CLASSIFYING 5-STAR HOTEL GUEST REVIEWS USING SENTIMENT ANALYSIS

Presented by: Gui Miow Wan (Mandy)

OUTLINE

- Data Description
- Data Exploration
- Data Cleaning and Pre-processing
- Sentiment Classification
- Feature Selection
- Data Modeling and Results
- Model Evaluation
- Feature Processing

DATA DESCRIPTION

Website Name	Tripadvisor.com
Source URL	https://www.tripadvisor.com.my/Hotels-g298570-zfc5-a_sort.POPULARITY-Kuala_Lumpur_Wilayah_Persekutuan-Hotels.html
Date Accessed	January 2023
Method	Python Requests Library

64583 rows of records
22 attributes

Attributes	Description
Id	Id of the review
hotelname	Name of the Hotel
comment_count	Total count of reviews received to date by the hotel.
rating	Rating number of the hotel, ranging from 1 to 5
rating_word	Rating in word format of the hotel
location_rating	Overall rating for hotel location by reviewer, ranging from 1 to 5
cleanliness_rating	Overall rating for hotel cleanliness by the reviewer, ranging from 1 to 5
service_rating	Overall rating for hotel service by the reviewer, ranging from 1 to 5
value_rating	Overall rating for hotel value by the reviewer, ranging from 1 to 5
Username	Login username of the reviewer
guestOriginLocation	The origin country of the reviewer
contributes	Total count of review contributions by the reviewer on the platform to date
helpfulvotes_guest	Total reviews of the reviewer flag as helpful by consumer
guest_rating_bubble	Overall rating by reviewer, ranging from 1 to 5
review_title	Review title by reviewer
review	Review by reviewer
dateofStay	Date of stay of the reviewer
trip_type	Type of trip of the reviewer when staying in the hotel
helpfulvotes	Total count of votes the review flag as helpful by consumer
rooms_rate	Rooms rating by the reviewer, ranging from 1 to 5
cleanliness_rate	Cleanliness rating by the reviewer, ranging from 1 to 5
service_rate	Service rating by the reviewer, ranging from 1 to 5

DATA EXPLORATION

Attributes	Data Type	Measurement variables	Non-Null Values counts	Missing Values (%)
id	int64	Nominal	64583	-
hotelname	Object	Nominal	64583	-
comment_count	Object	Discrete	64583	-
rating	float64	Ordinal	64583	-
rating_word	Object	Ordinal	64583	-
location_rating	Object	Ordinal	64583	-
cleanliness_rating	Object	Ordinal	64583	-
service_rating	Object	Ordinal	64583	-
value_rating	Object	Ordinal	64583	-
Username	Object	Nominal	64583	-
guestCountryOfOrigin	Object	Nominal	52637	18.47
contributes	float64	Discrete	61378	4.97
helpfulvotes_guest	float64	Discrete	49345	23.58
guest_rating_bubble	int64	Ordinal	64583	-
review_title	Object	Nominal	64583	-
Review	Object	Nominal	64583	-
dateofStay	Object	Nominal	64450	0.21
trip_type	Object	Nominal	215	99.67
helpfulvotes	float64	Discrete	8545	86.75
rooms_rate	Object	Ordinal	134	99.79
cleanliness_rate	Object	Ordinal	132	99.80
service_rate	Object	Ordinal	128	99.80

DATA CLEANING AND PRE-PROCESSING

Merge *review_title* and *review* columns

Remove non-English reviews

- 503 reviews written in non-English languages.

Remove reviews of equal or less than 5 words

Drop high null value count column

- 5 attributes ***trip_type***, ***helpfulvotes***, ***rooms_rate***, ***cleanliness_rate***, and ***service_rate***, with more than 85% of the missing value

DATA CLEANING AND PRE-PROCESSING

Imputation of the guest country of origin

- "pycountry" library to extract Country name only and store in a new column ***guestCountry***.

Imputation: rating

- The 4 attributes ***location_rating, cleanliness_rating, service_rating, value_rating***, and ***guest_rating_bubble*** format is difficult to interpret.
- Example: ui_bubble_ratingbubble_50 (Equivalent to rating 5.0)

Imputation of missing value

- ***guestCountry*** imputed with "Not Mentioned".
- ***helpfulvotes_guest***, ***contributes***, and ***dateofStay*** being filled with the number zero.

Drop attributes not in used

- The attributes ***username*** and ***guestCountryOfOrigin*** were being dropped

DATA CLEANING AND PRE-PROCESSING

Convert Data Type

- The attribute ***comment_count*** consists of "," were removed.
- Data type is converted from object to integer.

Label Encoding

- ***Hotelname, guestCountry***
- ***rating_word****: 'Excellent'* → *'2' and 'Very good'* → *'1'*
- ***dateofStay****: e.g.* January 2021→ "012021:

Text Pre-processing

- Lowercasing
- Remove punctuation
- Remove numbers
- Tokenization
- Remove stop words
- Lemmatization

SENTIMENT CLASSIFICATION

Method	Dataset Name
VADER Sentiment Analysis	VADER dataset
RoBERTa Pretrained Model	RoBERTa dataset

Sentiment Score Threshold	Sentiment
>= 0.05	Positive
In between -0.05 and 0.05	Neutral
<= -0.05	Negative

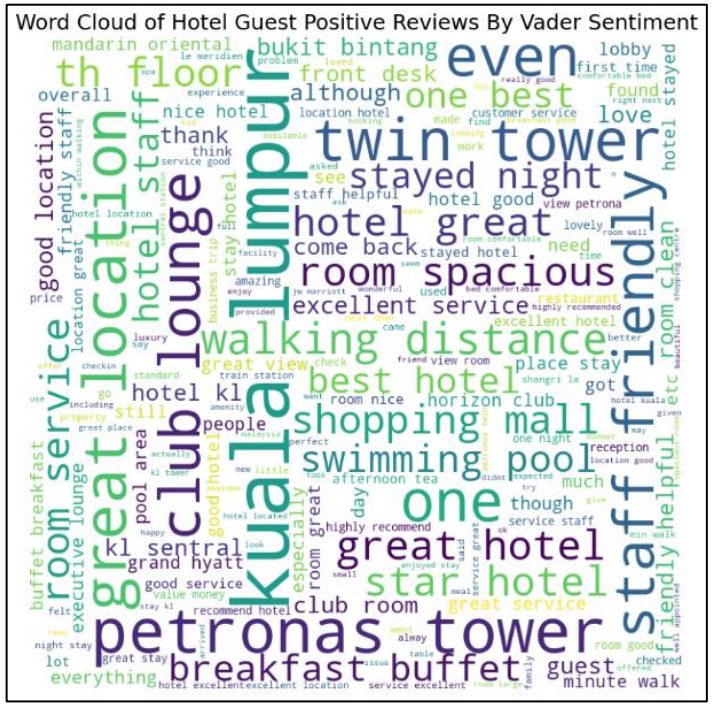
SENTIMENT CLASSIFICATION

Vader Sentiment Analysis

	vader_neg	vader_neu	vader_pos	vader_compound	polarity_vader
55890	0.043	0.621	0.336	0.9850	positive
56934	0.027	0.677	0.296	0.9782	positive
47966	0.180	0.735	0.085	-0.5198	negative
57437	0.083	0.714	0.204	0.8885	positive
57239	0.025	0.624	0.351	0.9872	positive

Sample Output of VADER Sentiment Analysis

Vader Polarity	Count
Positive	60969
Negative	2360
neutral	223



Word Cloud of Positive Review

Word Cloud of Negative Review

SENTIMENT CLASSIFICATION

RoBERTa Pretrained Model

$$Score\ Difference = Positive\ Score - Negative\ Score$$

$$Combined\ Score = Score\ Difference * (1 - Neutral\ Score)$$

	roberta_neg	roberta_neu	roberta_pos	polarity_roberta
11580	0.003547	0.077429	0.919024	positive
56417	0.002043	0.009727	0.988231	positive
1644	0.002357	0.016116	0.981527	positive
48697	0.128247	0.353720	0.518033	positive
55228	0.576567	0.343065	0.080369	negative

Sample Output of RoBERTa Pretrained Model

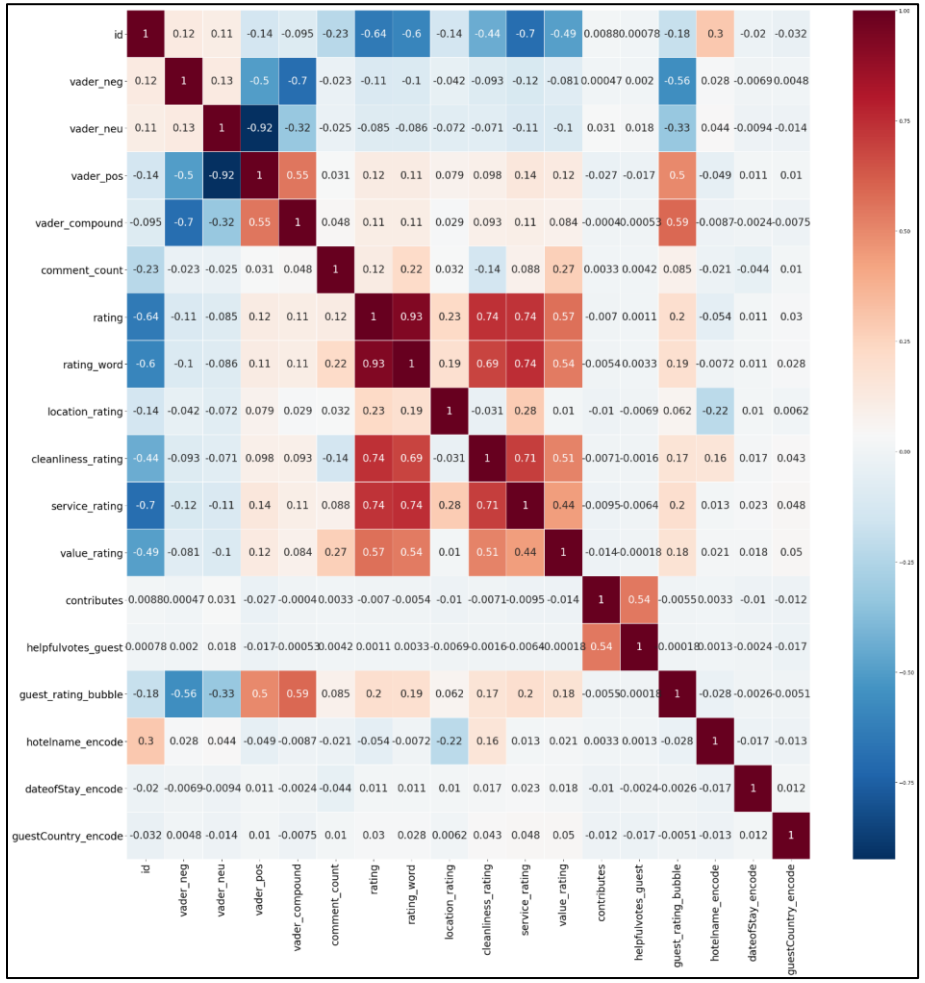
RoBERTa Polarity	Count
positive	58520
negative	3863
neutral	1169

Word Cloud of Positive Review

Word Cloud of Negative Review

FEATURE SELECTION - CORRELATION MATRIX HEATMAP

Vader Dataset



Attributes with Correlation > 0.8

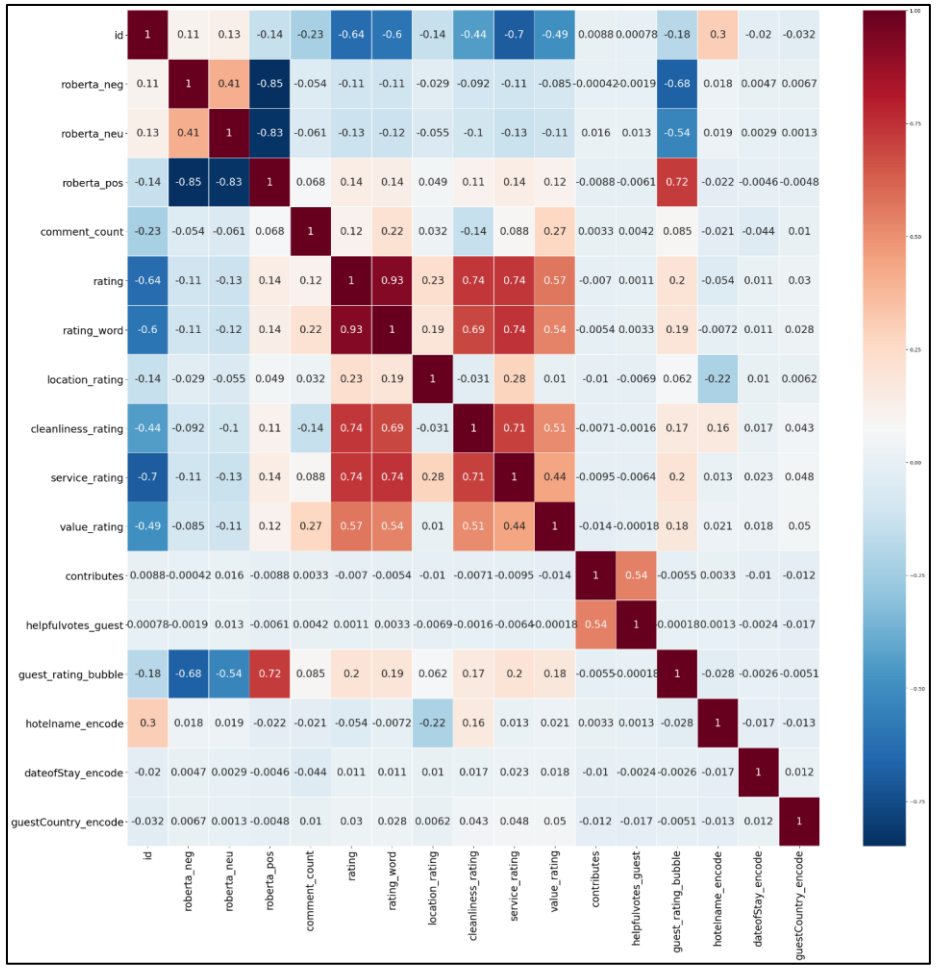
Attributes	Correlation
rating	0.9266
rating_word	0.9266
vader_neu	0.9243
vader_pos	0.9243

Attributes
hotelname_encode
comment_count
rating
location_rating
cleanliness_rating
service_rating
value_rating
contributes
helpfulvotes_guest
guest_rating_bubble
dateofStay_encode
guestCountry_encode
polarity_vader

The final dataset of VADER, named *vader_dataset_ml*

FEATURE SELECTION - CORRELATION MATRIX HEATMAP

RoBERTa Dataset



Attributes with Correlation > 0.8

Attributes	Correlation
rating	0.9266
rating_word	0.9266
roberta_neg	0.8493
roberta_pos	0.8493
roberta_neu	0.8301

Attributes
hotelname_encode
comment_count
rating
location_rating
cleanliness_rating
service_rating
value_rating
contributes
helpfulvotes_guest
guest_rating_bubble
dateofStay_encode
guestCountry_encode
polarity_roberta

The final dataset of RoBERTa, named *roberta_dataset_ml*

DATA MODELLING AND RESULTS

		VADER			RoBERTa		
Model		Random Forest	Decision Tree	Support Vector Machine (SGD)	Random Forest	Decision Tree	Support Vector Machine (SGD)
Accuracy	Train	0.99	0.99	0.96	0.99	0.99	0.95
	Valid	0.96	0.94	0.96	0.94	0.91	0.95
F1 Score	Train	0.99	0.99	0.96	0.99	0.99	0.93
	Valid	0.95	0.94	0.96	0.93	0.91	0.93
ROC AUC	Train	0.99	0.99	0.92	0.99	0.99	0.92
	Valid	0.90	0.67	0.92	0.91	0.73	0.92
Recall	Train	0.99	0.99	0.96	0.99	0.99	0.95
	Valid	0.96	0.94	0.96	0.94	0.91	0.95
Precision	Train	0.99	0.99	0.95	0.99	0.99	0.93
	Valid	0.95	0.94	0.95	0.93	0.91	0.93

MODEL EVALUATION

- Using 15% of the test set from the VADER dataset

	Accuracy	F1 Score	ROC AUC	Recall	Precision
Test Data	0.96	0.96	0.85	0.96	0.95

FEATURE PROCESSING – TF-IDF

✓ Using VADER dataset

- IDF Value for Top 10 Feature Words in negative reviews

Feature Words	IDF Value
room	143.73
hotel	130.29
service	79.57
staff	65.85
stay	59.80
time	57.43
check	52.70
bad	52.14
breakfast	51.23
star	50.67

FEATURE PROCESSING – TF-IDF

- IDF Value for Top 5 Feature Words in negative reviews for each hotel

Feature Words	IDF Value	Feature Words	IDF Value	Feature Words	IDF Value	Feature Words	IDF Value
8 Kia Peng Suites		Alila Bangsar Kuala Lumpur		Four Seasons Hotel Kuala Lumpur		Grand Hyatt Kuala Lumpur	
refund	0.54	room	2.44	room	2.18	room	7.97
money	0.40	hotel	1.83	hotel	1.93	hotel	6.01
pm	0.39	check	1.39	season	1.19	service	4.36
small	0.34	pm	1.31	day	1.18	staff	4.10
quite	0.31	service	1.15	service	1.15	check	3.69
Ascott Kuala Lumpur		Banyan Tree Kuala Lumpur		Grand Millennium Hotel Kuala Lumpur		Hilton Kuala Lumpur	
room	3.90	hotel	1.51	hotel	12.88	room	12.41
service	2.91	room	1.49	room	12.62	hotel	9.02
hotel	2.49	tea	1.18	service	7.28	service	6.80
time	2.21	tree	1.11	check	7.09	hilton	6.66
bad	2.15	banyan	1.11	staff	7.05	time	5.68
E&O Residences Kuala Lumpur		EQ Kuala Lumpur		InterContinental Kuala Lumpur, an IHG Hotel		JW Marriott Hotel Kuala Lumpur	
room	3.43	hotel	0.56	hotel	7.94	room	8.20
bad	2.25	blue	0.38	room	6.95	hotel	7.37
hotel	1.94	booked	0.38	service	4.43	service	4.17
dirty	1.78	sky	0.38	staff	3.60	staff	4.07
apartment	1.51	kl	0.32	star	3.22	marriott	3.94

FEATURE PROCESSING – TF-IDF

- IDF Value for Top 5 Feature Words in negative reviews for each hotel

Feature Words	IDF Value	Feature Words	IDF Value	Feature Words	IDF Value	Feature Words	IDF Value
Lanson Place Bukit Ceylon Kuala Lumpur		Le Meridien Kuala Lumpur		Renaissance Kuala Lumpur Hotel & Convention Centre		Royale Chulan Kuala Lumpur	
room	1.20	room	8.00	room	20.15	room	8.81
place	0.83	hotel	7.82	hotel	17.73	hotel	8.18
stay	0.68	staff	4.03	service	9.61	service	4.37
bad	0.55	service	3.82	wing	9.01	bad	3.86
smell	0.53	stay	3.82	staff	7.65	staff	3.53
Mandarin Oriental, Kuala Lumpur		Pavilion Hotel Kuala Lumpur Managed by Banyan Tree Kuala Lumpur		Shangri-La Kuala Lumpur		Sofitel Kuala Lumpur Damansara	
room	10.50	hotel	3.20	room	11.36	room	4.49
hotel	10.13	room	3.11	hotel	10.08	hotel	3.34
service	7.45	service	1.47	service	6.15	service	2.59
stay	5.56	stay	1.25	stay	5.54	staff	2.00
mandarin	5.15	experience	1.21	staff	5.02	sofitel	2.00
Pullman Kuala Lumpur Bangsar		Pullman Kuala Lumpur City Centre Hotel & Residences		The Face Suites		The Gardens-A St Giles Signature Hotel & Residence	
hotel	6.89	room	2.06	pool	2.61	hotel	8.90
room	6.07	hotel	1.59	room	2.43	room	7.23
service	3.04	breakfast	1.52	check	2.12	service	4.12
staff	2.75	staff	1.23	service	1.96	mall	3.73
stay	2.62	night	1.08	hotel	1.77	bad	3.42

FEATURE PROCESSING – TF-IDF

- IDF Value for Top 5 Feature Words in negative reviews for each hotel

Feature Words	IDF Value	Feature Words	IDF Value
The Majestic Hotel Kuala Lumpur		The Ritz-Carlton, Kuala Lumpur	
hotel	6.06	hotel	4.25
room	5.87	room	3.92
tea	4.28	service	3.24
service	4.12	ritz	3.02
food	3.28	time	2.04
The RuMa Hotel and Residences		The St. Regis Kuala Lumpur	
hotel	2.32	hotel	1.38
room	2.07	service	1.27
time	1.97	room	1.21
staying	1.84	st	1.00
food	1.71	regis	0.99
The Westin Kuala Lumpur		W Hotels Kuala Lumpur	
room	12.56	customer	0.86
hotel	9.56	hotel	0.69
westin	6.00	service	0.68
staff	5.32	room	0.63
time	5.27	worst	0.51

THANK YOU