PRE-COVID and POST-COVID Analysis on Hotel Review

by TF-IDF and Sentiment Analysis

Meihong Li, Hang Zhou, Siangling Hsu, Stephanie Zhao

BYGB-7978-002

Professor Yilu Zhou

Dec. 11, 2021

Content

Executive Summary

As we know, the pandemic has affected every sector across the globe, and the hotel industry is among the hardest hit. Many factors affect the travelers' intentions to travel, including safety and security, space accessibility, travel costs, quality issues, sanitation risks, hygiene, and destination trust.

As stated in the report (McKinsey. 2020 Hospitality and COVID-19: How long until no vacancy for US hotels?), it suggests that recovery to Pre-COVID-19 levels could take until 2022 or later. To help hotels survive in a pandemic, we compare the Pre-COVID and Post-COVID reviews to dig into the customer's thoughts. The more quickly hotels adapt to the pandemic the more chance to create a new business model.

Our outline of analyzing hotel's brand attributes is based on Feng Hu, Rohit H. Trivedi, 2020. Taking 11,627 hotel reviews obtained from Tripadvisor.com as samples, this report focuses on the hotel located on Times Square. Meanwhile, we explore customer satisfaction and the attributes that customers care about Pre-COVID-19 and Post-COVID-19 using TF-IDF.

To analyze customers' attitude, we use predefined factors from Nadeem Akhtar, Nashez Zubaira, Abhishek Kumara, Tameem Ahmada 2017 to do sentiment analysis. According to sentiment analysis results, this project will suggest operation strategies based on customers' satisfaction and dislike in the Post-COVID.

System Design

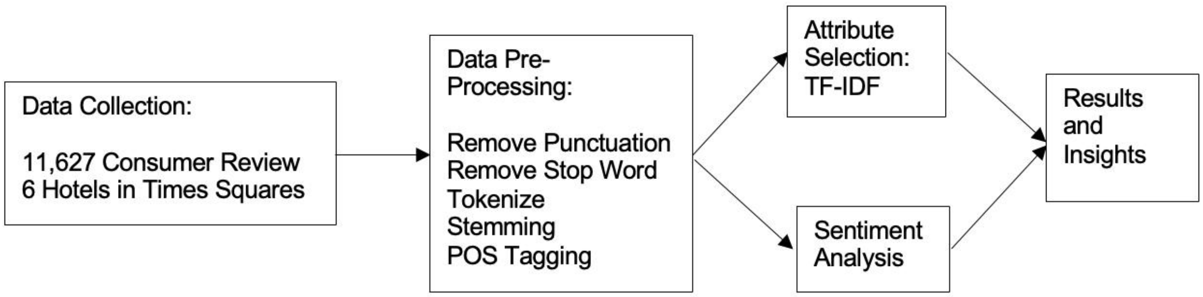


Fig1. Project Design

First, we crawl reviews by selenium and Beautifulsoup from 6 hotels located in Times Squares from the TripAdvisor Website. After crawling, we apply basic text cleaning skills to the data set, including removing punctuation, removing stop words, stemming words, and doing POS tagging. For POS tagging, we generate verbs, nouns, and adjectives' word clouds to see which one can show attributes of hotels.

Secondly, We select noun and adjective terms to calculate TF-IDF. After we calculate the TF-IDF in Pre-COVID and Post-COVID, we rank terms by TF-IDF and find the top 20 terms as hotel's attributes in each hotel. However, TF-IDF results can not reflect customer attitude, so we further apply sentiment analysis to identify negative or positive reviews.

Thirdly, we keep each sentence's verbs, nouns, adjectives, and adverbs in sentiment analysis. We predefine six aspects- meal, service, staff, room, facility, and quality. We use predefined aspects to categorize sentences and apply the VADER package to sentiment analysis. We use the score to differentiate positive or negative sentences.

Dataset Description

We use BeautifulSoup and Selenium to scrape Author, Ratings, Title of Review, Review Tet, and Date of Stay for each hotel review in TripAdvisor.com. To limit our variable and improve research accuracy, we mainly search the following six hotels in Times Square District in New York. We choose the Times Square because it is the hottest attraction in New York City and many tourists go there.

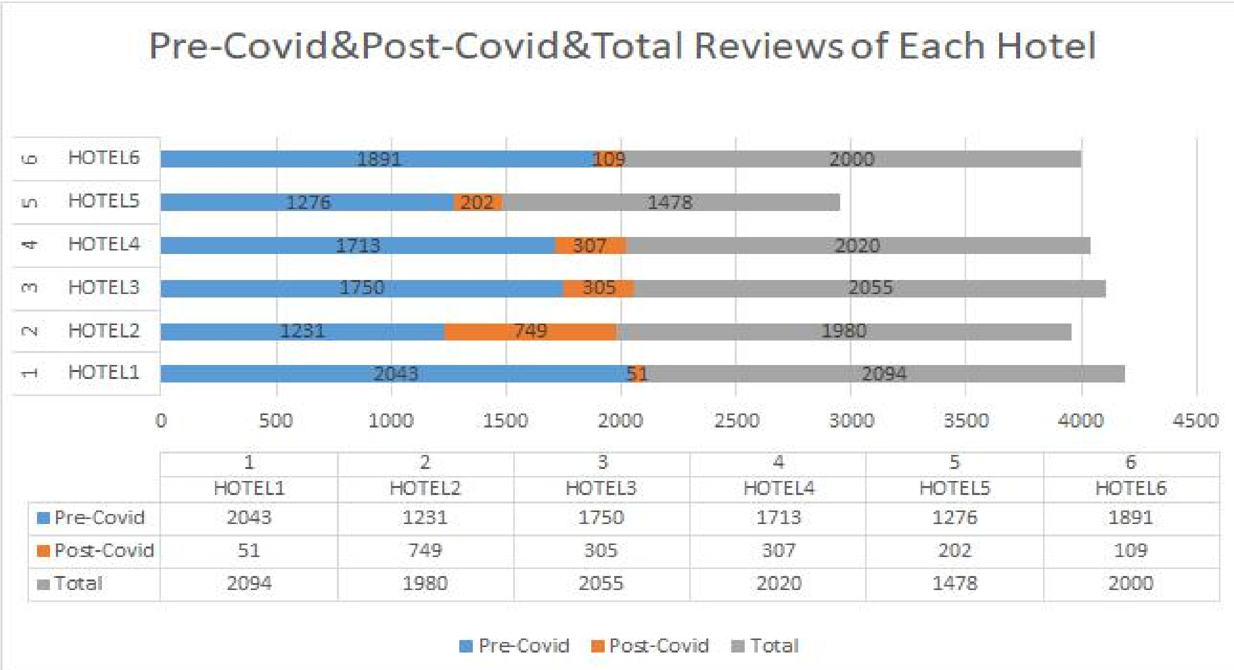
Sanctuary Hotel New York (H1), Hotel Edison (H2), DoubleTree by Hilton Hotel New York Times Square West (H3), Millennium Hotel Broadway Times Square (H4), Hampton Inn Manhattan / Times Square Central (H5), and Warwick New York (H6). In addition, we focus on English reviews and split the date of stay into the month and years to filter Pre-COVID and Post-COVID data.

The dataset overall contains 11,627 hotel customer reviews. We split the data based on January 2020 into Pre-COVID and Post-COVID. Table 2 and Figure 2 shows Our dataset's detailed statistics.

	A	B	C	D	E	F	G
1	reviews_title	reviews_text	ratings	author	stay_date	year	month
53	Fanstastic Location	My boyfriend and	5	Curious26531	Dec-19	2019	12
54	Lovely hotel would	Myself and my fia	4	Bexyboo13	Dec-19	2019	12
55	Staycation	Had a staycation t	4	Trinicz	Dec-19	2019	12
56	Our 3rd Visit - Supe	Our 3rd visit back	5	jac435	Dec-19	2019	12
57	Great stay!	My husband and I	5	Brooke D	Dec-19	2019	12
58	Great Location	We booked a cozy	5	Jayda D	Dec-19	2019	12
59	Good hotel to spen	My wife and I spe	4	Antoine H	Dec-19	2019	12
60	Wonderful for a we	The services and l	5	Natty M	Dec-19	2019	12
61	5th time staying at	This was our fifth	5	daniel m	Dec-19	2019	12
62	Truly a sanctuary. A	I was pleasantly su	5	Art M	Dec-19	2019	12
63	We visited to enjoy	We had a great ex	5	Melissa C	Dec-19	2019	12
64	Wonderful Staff	Our stay at The Sa	5	wos33	Dec-19	2019	12
65	Worst wait time eve	I arrived at the ho	1	Tori	Dec-19	2019	12
66	Don't Let the Exteri	I must say, I gener	1	Sabrina R	Dec-19	2019	12
67	I must have gotten	I travel to ny a lot	3	VNORVI	Dec-19	2019	12
68	Great location, staff	I鈥檝e been comi	5	grant11955	Nov-19	2019	11
69	Fantastic	Had a fabulous sta	4	a_fittis602	Nov-19	2019	11
70	Wonderful 5 nights	This hotel is super	5	menace22	Nov-19	2019	11
71	Superb stay	Monday 25 Nov t	5	Suzanne P	Nov-19	2019	11
72	Great hotel and stu	I don't typically lea	5	Matthew C	Nov-19	2019	11
73	Return visit	We returned for m	5	Doreen T	Nov-19	2019	11
74	Cool entrance and	Cool entrance and	3	AZbusinesstra	Nov-19	2019	11
75	Small rooms but far	The rooms are sm	5	gmlomas	Nov-19	2019	11
76	Friendly staff, conve	I stayed here on a	5	Lee Ann W	Nov-19	2019	11
77	Dreadful - avoid at	The 'guaranteed'	1	Howard Firkin	Nov-19	2019	11

Sanctuary_total | Sanctuary_Pre | Sanctuary_Post

Table1. Sample dataset of Sanctuary Hotel New York (H1)



	1 HOTEL1	2 HOTEL2	3 HOTEL3	4 HOTEL4	5 HOTEL5	6 HOTEL6
Pre-Covid	2043	1231	1750	1713	1276	1891
Post-Covid	51	749	305	307	202	109
Total	2094	1980	2055	2020	1478	2000

Table2. the number of Pre-COVID, Post-COVID, and Total reviews we obtained from each hotel

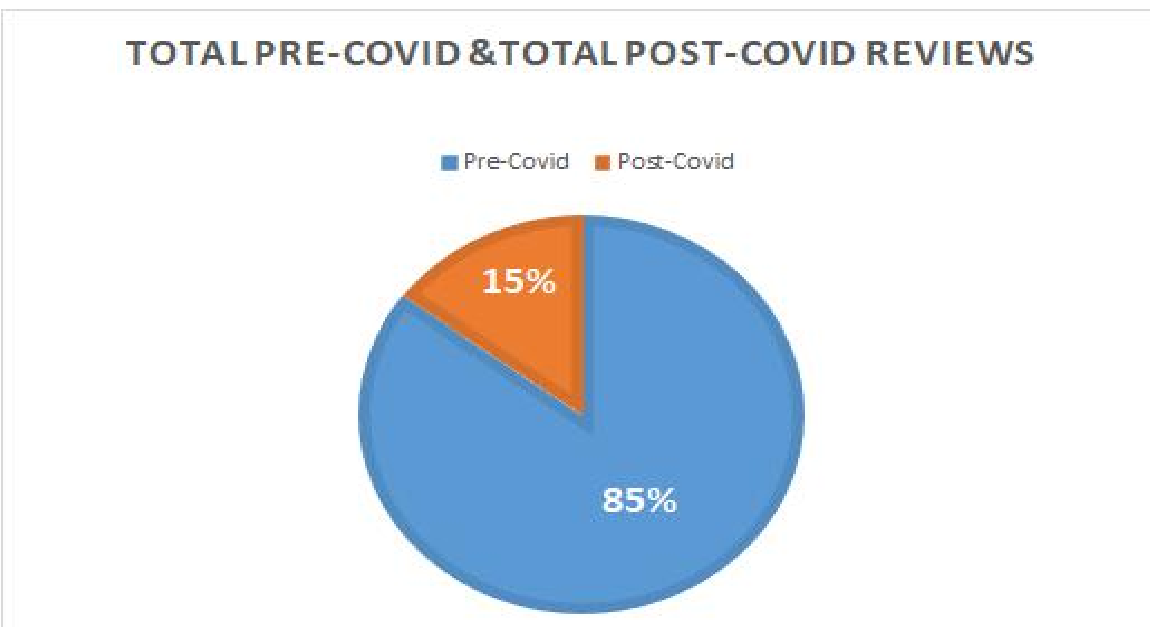


Fig2.Total Pre-COVID and Post-COVID reviews we obtained from 6 hotels

Project Implementation and Data Analysis

1. Data Pre-Processing

First, we remove non-alphabetical words and use the nltk word_tokenize package to tokenize each review. After that, we use nltk corpus to remove stop words and apply nltk stem porter to stem each review. In the end, we apply Part-Of-Speech (POS) tagging by WordNetLemmatizer.

2. Word Cloud

According to Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content' Feng Hu Rohit H.Trivedi 2020, hotel's attributes originated from nouns. We use word clouds to show words in verbs, nouns, and adjectives to see what kind of word can utilize our analysis.

Fig3.Verb(H2 Pre-COVID reviews)

From the Verb Word Cloud map, we can see that those words are not very useful. For example, the word got, go, made, recommend can not provide specific information about customers' attitudes and sentiments and can not show the characteristics of the six hotels.

Fig4. Noun (H2 Pre-COVID reviews) *Fig5. Adjective(H2 Pre-COVID reviews)*

However, from figure 4- noun word cloud and figure 5- adjective word cloud, most of the words are related to the hotels' attributes: room, staff, breakfast, location, bathroom, etc. Some words are more related to the customer experience with the hotel stay: perfect, helpful, clean, old, etc. As a result, we decide to use nouns and adjectives to analyze what customers care about Pre-COVID and Post-COVID.

3.TF-IDF

As we refer to above, we add the TF-IDF (Term Frequency - Inverse Document Frequency) values for every word. However, why not count how many times each word appears in every document? The problem with this method is that it does not consider the relative importance of words in the texts. A word that appears in almost every text would not likely bring helpful information for analysis. On the contrary, rare words may have a lot more meanings. The TF-IDF metric solves this problem: TF computes the times the word appears in the text; IDF computes a word's relative importance, which is the inverse number of how many texts the word can be found. Take Hotel Edison (H2) as an example. As figure 7 shows, the top 10 TF-IDF terms are Room, Staff, Great, Edison, Location, Stay, Times, Night, Clean and Good (descending order).

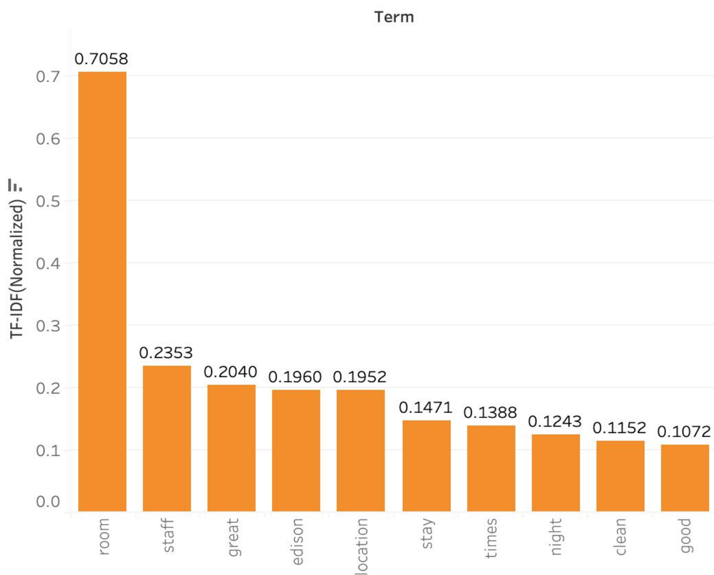


Fig6. TF-IDF - Term in H2 (Hotel Edison)

Next, we rank the TF-IDF for six hotels in Pre-COVID and Post-COVID, respectively. We choose the top 20 TF-IDF terms in each hotel and combine them into a table(table 3). During combination, we remove specific terms, such as Edison, times, square, etc., because of not represent hotels' attributes. The "-" in table 3 shows the ranking is behind 23.

term	H1_Pre	H1_Post	H2_Pre	H2_Post	H3_Pre	H3_Post	H4_Pre	H4_Post	H5_Pre	H5_Post	H6_Pre	H6_Post
room	1	1	1	1	1	1	1	1	1	1	1	1
breakfast	4	2	-	9	-	-	-	-	-	2	20	2
staff	2	3	2	3	2	2	2	5	3	4	2	4
stay	9	4	-	6	11	9	11	4	8	7	6	7
front	-	5	-	-	-	11	-	8	-	18	-	18
desk	-	14	-	-	18	10	18	6	20	16	-	-
good	11	6	10	12	8	14	8	15	6	9	7	9
great	3	7	3	4	3	6	3	3	4	5	5	5
nice	-	9	17	20	19	-	19	19	17	12	12	12
helpful	13	-	15	11	-	-	-	-	19	10	22	-
location	5	11	5	5	7	6	7	2	2	4	-	4
small	8	14	18	-	6	12	6	-	-	-	-	-
clean		15	9	7	13	13	13	13	9	6	17	6
night	10	16	8	10	9	8	9	10	10	11	15	-
restaurant	18	18	14	16	-	-	-	-	-	-	-	-
bar	6	-	-	-	12	19	12	-	-	-	18	-
day	14	-	12	17	10	4	10	17	12	16	19	16
new	15	-	11	19	16	-	16	21	13	-	13	-
service	17	-	-	14	15	18	15	7	14	23	8	23
floor	-	-	16	-	5	7	5	12	11	23	-	23

Table3. the ranking change of top 20 terms in Pre-COVID, Post-COVID

The ranking of the term [breakfast] improved obviously in Sanctuary Hotel New York (H1), Hotel Edison (H2), Hampton Inn Manhattan / Times Square Central (H5), and Warwick New York (H6). For instance, the term [breakfast] rose by 18 places in Warwick New York (H6) between Pre-COVID and Post-COVID which means customers pay more attention to the eating space and food quality. The ranking of the term [clean] also shows an upward trend in Sanctuary Hotel New York (H1), Hotel Edison (H2), Hampton Inn Manhattan / Times Square Central (H5), and Warwick New York (H6). The term [clean] improves by three places in Hampton Inn Manhattan / Times Square Central (H5), 11 places in Warwick New York (H6), and two places in Hotel Edison (H2) which means customer emphasis on sterilization and sanitation after COVID-19 very much.

However, the ranking of the term [service, elevator, and floor] has a slight decrease. Sanctuary Hotel New York (H1), Hotel Edison (H2), and Warwick New York (H6) do not show those terms in the top 20 attributes list, which shows that customers pay less attention to service and more details (service, elevator, floor) after COVID-19 situation.

term	H1_Pre	H1_Post	H2_Pre	H2_Post	H3_Pre	H3_Post	H4_Pre	H4_Post	H5_Pre	H5_Post	H6_Pre	H6_Post
breakfast	4	2	-	9	28	-	-	-	-	2	20	2
clean	-	15	9	7	13	13	13	13	9	6	17	6
bar	6	-	-	-	12	19	12	-	-	-	18	-
stay	9	-	6	-		-	-	-	8	-	6	7
service	17	-	-	14	15	18	15	7	14	23	8	23
elevator	-	-	-	-	4	3	4	-	19	-	-	-
floor	-	-	16	-	5	7	5	12	11	23	-	23

Table 4. TF-IDF - Term in 6 hotels

4. Sentiment Analysis

We keep nouns, verbs, adverbs, and adjectives in a sentence and break each review into several sentences. We apply sentiment analysis in user reviews to further understand customers' feedback. First we classified each sentence into predefined aspects(Nadeem Akhtar, Nashez Zubaira, Abhishek Kumara, Tameem Ahmada 2017).

Meal	Quality	Service	Facility	Staff	Room
breakfast	satisfactory	desk	rooftop	good	bed
bar	ample	front	bar	nice	bedroom
drik	hygienic	check-in	spa	polite	dirty
food	proper	check- out	wifi	friendly	clean
spicy	ambliance	reliable	pool	helpful	toilet
tasty	odour	fast	gym	relable	bathroom
buffet	smell	convenient	elevator	quick	shower
restaurant	sound	service	internet		dryer
dinner			floor		fridge
lunch			parking		view
brunch			wireless		water
delicious			broken		

Table5. Manually Predefined Aspects

We classify the sentence into the meal aspect if this sentence includes breakfast.It will be in the service category if it also consists of the service term. Secondly, we apply VADER sentiment analysis. This method will generate four parameters - positive, negative, neutral, and compound. The first three parameters are ratios. Take the first sentence in the table2 for example; this sentence contains 61.8% percentage positive words, 38.2% percentage neutral words. Therefore, its score is 0.9729. We get the score of each sentence, and if the score exceeds 0.05, we will treat it as a positive sentence. If the score of a sentence is less than -0.05, we will treat it as a negative sentence. Others will become a neutral sentence.

new_senc	Tag	Positive	Negative	Neutral	Compound	Result
Great hotel superb Times Square Friendly knowledgeable staff available hours Room clean spacious Particularly enjoyed Friday afternoon resident lobby food entertainment Definitely stay	Room	0.618	0	0.382	0.9729	Positive
desk reception area found connected hotel needs service training staff	Service	0	0.575	0.425	-0.6979	Negative

Table 6. Example Sentences of Sentiment Results

To understand the change of customers' attitudes, we calculate the percentage of negative sentences in each hotel's total review. For example, the number of Hotel New York (H1) Pre-COVID negative review sentences in the Meal aspect is 133, and the total number reviews of Hotel New York (H1) Pre-COVID is 10839. We divide 133 with 10839. We do the same calculation for the positive sentences. For figure 10, light blue shows the Pre-COVID, and dark blue shows the Post-COVID. For figure 11, light red indicates the Pre-COVID, and dark red shows the Post-COVID.

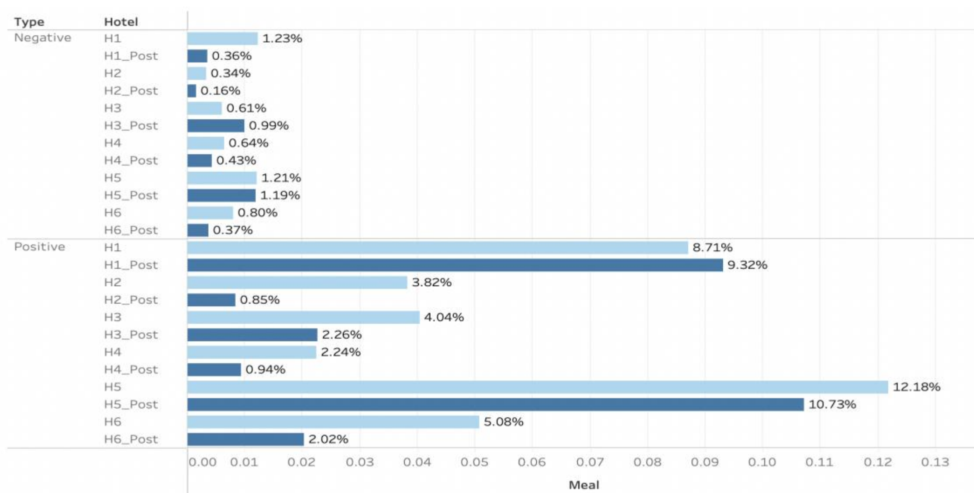


Fig 7. Negative and Positive Sentence Percentage of Meal Aspect

Most of the hotel's negative and positive percentages decrease in the meal aspect. In other words, customers did not complain more but also were not satisfied with the meal. However, Hotel1 is the only hotel whose positive percentage rose and negative share declined, which means that Hotel New York (H1) performs better in the meal aspect than others.

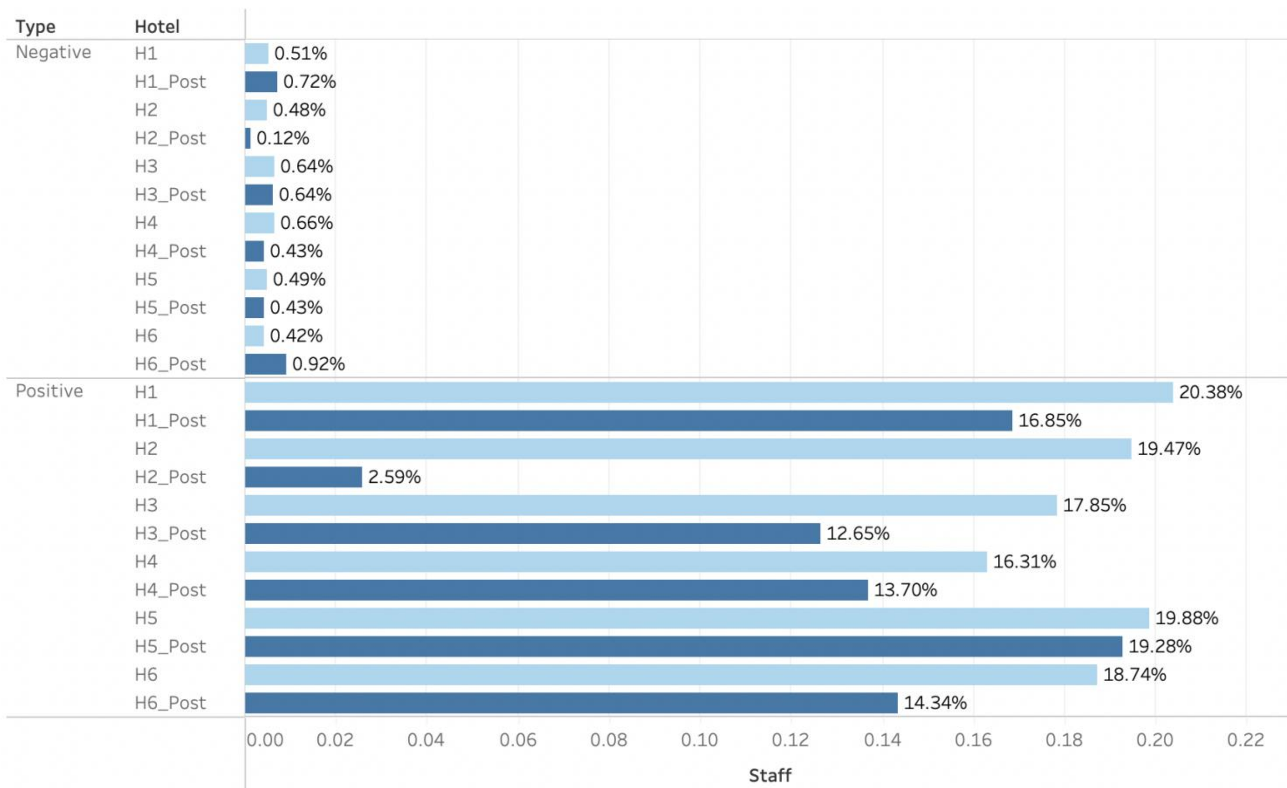


Fig 8. Negative and Positive Sentence Percentage of Staff aspects

According to figure 11, the negative part does not show a significant difference between Pre-COVID and Post-COVID. However, the positive percentage part has a significant decline from Pre-COVID to Post-COVID, especially Edison Hotel (H2), which declines 16.88%.

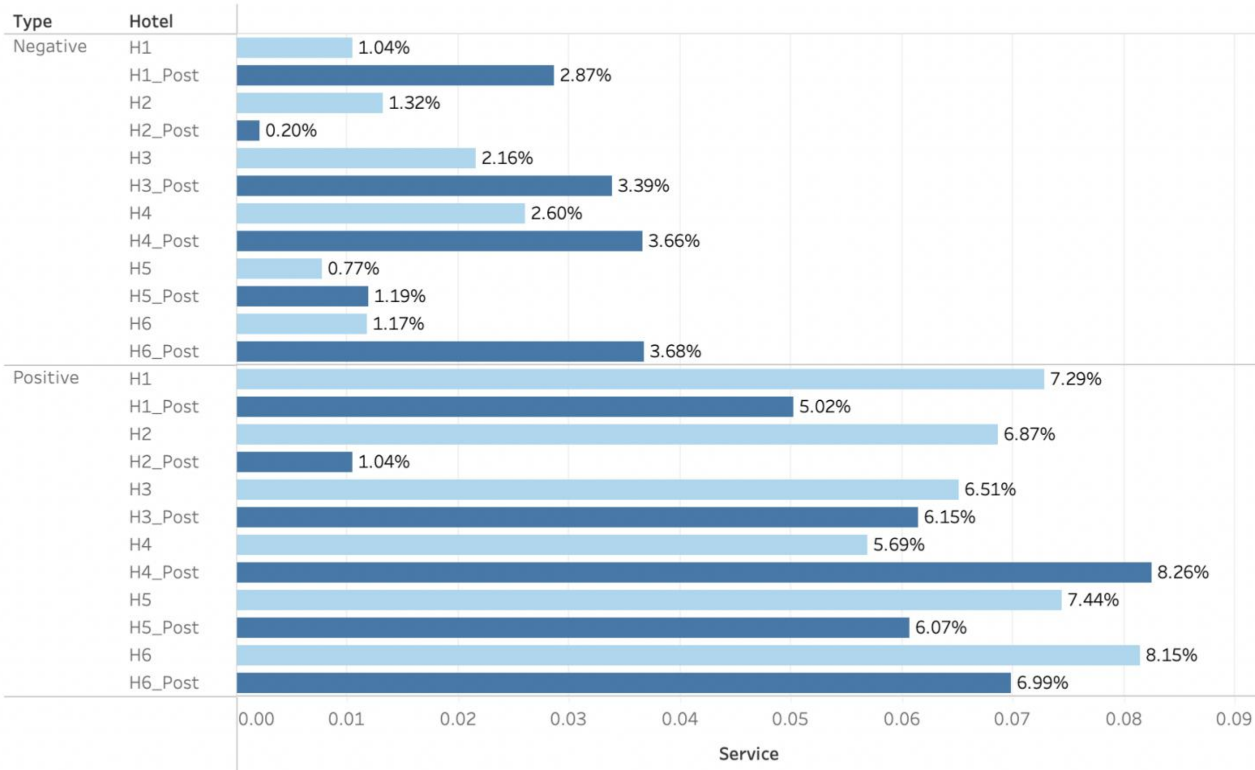


Fig 9. Negative and Positive Sentence Percentage of Service aspects

According to figure 12, customers had a more negative attitude on service during COVID-19, since hotels' negative sentences percentage rose, except for Edison Hotel (H2), and their positive sentences percentage also declined. The increase in negative sentence percentage is more significant than other aspects; many of them increase by over 1%.

Results and Insights

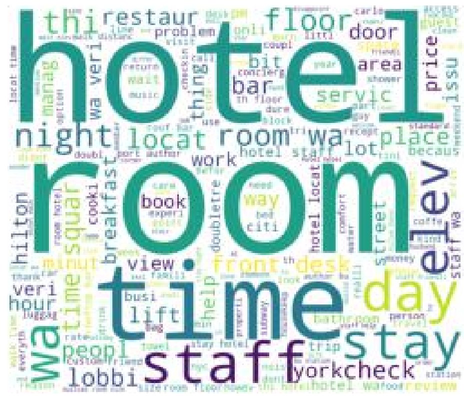
When we take a look into the customers reviews, we find people mentioning breakfast and the elevator for complaining about waiting too long. During the pandemic, fewer customers visit hotels and thus relieve the crowded situation. We believe that negative sentence percentage in meal drops attribute to the drop in crowd.

We originally expected that reviews in Post-COVID would appear in more specific terms, such as mask, sanitary, sanitizer, COVID, disinfect, etc. However, the result does not show the situation, our analysis result only shows [clean] term's TF-IDF rank moves up (Table 4). This indicates that customers take the hotel's precaution of pandemic as granted and hotels do not have any significant reformation operation.

The key to surviving in Post-COVID is to decrease contact with others. 'Contactless' operation is a new era in hotel management. Automatic check-in/check-out, digital payment, voice-controlled elevator and keyless entry not only can reduce the chance to contact but also can enhance the efficiency of operation. It is clear that hospitality was frustrated in the pandemic, and recovery is still not approaching. Nevertheless, now is a good time to update their facilities, to implement new technology and to transform.

Appendix

Appendix A. Hotel 3 Pre-COVID Adj.



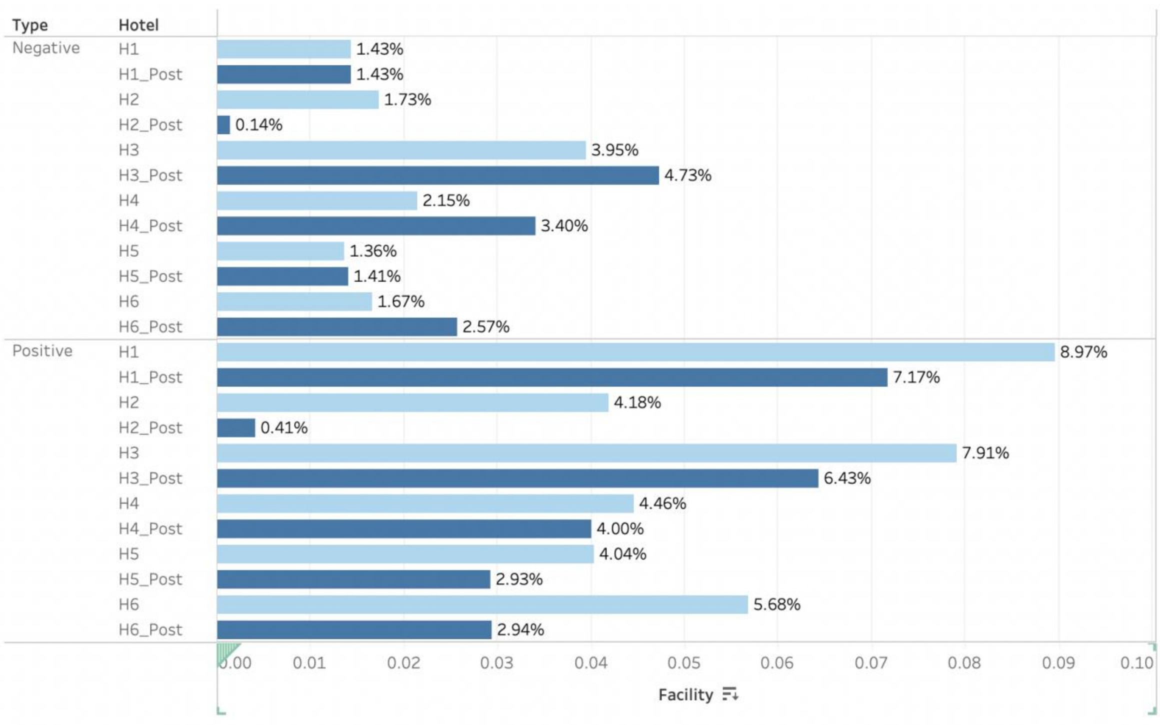
Appendix B. Hotel 3 Pre-COVID Noun

Appendix C. Hotel 3 Pre-COVID Verb

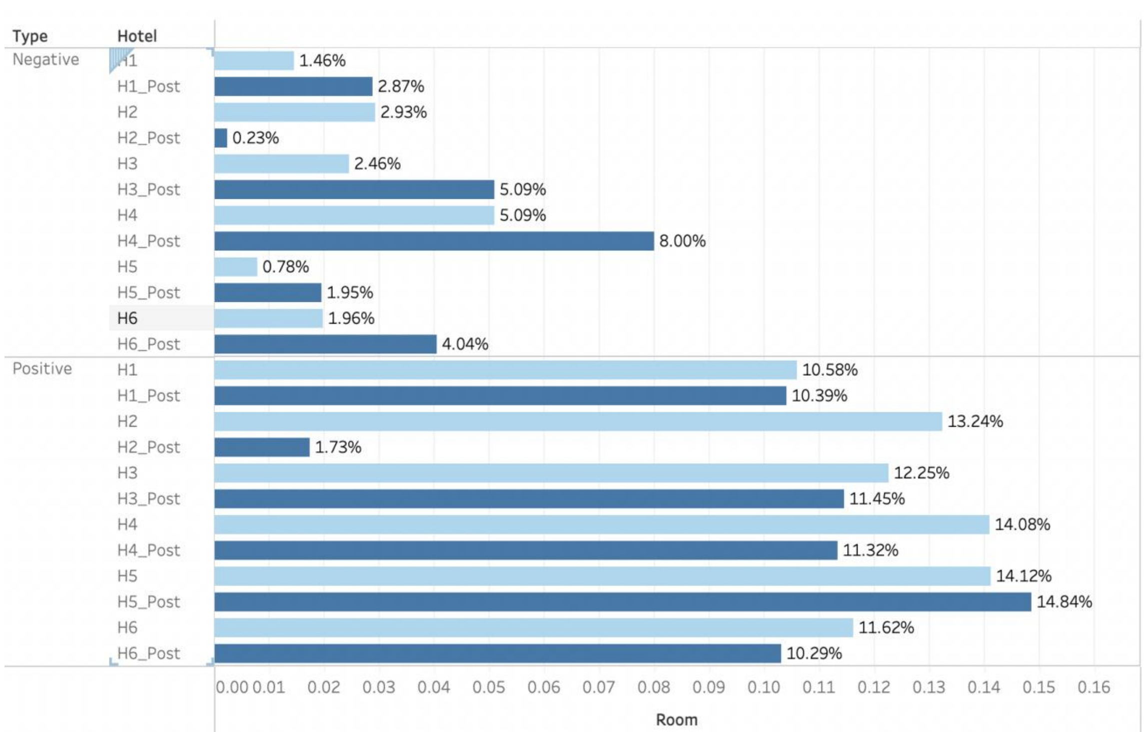
Appendix D. Hotel 4 Pre-COVID Adj.

Appendix E. Hotel 4 Pre-COVID Noun

Appendix F. Hotel 4 Pre-COVID Verb



Appendix G. Negative and Positive Sentence Percentage of Facility aspects



Appendix H. Negative and Positive Sentence Percentage of Room aspects

References

Akhtar, Nadeem, et al. "Aspect based sentiment oriented summarization of hotel reviews." *Procedia computer science* 115 (2017): 563-571.

Bajaj, A., 2021. *VADER Sentiment Analysis | NLP Sentiment Analysis Using VADER*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/> [Accessed 22 December 2021].

Hu, Feng, and Rohit H. Trivedi. "Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content." *International Journal of Hospitality Management 84* (2020): 102317.

Krishnan, Vik, et al. "Hospitality and COVID-19: How long until 'no vacancy'for US hotels?." (2020).

MMA. "Bag-of-Words and TF-IDF Tutorial. " *Mustafa Murat ARAT.*Apr. 3rd, 2020. https://mmuratarat.github.io/2020-04-03/bow_model_tf_idf