Unveiling the Shift:

A Comprehensive Analysis of Hotel Reviews Pre and Post-COVID Era

by

Ammar Bagharib, Aravinth Adarsh, Luong Hien Nga,

Nuzzul Islam Nurhaqim, Png Chen Wei, Yuen Pin Xuan Tammy

A group project completed as part of the Mining Web Data for Business Insights module

Supervised by Dr Wang Qiuhong

National University of Singapore AY2023/24 Semester 1

Contents

Abstract

This project tackles a pressing business challenge within the hotel industry by employing machine learning techniques to analyse and categorise customer reviews, with a specific focus on the impacts of the COVID-19 pandemic. The primary business issue revolves around the necessity for hotels to adapt to evolving customer preferences, particularly in the context of the global pandemic. The corresponding machine learning problem involves the implementation of advanced natural language processing techniques, specifically Latent Dirichlet Allocation (LDA) and Dependency Parsing. These techniques are employed to categorise customer reviews into distinct aspects and sentiments, providing valuable insights into customer satisfaction dynamics. We also compare feature importance of topics to conduct Pre-Post COVID analysis.

The dataset utilised for this analysis is a meticulously curated collection, focusing on the top 5 hotels in each star rating category (3-star, 4-star, 5-star) sourced from TripAdvisor. This dataset serves as a representative sample, ensuring a comprehensive analysis of customer reviews across different hotel tiers. The adapted models include LDA for aspect-based categorization and dependency parsing for sentiment extraction. Additionally, traditional machine learning classifiers such as SGD Classifier and Random Forest are employed to analyse sentiment changes before and after the COVID-19 pandemic.

Achievements and Highlights:

Temporal Analysis: The project introduces a creative temporal analysis by dividing the study into pre and post-COVID periods, providing unique insights into how customer sentiments have shifted in response to the pandemic.

Strategic Hotel Segmentation: Application of machine learning techniques to segment hotels into different star categories ensures a more tailored and accurate analysis, acknowledging the diverse expectations of customers across various hotel tiers.

Feature Importance Analysis: Beyond sentiment analysis, the project conducts a feature importance analysis using classifiers, offering nuanced insights into how different aspects contribute to overall customer satisfaction, both before and after COVID.

Section 1 Proposal Review

Proposed Contributions	Achieved Outcomes	Satisfaction Levels
Aspect Classification Methods	Evaluation of existing methods for aspect classification	Satisfactory
Temporal Analysis	Analysis of hotel reviews pre and post-COVID	Exceeds Expectations
Cross-Aspect Analysis	Attempted and results were very skewed and not useful to make any conclusions	Not Achieved
Hotel Segmentation (3-star, 4-star, 5-star)	Successful implementation of hotel segmentation based on star ratings.	Exceeds Expectations

Table 1. Contribution and Outcome Table

1.1 Justification (Cross-Aspect Analysis)

When conducting aspect-based analysis, we were met with challenges during the project such as having a limited number of aspects after categorising reviews by hotel stars. For example, looking at 3 star hotels, with only 4 aspects in categorising reviews, it is way too little to feature engineer any meaningful cross-aspect analyses, as our findings might actually prove to be misleading instead. Hence, this limitation resulted in cross-aspect analysis to fall under our expectations and impacted the meaningfulness of cross-aspect analyses, preventing the achievement of the initially proposed depth.

If given more time, we will consider revisiting the aspect categorization phase to extract a more granular set of aspects or combining related aspects to enhance the depth of cross-aspect analysis. In addition, we would also research more on potential solutions or strategies employed by other researchers facing similar challenges in aspect-based sentiment analysis.

Section 2 Data Description

Our dataset consists of 15 different hotel reviews scraped from TripAdvisor and ordered by the number of reviews, with the top 5 hotels for each 3-star, 4-star and 5-star hotels. Each dataset consists of 10 features before data processing. These features are: date_of_stay, traveller_username, review_title, review_text, travel_type, traveller_country_origin, traveller_total_contributions, traveller_total_helpful_contributions, rating1, rating2. The total number of data points are as shown:

	3-star Hotels	**4-star Hotels**	**5-star Hotels**
Number of Data Points	14885	14607	37604

Table 2. Number of data points for each hotel star rating.

After performing data processing and feature engineering, we have 8 features. We created a feature 'covid' based on the date_of_stay. Cleaned_review, stem_review, and lem_review were also created for ease of model training further on.

Feature Name	**Feature Description**	**Value Type**	**Statistics**
travel_type	The type of traveller the reviewer belongs to, classified by TripAdvisor.	Categorical	Couple, Family, Business, Friends, Solo
rating	The rating provided by the reviewer regarding their stay at the hotel.	Numerical	Mean: 4.294141 Min: 1 Max: 5 Median: 5 Standard Deviation: 0.957172 Missing Data: 0
label	Variable that indicates if the review is positive (4, 5), negative (1, 2) or neutral (3)	Categorical	Positive, Neutral, Negative
covid	Variable that identifies if the review is given before Covid (29/01/2020) or after Covid (1/4/2022)	Categorical	PreCovid, PostCovid
is_local	Variable that identifies if the reviewer is from Singapore.	Categorical	0, 1
cleaned_review	Reviews that are processed to remove non-english words, stop words, punctuation and numbers.	Text	NA
stem_review	cleaned_reviews that are stemmed using Porter Stemmer.	Text	NA
lem_review	cleaned_reviews that are lemmatized using WordNetLemmatizer.	Text	NA

Table 3. Feature Table

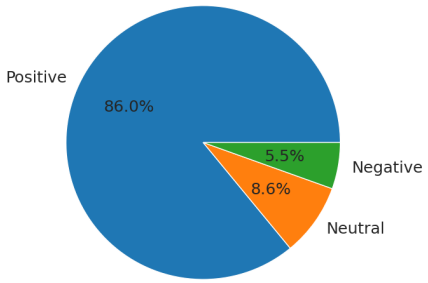

Fig 1. Pie Chart distribution of 'Label' variable

As seen in figure 1, there is data imbalance for the feature 'Label'. There are far more positive labels than negative labels. This will affect the performance of our models in our pre and post covid analysis. Hence, undersampling for the positive reviews is performed later on to reduce bias.

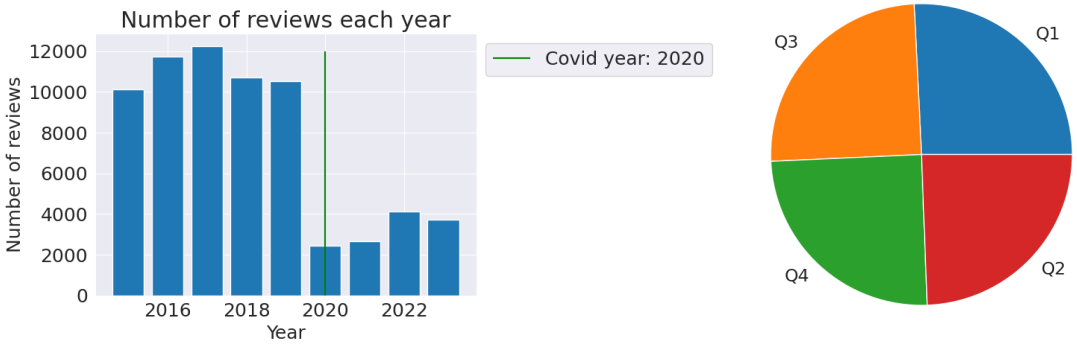

Fig 2. Bar chart of yearly reviews, and pie chart showing proportion of reviews each quarter

Figure 2 shows the number of hotel reviews in each year and quarter. We notice significantly lesser reviews written after Covid, which is expected. The pie chart shows that the number of reviews written in each quarter is quite consistent.

Section 3 Models and Performance

3.1 Sentiment Analysis

3.1.1 LDA

We opted for LDA to identify, classify and split hotel reviews into different aspects. LDA was chosen due to both its ability to categorise words and sentences into aspects, as well as its ease in understanding these aspects. As LDA uses conditional probability to identify the words belonging to the aspect, we are able to use the top 20 words for each topic in LDA to analyse and provide meaning for business users when conducting aspect-based sentiment analysis.

To identify the ideal number of aspects that is needed in our analysis, we opted to use coherence score as a scoring metric. The use of coherence helps to measure the degree of semantic similarity between high scoring words of each aspect, using a combination of normalised pointwise mutual information (NPMI) and cosine similarity. NPMI measures the probability that two words are associated with each other as compared to the probability of two words being independent. Thus, in our case, NPMI can tell us how likely two words will appear in the same aspect. On top of that, cosine similarity helps to further support the NPMI by calculating the similarity of the words in the aspect to new words, with a lower cosine similarity score indicating that two words are semantically closer to each other. Thus, to ensure that the words within our aspects are semantically similar to each other, we will pick the number of topics for each hotel review with the highest coherence score.

	3-star	**4-star**	**5-star**
Best number of aspects	4	5	5
Coherence Score	0.473	0.477	0.437
Aspects	Room, Food, Experience, Staff	Hotel Amenities, Room, Service, Location, Staff	Nearby Amenities, Food, Room, Service, Staff

Table 4. Best number of aspects for different star hotels

3.1.2 Dependency Parsing

Extraction of topics and their corresponding sentiment is done using dependency parsing. Dependency parsing is a process which uncovers the grammatical relationship between words in a sentence. After that, we extracted the topics and their sentiments via a 2-step rule-based approach. Firstly, we find a noun and adjective pair with an adjectival modifier[1] relationship. Secondly, we find an adverb that has an adverbial modifier relationship[2] with the adjective identified. We also labelled the topic number based on which topic the noun appears in our LDA topic modelling.

[1] Adjectival modifier of a noun is an adjectival phrase that serves to modify the noun.
[2] Adverbial modifier of a word is an adverbial phrase that serves to modify a predicate or modifier word.

3.1.3 Textblob Polarity

[topic, sentiment, aspect]	aspect_0	aspect_1	aspect_2	aspect_3	aspect_4
[[room, small, 2], [staff, great, 1]]	NA	0.70	-0.39	NA	NA

Table 5: Percentage of sentiment polarity by topics

For sentiment analysis, our project employed TextBlob, a python library capable of processing textual data and performing sentiment analysis on it. We utilised TextBlob upon our topic-sentiment pairs, that we extracted from the previous step, to output a sentiment score ranging from -1 to +1 for that aspect, denoting highly negative to highly negative. As seen in Table 5, we assign the polarities to the corresponding aspects. If an aspect occurs more than once within a single review, we extract a mean polarity score for that aspect.

3.1.4 PyABSA

PyABSA is an aspect-based sentiment analysis model that uses pre-tagged datasets and machine learning models to generate new aspect-based sentiment for new sentences (Yang, Zhang, and Li, 2022). The robustness of its model, as well as its reproducibility of aspect-based sentiment analysis (ABSA) makes it an ideal candidate for our evaluation model. Despite this, we elected not to use pyABSA due to its lack of hotel reviews in its dataset, long runtime to generate ABSA, and the high granularity generated from its aspects, making it difficult to generate insights using pyABSA's aspects. Thus, there is a need to manually categorise pyABSA's aspects into broader categories for further sentiment analysis, being unsuitable for our objectives in conducting ABSA.

3.1.5 Model Evaluation

	Negative		**Neutral**		**Positive**	
Topic No.	**pyABSA (%)**	**TextBlob (%)**	**pyABSA (%)**	**TextBlob (%)**	**pyABSA (%)**	**TextBlob (%)**
0	14.27	5.64	22.04	5.06	63.69	89.30
1	12.00	6.92	27.42	11.28	60.58	81.79
2	30.64	6.46	11.16	11.00	58.19	82.54
3	22.32	10.04	17.82	10.18	59.85	79.78
4	18.43	7.49	14.12	21.68	67.45	70.83

Table 6: Percentage of sentiment polarity by topics

Table 6 compares the difference in the percentage of negative, neutral and positive sentiments classified for each topic by both models. We use the extensively trained pyABSA model as a ground truth proxy to determine the accuracy of the TextBlob model. From Table 6, we concluded that the model performance of TextBlob is positively skewed. This could be attributed to several factors in the pipeline of our aspect-based sentiment analysis process. Firstly, the extraction of topic-sentiment pairs after the dependency parsing stage requires some level of proficiency in English linguistics in order to implement the appropriate sentence structure rules. Our extraction process is only governed by a

single condition of noun-adjective pairs with adjectival and adverbial modifiers. Our inexperience in English linguistics and the fact that most hotel reviews are informal and conversational-style makes it hard for us to implement the appropriate rules for topic-sentiment extraction. Secondly, although pyABSA is an extensively trained model, the model was trained on restaurant reviews which may result in some inaccuracy when testing on our hotel review data. This is one of the biggest limitations when embarking on this project as we lack an accurate target label to compare the performance of our model to.

3.2 Pre vs Post Covid analysis

3.2.1 Performance on test sets

Two classifiers were trained on the Pre-Covid and Post-Covid dataset separately. To ensure fairness between 2 classifiers, downsampling was performed to balance the Pre-Covid and Post-Covid data, as well as sentiment classes, resulting in approximately 800 data samples for each dataset. Linear SVM and Random Forest Classifier were employed for classification. GridSearchCV was used to optimise hyperparameters.

Random Forest Grid Search			**LinearSVM Grid Search**		
n_estimators	max_features	max_depth	C	dual	penalty
100, 200	‘sqrt’, ‘log2’	5, 10, 15	0, 0.01, 0.1, 0.5, 1.0, 10.0	True, False	‘l1’, ‘l2’

Table 7: Hyperparameters for GridSearchCV tuning

The differences between Pre-Covid and Post-Covid classifiers’ performance are negligible. Furthermore, the Pre-Covid Classifiers performed minimally better on the Post-Covid test set. This could be attributed to the lack of data after downsampling, which failed to capture the difference between Pre-Covid and Post-Covid reviews.

Random Forest Classifier			**LinearSVM Classifier**		
	Pre-Covid Classifier	Post-Covid Classifier		Pre-Covid Classifier	Post-Covid Classifier
Pre-Covid test set	0.88	0.87	Pre-Covid test set	0.89	0.91
Post-Covid test set	0.91	0.92	Post-Covid test set	0.91	0.92

Table 8: F1 scores obtained during testing Pre/Post-Covid classifiers on Pre/Post-Covid test sets

An alternative method could be exploring feature importance of two classifiers, which enables us to utilise all the data for training as dataset balance is not crucial for finding feature importance.

3.2.2 Feature Importance

To analyse feature importance, we fitted an SGDClassifier and Random Forest classifier to the Pre-Covid dataset and Post-Covid dataset. After obtaining the classifiers, we can get the importance of

each topic. Firstly, we obtain our features of interest, which are the words that make up each LDA topic. Then, we obtain a sum of the importance scores of these words within each respective LDA topic to obtain the overall importance score of the topic. After which we rank the topics based on their summed importance scores. Below is an example of the overall importance score comparison across all topics for 5-star hotels:

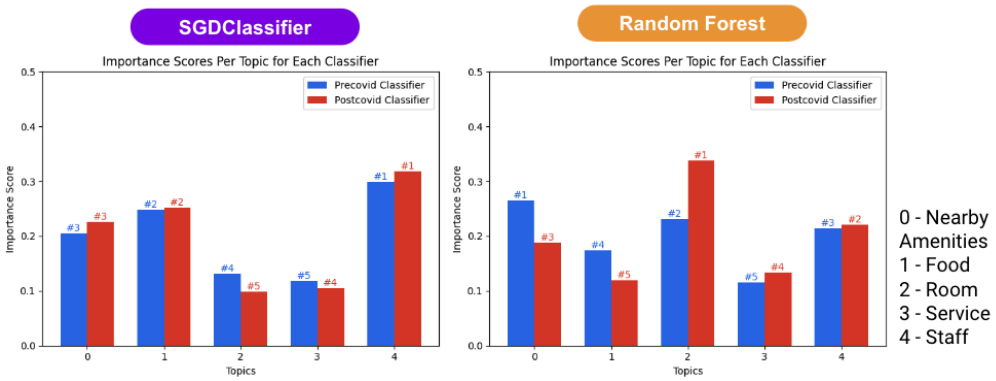

Fig 3. Graph displaying rankings of topics Pre vs Post covid - for SGDClassifier and RandomForest Classifier

From the two graphs above, we can observe that for the SGDClassifier, the ranking of each topic had no changes to the ranking of topics 0, 1, and 4, but topic 2 decreased in importance while topic 3 increased in importance. Thus the SGD classifier suggests that topic 'Room' played lesser significance in determining the positivity of rating, while 'Service' had higher significance.

The 2 classifiers offered different importance scores for the topics, as seen in the example where the Random Forest graph showed more drastic changes in topic rankings pre and post covid. This could be attributed to the innate difference in linearity of the models – where the Random Forest model observes non-linear relationships, and the SGD classifier observes linear relationships.

Section 4 Contribution and Justification

Contribution		Low	Medium	High
Use valuable and high-quality new datasets, including integrating existing datasets, scrawling or retrieving data, etc.	Effort			**X**
	Effectiveness			**X**
Design or adaptation of new ML methods/architecture or the integration of existing methods with a balance of resource and cost	Effort			**X**
	Effectiveness		**X**	
Creativity or insights in understanding or further explaining the prediction results and performance	Effort			**X**
	Effectiveness		**X**	

Table 9: Contributions and Efforts

4.1 Valuable Dataset Integration and Collection:

Our first contribution lies in the meticulous curation of high-quality datasets, specifically tailored for the hotel industry. Unlike existing studies that either focus on individual hotels or aggregate diverse hotel ratings, our research is distinguished by the strategic collection of the top 5 hotels from each star rating category (3, 4, and 5 stars) on TripAdvisor (Sodanila, 2016; Zhang, Lu, & Liu, 2021; Wulandari, Nuri, & Kurniasari, 2021). This approach ensures a nuanced understanding of the hotel landscape, acknowledging the preferences of diverse consumer bases (Juliana, Hubner, Sitorus, & Pramono, 2021). Instead of fetching widely-available datasets on Kaggle, we scraped these hotel reviews using Octoparse, a web scraping tool, after which we meticulously cleaned the datasets for modelling purposes. We strongly believed there was an overall gap in analysing Singapore based hotels tripadvisor reviews, explaining our choice of such datasets.

4.2 Machine Learning Methodology:

In terms of methodology, we introduced a thoughtful integration of machine learning techniques to derive meaningful insights. The adaptation of Latent Dirichlet Allocation (LDA) enabled us to extract five distinct topics for each hotel star rating while dependency parsing allowed us to identify the key adjectives describing the key aspects of each hotel review, providing a more granular understanding of customer sentiments. Our project's architecture reflects a deliberate choice to integrate LDA and DP instead of using PyABSA for aspect-based sentiment analysis. This decision was guided by a balanced consideration of resource efficiency and cost-effectiveness. While PyABSA offers a powerful ABSA model, its limitations in terms of dataset applicability and processing times prompted

us to leverage the capabilities of LDA and DP. This integration not only streamlines the analysis process but also ensures a practical and resource-conscious approach to achieving our research objectives.

4.3 Creative Temporal and Stratified Analysis:

Recognizing the scarcity of studies that explore temporal changes in hotel customer reviews, we innovatively divided our analysis into pre and post-COVID periods. This temporal lens provides a dynamic perspective on how aspects of hotel reviews have evolved in response to the pandemic. The COVID-19 pandemic has had an unparalleled impact on the global hospitality industry. Travel restrictions, lockdowns, and heightened health concerns have significantly altered the landscape of hotel services (Foroudi et al., 2020). Understanding how these changes manifest in customer reviews is crucial for both academic inquiry and practical implications for the hotel industry.

Additionally, obtaining different topics for the differently rated hotels, acknowledges the business reality that differentiates consumer expectations across hotel tiers (Hargreaves, 2015). This nuanced approach ensures that our findings are not only relevant but actionable for businesses catering to distinct market segments.

4.4 Insightful Interpretation of Results:

5 Star Hotels	Classifier	
	SGD	Random Forest
Nearby Amenities	No Change	Decrease
Food	No Change	Decrease
Room	Decrease	Increase
Service	Increase	Increase
Staff	No Change	Increase

Table 10: Feature Importance Differences Pre vs Post Covid

A key aspect of our contribution is the insightful interpretation of prediction results and performance metrics. By comparing feature importances across different classifiers, we uncovered nuanced variations in aspect importance pre and post-COVID. SGD classifiers perform well in identifying the linear relationships between the aspects and the sentiment, assuming they're independent of other aspects. Random forest on the other hand, is able to capture the non-linear relationships or the cross-interactions of aspects which the former model could not capture. For instance, as seen in Table 10, the differential impact observed with SGD and RF classifiers suggests that non-linear relationships and interactions among aspects may have become more pronounced post-COVID. This 2-pronged

approach allows us to better understand the importance of non-linear relationships between our aspects pre and post-covid.

Another example, for aspects which demonstrated similar types of differences by both classifiers in feature importance pre and post covid, e.g., ‘Service’ for 5 star hotels. From a hotel’s perspective, they can then analyse the new initiatives that they’ve implemented with regards to these aspects pre and post covid. This interpretation goes beyond conventional analyses and opens avenues for understanding the dynamic nature of consumer preferences in the wake of global events. Additionally, our recommendation for hotels to analyse new initiatives based on classifier-generated changes in feature importance adds a practical dimension to the interpretability of our results.

In summary, our research helps the field especially from a local lens, by employing holistic machine learning methods, and offering creative insights through temporal and stratified analyses. These contributions collectively contribute to a more comprehensive understanding of Singapore’s top-performing hotel customer reviews, particularly in the context of evolving consumer preferences during and after the COVID-19 pandemic.

Section 5 References

Foroudi, P., Tabaghdehi, S. A. H., & Marvic, R. (2020, November 3). The gloom of the COVID-19 shock in the hospitality industry: A study of consumer risk perception and adaptive belief in the dark cloud of a pandemic. International Journal of Hospitality Management. Advance online publication. https://doi.org/10.1016/j.ijhm.2020.102717

Hargreaves, C. A. (2015). A Comparative Analysis of Hotel Ratings and Reviews: An Application in Singapore. American Journal of Marketing Research, 1(3), 118-129. http://www.aiscience.org/journal/ajmr

Juliana, J., Hubner, I., Sitorus, N. B., & Pramono, R. (2021). The Influence of Hotel Customer Demographics Differences on Customer Perceptions. African Journal of Hospitality Tourism and Leisure, 10(3), 863-880. https://doi.org/10.46222/ajhtl.19770720-137

Sodanila, M. (2016). Multi-Language Sentiment Analysis for Hotel Reviews. MATEC Web of Conferences, 75, 03002. https://doi.org/10.1051/matecconf/20167503002

Wulandari, N. D., Nuri, M. H. Z., & Kurniasari, L. (2021). Customers' Satisfaction and Preferences Using Sentiment Analysis on Traveloka: The Case of Yogyakarta Special Region Hotels. Journal of Tourism Research, 8(2), 45-62. DOI: 10.2991/assehr.k.211020.058

Yang, H., Zhang, C., & Li, K. (2022). *Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis* (arXiv:2208.01368). arXiv. https://doi.org/10.48550/arXiv.2208.01368

Zhang, J., Lu, X., & Liu, D. (2021). Deriving customer preferences for hotels based on aspect-level sentiment analysis of online reviews. Electronic Commerce Research and Applications, 49(C), Article 101094. https://doi.org/10.1016/j.elerap.2021.101094