CHAPTER 3

RESEARCH METHODOLOGY

3.1 Overview

In this chapter, introduce a complete emotion prediction for a collected dataset. The process including 4 stages, that is data collection, data preprocessing, emotion classification and XAI model interpretation. Each stage describes in detail in the following. This is to improve the performance and accuracy of the prediction.

3.2 Proposed Methodology

The methodology process for this project has 4 main phases. Before starting the process must define the project problem and objective. This is to ensure the project process did not run away from the project objectives. After define the project problem and objectives, the first step fort this project is data collection. In this project, the data was collected through web scrapping. The data scrape from the social media platform Reddit. The second phase of the project is data preprocessing. In this phase, the data will do the normalization, remove unwanted content, remove special characters and whitespace and last tokenization. This phase is to make sure the data was ready to the next phase. The next phase is emotion classification. In this phase, each text will be process and a prediction emotion will be given. In this phase DistilBERT model was used. Last phase for the project is XAI model interpretation. This phase is to interpret the emotion prediction process done by the model. This help to increase the trusty of the result. Figure 3.1 show the proposed research framework.

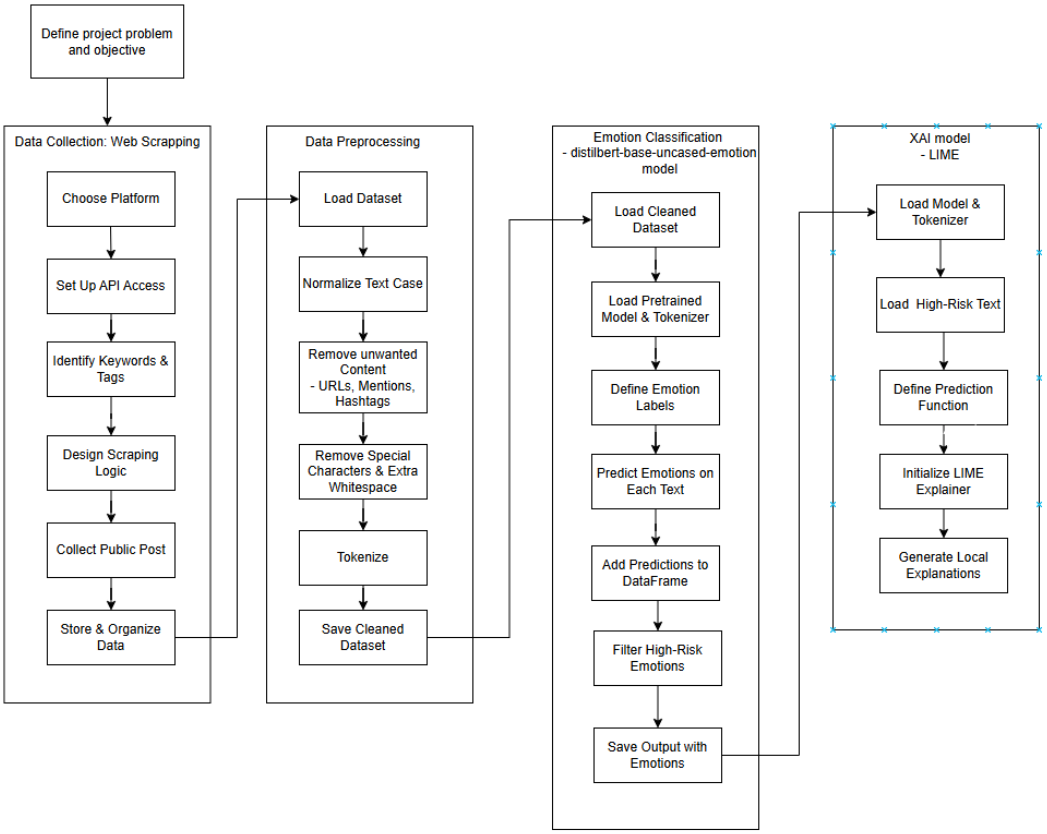

Figure 3.1 Proposed research framework

3.3 Data Collection

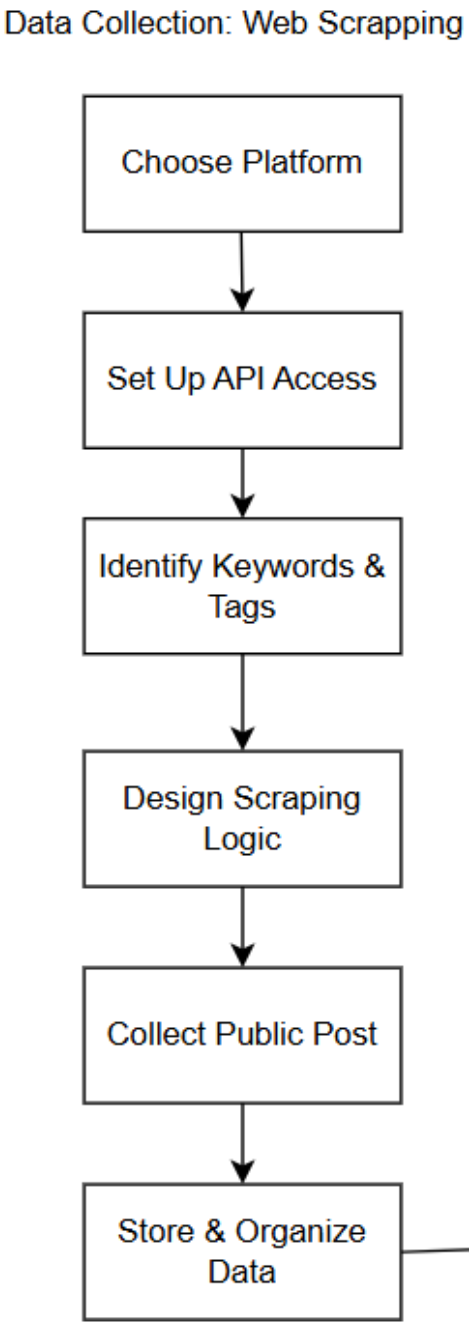

Figure 3.2 Data Collection

This project will gather the dataset by web scraping from social media platform like Reddit. This is due to the current dataset about mental health was mostly did not collect for Malaysia thus in this project use web scraping to gather the dataset that is from Malaysia users. The goal was to gather 100, 000 text samples that is related to

emotional distress, depression, anxiety and suicidal ideation, with focus on the content that is relevant to Malaysian users.

Reddit was chosen for data collection is due to its widespread use in Malaysia, its availability of open APIs and relevance to mental health discussions in Malaysia. Reddit is a popular discussion-based platform that consisting communities called subreddits which users will discuss different topic. For this project, r/Malaysia will be the subreddit use to collect the discussion for local users. r/mentalhealth, r/depression and r/anxiety will be use for more direct conversations about emotional well-being. r/suicidalthoughts will be use to collect the data that might show signs of serious distress. While Reddit let user can post by anonymously that encourage user can honest and feel free to share and express their feel.

Before starting to collect the data, it must set up an app on the Reddit Developer Portal to get the permission to use their API. After get the permission, the scraping we done by using a tool called Python Reddit API Wrapper (PRAW) to get the post text for selected subreddits. PRAW provides a straightforward interface for interacting with Reddit's API which allow to retrieve posts from selected subreddits efficiently. The data for each post collected will include post text, timestamp, and author. All of the data will save into a CSV file for easier handling. Each row in the CSV file will present a single post collected from Reddit which include text, timestamp and author.

3.4 Data Preprocessing

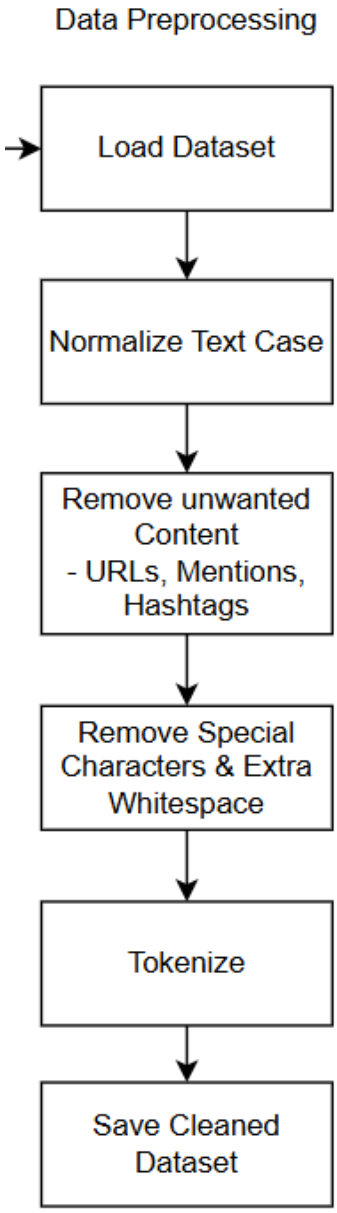

Figure 3.3 Data Preprocessing

The raw data collected from Reddit contained a mix of text samples related to emotional distress, depression, anxiety and suicidal ideation among Malaysian users. But the raw data must do some preprocessing before analysis. This is to ensure the data consistency, readability and suitability.

3.4.1 Load Raw Dataset

The first step involved load the raw dataset that was collected during the data collection phase. The dataset includes the textual content that gathered from selected subreddits that are related to mental health and Malaysian discussions. Each entry included the original post text with the timestamp and author.

3.4.2 Normalize Text case

Normalization is a data preprocessing technique that use to ensure uniformity across all text where all letters will convert into lowercase. This helped to reduce the variability caused by differences in capitalization and improved the model consistency during later stages. In this project, it needs to import libraries to do the normalization like nltk and re, which nltk is a standard python library and re is regex library.

3.4.3 Remove unwanted content

The text data collect from social media and for social media post often contain some element like URL link, user mentions and hashtags. Thus, in this phase, these elements that are not relevant to the emotional content of the posts have to be remove. A rule-based filtering approach was used to remove such patterns using regular expressions. For example, using ‘r’https?://\S+|www\.\S+’ to remove the URL link.’.

3.4.4 Remove special characters and extra whitespace

Next, the original post text data must consist some special character like !, @, ? and others characters. Other than that, excessive whitespace also has to remove. This is to enhance the clarity and reduce the complexity of the input data. To remove punctuation and special characters, ‘re.sub(r’[^a-zA-Z0-9\s’, ‘ ’, text]’ was used.

3.4.5 Tokenization

Tokenization refers to the process of splitting the cleaned text into individual units for the use of further processing by NLP models. This step transformed each post into a sequence of tokens, which later can be encoded numerically for model input.

3.4.6 Save cleaned dataset

Once all the preprocessing steps were completed, the cleaned dataset was stored in a CSV file for ease of access and future use in the emotion classification and explainable AI phase. Each of the row in the dataset included the original post, cleaned text, and tokenized text.

3.5 Emotion Classification

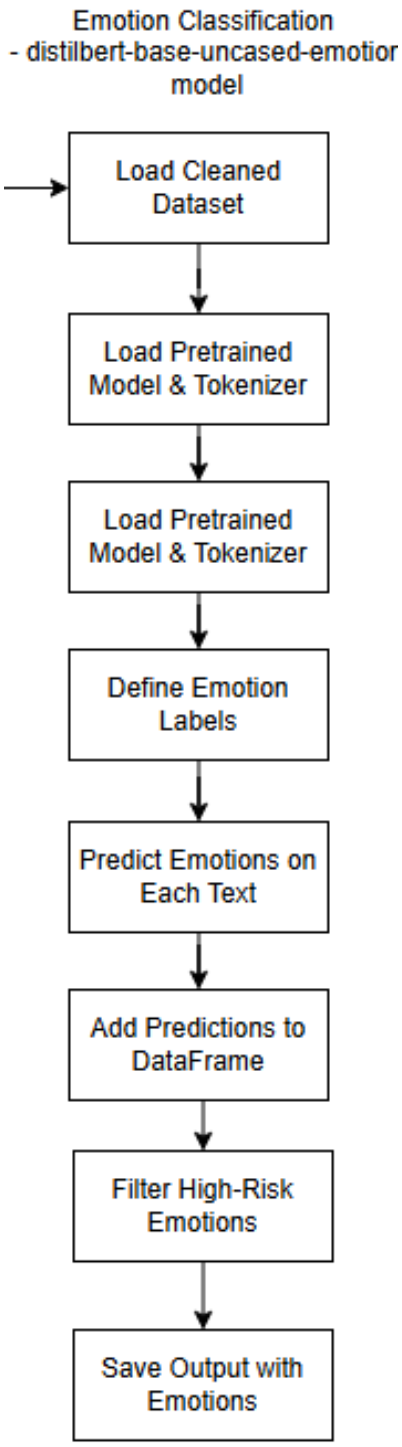

Figure 3.4 Emotion Classification

The emotion classification phase of this project is the main part of this project. This phase is to identify emotional states expressed in the cleaned social media texts collected from Reddit. The goal was to detect the emotional cues that may indicate potential mental health problem such as sadness, fear, and suicidal thoughts. Unlike the traditional sentiment analysis, which typically categorizes text into positive, neutral or negative, this project use a fine-tuned DistilBERT model trained on emotion-labeled conversational data. This allowed for a more detailed understanding of the emotional tone within the user generated content that related to mental health.

3.5.1 Model Selection

For the emotion classification task, the model selected to use is joeddav/distilbert-base-uncased-go-emotions-student model. This is because this model has strong performance on multi-label emotion classification. This model is based on DistilBERT, a compact and efficient variant of BERT. It has been fine-tuned on the GoEmotions dataset, which includes labels for 27 distinct emotional states such as ‘sadness’, ‘fear’, ‘anger’, ‘anxiety’, ‘joy’, ‘love’, and others emotional states. This model has the ability to assign multiple emotion labels to a single text; this makes it suitable for analyzing open-ended discussions about the emotional well-being found in subreddits.

3.5.2 Input Preparation

Before applying the model, the cleaned text samples have to tokenized and formatted to meet the input requirement of the DistillBERT architecture. Although the text had already undergone data cleaning and normalization during the data preprocessing phase, there is still have additional input formatting required before applying the emotion classification model. This included tokenizing the text using the DistilBERT tokenizer and applying or truncation to ensure uniform input length across samples. These steps were necessary to match the input requirement of the transformer-based model.

In this stage, each text was encoded into a sequence of tokens that is compatible with the model's vocabulary. Inputs are padded or truncated into a maximum of 512 tokens; this is to ensure the consistency across all samples. This step did not involve the manual feature engineering; this is because the transformer-based model will automatically capture semantic relationships between words and phrases.

3.5.3 Prediction Process

The model generated raw output values called logits, which corresponded to each of the 27 emotion classes. To convert these logits into interpretable probabilities, the sigmoid function was applied:

$$P(y_i) = \frac{1}{1 + e^{-z_i}}$$

In the formula, z_i represent the logit value for the emotion i. $P(y_i)$ is the probability that emotion i is present in the text. This transformation allowed for the independent evaluation of each emotion, enabling the model to assign the model to assign multiple emotions to a single post. Each emotion will have it own probability.

3.5.4 Filter high-risk emotions

To flag the potentially concerning posts, a threshold was applied to the predicted probabilities, a commonly used threshold value of 0.3 was chosen to balance the sensitivity and specificity.

$$\text{If } P(y_i) > 3, \text{ predict emotion } i \text{ is present}$$

This is to ensured that only emotions with reasonably high confidence scores were considered. In this stage, the attention was also be given to high-risk emotions such as 'sadness', 'fear', 'suicidal thoughts' and 'hopelessness', which these emotions are related to mental health crises.

3.6 XAI Model Interpretation

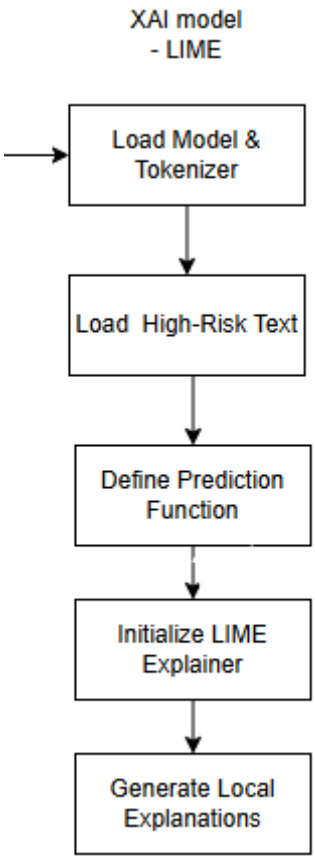

Figure 3.5 XAI model interpretation

The Explainable AI (XAI) stage of this project is to interpret the predictions made by the DistilBERT-based emotion classification model. Due to the dataset is scraped from social media platform and it is no labeled data; it cannot use traditional evaluation metrics such as accuracy and F1-score. Thus, interpretability become important to validate model decisions and understand which text contributed most to the predicted emotional states. In this project, Local Interpretable Model-Agnostic Explanations (LIME) was chosen to interpret the prediction outcome. LIME provides local explanation by approximating the behavior of complex models with simpler, interpretable models around individual predictions. This method enabled to highlight the key phrases that influenced the model's decision-making process, especially those associated with high-risk emotions.

3.6.1 Load Pretrained Model & Tokenizer

The DistilBERT-based emotion classifier and its corresponding tokenizer were load into memory. This is to ensure compatibility with the input format expected by the model. This allowed the system to pass cleaned social media text through the model and retrieve the predicted emotion probabilities.

3.6.2 Selected High-Risk Texts

From the emotion classification stage, there was a subset of text the containing high-risk emotions such as ‘sadness’, ‘fear’, ‘anxiety’ and ‘suicidal thoughts’ was selected for deeper interpretation using LIME. These texts represented potential concerning expressions of emotional problem and were prioritized for explanation due to their relevance to mental health crises detection.

3.6.3 Define Prediction Function

A wrapper function was created to convert the raw text input into a model output which is the emotion probability scores. This function includes tokenized the input text, applies padding and return emotion probabilities from DistilBERT model. This function enabled LIME to simulate how small changes in the text that affected the model’s output.

3.6.4 Initialize LIME Explainer

An instance of Lime Text Explainer was initialized using the full list of emotion labels from the model configuration. This is to ensured the explanations aligned directly with the emotion categories being predicted. The explainer was configured to return explanation in a human-readable format, highlighting which words is contributed positively or negatively to each prediction.

3.6.5 Generate Local Explanations

For each selected post the LIME generated a local explanation by perturbing the original text, observing how the model's prediction changed and last fitting a simple, interpretable model to approcimate the DistilBERT model's behaviour. The process can be summarized as follows:

$$explanation\ (x) = \arg\min_{g \in G} L(\hat{f}, g, \pi_x) + \Omega(g)$$

The explanation model for instance x is the model g that minimizes loss L, which measures how close the explanation is to prediction of the original model $\hat{f}$, while the model complexity $\Omega(g)$ is kept low. G is the family of possible explanations. The proximity measure π_x defines how large the neighborhood around instance x is that consider for the explanation.

In this project, LIME focus on minimizing the loss function L. Constraints such as maximum number of features to include in the explanation have define. This approach allowed the system to generate local feature importance weight, showing which words most strongly influenced the prediction of the high-risk emotions such as 'sadness' or 'anxiety'.

3.7 Tool and Platforms

Algorithms	• DistilBERT model - LIME
Software	• Operating System: Windows 11 • Language : Python 3.8 • Software: Jupyther Notebook - Dependency : Numpy, Pandas, re, nltk, PRAW
Hardware	• CPU : Intel Core Ultra 7

	• GPU : Intel ARC Graphics • Storage : 16GB

3.8 Summary

In this chapter describes the complete process of mental health crises prediction based on the text data collected from social media. The process is divided into 4 phase which is data collection, data preprocessing, emotion classification and XAI model interpretation.

For the data collection phase, dataset collected by do the web scraping from the social media platform which is Reddit. The feature collected will include the post text, timestamp and the author.

The data preprocessing phase is to ensure the data is ready for the next stage thought normalize text, remove unwanted content, remove special characters and whitespace, and tokenize.

The next phase is emotion classification; this phase is to define the emotion for each of the social media text. In this phase, a model called DistilBERT model was used to do the emotion prediction of each text. This is a fine-tuned model that can predict 27 emotions.

The last stage is XAI model interpretation. This stage is important due to the collected dataset did not have label because that cannot see the accuracy of the prediction. Thus, for this project XAI model was used to interpret the model decision making process. This is to understand how the prediction is done and let the prediction can be trust.