How the Pandemic Affected Airbnb and Hotel Consumer Sentiment

Introduction/Problem Definition

Since its creation in 2008, Airbnb has grown to become an alternative to traditional hotel networks, and in many cases, even the preferred way of travelling. However, the ongoing pandemic may have disrupted this growth trajectory, as factors such as social distancing, travel limitations, and hygienic prioritizations have had a heavy influence on the travel industry.

The COVID-19 pandemic has challenged the business community to provide consistent and quality service in the middle of a public health crisis. Consumer sentiment is critical in determining whether a business survives or fails during such a tumultuous time; using natural language processing techniques, we can analyze the reviews of Airbnb and hotels to determine the level of satisfaction and key topics within those reviews both pre- and post-pandemic. As far as we can tell, no study has been performed on consumer sentiment evolution throughout the pandemic as it pertains to the lodging sector of the hospitality industry. The novelty of this study should form the basis for success.

Literature Survey

Sentiment analysis and opinion mining can be traced back to the early 20th century when it was mainly focused on public opinion analysis. However, it has only begun to flourish in the 21st century as more subjective text has appeared on the web and social media has rapidly grown [1, 2]. Sentiment analysis can be used as a method to extract and infer someone's opinion or attitude toward a subject through the syntax and diction of their writings. There are several techniques and methodologies that can be exploited in this realm, each having their own advantages or issues [3, 4].

Three main categories of sentiment classification techniques are machine learning (ML) based: applying ML algorithms to the linguistic features, lexicon based relying on a predefined dictionary or a combination of both methods [5]. Lexicon is a dictionary of language or a specific knowledge. One of the few NLP-based sentiment analysis algorithms is Valence Aware Dictionary for sEntiment Reasoning (VADER), which blends a lexicon sentiment approach and grammatical based rule [6]. One such application assessed a variety of sentiment analysis tools and determined VADER to be the most accurate, on average [7].

Other researchers have recently applied sentiment analysis and natural language processing techniques to Airbnb and hotel data [8, 9]. In their research, Lawani et.al, used both sentiment analysis and a hedonic spatial autoregressive model, to show that customer reviews are one of the key factors playing a role in determining the rental price on Airbnb website [8]. In another study performed by Geetha et.al., the Naïve Bayes classification technique has been used to define polarity of each review based on the match between the reviews and a lexicon [9]. A textual analysis of Airbnb reviews near Sydney, Australia discovered that location, amenities, and the host interaction were more important factors than price in user experience [10]. Another analysis applied latent Dirichlet allocation (LDA) to over one million Airbnb reviews in New York City to distinguish 43 different subjects, followed by a hierarchical clustering algorithm [11]. Other researchers combine LDA with different statistical methods, such as stepwise regression, perceptual mapping and analysis of variance (ANOVA), to examine factors affecting consumer sentiment [12, 8, 7]. One study in India concluded a positive effect of customer sentiment polarity on hotel rating [9], while another compared TripAdvisor ratings to those of independent and chain hotels across the globe in order to glean consumer preferences [13].

The COVID-19 pandemic impacted many businesses who have faced closure, mass layoffs, and other financial issues [14]. The travel industry has especially been affected due to a decline in demand, causing a loss in spending, output, and jobs globally [15, 16]. Some authors already suggest that Airbnb will recover from its initial drop-in activity, but not to its pre-COVID level due to an implicit supply limit [17]. Applying natural language processing techniques to Airbnb and hotel data could give us further insight into consumer expectations during this crisis, highlight avenues for recovery and identify any differences between hotel and Airbnb sentiment.

Methodology

Data Collection

Airbnb collects robust data that details the hosts, locations, and reviews of their listings, which are publicly available. The independent organization, Inside Airbnb, has made reviews available for 28 cities or metropolitan areas in US. The data has been archived roughly on a monthly basis and is in the compressed CSV format. Hotel reviews have been scraped from Tripadvisor using Apify [18] and require processing. Tripadvisor data was collected in the json format and downloaded from the Apify website. An account was created by each team member to take advantage of the free trial and avoid financial costs. For both sets of data, the data collection process was split equally between group members. Data sets were downloaded, renamed via a standardized convention, then uploaded to a central repository. A Microsoft OneDrive cloud-storage repository was used for storage of all gathered data.

Data Cleaning & Integration

Neither datasets are flawless and flaws in the Airbnb data have also been identified by another researcher [19]. Pre-processing and data wrangling were performed on the data prior to analysis. For the Airbnb data, a Python script was written to filter specific columns including reviews, review date, listing neighborhood and zip code (when available) along with the price. This was followed by dropping missing/null data values, and de-duplicating records. The scraped TripAdvisor data included extraneous data out of which only price level, zip codes of the hotels, date stayed at hotel and review text was kept. The review titles and review text were concatenated to ensure accurate sentiment capture. Hotels with less than 20 reviews were not included. Additionally, multiple files were concatenated and merged to create master data files for each city. Finally, since online reviews can be submitted in any language, a filter was applied to retain only English text, using Python's “langdetect” module [20]. Since the analysis intends to compare consumer sentiment both pre- and post-COVID, records were filtered beginning January 1, 2018, then annotated based on

March 11, 2020, which is the date that the World Health Organization (WHO) declared the disease a global pandemic [21]. This division resulted in a balance of approximately 12% of the reviews Post-COVID. In total, over 10 million lodging reviews were cleaned and prepared for analysis.

Analytics

Textual analysis can measure various aspects of the lodging reviews. The analytical purpose is to determine both the overall sentiment (tone) of the review and extract specific topics that are discussed. Once the reviews are analyzed in this manner, aggregations, comparisons, and assessments can be made. Reviews for specific cities or lodging-type (hotel vs private owner) can be collectively compared. Do certain cities receive more positive reviews than others? Sentiment both pre- and post-COVID can be evaluated, along with topical analysis. Do health-related topics, such as "cleanliness" and "sanitation" become more prominent post-COVID? These examples are only several of the questions that we seek to answer with our analysis. To achieve these analytical goals, two different techniques are required. For sentiment analysis, Valence Aware Dictionary and sEntiment Reasoner (VADER) is applied. Developed in 2014, VADER compares the text to a rule-based lexicon dictionary to calculate a positive, negative, neutral, and a compound (normalized) polarity score [6]. The compound result is generally used to classify the entire text as positive, negative, or neutral, and scores range between -1 and +1. The strength of the score itself also indicates the intensity of the sentiment, so scores may be compared and ordered. VADER is preferred to other techniques because it includes acronyms, emoticons, slang, and punctuation common in online contexts, and it outperformed other baseline methods [22,7]. For the topic extraction, latent Dirichlet allocation (LDA) is applied. Developed in 2003, LDA employs a tri-level Bayesian mixture model to the reviews in order to generate a set of underlying topics [23]. Prior to executing the algorithm, LDA requires that the text is lemmatized, where stopwords are removed and word variations are standardized. One parameter set by the user is the number of topics to generate–this value may affect how much clarity and/or overlap occurs between extracted topics. LDA then scores the relevancy of particular words to that topic–so it is left for the researcher to define the topic, based on the words that comprise its distribution [24].

Experiments

Computational Resources

The experiment ran on a virtual machine of the Linux distribution Ubuntu 18.04.5. The machine has 8 CPU cores at its disposal, as well as 16GB of RAM. Between the two NLP kits, LDA and VADER, LDA is the more resource intensive. Fortunately, Gensim LDA [25] has the option of multicore processing. LDA uses what are called "worker" processes to compute the work of the algorithm [25]. The amount of worker processes a machine allows is one less than the total cores available; this may be scaled down, but in our case using all 8 CPU cores is ideal. The cleaned, pre-processed AirBnB data has a total size of 722 MB, while the total size of the TripAdvisor data has a total size of 383 MB, for a combined total of just over 1.1 GB. The two main jupyter notebooks that were created for data processing executed VADER and LDA, separately. Processing time during the VADER notebook took around 2 hours, while the LDA notebook took around 7.75 hours. The total size of the data after the VADER notebook is 1.13 GB and the total size for the data after the LDA notebook is 1.3 GB.

Evaluations and Results

The compound score reported in VADER analysis was the standard for categorizing reviews. Negative labels were assigned to compound scores below zero, positive labels were assigned to the scores above zero, and any review with a compound score of zero was classified as neutral. To assess VADER's performance, 100 reviews were selected randomly (50 from positive, 30 from negative and 20 from neutral predicted classes). The actual labels (per human inference) were compared to the predicted class. The model predicted positive and negative labels very well; however, some reviews (for example, "the place is close to the beach"), are classified as neutral. Since the neutral predicted class is only 1.6% of the total records, the accuracy (96.25%), precision (97.96%), sensitivity (97.96%) and specificity (93.55%) were calculated based only on negative and positive classes. The confusion matrix is shown below.

Table 1: VADER model confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	48	1
Actual Negative	2	29
Total	50	30

Since the LDA algorithm falls in the realm of unsupervised learning, some commonly used concepts such as accuracy and precision are not applicable for model evaluation. For this reason, we used the u_mass topic coherence metric in order to validate the model. This metric measures the co-document frequency of words within each topic and compares them to the total frequency of that given word [26]. Gensim's topic coherence implementation requires the entire corpus to be loaded into memory in order to generate the output [27]. This was not possible given the size of our data and computational limits. To address this, we took 40 random samples of 255K reviews (10.2M total reviews) and computed the topic coherence for each sample. The coherences had minimal variation, ranging from -3.16 to -3.05. Although not bad, the score implies that there are some mismatched words within each of the topics and there is room for improvement of the algorithm.

The LDA model was set to generate ten topics. Three topics were fairly general, and conveyed an overall positive experience. For purposes of analysis and visualization, we combined these into one macro-topic, "Overall Positive Stay". Other topics highlighted the location of the stay, any issues that occurred, and the quality of accommodation. Below is a table showing the top two topic definitions, along with the human-readable name our team assigned to it. The complete topic list is shown in Appendix A.

Table 2: Topics generated using LDA model

Assigned Topic Name	Key Descriptive Words
Overall Positive Stay	host place stay great clean beautiful space recommend apartment everything
Issue Related	night room place get door issue bit house time stay

Results for the VADER algorithm show 2.9129% of reviews being negative during COVID-19 and 2.9154% reviews being negative prior to COVID-19. To assess a more macro overview of review positivity over time, we grouped the VADER output on a quarterly basis from 2018 Q1 to 2021 Q1 for each city; through ranking each city's positivity by quarter and aggregating a cumulative positivity score, there's a distinguishable drop in positivity during 2019 Q4 that steadily improves through 2020 Q2 (Appendix B-Figure 2). However, the overall positivity percentage for city-wise aggregation across quarters suggests minimal variance from the expected norm during 2020 Q2, which is when we'd expect reviews to be the worst due to the beginning of COVID. This is likely due to an issue of data availability – when looking at the frequency of reviews per year, there's a clear decline after COVID's arrival: the average amount of reviews in 2018 and 2019 is 4,189,528 compared to just 1,638,752 in 2020 (Appendix B-Figure 3). Additionally, 10 of our 29 cities have one of their 3 worst-reviewed-quarters during 2020 Q2, with 8 of them having this as their worst quarter (Appendix B-Figure 4). In contrast, some differences appear between Airbnb (private owners) and TripAdvisor (commercial hotels) lodging: Airbnb recorded 96.29% positive reviews while TripAdvisor logged 91.59% positive. More striking, however, were the negative review rates: Airbnb registered only 1.74% while TripAdvisor posted 8.29% negative. Remaining reviews were classified as neutral. These results suggest that, while overall positivity towards lodging experiences remained the same in spite of the pandemic, there is a prominent difference between private owners and commercial hotels–consumers are more likely to criticize hotels with a negative review. Similar to sentiment, there were only minor shifts in review topics before and after the pandemic began. The largest changes were with "Great Value Location" reviews which increased by 2.25 percentage points, "Issue Related" reviews decreased by 2.24 percentage points, and "Overall Positive Stay" reviews, which increased by 2.05 percentage points.

Visualization

In order to distill our analysis of these 10-million lodging reviews, our team built an interactive website using D3.js to present and visualize the analysis results. Post-analysis processing consisted of aggregating the results at the city/county and quarterly level. Additional consolidation of adjacent localities was done, resulting in 21 locations and 13 quarters (first quarter, 2018 through first quarter, 2021). The visualization depicts three maps where the user can compare review statistics between different locations and time periods. The first map allows the user to conduct this comparison based on percentage of review sentiment, the second is oriented around review quantities, and the third displays the top five topics in a word-cloud format. The user gains additional information through tool-tips by hovering over specific locations, along with an interactive slider-bar to select the quarter in time (Figure 1). Further details about the maps are provided as descriptions in the visualization.

As stated above, the visualization used D3.js, specifically version 5. Version 5 was used because of the familiarity gained through the homework assignments conducted throughout the semester. In addition to the D3.js base code, various plugins were included. d3-legend.js simplifies the creation of legends and is primarily used on the first map. d3-tip.js simplifies the creation of tooltips and is utilized for all locations on all three maps. d3.layout.cloud.js generates word clouds easily and is specifically used on the third map. Unfortunately, the word cloud plugin and code isn't compatible with Firefox browser. In addition to plugins, some code was adapted from the internet, so the links associated with the code can be found in the comments located in the index.html file.

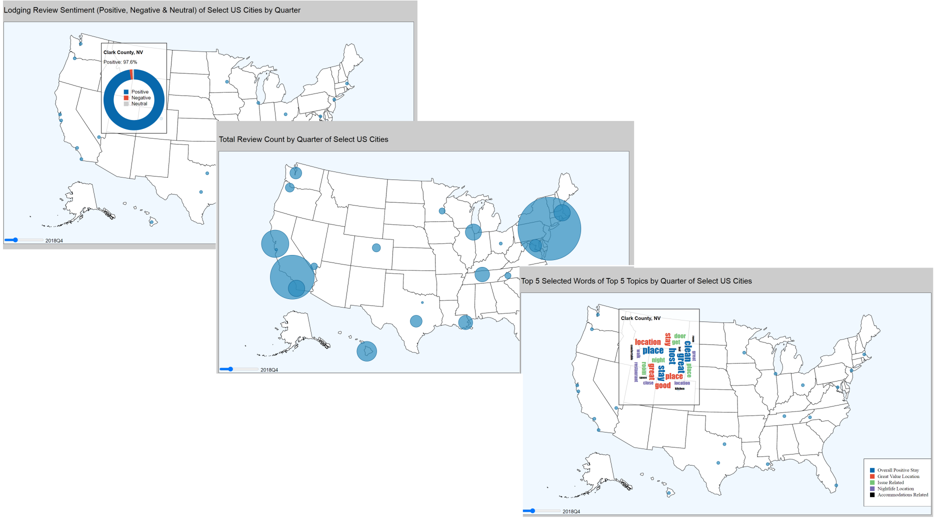


Figure 1: Visualization screenshots for the three maps.

Conclusion and Discussion

Although we were expecting changes in sentiment and topics before and after the pandemic began, the analysis has shown negative results. We found minimal differences in pre- and post-COVID positivity and topic content. This may be due to a lack of diversity in review content as compared to other analyses that sentiment analysis and topic modeling is performed on (i.e. news articles, social media posts, etc.). Still we think this methodology can be applied to these other corpora and domains outside of hospitality to gain a broader understanding of how the pandemic shifted discourse in our society. Our primary finding was that review quantity did drop significantly due to the pandemic.

We consider the following to be our innovations: creating scripts to run both textual analysis algorithms, and the interactive visualization (website) for comparing lodging reviews pre- and post-COVID, by location. This combination of analysis and visualization exceeds the state-of-the-art, and demonstrates a contribution to COVID effects' research.

Accumulation of Activities

All six team members have made significant contributions to the following tasks: research topics, datasets, literature and user interfaces for the project proposal. Create a Github repo for the project and include initial demo files related to scraping and NLP. Prepare the presentation slide deck, along with a presentation video. Download AirBnB and scrape TripAdvisor data with Apify. Clean data, detect language, and integrate data. Research technology topics (e.g. software libraries and products), as well as NLP models and methods. Establish One Drive repository. Script NLP jupyter notebooks for processing cleaned data. Setup Virtual Machine, run and monitor NLP jupyter notebooks. Implement project management tools to keep everyone on track. Update and complete project proposal and progress report. Analysis and evaluation of the results of the models. Detail, design, and code an interactive visualization. Write the final project report and the final poster for presentation.

References

[1] R. Yanguni, Q. Li, X. Mao and L. Wenyin, “Sentiment topic models for social emotion mining,” Information Sciences, vol. 266, pp. 90-100, 2014.

[2] M. V. Mäntyläa, D. Graziotin and M. Kuutilaa, “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,” Computer Science Review, vol. 27, pp. 16-32, 2018.

[3] E. Cambria, D. Das, S. Bandyopadhyay and A. Feraco, A Practical Guide to Sentiment Analysis, Springer International Publishing, 2017.

[4] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishing, 2012.

[5] W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithms and applications: A survey.,” Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, 2014.

[6] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” in Proceedings of the Eighth AAAI Conference on Weblogs and Social Media, 2014.

[7] S. Al-Natour and O. Turetken, “A comparative assessment of sentiment analysis and star ratings for consumer reviews,” International Journal of Information Management, vol. 54, pp. 102-132, 2020.

[8] A. Lawani, M. Reed, T. Mark and Y. Zheng, “Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston,” Regional Science and Urban Economics, vol. 75, pp. 22-34, 2019.

[9] M. Geethaa, P. Singha and S. Sinha, “Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis,” Tourism Management, vol. 61, pp. 45-54, 2017.

[10] M. Cheng and X. Jin, “What do Airbnb users care about? An analysis of online review comments,” International Journal of Hospitality Management, Vols. 76- Part A, pp. 58-70, 2019.

[11] I. Sutherland and K. Kiatkawsin, “Determinants of Guest Experience in Airbnb: A Topic Modeling Approach Using LDA,” Sustainability , vol. 12, no. 8, 2020.

[12] Y. Guo, S. J. Barnes and Q. Jia, “Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation,” Tourism Management, vol. 59, pp. 467-483, 2017.

[13] S. Banerjee and A. Y. Chua, “In search of patterns among travellers’ hotel ratings in TripAdvisor,” Tourism Management, vol. 53, pp. 125-131, 2016.

[14] A. W. Bartik, M. Bertrand, Z. Cullen, E. L. Glaeser, M. Luca and C. Stanton, “The impact of COVID-19 on small business outcomes and expectations,” Proceedings of the National Academies of Science (PNAS), vol. 117, no. 30, pp. 17656-17666, 2020.

[15] A. Abu-Rayash and I. Dincer, “Analysis of mobility trends during the COVID-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities,” Energy Research & Social Science, vol. 68, 2020.

[16] M. Škare, D. R. Soriano and M. Porada-Rochoń, “Impact of COVID-19 on the travel and tourism industry,” Technological Forecasting and Social Change, vol. 163, February 2021.

[17] S. Dolnicar and S. Zare, “COVID19 and Airbnb - Disrupting the Disruptor,” Annals of Tourism Research, vol. 83, 2020.

[18] “Web Scraping, Data Extraction and Automation · Apify,” [Online]. Available: https://apify.com/. [Accessed 2021].

[19] A. Alsudais, “Incorrect data in the widely used Inside Airbnb dataset,” Decision Support Systems, vol. 141, February 2021.

[20] “langdetect,” 2021. [Online]. Available: https://pypi.org/project/langdetect/.

[21] “World Health Organization,” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.

[22] X. Fang and J. Zhan, “Sentiment analysis using product review data,” Journal of Big Data, vol. 2, no. 5, 2015.

[23] M. D. Hoffman and D. M. B. F. Blei, “Online Learning for Latent Dirichlet Allocation,” NIPS’10: Proceedings of the 23rd International Conference on Neural Information Processing Systems, vol. 1, 2010.

[24] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation,” The Journal of Machine Learning Research, vol. 3, pp. 993-1022, March 2003.

[25] “GENSIM - topic modeling for humans,” [Online]. Available: https://radimrehurek.com/gensim/models/ldamulticore.html.

[26] D. Mimno, H. Wallach, E. Talley, M. Leenders and A. Calum, “Optimizing Semantic Coherence in Topic Models,” in Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011.

[27] “Topic Coherence Pipeline,” [Online]. Available: https://radimrehurek.com/gensim/models/coherencemodel.html.

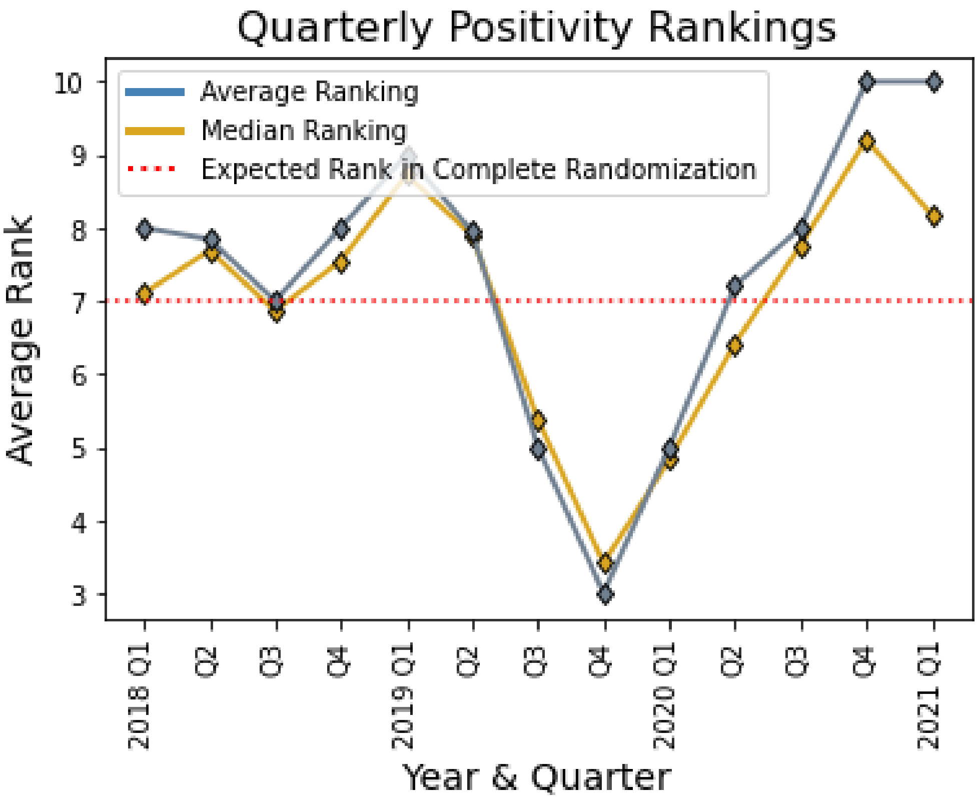
Appendix

Appendix A

Table 3: Topics generated using LDA model

Assigned Topic Name	Key Descriptive Words
Overall Positive Stay	host place stay great clean beautiful space recommend apartment everything
Issue Related	night room place get door issue bit house time stay
Nightlife Location	great walk restaurant location close downtown walking distance place minute
Accommodations Related	kitchen space room bed comfortable clean water nice bedroom bathroom
Great Value Location	great location place good stay nice clean value price really
Scenic Location	boston hill exploring garden bay ferry federal east rent trail
Communication Related	check question communication respond arrival day response communicate message reservation
Relaxing Vacation	beach perfect weekend cottage house getaway family day night stayed

Appendix B



Note: Lower rankings indicate less positive reviews

Figure 2: Quarterly positivity ranking

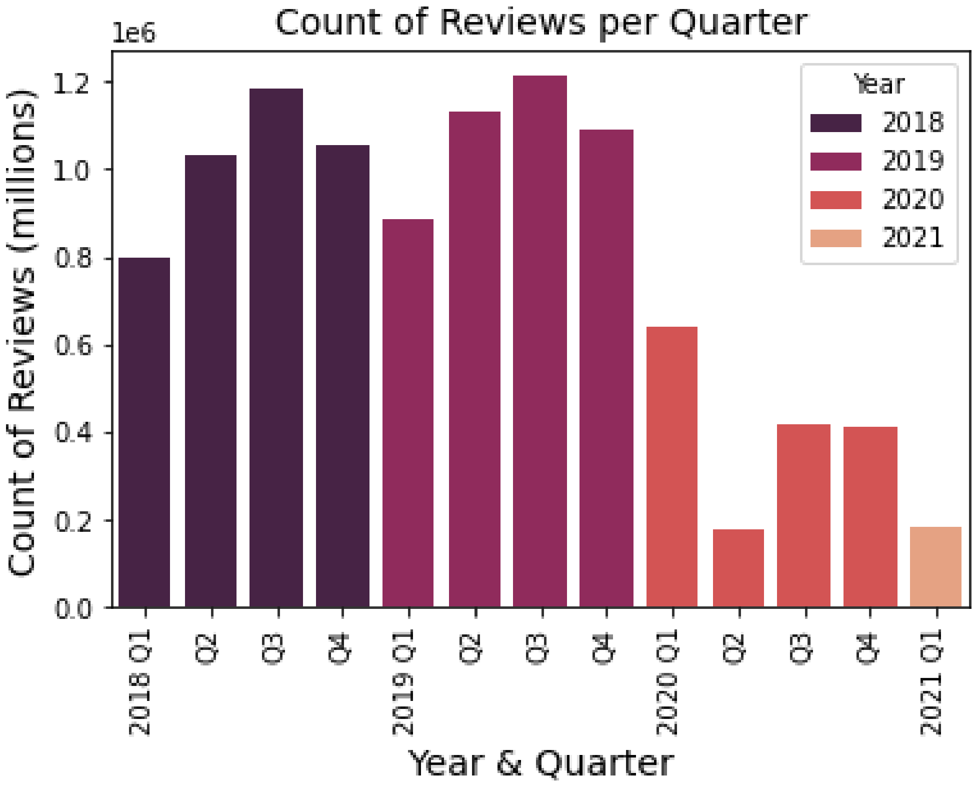


Figure 3: Count of reviews per quarter

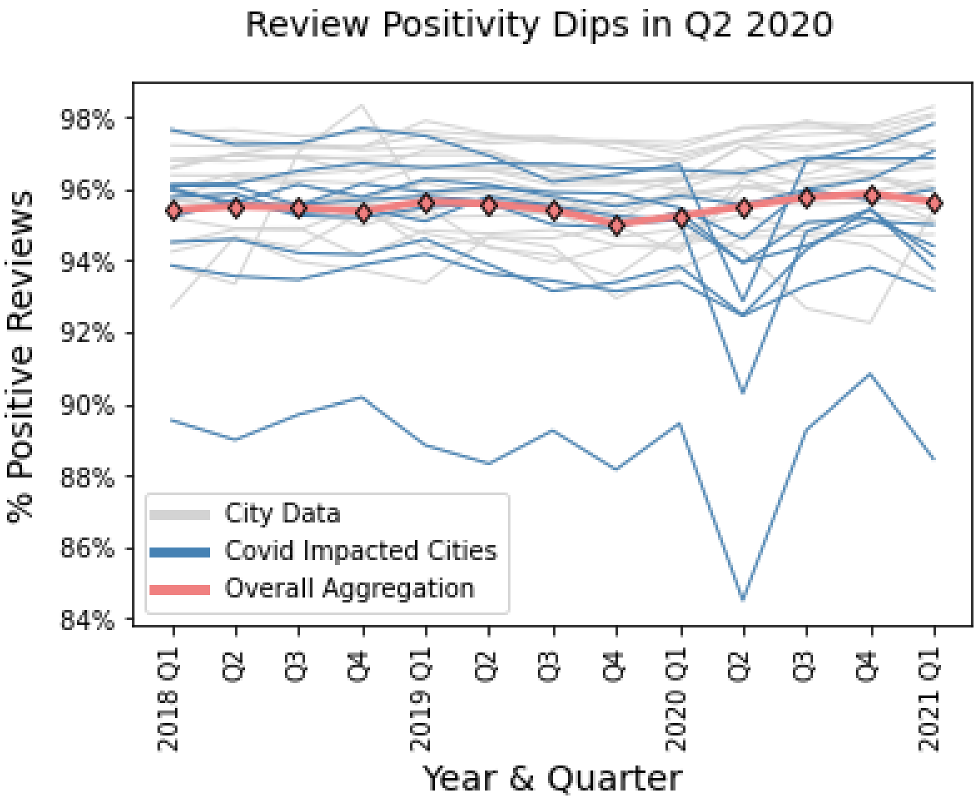


Figure 4: Review positivity change in each city over time - Note: Covid impacted cities are defined as cities where 2020-Q2 is one of their worst 3 positivity scores. These cities include: Boston, Cambridge, Chicago, Dallas, Denver, New Orleans, New York, Oakland, Pacific Grove and Santa Clara County.