amsterdam? A Sentiment Analysis. *International Journal of Advanced Computer Science and Applications* 12 (01 2021). https://doi.org/10.14569/IJACSA.2021.0120628

[13] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[14] Muhammad Raheel Raza, Walayat Hussain, and Asaf Varol. 2022. Performance Analysis of Deep Approaches on Airbnb Sentiment Reviews. In *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*. 1–5. https://doi.org/10.1109/ISDFS55398.2022.9800816

[15] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. 2021. Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In *Machine Learning and Knowledge Extraction*, Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Cham, 173–184.

[16] Gustavo Santos, Vinicius F. S. Mota, Fabricio Benevenuto, and Thiago H. Silva. 2020. Neutrality May Matter: Sentiment Analysis in Reviews of Airbnb, Booking, and Couchsurfing in Brazil and USA. arXiv:2005.06591 [cs.SI]

[17] A. Vassilikopoulou, I. Kamenidou, and C.-V. Priporas. 2022. Negative Airbnb reviews: an aspect-based sentiment analysis approach. *EuroMed Journal of Business* ahead-of-print, ahead-of-print (2022). https://doi.org/10.1108/EMJB-03-2022-0052

[18] Giampaolo Viglia, Roberta Minazzi, and Dimitrios Buhalis. 2016. The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management* 28, 9 (2016), 2035 – 2051. https://doi.org/10.1108/IJCHM-05-2015-0238 Cited by: 217; All Open Access, Green Open Access.

[19] Jianghui Ying, D. Gracanin, and Chang-Tien Lu. 2004. Web visualization of geo-spatial data using SVG and VRML/X3D. In *Third International Conference on Image and Graphics (ICIG'04)*. 497–500. https://doi.org/10.1109/ICIG.2004.147