Text Mining and Sentiment Analysis for examining factors of hotel guest experience and satisfaction in Wildlife tourism

Presenter: Manisha Panta and Md Wasi Ul Kabir
Advisor: Professor Shaikh Arifuzzaman

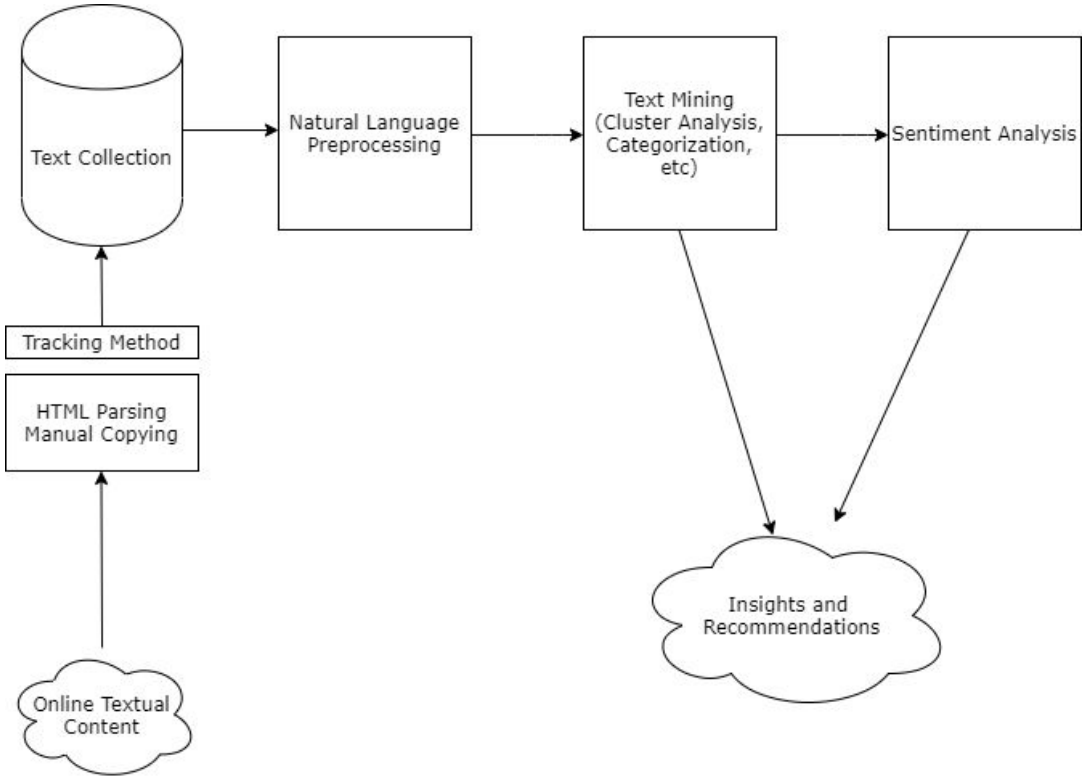
Research Questions

- What is the underlying structure of hotel guest experience represented in customer reviews?
- **What are the factors affecting guest satisfaction in hotels located in and around National Parks?**

Technical Challenges

- Identifying on the basis of how we should select national parks in South East Asian region.
- Unavailability of all data required for better predicting the guest satisfaction (eg. Demographic data of reviewers)
- To collect large amount to data by scraping from Tripadvisor website.
- Text Mining took a lot of time and effort.
- Select relatable words to our study by a domain experts took a lot of time. - Iterative coding process

Implementation



Solution Strategy

Language : Python

Web Scraping : BeautifulSoup

Natural Language Processing : NLTK package

Statistical Analysis tool : SPSS

Machine Learning Algorithm: Logistic Regression

Platform: Hoque Servers, smcluster

Dataset

- Collected data from four South East Asian region (Bangladesh, India, Nepal and Sri Lanka)
- Focused on popular wildlife tourist destination site, rated most visited in government site.
- Title of the review, review body, date of the review and overall rating are extracted from TripAdvisor website.

Country	National Park	Number of Hotels	Total Number of Reviews
Bangladesh	Lawachara National Park	13	4398
	Bangabandhu Safari Park	17	
India	Ranthambore National Park	25	23842
	Jim Corbett National Park	53	
Nepal	Bardia National Park	20	4613
	Chitwan National Park	44	
Srilanka	Yala National Park	44	21350
	Udawalawe National Park	50	
Total		266	54203

Data Preprocessing

Tokenization - process of splitting reviews into words

Removed **Punctuations**

Removed **Stop Words** (pronouns, adverbs, and conjunctions)

Checked Spellings

Lemmatization - remove inflectional endings and return the base or dictionary form of a word, which is known as the lemma

Word frequency

Total Words: **959344**
Selected Words: **538164**

Cut-off point for word selection: **0.03** freq. Per review

Rank	Words	Freq.	Pct.	Freq. per Review	Rank	Word	Freq.	Pct.	Freq. per Review
1	room	45208	4.71	0.83	41	leopard	3845	0.40	0.07
2	staff	39636	4.13	0.73	42	tea	3828	0.40	0.07
3	food	39526	4.12	0.73	43	city	3693	0.38	0.07
4	service	25327	2.64	0.47	44	lodge	3648	0.38	0.07
5	safari	21838	2.28	0.40	45	reception	3589	0.37	0.07
6	experience	17596	1.83	0.32	46	lunch	3500	0.36	0.06
7	pool	16164	1.68	0.30	47	chef	3366	0.35	0.06
8	restaurant	14295	1.49	0.26	48	courteous	3309	0.34	0.06
9	jungle	12845	1.34	0.24	49	road	3109	0.32	0.06
10	friendly	12466	1.30	0.23	50	luxury	2875	0.30	0.05
11	park	12250	1.28	0.23	51	game	2854	0.30	0.05

What is the underlying structure of hotel guest experience represented in customer reviews?

Exploratory Factor Analysis (EFA) :

Exploratory Factor Analysis is a statistical method used to uncover the underlying structure of a relatively large set of variables.

Final words : 80

Result = 10 factors

Name **(Eigenvalue, % Variance)**	Words
Guestroom (C1) **(5.66, 7.075)**	shower water toilet bathroom
Guide (C2) **(3.238, 4.047)**	guide knowledgeable
View (C3) **(1.87, 2.015**	river view
Coffee Shop (C4) **(1.612, 2.015)**	coffee shop
Employee Interaction (C5) **(1.564, 1.955)**	staff helpful friendly
Food Service (C6) **(1.413, 1.766)**	dinner breakfast lunch buffets
Tiger (C7) **(1.413, 1.766)**	tiger
Hotel Amenities (C8) **(1.340. 1.675)**	bar pool
Food (C9) **(1.304, 1.630)**	food
Hotel Attribute (C10) **(1.243, 1.553)**	airport city

What are the factors affecting guest satisfaction in hotels located in and around National Parks?

We used Regression analysis to identify the factors affecting hotel guest satisfaction

Logistic Regression Results

Model summary using only factor score variables

Coefficients[a]

Constants	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
C1	-0.173	0.005	-0.163	-37.041	0.000
C2	0.096	0.008	0.053	12.182	0.000
C3	0.052	0.005	0.041	9.732	0.000
C4	-0.056	0.010	-0.024	-5.582	0.000
C5	0.034	0.003	0.046	10.807	0.000
C6	-0.071	0.004	-0.085	-19.211	0.000
C7	0.040	0.006	0.027	6.321	0.000
C8	-0.025	0.005	-0.022	-5.018	0.000
C9	0.011	0.005	0.010	2.318	0.020
C10	-0.089	0.009	-0.041	-9.779	0.000

Model summary using only sentiment score variables

Coefficients[a]

Constants	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
C1_Pos	0.121	0.035	0.013	3.450	0.001
C1_Neg	-4.370	0.077	-0.216	-56.420	0.000
C2_Pos	0.684	0.044	0.059	15.476	0.000
C2_Neg	-2.690	0.260	-0.040	-10.344	0.000
C3_Pos	0.454	0.029	0.058	15.469	0.000
C3_Neg	-2.647	0.188	-0.053	-14.109	0.000
C4_Pos	0.442	0.070	0.024	6.282	0.000
C4_Neg	-1.986	0.217	-0.035	-9.132	0.000
C5_Pos	0.418	0.012	0.129	34.569	0.000
C5_Neg	-4.329	0.071	-0.229	-60.795	0.000
C6_Pos	0.237	0.028	0.032	8.559	0.000
C6_Neg	-3.537	0.105	-0.128	-33.805	0.000
C7_Pos	0.333	0.054	0.024	6.221	0.000
C7_Neg	0.086	0.199	0.002	0.431	0.667
C8_Pos	0.167	0.024	0.026	6.826	0.000
C8_Neg	-3.135	0.118	-0.100	-26.523	0.000
C9_Pos	0.429	0.016	0.102	27.387	0.000
C9_Neg	-3.386	0.074	-0.172	-45.738	0.000
C10_Pos	-0.087	0.070	-0.005	-1.240	0.215
C10_Neg	-1.570	0.214	-0.028	-7.345	0.000

Logistic Regression Results

Model summary using both factor score and sentiment score variables

Coefficients[a]

Constants	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
C1	-0.058	0.005	-0.055	-11.567	0.000
C2	0.070	0.010	0.038	7.023	0.000
C3	0.020	0.007	0.016	2.946	0.003
C4	-0.093	0.012	-0.040	-7.686	0.000
C5	0.002	0.004	0.003	0.467	0.640
C6	-0.047	0.004	-0.056	-11.524	0.000
C7	0.062	0.008	0.042	8.020	0.000
C8	0.000	0.006	0.000	-0.023	0.982
C9	0.002	0.005	0.002	0.444	0.657
C10	-0.068	0.010	-0.032	-6.887	0.000
C1_Pos	0.363	0.039	0.039	9.222	0.000
C1_Neg	-3.917	0.084	-0.194	-46.732	0.000
C2_Pos	0.400	0.059	0.034	6.723	0.000
C2_Neg	-3.281	0.273	-0.048	-12.026	0.000
C3_Pos	0.391	0.040	0.050	9.715	0.000
C3_Neg	-2.687	0.193	-0.054	-13.935	0.000
C4_Pos	0.940	0.089	0.052	10.608	0.000
C4_Neg	-1.479	0.224	-0.026	-6.594	0.000
C5_Pos	0.415	0.018	0.128	23.687	0.000
C5_Neg	-4.234	0.074	-0.224	-57.087	0.000
C6_Pos	0.468	0.033	0.064	14.336	0.000
C6_Neg	-3.146	0.108	-0.114	-29.095	0.000
C7_Pos	0.029	0.065	0.002	0.452	0.651
C7_Neg	-0.618	0.216	-0.012	-2.857	0.004
C8_Pos	0.193	0.030	0.030	6.470	0.000
C8_Neg	-3.053	0.123	-0.097	-24.902	0.000
C9_Pos	0.415	0.018	0.099	22.790	0.000
C9_Neg	-3.330	0.076	-0.169	-43.747	0.000
C10_Pos	0.224	0.080	0.012	2.783	0.005
C10_Neg	-1.139	0.222	-0.020	-5.123	0.000

a. Dependent Variable: Rating

Factors affecting Hotel Guest Satisfaction

C6 : **Food Service** (B = -0.056, c < 0.001)

C1: **Guestroom** (B = -0.055, p < 0.001)

C7: **Tiger** (B = 0.042, p < 0.001)

Findings / Implications

- Hotel managers should emphasis on improving employee interaction, guestroom experience and quality of food

Limitations

Data may have been biased

Includes only one source of social media review data

Future Enhancements

- Addition of hotel reviews
- Including reviews from other social media platforms
- Applying more data pre-processing techniques
- Further cleaning data for statistical analysis

Acknowledgment

Special Thanks
Dr. Smrittee Kala Panta
Kathmandu University School of Management
Specialization: Recreation Parks and Tourism