A comparative analysis of travelers' online reviews among China, South Korea, and USA using sentiment analysis in the era of the COVID-19 pandemic

Junwoo Hong, Jonggwan Won, and Taeho Hong

School of Business, Pusan National University

22nd International Conference on Electronic Commerce

CONTENTS

1. Introduction

Hotel industry after COVID-19 pandemic outbreak

- As spread of COVID-19 and travel restriction from each government has continued, all industries related to **tourism** including hotels, restaurants, and shopping malls are "**already facing collapse**" or "**in a fight for survival**".(Jiang and Wen, 2020)
- There are differences in the number of COVID-19 confirmed patients by country, which may result in **different effects in each country** at the same time. (Fanelli and Piazza, 2020)

Online review usefulness

- Online reviewers can **share their experiences and emotions** through the reviews written on online platforms, influencing potential consumers' exploration of information and purchase.(Yao et al., 2020)
- Utilizing online reviews written in e-commerce helps **customer make better decisions** such as search and purchase processes.(Cheng and Ho, 2015)

- As the hotel industry has difficulty in making profit after the COVID-19 pandemic outbreak, it is more important to analyze OCRs(online customer reviews) properly to help customers make better decisions and hotels make effective strategies.
- In this study, OCRs are analyzed to extract new insights about hotel industry and the results are compared by period and country.

2. Literature Reviews

LDA(Latent Dirichlet Allocation)

- LDA is a widely used **method of Topic modeling** and by using LDA, topics of reviews can be extracted and keywords can be classified by related topics.(Blei et al., 2003)
- **Finding optimal number of topics is crucial** for performance of LDA.(Zhao et al., 2015)
- The concept of perplexity are used to determine the optimal number of topics by some researchers. And the concept of perplexity combined with 5-fold cross validation can be useful when analyzing the hotel reviews using LDA.(Hong et al., 2019)
- Studies using LDA
  - ❑ Poria et al(2016) applied LDA to extract three perspectives of the hotel industry('location', 'service', 'value').
  - ❑ Priyantina et al.(2019) use LDA to determine the hidden topics of a term list and use this results in predicting sentiment scores.
  - ❑ Taecharungroj and Mathayomchan(2019) to help determine the dimensions of each type of attraction and find four dimensions for beaches and island, three dimensions for pedestrian street and temples and two dimensions for markets.

2. Literature Reviews

Sentiment Analysis

- Sentiment analysis is called opinion mining and it is a method to **extract people's opinions, attitudes, and emotions** created from their personal experiences.

(Liu., 2012)

- There are some kinds of methodologies to calculate sentiment scores such as lexicon based, machine learning based, and hybrid methods.

(Alamoodi et al., 2020)

- In some researches, lexicon based methods were used for calculate sentiment scores because **it is easy to apply and widely used** in other research.

(Ma et al., 2018)

- Sentiment analysis techniques are used to build recommendation system or to find inconsistent reviews
  - ❑ Shafqat & Byun(2020) analyzed tourist reviews and developed a recommendation system that recommends tourist destinations with high ratings and positive sentiment scores.
  - ❑ Yi & Xiaowei(2021) calculated sentiment scores and used scores to find the inconsistent reviews using the method that finds the large gap between star rating and sentiment scores.

3. Research Frameworks

Phase1: Data Collection and Preprocessing

- Source of reviews
  - ❑ Ctrip.com(China)
  - ❑ Tripadvisor.com(South Korea, USA)
  - ❑ Booking.com(South Korea, USA)
- Reviews' places
  - ❑ Bejing, Seoul, New York(City)
  - ❑ Sanya, Jeju island, Hawaii(Attraction)
- Preprocessing process
  - ❑ Removing stopwords
  - ❑ Missing value processing

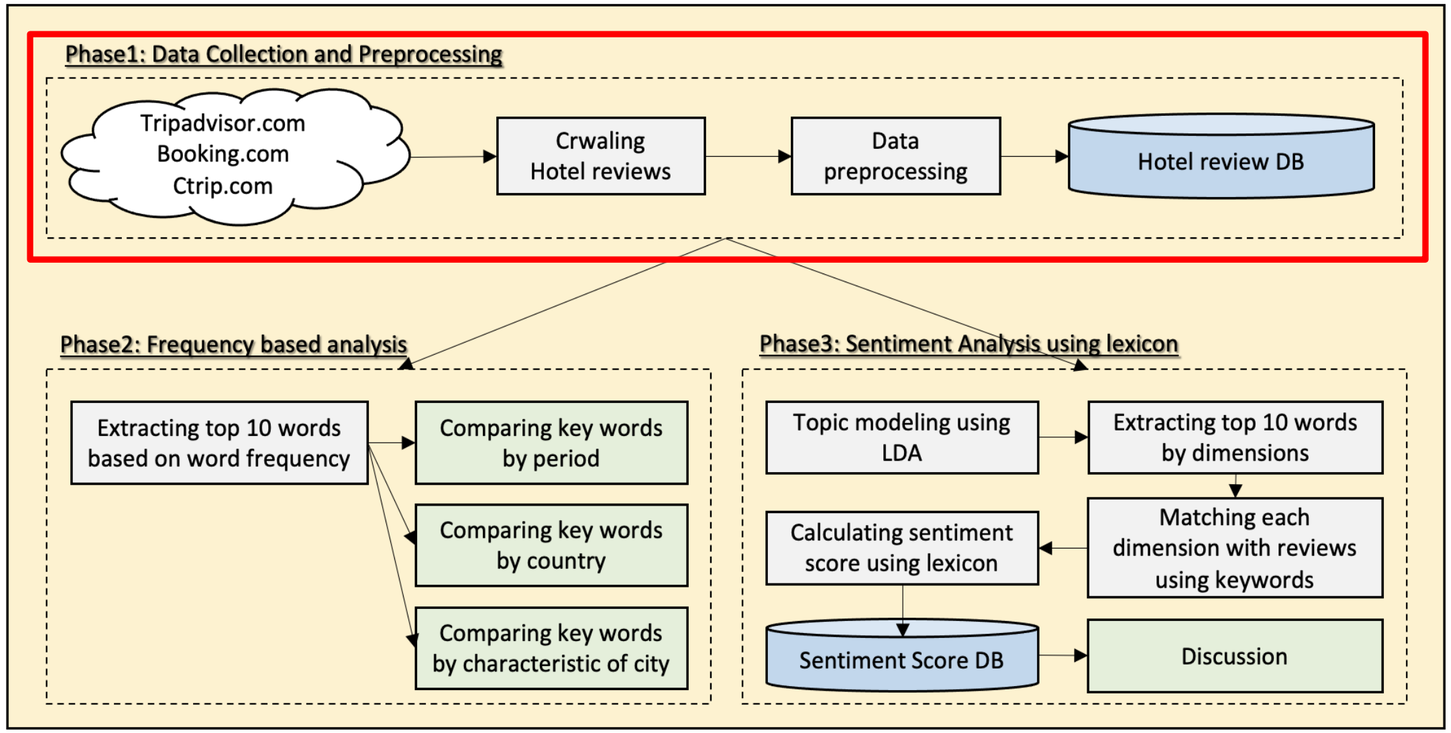

Figure 1. Research Frameworks

3. Research Frameworks

Phase2: Frequency based analysis

- Top 10 key words were extracted by period. (Before / After COVID-19)
- Key words are compared by period, country and characteristics of place.

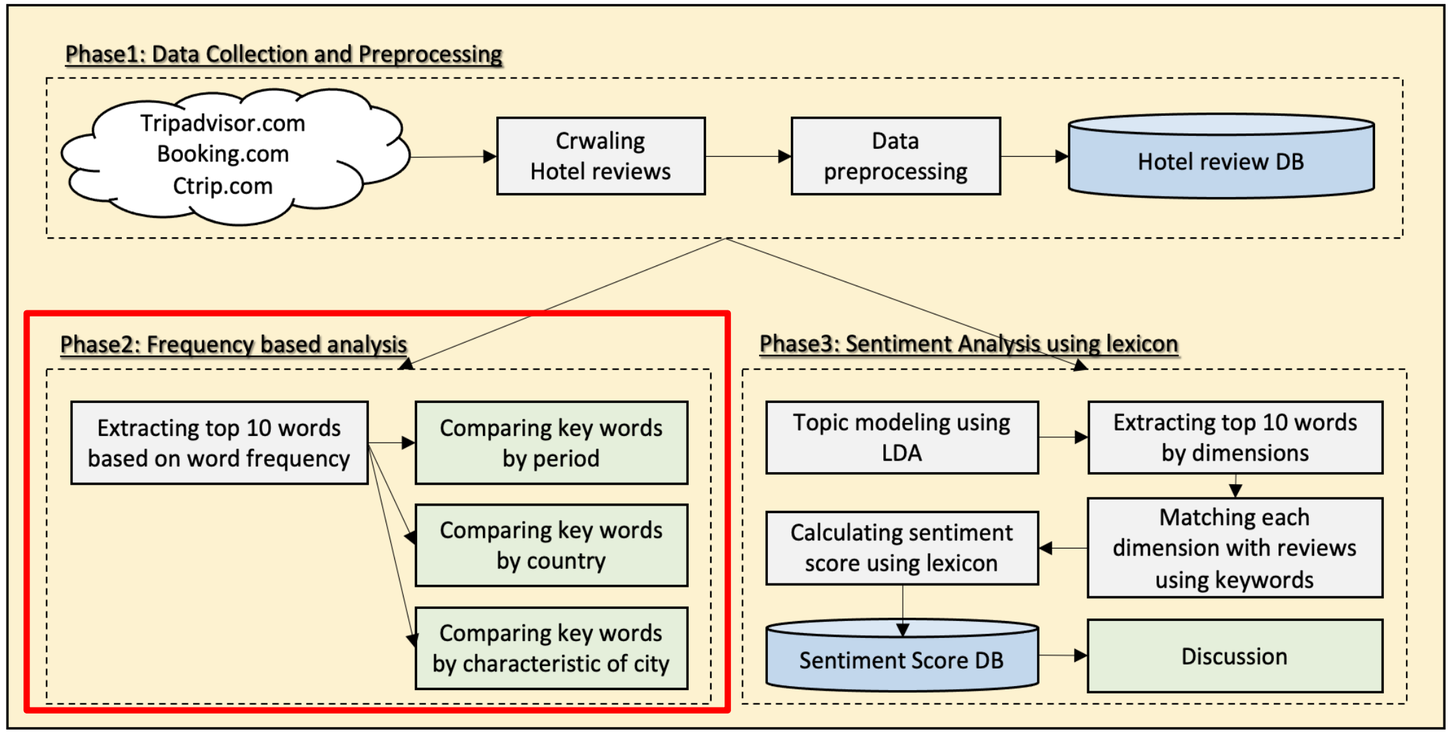

Figure 1. Research Frameworks

3. Research Frameworks

Phase3: Sentiment analysis

- Extract several topics and keywords using LDA.
- Build keywords table by six dimension of hotel.
- Calculate sentiment score using keywords and lexicon.
- Discussion.

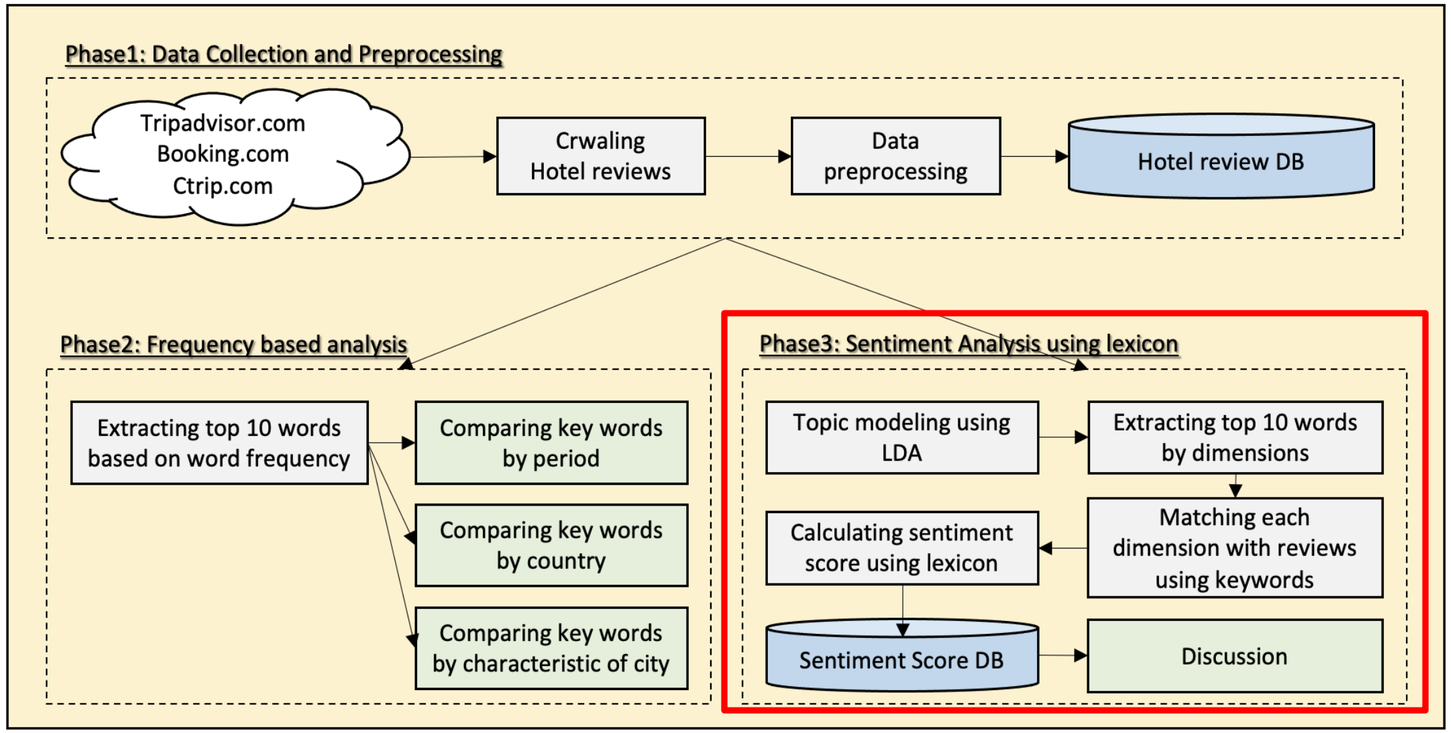

Figure 1. Research Frameworks

4. Experiments and results

4.1 Data description

- "Grey zone" is set between the before and after COVID-19 outbreak because that period could reflect both periods.

- The number of hotel reviews after the COVID-19 has significantly decreased in the USA.
- It could indicate that the number of hotel guests has decreased in the circumstance of national lockdowns and restrictions on human mobility in USA.

(Yi & Xiaowei., 2021)

Table 1. Descriptive statistics of extracted reviews

Country	China				South Korea				USA			
City	Bejing		Sanya		Seoul		Jeju island		New York		Hawaii	
Period	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Average rating	4.74	4.76	4.71	4.68	4.74	4.68	4.57	4.5	4.58	4.34	4.19	4.31
Number of reviews	37,941	19,515	84,165	80,382	1,328	1,597	1,035	1722	9,661	1,691	4,201	572
Total	57,456		164,563		2,925		2,757		11,352		4,773	

*Before: 2019.01.01 – 2019.11.31(before COVID-19); After: 2020.04.01 – 2021.02.28(after COVID-19)

4. Experiments and results

4.2 Frequency based analysis

- Ranking of keyword 'clean' has risen after the COIVD-19 outbreak except Sanya
  - Bejing(9th → 5th)
  - New York(6th → 4th)
  - Hawaii(12th → 7th)
- Keword 'covid' appeared after COVID-19 outbreak in USA
  - New York(8th)
  - Hawaii(6th)
- Ranking Top 3 keywords in each city have changed very little

Table 2. Top 10 keywords extracted from reviews of each city

City	Bejing		Sanya		New York		Hawaii	
Period	Before	After	Before	After	Before	After	Before	After
1	hotel	hotel	hotel	hotel	room	room	room	room
2	service	service	service	service	staff	location	staff	staff
3	room	room	room	room	location	staff	resort	beach
4	breakfast	breakfast	breakfast	breakfast	service	**clean**	pool	view
5	convenience	**clean**	convenience	check in	friendly	breakfast	beach	pool
6	location	convenience	location	convenience	clean	view	ocean	**covid**
7	stay	front desk	stay	stay	time	small	service	**clean**
8	surroundings	location	surroundings	surroundings	helpful	**covid**	view	location
9	clean	surroundings	**clean**	front desk	breakfast	time	property	ocean
10	check in	check in	check in	swimming pool	comfortable	friendly	time	service

4. Experiments and results

4.3 Sentiment Analysis - LDA

- First, LDA is applied to out dataset to extract keywords for six dimensions.
- We calculated the perplexity by the number of topics combined with 5-fold cross-validation of Hawaii dataset.
- The number of topics can be selected where the slope of the graph becomes smoothed.

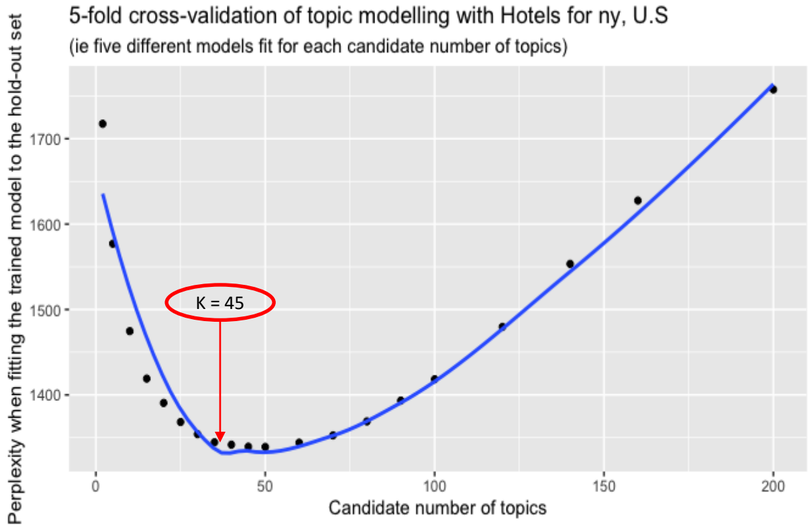

Figure 2. Perplexity with 5-fold cross validation(New York)

Table 3. The number of Topics for each city

Country	China		USA.	
City	Bejing	Sanya	New York	Hawaii
Number of topics	40	50	45	30

4. Experiments and results

4.3 Sentiment Analysis – Keywords extraction

- Key words for each topic can be extracted by applying LDA on datasets.
- We can name each topic using top 10 keywords and then we summarized the key words by six dimensions(value, rooms, location, cleanliness, sleep quality, service)
- After keywords are distributed to each dimension, Each country's keywords table was used to calculate sentiment score by dimension when we find the corresponding sentence of each dimension.

Table 4. Topic and words extraction results (Hawaii of USA)

Topic ID	Dimension	Topic	Top 10 words
1	service	meal	breakfast, free, buffet, bar, drinks, restaurant, food, dinner, live
2	room	room	room(s), balcony, view(s), door, outdated, bed, pool, bedroom
	…	…	…
29	Sleep quality	noise	relax(ing), hear, loud, quiet, close, air, partial, noise, night
30	location	accessibility	parking, walk, easy, drive, far, bay, short, main, town, street

Table 5. Key words by dimension(USA)

City	Dimension	Key words
Hawaii	value	free, easy, worth, fee(s), value, price, pay, extra, rental
	room	sunset(s), warm, view(s), room(s), small, space, pool(s), chair(s), facilities, outdated, suite, old, options, main, private, space, daily
	location	time, spent, location, close, park, easy, access, oceanfront, rays, bay, sunsets, forward, construction, looking, seaside
	cleanliness	covid, clean, dirty, negative, bad, cleaning, evening, dirt, issue, full
	sleep quality	relax(ing), hear, loud, quiet, close, air, partial, noise, night
	service	lobby, reviews, buffet, bar, drinks, happy, restaurant(s), food, dinner, live, breakfast, spa, service(s), amenities, accommodations, options, front desk, people, help, care, housekeeping, shuttle, check, welcoming, professional, housekeeping, concierge, attentive, busy, enjoy, kids, children, families, available, taken, working, quick, job

4. Experiments and results

4.3 Sentiment Analysis

- The sentiment scores of 'cleanliness' of each city are different by the characteristic of city. The scores in tourist attraction has increased after the COVID-19 pandemic declaration, and in one of the biggest cities in each country, the scores have decreased.
- Due to cross-border travel restrictions, travelers were unable to go to tourist attractions such as sanya and hawaii in the early stages of the COVID-19 outbreak. So people vaccinated or have been able to travel were more satisfied to the things related to tourism.

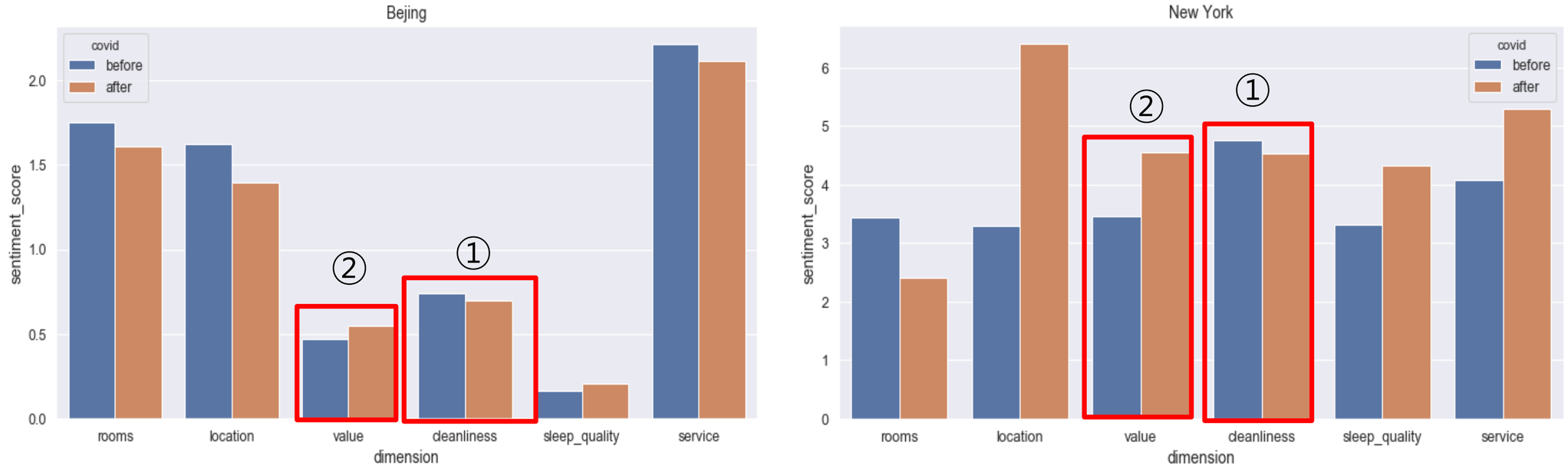

Figure 3. Sentiment scores of Cities by dimension(left: Bejing, right: Sanya)

4. Experiments and results

4.3 Sentiment Analysis

- The sentiment scores of 'value' have increased after the declaration of a pandemic, regardless of the characteristics of the place or country.
- Some hotels use sales promotion through discount vouchers/deals to overcome the crisis by the COVID-19. This strategy could have influenced the sentiment scores of 'value'.

(Le & Phi, 2021)

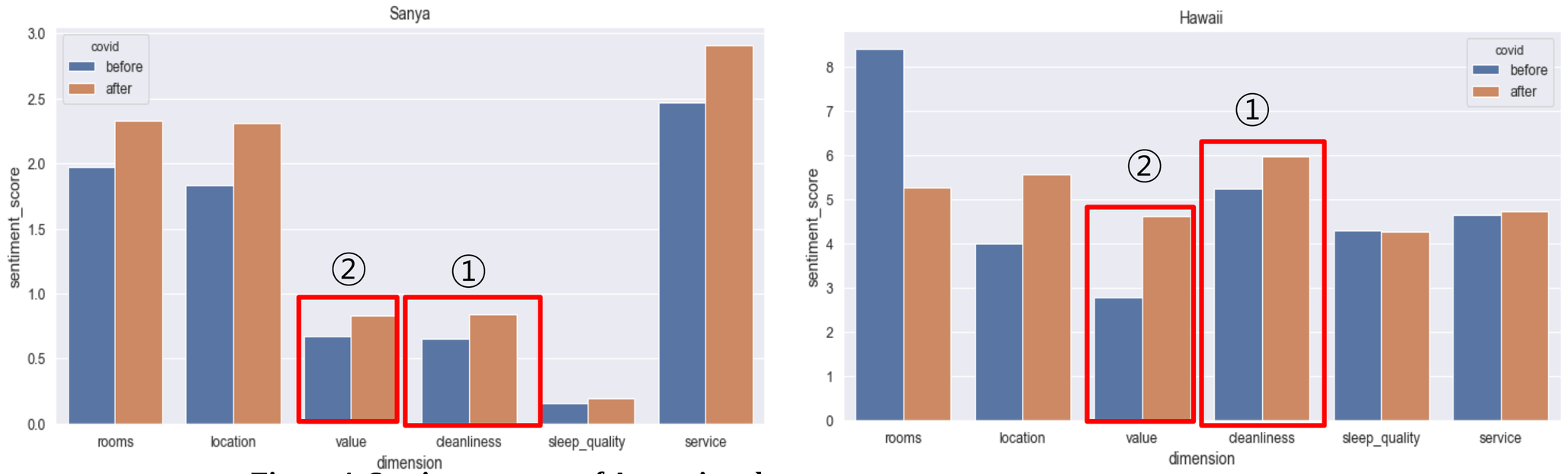

Figure 4. Sentiment scores of Attractions by dimension(left: New York, right: Hawaii)

5. Conclusions

Contribution

- This study comparatively analyzed online hotel reviews of China, South Korea, and USA using frequency based and sentiment analysis techniques.
- Top 10 most frequently used words in OCRs were presented, and those words could be used to build proper strategy for hotel industry.
- Sentiment analysis technique was applied to our datasets and sentiment scores are compared by country, period, and characteristics of place. With this method, hotels can find new insights and proper strategies after the COVID-19 pandemic outbreak.

Further research direction

- OCRs of South Korea were not analyzed because there is few trustworthy stopword dictionary or lexicon to be applied to sentiment analysis. So in the further research, we have to find the method effectively treat the stopwords of South Korea and the method to calculate sentiment scores.
- Sentiment scores are calculated by lexicon based method. One of the disadvantage of using lexicon on sentiment analysis is that lexicon sometimes could not reflect the specific domain(Ma et al., 2018). So we have to apply various methods such as machine learning methods or hybrid methods.

22nd International Conference on Electronic Commerce

Thank you for Listening