Topic-Based Tourism Sentiment Analysis using BERTopic and Deep Learning

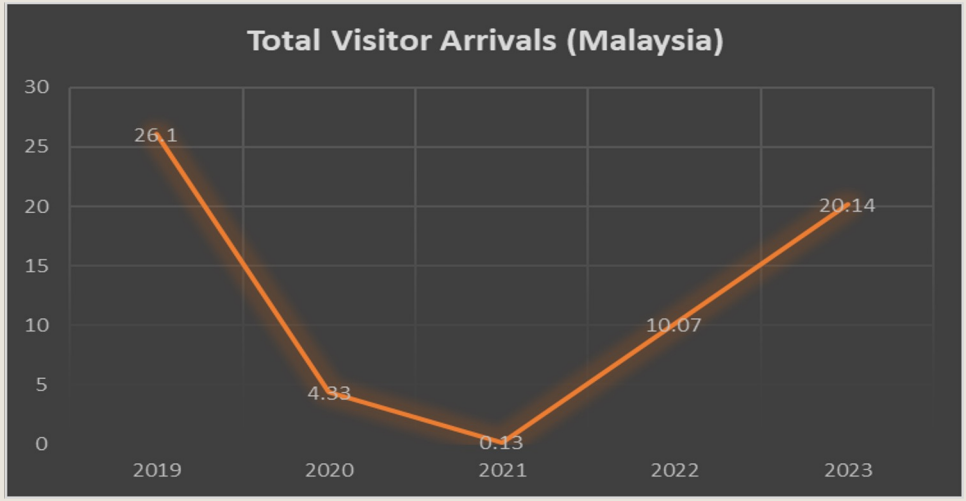
SOLEHAH NAJIIHAH BINTI ABD JAMAL
MCS231035

Outline

- Background
- Problem Statement
- Objective
- Literature Review
- Research Methodology
- Initial Findings (EDA)
- References

Background

- Tourism Industry: one of contributing factors for Malaysia's economic growth
- Covid19: Heavily impact on tourism industry
- User Generated Contents (UGC)
  - ➔ Sentiment Analysis
  - ➔ Travelers experience influence other travellers perceptions

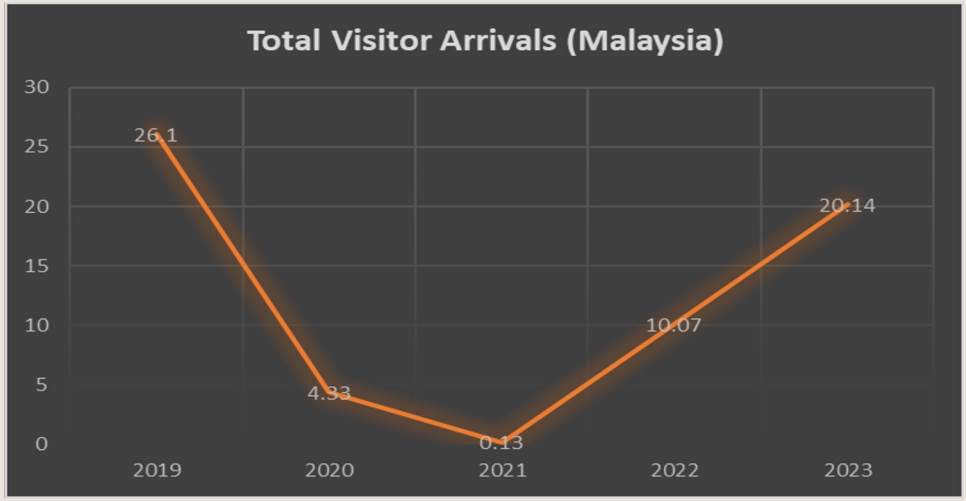


Source: LADA

Background

Langkawi:

- Langkawi Tourism Recovery Plan (T-REC) 2021-2022 - LADA
- Geographically strategic (Kuah Jetty & Langkawi International Airport)
- Nature (Beach)
- Tourism surpassed agriculture & fishing
- Market Report by Horwath HTL



Source: LADA

Problem Statement

- Tourism Industry is affected by COVID19, resulting in drop numbers of tourists.
- Travellers reviews have significance influence on travelers' behaviour and perceptions on destinations.
- Identify the aspects that drives a person's feeling and the sentiment in reviews.
- Leverage deep learning methods for sentiment analysis.

Objective

- To determine the key topics from user reviews on the tourism industry in Langkawi, Kedah.

- To develop LSTM algorithm and fine-tune BERT for sentiment analysis on chosen topics.

- To evaluate the performance of LSTM and BERT in sentiment analysis.

SCOPE OF WORK

- Online review sentiment analysis specifically in Langkawi, Kedah.
- Web scraping from TripAdvisor and Google Map website limited to English language.
- Employing LSTM and BERT for sentiment analysis.
- Performance evaluation: Compare the performance of LSTM and BERT in sentiment analysis using precision, recall, F1-score and AUROC

Literature Review

Tourism Sentiment Analysis

Reference	Techniques	Results
Cao, Z., Xu, H., & Teo, B. S. X. (2023). Sentiment of chinese tourists towards malaysia cultural heritage based on online travel reviews. Sustainability, 15(4), 3478.	- Chinese tourists sentiment towards Malaysia's cultural heritage - CharCNN, LSTM, BiLSTM, and BERT	- BERT outperforms other models - 9 scenic spots identified - Positive sentiments: cultural atmosphere, material culture, and scenic landscapes. - Negative emotions: lack of cultural experiences, leading to feelings of boredom
Mehra, P. (2023). Unexpected surprise: Emotion analysis and aspect based sentiment analysis (ABSA) of user generated comments to study behavioral intentions of tourists. Tourism Management Perspectives, 45, 101063.	- investigate emotions and sentiments derived from these comments may affect post travel behaviour - ABSA and emotion analysis	- Sad feelings are mostly caused by things like food and bathrooms in China, women's empowerment and alcohol in the UAE, traffic, hygiene, time, and poverty in India.

Literature Review

Topic-Based Sentiment Analysis

Reference	Techniques	Results
Ounacer, S., Mhamdi, D., Ardchir, S., Daif, A., & Azzouazi, M. (2023). Customer sentiment analysis in hotel reviews through natural language processing techniques.	- Significance of customer reviews in influencing decisions in the tourism sector. - Topic Modeling: Correlation Explanation - Sentiment Analysis: LR, RG, NB, DT, KNN, SVM, ET (Extratree), AB,GB (Adaboost and Gradient Boost)	- Logistic Regression + CountVectorizer (precision (82%), recall (69.59%), accuracy (91%) and F1-score (73.27%)_ - RandomForest + TF-IDF (precision (81.01%), recall (74.78%), accuracy (86%) and F1-score (76.59%).
Syamala, M., & Nalini, N. J. (2019, July). LDA and deep learning: a combined approach for feature extraction and sentiment analysis. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.	- Identify the exact preferences of good or bad based on the feature - Topic Modeling: Latent Dirichlet Allocation (LDA) - Sentiment Analysis: VADER, TextBlob (lexicon-based approach)	- 4 topics identified - TextBlob: 77.3% - VADER : 72.6% - Joint Sentiment/Topic: 69.6%

Literature Review

Tourism Sentiment Analysis

Reference	Techniques	Results
Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental study on BERTopic technique. Procedia computer science, 189, 191-194.	- Topic Modeling: LDA, NMF, BERTopic - Measurement:Normalized Pointwise Mutual Information (NPMI) - pre-trained Arabic language models as embeddings for the BERTopic technique, including AraBERTV2.0, ARBERT, QARiB, and XLM-R	- BERTopic were more closely related and relevant to each other, as evidenced by the higher NPMI scores.
Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology, 7, 886498.	- Comparative analysis of the results produced by each algorithm on Twitter post - Topic Modeling: LDA, Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic	- LDA: failing to capture the nuances of the data effectively. - Top2Vec: overlapping themes, less effective for clear topic differentiation. - BERTopic: aligned with air travel and issues faced during the pandemic - BERTopic and NMF: able to capture relevant insights

Research Methodology

- **Data Preparation (Collection and Cleaning)**

- Web scraping from TripAdvisor using Instant Data Scraper, SimpleScraper and Apify
- 6 destinations: Langkawi Sky Bridge, Crocodile Adventureland Langkawi, Kilim Geoforest Park, Cenang Beach, Telaga Tujuh Waterfalls and Underwater World Langkawi
- Datasets: 1,187

Research Methodology

(1187, 24)

	placeInfo/address	placeInfo/addressObj/city	placeInfo/addressObj/country	placeInfo/addressObj/postalcode	placeInfo/addressO
0	Jalan Telaga Tujuh, Langkawi 07000 Malaysia	NaN	Malaysia	7000	
1	Jalan Telaga Tujuh, Langkawi 07000 Malaysia	NaN	Malaysia	7000	
2	Jalan Telaga Tujuh, Langkawi 07000 Malaysia	NaN	Malaysia	7000	
3	Jalan Telaga Tujuh, Langkawi 07000 Malaysia	NaN	Malaysia	7000	

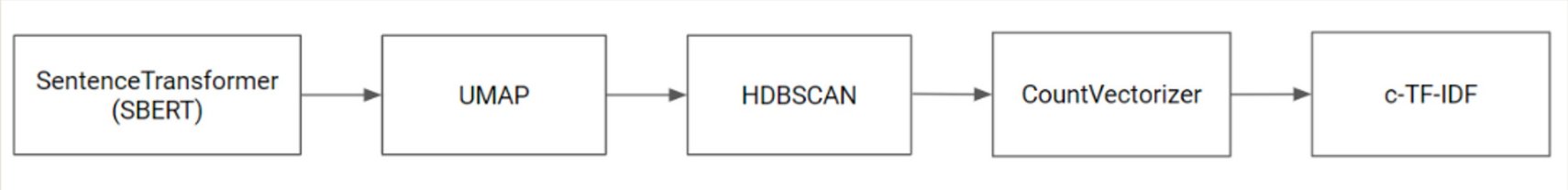
	Destination	TravelDate	Rating	Review
0	Telaga Tujuh Waterfalls	2024-07	4	7 Wells, LangkawiGo early when it's cooler and...
1	Telaga Tujuh Waterfalls	2024-05	4	Beautiful 7 wells waterfall Telaga TujuhIt was...
2	Telaga Tujuh Waterfalls	2024-04	5	Incredible views, well worth the walk!Had an a...
3	Telaga Tujuh Waterfalls	2024-01	5	Highly recommendAmazing experience but make su...
4	Telaga Tujuh Waterfalls	2023-08	5	Amazing waterfall and the Seven Wells waterfal...

1. Removing unwanted columns.
2. Renaming columns
3. Identify missing and duplicate rows, remove them

Research Methodology

❖ Modeling

- BERTopic: Identify topics in corpus



- Default BERTopic model.

- CountVectorizer and c-TF-IDF responsible for topic representations.
  - Hyperparameters:

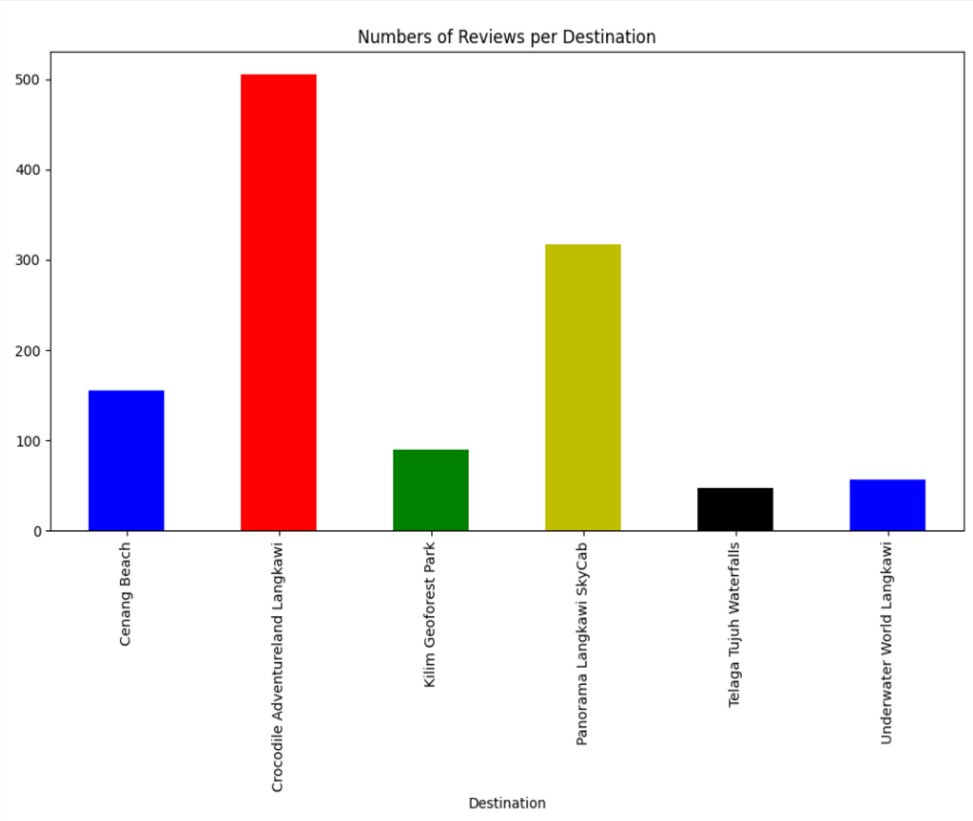
Research Methodology

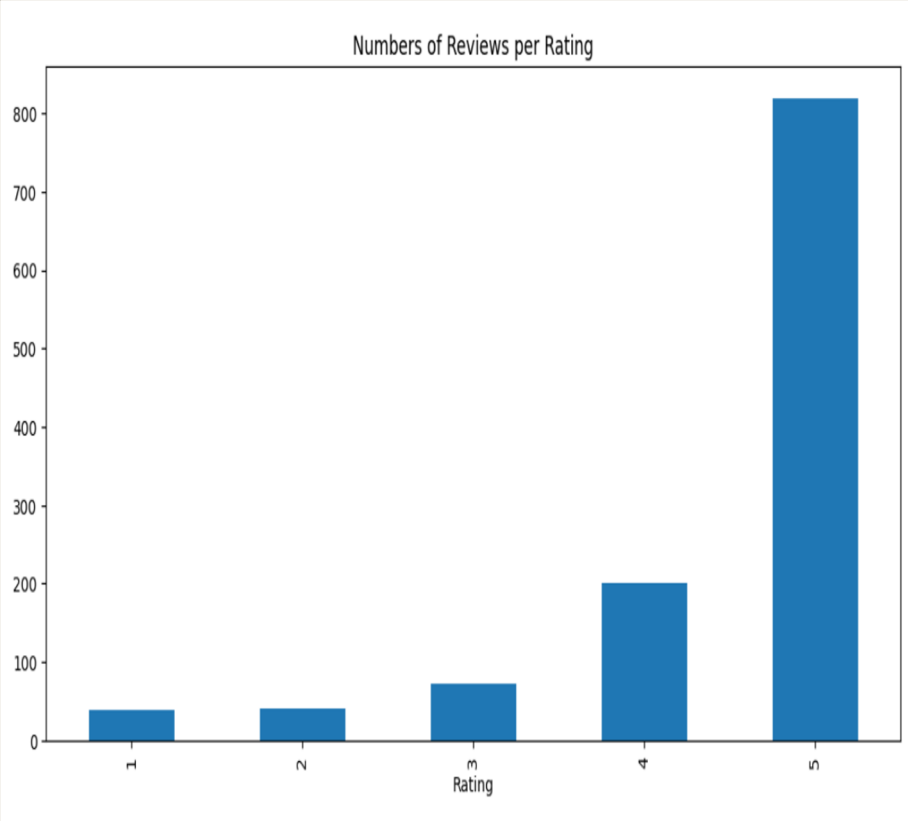
❖ **Modeling**

- BERT
  - bert-based-uncased
  - bertTokenizer
- LSTM

Parameter	Value
Word vector dimension	100
Kernel_size	3
Filters	100
Loss	Binary_crossentropy
Optimizer	adam
Activation	relu

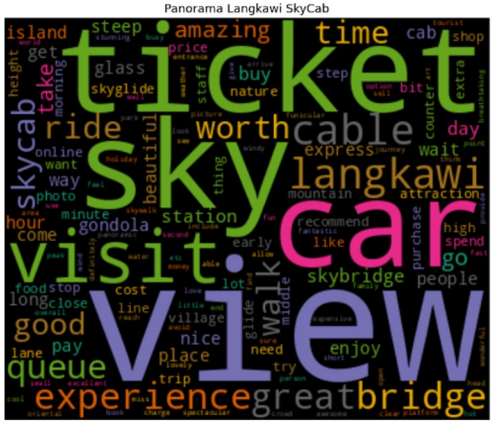
Initial Findings



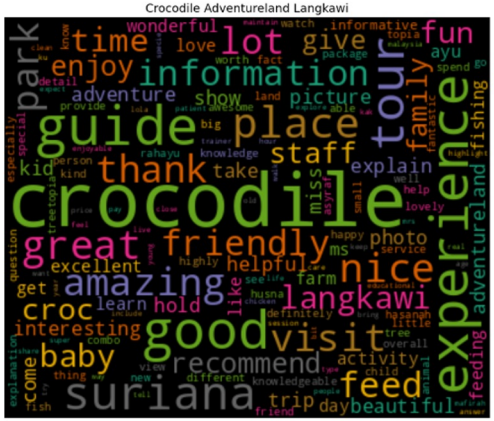


Cenang Beach

Panorama Langkawi SkyCab

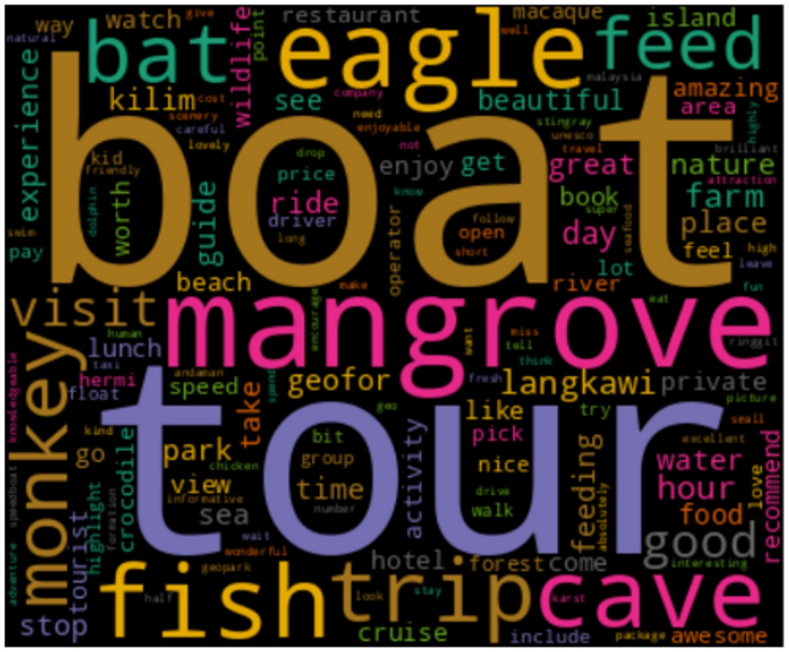


Crocodile Adventureland Langkawi



Telaga Tujuh Waterfalls

Kilim Geoforest Park



Underwater World Langkawi