A visual representation of the project framework is shown in Figure 1. It contains the flow of the project in conducting topic-based sentiment analysis on Langkawi, Kedah. The entire process can be categorized into six phases:

1. Problem Formulation
2. Data Collection
3. Data Pre-processing
4. Modeling
5. Testing and Validation
6. Performance Evaluation

## 

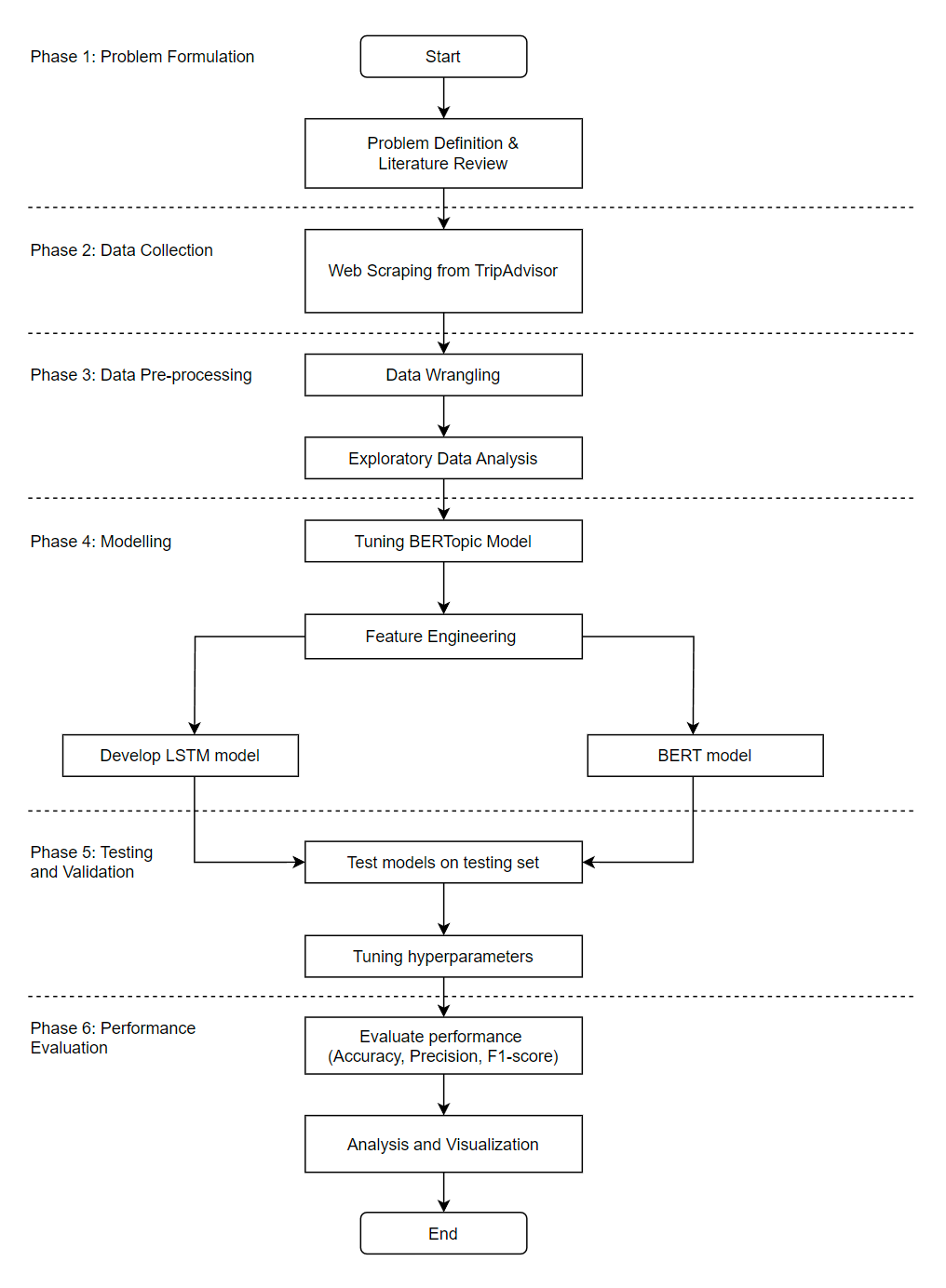


Figure 1: Research Framework for Tourism Sentiment Analysis

**Problem Formulation**

The primary aim of this project is to identify the key topics that influence the tourists in Langkawi, Kedah using BERTopic and using deep learning methods for sentiment classification on chosen topics. Understanding tourist sentiments in Langkawi, Kedah, is critical for improving services and visitor experiences. Traditional sentiment analysis approaches may not fully capture the nuances of user feedback. Advanced models such as BERTopic, LSTM, and BERT could increase sentiment analysis accuracy.

**Data Collection**

Tourists' reviews are crawled from TripAdvisor websites using Instant Data Scraper, SimpleScraper and Apify. The data collection is done by restricting the review’s date starting from January 2020. 5 popular destinations are chosen based on TripAdvisor Travellers Choice; Langkawi Sky Bridge, Crocodile Adventureland Langkawi, Kilim Geoforest Park, Cenang Beach, Telaga Tujuh Waterfalls and Underwater World Langkawi. 1,187 reviews are extracted from TripAdvisor website. The datasets with their description are summarized in Table 1. TripAdvisor has a filter section where the language of the review can be chosen. Since this project scope focuses on English dataset, only English reviews are extracted. However, Google Map website does not have filter features, hence the dataset consists of multiple languages such as Chinese, Arabic, Malay, English, German. The data cleaning process will be discussed in the next sub section.

Figure 2: Name of columns of the dataset.

**Data Pre-processing**

Preliminary analysis is done to understand the data and identifies the important and related features needed for the project. This dataset contains 1187 rows and 24 columns. The first step is removing unwanted columns from the dataset to get a better view of the dataframe. After deleting unwanted columns, the remaining columns are renamed into more understandable formats and re-arrange to four columns; Destination, TravelDate, Rating and Review. Review column is the concatenation of text (the reviews) and title (title of the reverie) columns. Data cleaning includes finding duplicate and missing rows and removing them. Then, both datasets are concatenated into a single dataframe for further data cleaning.

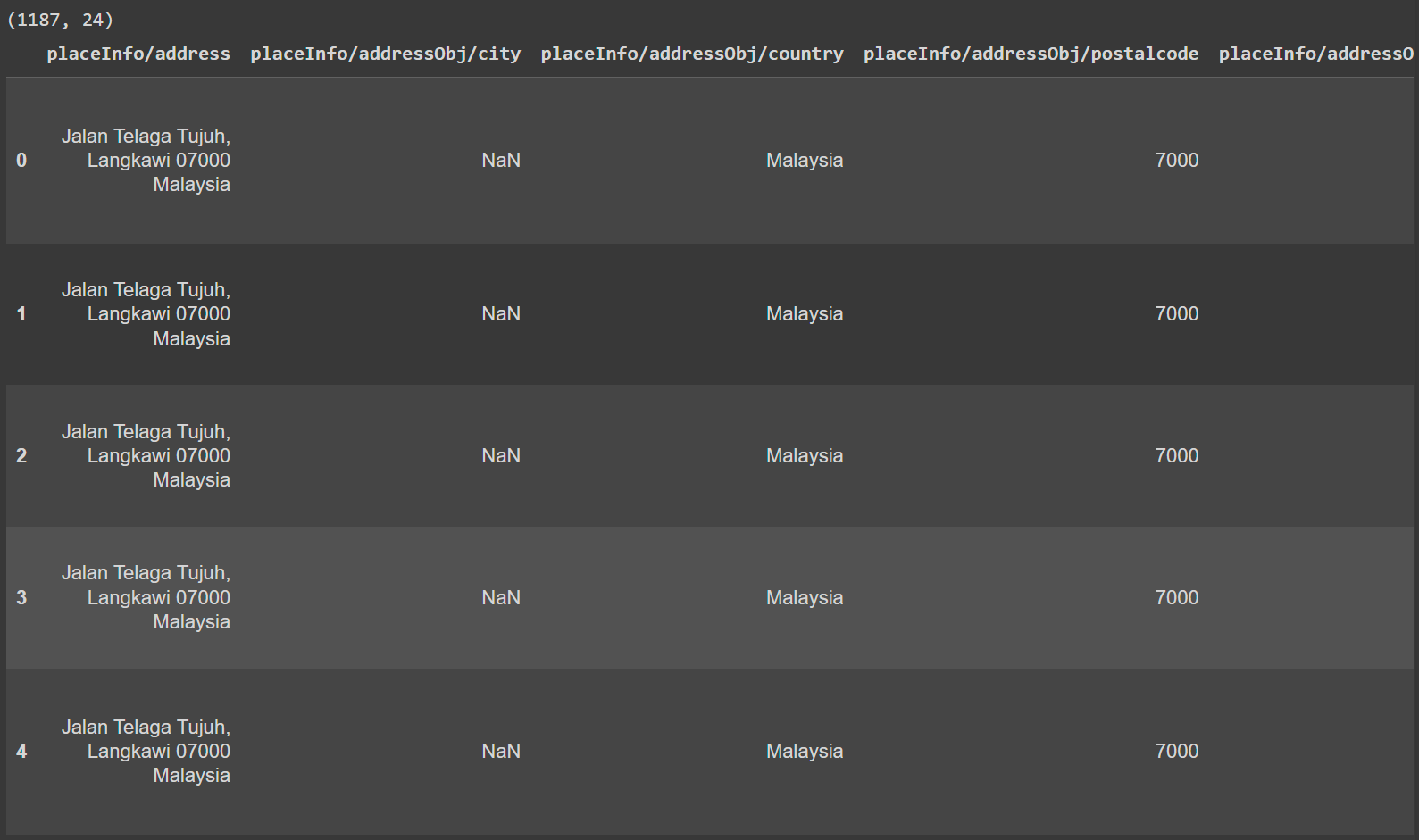


Figure 3: Dataset before cleaning.

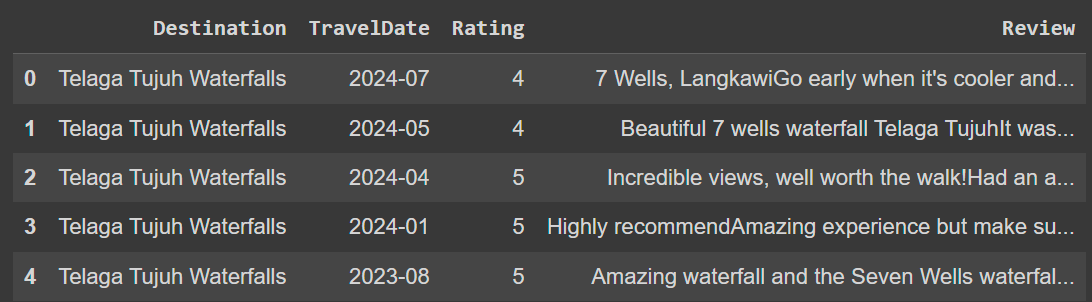
Figure 4: Dataset after cleaning.



Figure 5: Summarisation of duplicate and missing rows.

**Modeling**

**BERTopic**

BERTopic is a topic modeling technique that uses transformers and c-TF-IDF to produce dense clusters. This allows for readily interpretable topics while maintaining significant words in the topic descriptions. It generally consists of three pillars; embedding, clustering and topic representation generation. In this project, the default version of BERTopic proposed by [43] is used and the hyperparameters are tuned to produce coherence topics based on the dataset. Embedding is a process of transforming non-vectorized data, such as tokens, into a vector space and allows the model to automatically understand the connections between words and their meanings. The output of the embedding step creates high dimensional output giving difficulty in clustering, hence UMAP is used to reduce the dimensionality. UMAP is the default in BERTopic because it can represent both the local and global high-dimensional space in lower dimensions. After dimensionality reduction, the input embeddings are clustered into groups of similar embeddings to get the topics. HDBSCAN is used as it is able to capture structure with different densities. CountVectorizer and c-TF-IDF are responsible for creating topic representation, hence the parameters must be fine-tuned to give coherence topics.

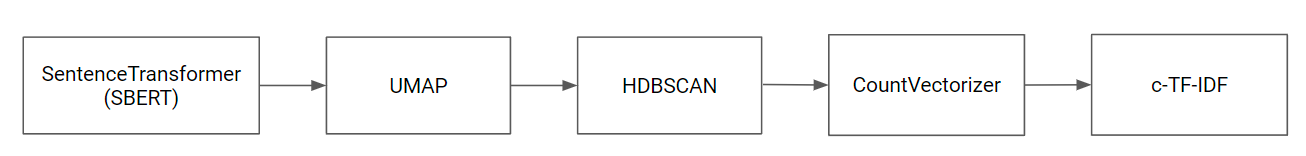


Figure 6: Default BERTopic model.

**BERT**

BERT stands for Bidirectional Encoder Representations from Transformers is a pre-trained model, trained on enormous general purpose language that can be fine-tuned on specific smaller datasets for specific tasks such as sentiment analysis. The previous language models looked at text sequence either from left-to-right or combined left-to-right and right-to-left, predicting the next possible word. BERT introduces a new approach which is bidirectionally trained, providing a deeper understanding of the language context. BERT is based on transformer model architecture where it applies an attention mechanism to understand the relationship between words in a sentence. In this project, BertTokenizer and BERT-Base, uncased models are used. The datasets are split into training and testing and feed into the BERT model.

**LSTM**

LSTM is an extended version of RNN that learns sequential (temporal) data and long-term connections with more precision than normal RNNs. LSTM is chosen as it is able to handle long-term dependencies in sequential data such as in text documents. The proposed LSTM architecture is shown in Table 1. Tokenizer is used to convert the input data into numerical values and padding to set the length of the review.

| **Parameter** | **Value** |
| --- | --- |
| Word vector dimension | 100 |
| Kernel\_size | 3 |
| Filters | 100 |
| Loss | Binary\_crossentropy |
| Optimizer | adam |
| Activation | relu |

Table 1: Proposed LSTM Architecture.

**Performance Evaluation**

The comparison performance between these models can be evaluated by calculating the Precision, F1-score, Recall and Area Under ROC Curve (AUROC).