Sentiment Analysis

Sentiment analysis is a method of natural language processing (NLP) which can be used to ascertain the sentiment present in a review. By focusing on key words present within the review, NLP may be used to determine general feeling of a review, whether it be good, bad or neutral. Quickly understanding the patterns present within reviews may provide valuable insight into how well a business is operating during a given period. Which could provide management of an organization with direction in creating operational plans based on the analysis and visualizations we provide on reviews to improve their business and retain a higher number of customers.

DATASETS

Dataset 1: TripAdvisor [1]

The dataset features hotel reviews scraped from TripAdvisor for 10 different cities. (Dubai, Beijing, Long, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, Chicago) each city containing a varying number of hotels where each hotel has text. The text fields featured within the dataset are date, review title and full review as shown in Fig 1. It contains approximately 259 000 reviews. For this project the focus was on 5 of the 10 cities London, Dubai, Delhi, Shanghai and New York.

Dataset 2: Booking.com[6]

•Found on kaggle, features 515k reviews from luxury hotels in Europe.

•Dataset was scraped from booking.com, it features 17 fields.

Learning Object

Jannatul:

- Familiarize myself with different NLP techniques. [2,3,5] Mostly I learned details of the concepts of tokenizer, autotokenizer and how to perform TF-IDF (Term Frequency - Inverse Document Frequency) in week 1. [7,8]
- In week 1, I familiarized myself with various NLP techniques, including the concepts of tokenizer, autotokenizer, and performing TF-IDF. Specifically, I delved into the details of these techniques and gained a good understanding of them.
- In the next week, I also learned about Vader and Bert pretrained models and the concept of a pipeline. [9,10] I was able to implement these models and their pipelines effectively.

- Moving on to the next week, I successfully implemented TF-IDF (Term Frequency - Inverse Document Frequency) on Dataset-1 and generated a list of the most relevant words for five cities.
- In the following 2 weeks I have learned and implemented Vader (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis. I used this technique to analyze the reviews of four hotels in Dataset-1 and achieved satisfactory results.
- After studying the Bert (Bidirectional Encoder Representations from Transformers) pre-trained model, I attempted to apply it to hotel reviews from Delhi. However, I encountered several challenges along the way. First, I discovered that the Bert model cannot accept input if the reviews exceed 256 tokens, and it does not work with non-English characters or languages other than English. Although most of the reviews in the dataset were in English, there were a few rows with reviews containing unfamiliar characters such as (? or !). Devanshi and I worked together to address this issue, I had to perform data cleaning again and remove those rows. In addition, I added a 'try except' loop to ignore rows with reviews that exceeded 256 words. Finally, I faced the challenge of requiring a GPU to run Bert, which was a new concept for me. I ended up paying for Colab Pro to use GPU and TPU to run the Bert model on the datasets of four cities (Dubai, Delhi, Shanghai, London), and obtained some useful results.
- However, another challenge we faced was that our dataset did not contain a rating column, so we could not compare the sentiments obtained from Bert and Vader with the original ratings.

SENTIMENT ANALYSIS WITH VADER AND BERT

Vader (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis

VADER sentiment analysis is a lexicon and rule-based sentiment analysis tool that is widely used to sentiments expressed in social media. It is fully open-sourced under the MIT License.[12] This model returns a sentiment score in the range -1 to 1, from most negative to most positive.
In order to use the VADER sentiment analysis tool, the transformer and TensorFlow were installed, followed by importing the pipeline and downloading the VADER sentiment model. Then the VADER model was run on the entire dataset to obtain the sentiments expressed in the dataset.

Bert (Bidirectional Encoder Representations from Transformers) Sentiment Analysis from HuggingFace

Bert is a pre-trained deep learning model which is developed by Google. This model is used for natural language processing (NLP) tasks, e.g. sentiment analysis. Hugging Face provides the platform for Bert and the Hugging Face transformers library contains lots of pre-trained BERT models.[9]
In a Google Colab notebook, the "cardiffnlp/twitter-roberta-base-sentiment" model was imported after importing AutoTokenizer. [10] Next, a dictionary was created to store sentiments. Finally, a try-except loop was implemented within a for loop to ignore longer reviews.

The outcomes of both the Vader and Bert models are displayed in Tables 1-4 for four cities: Delhi, Dubai, Shanghai, and London.

Table 1: Vader and Bert Sentiment Scores on Hotel Reviews of Delhi

id	vader_neg	vader_neu	vader_pos	vader_compound	bert_negative	bert_neutral	bert_positive	Reviews	polarity	subject	Dates_extracted	Name
1	0.000	0.617	0.383	0.9867	0.062750	0.234247	0.703003	Not bad I expected If compare American standar...	0.205655	0.459524	Nov 4 2009	india_new delhi_airport_hotel.csv
2	0.057	0.860	0.083	0.5067	0.356102	0.467573	0.176326	Don't stay !! Wrote mail got reservation inclu...	-0.114876	0.501928	Jun 6 2009	india_new delhi_airport_hotel.csv
3	0.000	0.779	0.221	0.7579	0.167867	0.471659	0.360474	Better stay Airport Due unavailability star ho...	0.291667	0.458333	May 4 2009	india_new delhi_airport_hotel.csv
4	0.197	0.755	0.048	-0.9569	0.943813	0.049174	0.007013	My worst experience ever! I stay one night nea...	-0.248148	0.642593	Feb 11 2009	india_new delhi_airport_hotel.csv
5	0.190	0.698	0.112	-0.9531	0.660240	0.284636	0.055124	poor stay decided stay one night layover airpo...	-0.187115	0.617627	Jan 5 2009	india_new delhi_airport_hotel.csv
6	0.066	0.868	0.066	0.2519	0.324600	0.502956	0.172444	Not Recommended Because unexpected layover Del...	0.196333	0.410250	Mar 1 2008	india_new delhi_airport_hotel.csv
7	0.226	0.774	0.000	-0.9266	0.948161	0.046036	0.005803	Stay Away!!!!!! Had transit via Delhi way Sura...	-0.200000	0.660000	Feb 1 2008	india_new delhi_airport_hotel.csv
8	0.559	0.441	0.000	-0.8225	0.954510	0.041048	0.004442	Worst ever High rates, bad rooms, complicated ...	-0.510000	0.801667	Dec 1 2007	india_new delhi_airport_hotel.csv
9	0.000	1.000	0.000	0.0000	0.126387	0.716171	0.157442	Bruyant	0.000000	0.000000	Jul 1 2008	india_new delhi_airport_hotel.csv
10	0.044	0.662	0.294	0.9382	0.005233	0.073866	0.920902	decent interesting part Delhi If wanting good ...	0.330513	0.535256	0.5352564102564102	india_new delhi_ajanta_hotel.csv

Table 2: Vader and Bert Sentiment Scores on Hotel Reviews of Dubai

id	vader_neg	vader_neu	vader_pos	vader_compound	bert_negative	bert_neutral	bert_positive	Reviews	polarity	subject	Dates_extracted	Name
1	0.065	0.651	0.284	0.8439	0.002237	0.039584	0.958179	- Situated Heart Dubai Damn good hotel, right ...	0.338413	0.375079	Feb 18 2009	are_dubai_royal_ascot_hotel.csv
2	0.039	0.724	0.237	0.9727	0.004337	0.038406	0.957257	!!!!Perfect!!!! A freind I stayed Rimal Rotan...	0.453571	0.688492	Mar 15 2007	are_dubai_rimal_rotana_dubai.csv
3	0.015	0.729	0.256	0.9954	0.004523	0.046525	0.948952	!Deluxe Dubai! Well well well....where start??...	0.173812	0.549249	Jul 3 2008	are_dubai_jumeirah_beach_hotel.csv
4	0.011	0.604	0.385	0.9954	0.002129	0.021916	0.975956	" Loved - Great Stay" Stayed May 4 nights husb...	0.417540	0.589286	Jun 5 2008	are_dubai_towers_rotana_dubai.csv
5	0.027	0.586	0.387	0.9822	0.001634	0.011911	0.986455	" Wonderful super friendly staff service" This...	0.192745	0.573922	Nov 7 2007	are_dubai_hilton_dubai_jumeirah.csv
6	0.111	0.775	0.114	0.0464	0.764026	0.196562	0.039412	" Worst Hotel I ever stayed at" This Hotel loo...	-0.057343	0.606294	Jul 18 2009	are_dubai_orchid_hotel.csv
7	0.082	0.736	0.182	0.9038	0.010035	0.123792	0.866173	"A Class" London. Where I begin....I stayed al...	0.154491	0.468426	Sep 3 2006	are_dubai_moevenpick_hotel_bur_dubai.csv
8	0.014	0.630	0.356	0.9966	0.002421	0.017497	0.980082	"ABSOLUTELY FANTASTIC....!!!" Where start....w...	0.261236	0.616835	Jun 17 2008	are_dubai_sheraton_jumeirah_beach_resort_tower...
9	0.000	0.718	0.282	0.9836	0.002965	0.036615	0.960420	"Best Hotel Dubai" My partner I stayed JBH Jun...	0.391741	0.461161	Sep 8 2006	are_dubai_jumeirah_beach_hotel.csv
10	0.000	0.468	0.532	0.9876	0.001735	0.011997	0.986268	"best I ever to" Most amazing I ever to! The I...	0.545000	0.660000	Feb 2 2008	are_dubai_habtoor_grand_resort_spa.csv

Table 3: Vader and Bert Sentiment Scores on Hotel Reviews of London

id	vader_neg	vader_neu	vader_pos	vader_compound	bert_negative	bert_neutral	bert_positive	Reviews	polarity	subject	Dates_extracted	Name
1	0.000	0.196	0.804	0.6249	0.005047	0.071278	0.923675	Great Hotel	0.800000	0.750000	Apr 1 2004	uk_england_london_best_western_the_delmere_hot...
2	0.393	0.446	0.161	-0.3818	0.242634	0.699878	0.057488	Avoid No elevator. Room top floor. Beds lumpy.	0.500000	0.500000	Apr 1 2004	uk_england_london_castleton_hotel.csv
3	0.000	0.712	0.288	0.9686	0.007351	0.087279	0.905370	They even watched grandmother!! Everyone entit...	0.292424	0.518182	Apr 1 2004	uk_england_london_concorde_hotel.csv
4	0.000	0.440	0.560	0.9184	0.001553	0.012232	0.986215	Warm helpful staff What gem! Fabulous large ro...	0.391071	0.582143	Apr 1 2004	uk_england_london_staunton_hotel.csv
5	0.000	0.000	1.000	0.3400	0.012220	0.488090	0.499690	Worthwhile	0.500000	0.500000	Apr 1 2004	uk_england_london_the_darlington_hyde_park.csv
6	0.039	0.575	0.386	0.9840	0.002348	0.015392	0.982260	Absolutely fabulous stay Thistle Victoria Stay...	0.209810	0.543478	Apr 1 2004	uk_england_london_the_grosvenor.csv
7	0.000	0.132	0.868	0.7650	0.004446	0.063912	0.931642	Good excellent location.	0.850000	0.800000	Apr 1 2004	uk_england_london_westbury_kensington.csv
8	0.000	0.645	0.355	0.9690	0.030539	0.266207	0.703255	Comfortable top class location Booked 2 night ...	0.289167	0.562500	Apr 1 2005	uk_england_london_athenaeum_hotel_apartments.csv
9	0.067	0.580	0.353	0.9876	0.040069	0.137173	0.822759	Great location - good value Can agree previous...	0.297901	0.630309	Apr 1 2005	uk_england_london_thanet_hotel.csv
10	0.033	0.478	0.489	0.9826	0.008431	0.037667	0.953902	A satisfied customer I pleased choice visiting...	0.325566	0.685417	Apr 1 2006	uk_england_london_arriva_hotel.csv

Table 4: Vader and Bert Sentiment Scores on Hotel Reviews of Shanghai

id	vader_neg	vader_neu	vader_pos	vader_compound	bert_negative	bert_neutral	bert_positive	Reviews	polarity	subject	Dates_extracted	Name
1	0.000	0.583	0.417	0.9847	0.001719	0.016303	0.981978	LOVELOVELOVE THIS HOTEL!!! Have stayed 88 Xian...	0.442361	0.835880	Nov 24 2009	china_shanghai_88_xintiandi_executive_residenc...
2	0.000	0.649	0.351	0.9823	0.002367	0.015850	0.981782	Fantastic Hotel comes premium I stayed Xintian...	0.238495	0.568903	Oct 12 2009	china_shanghai_88_xintiandi_executive_residenc...
3	0.019	0.632	0.348	0.9975	0.001856	0.028344	0.969800	Best area Shanghai We spent five days Shanghai...	0.380028	0.523350	Oct 11 2009	china_shanghai_88_xintiandi_executive_residenc...
4	0.035	0.658	0.307	0.9886	0.002412	0.030429	0.967159	great location great service Have stayed 88 Xi...	0.356875	0.518264	Sep 9 2009	china_shanghai_88_xintiandi_executive_residenc...
5	0.017	0.661	0.322	0.9941	0.011497	0.177488	0.811016	Upscale Alternative This property much acclaim...	0.326357	0.614000	Aug 19 2009	china_shanghai_88_xintiandi_executive_residenc...
6	0.000	0.770	0.230	0.9885	0.005853	0.244569	0.749577	An Oasis Calm Elegance bustling Shanghai This ...	0.261000	0.540000	Oct 19 2008	china_shanghai_88_xintiandi_executive_residenc...
7	0.051	0.482	0.467	0.9413	0.006895	0.039839	0.953266	great fantastic - rooms awesome - ask lake fac...	0.533333	0.629167	Aug 17 2008	china_shanghai_88_xintiandi_executive_residenc...
8	0.032	0.635	0.334	0.9808	0.004534	0.043971	0.951495	Great modern boutique This really nice place s...	0.235982	0.573571	Jun 26 2008	china_shanghai_88_xintiandi_executive_residenc...
9	0.081	0.738	0.181	0.9827	0.368379	0.487340	0.144281	It OK Maybe I higher expections I have.I arran...	0.197083	0.496429	Jun 23 2008	china_shanghai_88_xintiandi_executive_residenc...
10	0.000	0.576	0.424	0.9918	0.003848	0.023846	0.972307	Exceeds expectations 88 Xintiandi truly wonder...	0.406607	0.724107	May 17 2008	china_shanghai_88_xintiandi_executive_residenc...

Then bar graphs were plotted to observe the comparison of sentiments of the reviews captured using two methods, this procedure was repeated for four cities (Delhi, Dubai, Shanghai, and London).

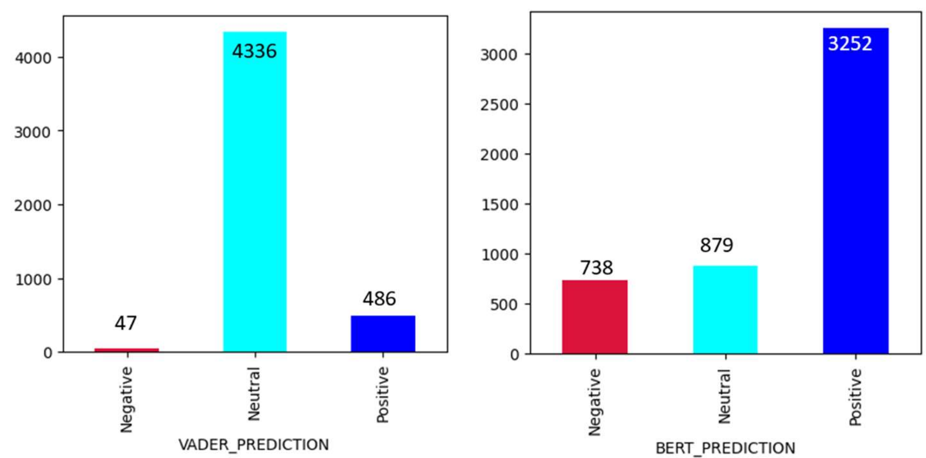

Fig 9: Sentiment Analysis on Hotel Reviews in Delhi

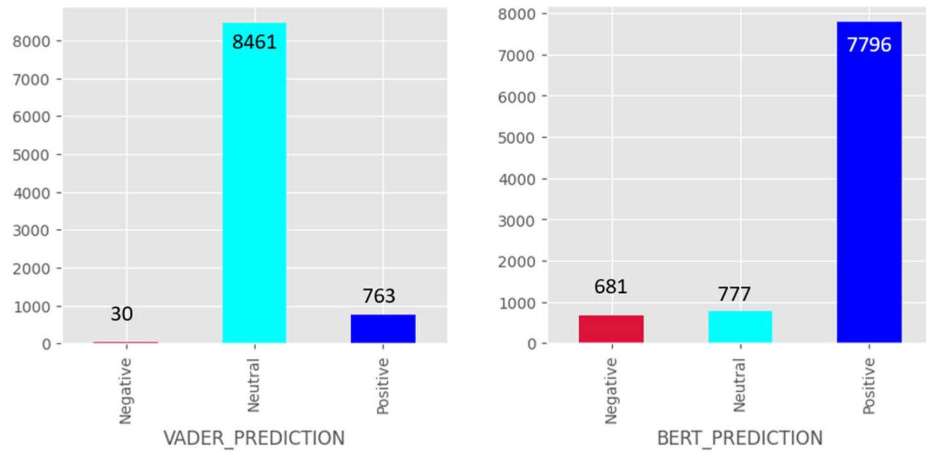

Fig 10: Sentiment Analysis on Hotel Reviews in Dubai

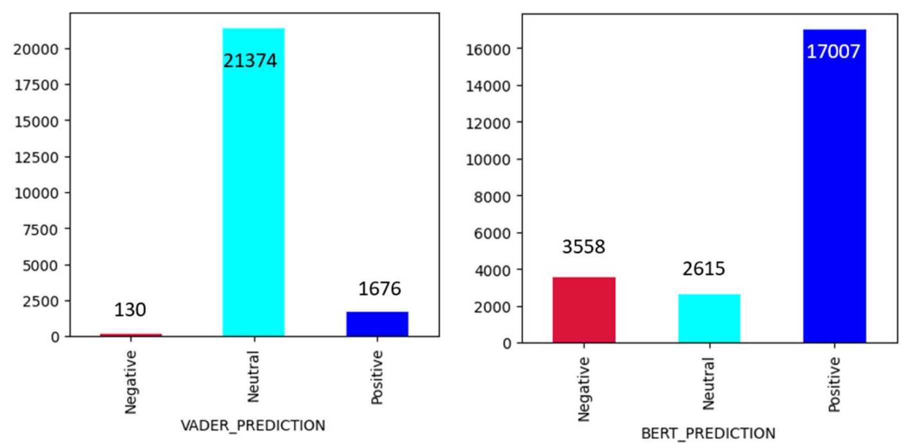

Fig 10: Sentiment Analysis on Hotel Reviews in London

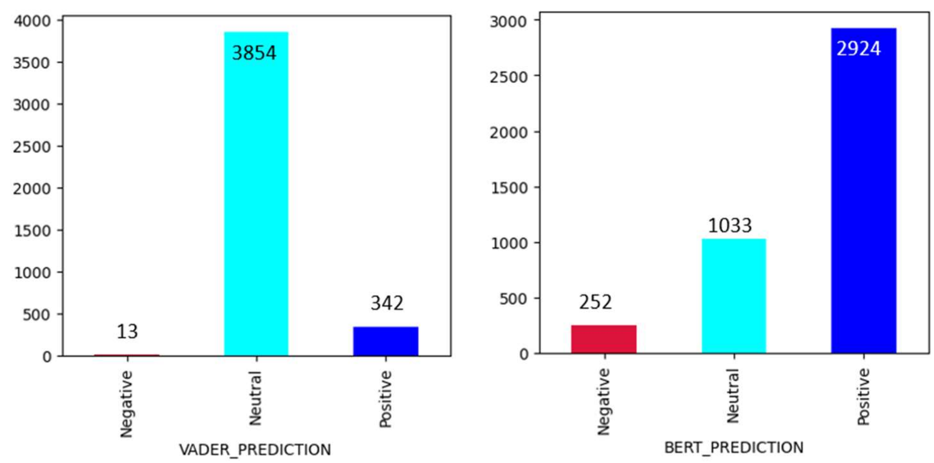

Fig 11: Sentiment Analysis on Hotel Reviews in Shanghai

Insights from Vader and Bert Model:

From fig 9-11, it is evident that the Bert model outperforms the Vader model in capturing sentiment. The Vader model tends to label a significant portion of sentiments as neutral, while the Bert model is more effective in distinguishing between different types of sentiments. As a result, when applied to the same dataset for each city, the Bert model detects more positive or negative sentiment compared to the Vader model. This suggests that the Bert model is a more reliable tool for sentiment analysis in this context.

	Dubai	Delhi	Shanghai	London
Positive	84.24%	66.79%	69.47%	73.37%
Negative	7.36%	15.15%	5.98%	11.28%

Fig 12: Positive and Negative Sentiments at Different Cities using Bert Model

Insights from Fig 12:

As it was already established that Bert model performed better, it was used to analyze and summarize the percentage of positive and negative sentiments in hotel reviews from four different cities: Delhi, Dubai, Shanghai, and London. The results of this analysis were presented in figure 12. The findings of figure 12 showed that Dubai had the highest percentage of positive sentiments (84.24%) in their hotel reviews, indicating that guests had a highly positive experience during their stay. In contrast, Delhi had the least percentage of positive sentiments, implying that guests were less satisfied with their stay. Moreover, Delhi had the highest percentage of negative sentiments (15.15%) in their hotel reviews, suggesting that guests were dissatisfied with their stay and London following closely behind. The finding that London had the second-highest percentage of negative sentiments was quite surprising and could indicate potential issues with hotels in the city.

REFERENCES

1. Trip Advisor Review Dataset. Retrieved from: http://kavita-ganesan.com/entity-ranking-data/#.XQESU9NKgWq on 5th March 2023.
2. What is natural language processing (NLP). Retrieved from: https://www.ibm.com/topics/natural-language-processing on 5th March 2023.
3. Natural Language Processing (NLP), What it is and why it matters. Retrieved from: https://www.sas.com/en_ca/insights/analytics/what-is-natural-language-processing-nlp.html on 5th March 2023.
4. Generating Word Cloud in Python. Retrieved from: https://www.geeksforgeeks.org/generating-word-cloud-Python/ on 5th March 2023.
5. F. Heimerl, S. Lohmann, S. Lange and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 2014, pp. 1833-1842, doi: 10.1109/HICSS.2014.231. Retrieved from: https://ieeexplore.ieee.org/document/6758829 on 5th March 2023.
6. Trip advisor Review Dataset on Kaggle. Retrieved from: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe on 19th March 2023.
7. TF-IDF [Tutorial] in Python. Retrieved from: https://www.kaggle.com/code/paulrohan2020/tf-idf-tutorial on 20th March 2023.
8. Sklearn feature extraction text and TfidfVectorizer. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html on 20th March 2023.
9. Hugging face pretrained models. Retrieved from: https://huggingface.co/models on 20th March 2023.
10. Hugging face transformer Bert models. Retrieved from: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertModel on 20th March 2023.

11. Bag of words (BoW) model in NLP. Retrieved from: ttps://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/ on 5th March 2023.
12. Vader Sentiment 3.3.2. Retrieved from: https://pypi.org/project/vaderSentiment/#:~:text=VADER%20(Valence%20Aware%20Dictionary%20and,on%20texts%20from%20other%20domains. on 5th March 2023.