DETECTING SIGNS OF MENTAL HEALTH CRISES IN MALAYSIAN SOCIAL MEDIA TEXT USING EMOTION CLASSIFICATION AND EXPLAINABLE AI

CHOONG ZI XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	CHOONG ZI XUAN			
Student's Matric No.	:	MCS241038	Academic Session	:	SEMESTER 2, 2024/2025
Date of Birth	:	09 JUNE 2001	UTM Email	:	choongzixuan@graduate.utm.my
Project Report Title	:	DETECTING SIGNS OF MENTAL HEALTH CRISES IN MALAYSIAN SOCIAL MEDIA TEXT USING EMOTION CLASSIFICATION AND EXPLAINABLE AI			

I declare that this project report is classified as:

☒ **OPEN ACCESS** I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the project report belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this project report for academic exchange.

Signature of Student:

Signature :

Full Name: CHOONG ZI XUAN
Date: 30/6/2025

Approved by Supervisor(s)

Signature of Supervisor I:

Full Name of Supervisor I
ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN

Date :

Signature of Supervisor II

Full Name of Supervisor II

Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

“I hereby declare that I have read this project report and in my opinion this project report is sufficient in term of scope and quality for the award of the degree of Master in Data Science”

Signature : ______________________________

Name of Supervisor I : ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN

Date :

Signature : ______________________________

Name of Supervisor II :

Date :

Signature : ______________________________

Name of Supervisor III :

Date :

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Choong Zi Xuan and University Teknology Malaysia (UTM)

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

DETECTING SIGNS OF MENTAL HEALTH CRISES IN MALAYSIAN SOCIAL MEDIA TEXT USING EMOTION CLASSIFICATION AND EXPLAINABLE AI

CHOONG ZI XUAN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Doctor in Data Science

School of Computing
Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this project report entitled *“Detecting Signs of Mental Health Crises in Malaysian Social Media Text Using Emotion Classification and Explainable AI”* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name : CHOONG ZI XUAN

Date : 30 JUNE 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, ASSOC. PROF. DR. MOHS SHAHIZAN BIN OTHMAN, for encouragement, guidance, critics and friendship. Without his continued support and interest, this thesis would not have been the same as presented here.

My fellow postgraduate student should also be recognised for their support. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

The increasing of mental health issues in Malaysia has highlighted the need for innovative approaches to identify early signs of emotional distress in online communication. This project explores the use of Natural Language Processing (NLP) techniques to detect potential mental health concerns from Malaysian social media posts collected from Reddit. A ttal of 13,726 posts with comments were gathered using Python Reddit API Wrapper (PRAW) based on keywords like ‘mental health’, ‘stress’, ‘depression’, and ‘anxiety’. After that some preprocessing step such as text normalization, stop word removal, and lemmatization was applied to the Reddit dataset. Emotion classification model based on DistilBERT ‘bhadresh-savani/distilbert-base-uncased-emotion’ was applied to classify each post with comment into one of six emotions like anger, fear, sadness, joy, love, and surprise. After that, emotions like sadness and fear were identified as high-risk indicators of mental health crises was flagged out which have 6,035 posts were classified with these emotions. To improve interpretability and trust in model decisions, the Local Interpretable Model-Agnostic Explanations (LIME) was used to explain why certain classifier were made. It highlighting the key words that contributed most to the emotion detection. To further assess model performance, a training and evaluation phase was introduced. The model was fine-tuned using the publicly available dataset ‘dair-ai/emotion’ and evaluated on the Reddit-based pseudo-labeled test set. Based on the outcome, it shows a steady decrease in training loss and validation loss increased over time, this was the signs of overfitting. The Evaluation metrics showed very low accuracy (7.5%) and F1-score (0.18), this shows the models trained on general English emotion dataset may not perform well when applied to Malaysian English expression. This is because Malaysian English often include mix language of Malay and English and have informal tone in the text. This project contributes to both research and practical application by developing an end-to-end pipeline for detecting emotional distress from social media content. It also demonstrates the value of Explainable AI (XAI) in validating model outputs when true labels are unavailable.

ABSTRAK

Peningkatan isu kesihatan mental di Malaysia telah menimbulkan keperluan pendekatan inocatif untuk mengesan tanda-tanda awal kesusaan emosi dalam komunikasi dalam talian. Project ini menyiasat penggunaan Teknik Prmprosesan Bahasa Semula Jadi (NLP) untuk menyiasat kemungkinan isu Kesihatan mental daripada hantaran media social dart pengguna Malaysia yang dikumpulkan dari platform Reddit. Sebanyak 13,726 post berserta komen telah dikumpulkan menggunakan Python Reddit API Wrapper (PRAW) berdasarkan kata kunci seperti ‘mental health’, ‘stress’, ‘depression’, and ‘anxiety’. Selepas itu, beberapa langkah menbersih seperti normalisasi, penyingkiran stop words, dan lematisasi telah dilaksanakan untuk memastikan kualiti data sebelum analisis lanjut. Sebual model klasifikasi emosi berdasarkan DistilBERT iaitu ‘'bhadresh-savani/distilbert-base-uncased-emotion’ digunakan untuk mengklasifikasikan setiap post dengar komen ke dalam salah satu daripada enam emosi (anger, fear, sadness, joy, love, dan surprise). Emosi seperti kesedihan (sadness) dan ketakutan (fear) dikenal sebagai indicator risiko tinggi yang berkaitan dengan krisis. kesihatan mental. Sebanyak 6,035 posts telah dikelaskan sebagai berisiko tinggi menghadapi ksisis kesihatan mental. Untuk meningkatkan keyakinan terhadap Keputusan model, Teknik Explainable AI (XAI) seperti LIME digunakan bagi menjelaskan bagaimana dan mengapa model membtal ramalan tersebut. LIME menbantu mengenalkan perkataan-perkataan utama yang memberi Kesan kepada pengesanan emosi, serta memvalidasi sama ada Keputusan model dibuat berdasarkan ciri linguistic yang bermakna atau tidak bermakna. Fasa tambahan melibatkan Latihan dan penilaian model, di mana model DistilBERT ditinfan menggunakan dataset emosi berlabel ‘dair-ai/emotion’ dan dinilai menggunakan Reddit-based pseudo-labeled test set. Hasil memnunjukkan kehilangan latihan berkurangan secara bertahap, tetapi kehilangan validasi meningkat. Ia menunjukkan tanda-tanda overfitting. Metrik penilaian menunjukkan kejituan yang renday (7.5%) dan F1-score (0.18), ia membuktikan bahawa model dilatih menggunakan dta emosi dalam Bahasa Inggeris umum mumgkin tidak relecan apabila digunakan pada ekspresi Bahasa Inggeris tempatan Malaysia yang sering kali merangkumi kodetokaran antara Bahasa Melayu dan Bahasa Inggeris serta gaya penulisan uang tidak formal. Projeck ini memberi sumbangan kepada kedua-dua aspek penyelidikan dan aplikasi praktikal dengan Pembangunan pipeline hujung ke hujung untuk mengesan kesusaan emosi daripada kandungan media social. Ia juga menunjukkan kepentinggan teknologi Explaniable AI (XAI) dalam memvalidasi output model Ketika label sebenar tidak tersedia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

XAI	-	Explainable Artificial Intelligence
CNN	-	Convolutional Neural Network
LIME	-	Local Interpretable Model-agnostic Explanations
SVM	-	Support Vector Machine
KNN	-	K-nearest neighbour
NB	-	naïve Bayes
RNN	-	Recurrent Neural Network
NLP	-	Natural Language Processing
RF	-	Random Forest
DT	-	Decision Tree
LG	-	Logistic Regression
XGBoost	-	eXtreme Gradient Boosting
ML	-	Machine Learning
PRAW	-	Python Reddit API Wrapper
SHAP	-	Shapley Additive Explanations

CHAPTER 1

INTRODUCTION

1.1 Introduction

Mental health crises are a critical issue that have affected over 970 million peoples globally. Among the mental health crises depression and anxiety were the most popular problems ((WHO), 2019). These issues not only impact individual life but also affect society and economic ((WHO), Mental disorders, 2022). Other than that, 720,000 lives also be taken by suicide due to mental problem ((WHO), Suicide prevention, 2021) . No matter male, female, adults or children, they may be suffering for mental health problem (Kamal, et al., 2020). Depression, bipolar disorder, autism spectrum disorder, schizophrenia and other psychoses are the most comment mental health crises (Zhang, Yang, Shaoxiong, & Ananiadou, 2023). According to the study, depression is a mental health crisis that cannot be detect by traditional clinical methods (Tahir, et al., 2025). Since Covid-19 people been lock down and quarantine, the use of social media such as Facebook, X, WhatsApp and Reddit has increased (Banna, Ghosh, Md. Jaber Al Nahian, Mahmud, & Taher, 2023). Increasing of social media used also increase the use of social media's users to express their mental problems or illness with other people who they did not know (Kim, Lee, Park, & Han, 2020). Explanation Artificial Intelligence (XAI) let people can easily understand the outcome of the machine learning (Hulsen, 2023). XAI such as Local Interpretable Model-agnostic Explanations (LIME) offer user friendly visualization for user can easily interpret the machine learning outcome (Gerlings, Shollo, & Constantiou, 2021). Thus, this study aims to use machine learning to do the mental health crises prediction for social media while using XAI methos like SHAP to interpret machine learning.

1.2 Problem Background

Social, economic and environmental are the factors that affect a person mental health. Mental health crises commonly appear for people who live in urban area due to people social disparity, social security problem, pollution and connection with nature is not enough (Antonio, Julio, & M., 2020). Estimate 20% to 47% of emerging adults have face mental health crises in the preceding year (Eric, Punyanunt-Carter, R.LaFreniere, S.Norman, & G.Kimball, 2020). Early intervention for mental health crises is important but people feel stigma to associated with depression thus more than 60% of people that face depression did not seek help from professional (Bao, Pérez, & Parapar, 2024).

Social media is a source for people to communication and interaction with other people. People will share their emotion, thoughts and opinions on the social media (Kamal, et al., 2020). This open opportunity for psychiatrists for early detection for mental health crises based on the data from social media platform (William & Suhartono, 2020). Growing of the social media platform such as Twitter, Reddit, Facebook, Instagram, and Weibo let researcher can use machine learning and deep learning method to analyse user behaviour patterns and text use for the post or comment to detect depression (Tahir, et al., 2025).

According to the study of Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han use XGBoost and convolutional neural network (CNN) to do the prediction of mental health based on social media post. To avoid the prediction, have multiple symptoms, they developed 6 independent models for each symptom. Based on their study, their classification the prediction into depression, anxiety, bipolar disorder, schizophrenia, and autism (Kim, Lee, Park, & Han, 2020).

Based on the study of Brian, et al., they used Natural Language Processing (NLP) to detect and interpret language patterns of the mental health crises (Bauer, et al., 2024). The study use sentence embeddings based on large language models to extract latent linguistic dimensions of user posts from several mental health-related subreddits, with a focus on suicidal tendencies. In this study they analysed 2.9 million posts extracted from 30 subreddits. The result of this study shows that the users who wrote about feelings disconnection, burdensomeness, hopeless, desperation, resignation and trauma have the trend of suicide.

Kernel support vector machine (SVM), random forest, logistic regression K-nearest neighbour (KNN), and complement naïve Bayes (NB) are the 5 machine learning models use by Kabir, et al. to do the detection of depression. In this study, the study conducted is using Bengali text-based data from blogs and open-source platforms. The result of this study shows that recurrent neural network (RNN) models have the highest accuracies while GRUs have 81% of accuracy. Kernal SVM have 78% accuracy on the test data. (Kabir, Islam, Kabir, Haque, & Rhaman, 2022).

Most of the recent work show effectiveness in deep learning methods to detect depression or mental health crises but most of it not provide explainable to the detection of depression (Zogan, Razzak, Wang, Jammel, & Xu, 2022). The studies focus on achieving better classification results and does not explain and interpret about the classification methods (Bao, Pérez, & Parapar, 2024). Lack of transparency and explainable of the decision made by machine learning has led to the introduction of XAI (Minh, Wang, Li, & Nguyen, 2021).

XAI show the process and explain how the machine learning made the decisions (Minh, Wang, Li, & Nguyen, 2021). XAI is the process of opening the “black box” of the machine learning because complex and “black box” type of machine learnings can lead to dangerous or fatal consequences (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021). SHAP and LIME is the most exploited XAI methods. SHAP explain the role of each feature for all instances and for a specific instance while LIME only explain specific instance in the machine learning (Salih, et al., 2024).

Based on the study of the Jo et al., they use several machine learning methods to do the detection of depression. The machine learning used include Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbours (KNN), Gradient Boosting, and Decision Tree classifiers. Among these machine learning methods, Random Forest have the highest accuracy rate which score 99.30% of accuracy. In this study they also use XAI models such as SHAP and LIME to interpret the predictions. By using SHAP and LIME they more understand how the machine learning does the predictions and get know about the key role of the feature that determining an individual's depression state (Jo, Raj, Vino, & Menon, 2024).

Study of the Tang, et al., they use the data that collected from Colombo South Teaching Hospital-Kalubowila in Sri Lanka. They first use NLP to do the text preprocessing. Based on their study, their find that the words like "physical disabilities", "mental disorders", "chronic disease", and "family disputes" have high frequency appear on the post of a person who suicide. They use various of machine learning do the prediction of the suicide risk. The machine learning used included Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Perceptron, and eXtreme Gradient Boosting (XGBoost). Their result show that Random Forest model has the best result. They also use XAI like SHAP to interpret the machine learning prediction. Their study shows that, anger issues, depression and lack of socialisation is the top issues that cause suicide (Tang, et al., 2024).

Based on recent study, there is not study significant in Malaysia. Thus, this study will focus on dataset from Malaysia and using DistilBERT model to do the emotion classification while flag out the post with emotion that have high risk in mental health crises. Lastly, using XAI method like SHAP to interpret the machine learning to increase the transparency and trustworthy of the prediction.

1.3 Problem Statement

Based on the recent study, the following statement of the problems are address:

(a) Social media posts contain early signs of mental health crises such as depression but these signals are not systematically analysed for intervention.

(b) The prediction of machine learning is not transparent and let people cannot trust the prediction outcome.

(c) Existing research lacks culturally and linguistically tailored models for Malaysia populations. No previous studies have use local datasets for crises prediction.

1.4 Research Aim

This study aims to design and evaluate a machine learning framework that prioritization of mental health crises in social media data while using Explanation Artificial Intelligence (XAI) to enhance the transparency and trustworthy of the predictions

1.5 Research Questions

This study finds the answer for the following question:

(a) What textual and emotional features in social media posts indicate mental health crises?

(b) How effective are machine learning models in prediction mental health crises from social media data?

(c) How can explainable artificial intelligence (XAI) techniques improve the interpretability and clinical utility of model predictions?

1.6 Research Objectives

Based on the recent study, the following research objectives are address:

(a) To collect and preprocess mental health related social media data from Malaysian Reddit users for emotion-based analysis.

(b) To apply a pretrained DistilBERT emotion classification model to detect high-risk emotional states in Malaysian social media text.

(c) To interpret the model predictions using Explainable AI (XAI) techniques.

1.7 Research Scope

The scope of this study includes following objectives:

(a) Data Sources: This study will do the web scrapping to get the real time dataset from Reddit.

(b) Target Conditions: This study aim to get the outcome of emotion classification with flag out the post with emotion that have high-risk for mental health crises.

(c) Technical Focus: This study will use DistilBERT model to do the prediction of mental health crises and LIME to interpret the machine learning.

1.8 Significance of Study

This study helps solve the important problems that related to mental health and social media. This study has highlighted the following significance:

(a) Better tools for early detection of mental health crises: social media is a platform for people to share their emotion and thinkings. Their sharing might include early sign of mental health crises. This study uses machine learning like Random Forest to detect the signs in real time.

(b) Making machine learning prediction easier to understand: Most machine learning models work like “black box” which the machine learning just give the result but did not explain why this result. This make the trustworthy of the prediction is low. In this study, XAI tools like SHAP use to show how the machine learning makes the decision and increase the trustworthy of the machine learning.

(c) Focus on Malaysia region: Based on the recent study there is no one have studied mental health crises in Malaysia using local social media data. By analysing Reddit posts in Malaysia, this study predicts Malaysian mental health crises.

1.9 Thesis Organization

This thesis is divided into 4 chapter. The first chapter about the introduction of the study which include the problem background, problem statement, research aim and objective, and research scope. For the second chapter which is literature review. This chapter is about the research that have done and the paper that have been read. Chapter 3 is research methodology and this chapter is about the method that use in this study such as data collection, feature engineering and model design. Chapter 4 is initial finding and analysis which in this chapter will show the outcome and the result from data collection, data preprocessing until the model evaluation. This chapter will also do the discussion for the model result. Chapter 5 is the conclusion for this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter provides an overview of existing research that related to the detection of mental health crises using social media data and machine learning techniques. It begins with an exploration of the role of social media in modern communication and its growing use as a platform for emotional expression. This chapter also reviews the key concepts of mental health while focusing on depression, anxiety and suicidality. Machine learning algorithms use in the detection or prediction like Natural Language Processing (NLP), Random Forest, SVM, Logistic Regression and more algorithms also be state in this chapter. This chapter introduces Explainable Artificial Intelligence (XAI) such as SHAP and LIME that help in improve transparency and trust of the model. Lastly, this chapter study about the recent study in the field and highlight the trends, performance metrics and limitation of the papers.

2.2 Social media

Social media have a great impact on the world today. Social media is a web or mobile platform that allows people to communicate and interact with others through virtual networks. People can share, create and exchange their emotions, thoughts and opinions in digital form (A.Naslund, Bondre, Torous, & A.Aschbremmer, 2020). According to the research in year 2021, more than 50% of the world population have account for social media and in a day people spend two and a half hours to use the social media (Braghieri, Levy, & Makarin, 2022).

According to the research, in 2021 YouTube and Facebook was the social media platform that have most widely used online platform, which reported 81% and 69% ever using these sites (Auxier & Anderson, 2021). But in the research in year

2022 shown that among teenagers, YouTube, TikTok and Instagram were the top 3 social media platform use by them while for Facebook only have 32% of teenagers say they ever using it (A.Vogels, Gelles-Watnick, & Massarat, 2022). However, Facebook still is a hot social media for people to share their lives.

2.3 Mental Health

Mental health allows people to remain emotionally strong ang resilient in the face of life's ups and downs. It helps people grow, learn, perform well in work, and communication with others. Mental health plays an important role for people lives and contribution to the world (WHO, 2025). Based on the study of Fusar-Poli, et al., 2020, a good mental health means people can handle the normal stress of daily life and complete the daily work. It helps people feel okay most of the time and keep people continue going even face problem (Fusar-Poli, et al., 2020). Mental health crises including depression, anxiety, and suicidality (Eleftheriades, Fiala, & Pasic, 2020).

2.3.1 Depression

First mental health crises is depression. Depression is a serious mood condition that causes continue sadness, low mood, cannot think clearly, and loss of interest in things used to matter (Dobrek & Glowacka, 2023). Persistent changes in sleep, appetite, energy, and thoughts about ending your life are also symptoms of depression, according to the National Institute of Mental Health (Sarno, Moeser, & Robison, 2021). Depression is a serious mental health issue that affecting people and is the top cause of disability linked to mental health. It is a very common mental health crises which about 4.4% of the global population experiencing it. Many people first get depression during 14-25 years old, with around 4 to 5 % of young people be effect by depression in these age (Marwaha, et al., 2023). Depression will affect people daily life such as school and work performance and relationship between people.

2.3.2 Anxiety

Anxiety is about the feeling of worry, fear or nervousness that make people feel uneasy. It will cause physical symptoms such as heart beat fast, sweating and leave people restless (Health, n.d.). It is normal to feel anxiety when feel stress but when the condition becomes worse which the anxiety did not go away and start to affecting daily life it become anxiety disorders. Estimated 4% of the global population affect by anxiety disorder ((WHO), n.d.). Even though there are very effective treatments for anxiety disorders but only small amount of people receives the treatment which around 1 in 4 people. Anxiety disorders are the most frequently diagnosed mental health crises in children and teenagers (Walter, et al., 2020).

2.3.3 Suicidality

Suicidality including a range of experience which start from having suicidal thoughts until making a plan, to actual suicide attempts and complete the suicide. For young people now, suicide thought may be frequently such as the time when feel like life is not worth living, they will have the thought to end their life (Becker & Correll, 2020). Suicide is a major cause of death around the world with almost 1 million lives lost due to suicide in a year. Suicide attempts are also most common occur for teenagers while the risk of dying by suicide rise with age for teenagers (Carballo, et al., 2020). The key factors that cause teenagers suicidality including external pressures like bullying, sleep problems and the use of antidepressants. There are also personal vulnerabilities play as a key role causing suicidality such as gender, mental health struggles, sexual orientation and history of previous suicidal thoughts or self-harming behaviors (Richardson, et al., 2024).

2.4 Natural Language Processing

Natural Language Processing (NLP) is a part of artificial intelligence that handle difficult and complex language-related tasks. It covers tasks like translating text between languages, answering questions and creating summaries. NLP focuses on developing algorithms, system and models that enable computers to understand and interact with human language (Lauriola, Lavelli, & Aiolli, 2022). The main goal of NLP is to simply processing text or speech as a string of characters or sentences. Other than that, NLP also treats language as complex data that carries structure, meaning and sound patterns. This allows NLP models to pick up on meaning and generate useful results in numerical form (Locke, et al., 2021). NLP let people can easily to collaborate and communication with computer. NLP also offers advantages across many industries and applications like improved data analysis and insights, improve content generation, enhance search, and automation of repetitive tasks (Stryker & Holdsworth, 2024).

2.4.1 NLP in mental health section

NLP have highly used as part for mental health detection. Pillai et al. have used NLP techniques like Word Frequency Analysis to identify the most common words or phrases that related to mental health (Pillai, Polimetla, Avacharmal, & Perumal, 2022). NLP was used to evaluate the volunteers' adherence to conversational techniques and formats, as well as to gain insights into different demographic user groups and their behaviors in expressing stress. (Liu, et al., 2021)

2.5 Emotion classification models

Emotion classification is used to detect the emotional content in the input text and determine what kind of emotional content is present based on different methods. There are several DistilBERT model use for emotion classification.

2.5.1 bhadresh-savani/distilbert-base-uncased-emotion

'bhadresh-savani/distilbert-base-uncased-emotion' model is a version pf DistilBERT model that has been fine-tuned on the Twitter-Sentiment-Analysis dataset for emotion classification. This model is 40% smaller and 60% faster than original BERT model. This model also retaining 97% of the original model language understanding capabilities. This model achieves accuracy of 93.8% and F1-score of 93.79% on the test set with can test for 398.69 samples per second which faster compared to other models. This model can classify 6 emotion label which are sadness, fear, anger. Joy, love, and surprise. (bhadresh-savani, 2025)

2.5.2 joeddav/distilbert-base-uncased-go-emotions-student

‘joeddav/distilbert-base-uncased-go-emotions-student’ model is distilled from the zero-shot classification pipeline on the unlabeled GoEmotions dataset. It was trained with mixed precision for 10 epochs. This model can classify 27 emotions, which include sadness, joy, anger, and other emotions. (joeddav, distilbert-base-uncased-go-emotions-student, 2025)

2.6 Machine Learning (ML)

Machine learning algorithms work by analyzing data and using the information to learn and get better at making the prediction. Different with traditional programming, ML not need to be told what exactly to do, ML will improve automatically through experience (Ha, Nguyen, & Stoeckel, 2024).

2.6.1 Random Forest

Random Forest is a tree-based model works by splitting the data into smaller groups over and over again, based on certain rules or conditions, until a stopping point is reached. Decision trees conclude with terminal nodes known as leaf nodes or leaves, which are the points for the final predictions or decisions are generated (Schonlau & Zou, The random forest algorithm for statistical learning, 2020). Proper model training requires setting 3 key hyperparameters in advance which are the minimum size of each node, the total number of trees in the forest, and the number of features randomly sampled at each split (IBM, n.d.). Random forest uses a straightforward analysis method to build the decision trees by selecting the nodes. Random forest selects the root nodes, internal nodes and leaf nodes based on the same set of attributes and information. This process remains consistent regardless of the specific criteria used for splitting the data (Wijaya & Rachmat, 2024).

2.6.2 Logistic Regression

Logistic Regression is a supervised learning technique used in machine learning to analyse data and model the relationship between one or more predictor variables and a binary response variable (Wijaya & Rachmat, 2024). It helps to understanding how the input features influence the likelihood of a particular outcome. In logistic regression, the outcome variable is binary, means that only have 2 possible categories which the occurrence of an event is represented by 1 while the non-occurrence event is assigned a value of 0 (Alves, et al., 2020). Other than that, logistic regression uses a logistic function to turn its outputs into probabilities, making it easier to interpret the likelihood of an event happening. One major advantage of logistic regression is that the model's coefficients are easy to understand. It shows how strongly each predictor influence the outcome and in what direction, often explained using odds ratios (Kumar & Gota, 2023). To assess how well the model performs, metrics like accuracy, sensitivity, specificity, they are under the ROC curve (AUC) are commonly used (Kumar & Gota, 2023) (Schonlau, Logistic Regression, 2023).

2.6.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful machine learning algorithm used for classification and regression (K & Wong, 2023) (Khanduja & Kaur, 2023). SVM work by transform the input data into a new higher-dimensional space by kernel function such as linear, polynomial, Gaussian which establishing the best decision boundary to separate classes. Ability of SVM to handle complex and non-linear relationships has made it become a popular tool across many domains which from document classification and drug design to image classification. One of the advantages of SVM is that it is versatility which it is applicable in binary classification and continuous outcome prediction, making it a useful tool for a great range of machine learning problems (Hasija & Chakraborty, 2021) (Feizi & Nazemi, 2022). In addition, SVM maximizes the margin between classes to reduce overfitting and enhance generalization on new and unseen data (Hasija & Chakraborty, 2021). SVM model performance is typically evaluated against the standard performance metrics such as accuracy, F-Score and ROC curves that provide insights into their predictive performance and reliability (Khanduja & Kaur, 2023).

2.6.4 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a powerful language representation model that has significantly advanced the field of the natural language processing (NLP) and frequently used in healthcare (Chang, et al., 2023). Its ability to understand context in text has led to remarkable improvements in tasks such as classification, prediction, and protocol selection. BERT modal was used for sentiment analysis on Twitter data and achieving 87% of accuracy with proficiency in handling Twitter's linguistic nuances (Renuka & Radhakrishnan, 2024).

2.6.5 Naïve Bayes

Naïve Bayes is a predictive model based on Bayesian analysis, used Bayes' Theorem to calculate the probability of different outcomes (Acito, 2023). Naïve Bayes is a widely used classification algorithm due to its simplicity and efficiency. The method assumes that all predictor variables are conditionally independent given the class label which is a strong assumption that, while often unrealistic, significantly simplifies the computation of probabilities and allows for fast model training. (Vishwakarma & Ganguly, 2023). It widely used in natural language processing, spam detection, and sentiment analysis (Kumar, Goswami, Mhatre, & Agrawal, 2024).

2.6.6 CNN

Convolutional Neural Networks (CNN) are a popular group of neural networks that are designed to efficiently process image data by considering local and global characteristic of the input data (Pinaya, Vieira, Garcia-Dias, & Mechelli, 2020). CNN use a unique architecture that includes convolutional layers, which automatically learn important features from input data using filters traversing the image, generating activation maps showing important patterns (Convolutional Neural Networks, 2022). This ability to directly learn hierarchical features from uncooked data makes CNN powerful for computer vision and image recognition tasks. Other than convolutional layers, CNN also typically include pooling layers which reduce the spatial size of the data but retain significant details like improving computational efficiency and generalization of the models (Das & Ahmed, 2023). Due to its high success rate, CNN are widely used in a number of high impact tasks like medical imaging and object detection, where it delivers accuracy and performance at scale.

2.7 Explainable AI (XAI)

Explainable AI (XAI) is a crucial field focused on making artificial intelligence system become more transparent and understandable to people. As AI model become more complex, it increasing the need to understand the decision-making process especially on the area or sector that the outcome that will impact human lives (Zodage, Harianawala, Shaikh, & Kharodla, 2024). XAI be used to enhance the trust and accountability between human and AI. It enhances the trust and accountability by explain the decision-making process; this ensure that AI system not just powerful but also reliable and fair. XAI also support better collaboration between AI and human by offering clear explanations that lead to more informed decisions. Besides that, XAI help in identifying biases and weaknesses of AI system. This let developers can continue improve the performance and ensure long-term reliability (Zodage, Harianawala, Shaikh, & Kharodla, 2024). SHAP and LIME are the 2 model of XAI.

2.7.1 SHAP

Shapley Additive Explanations (SHAP) is a framework designed to enhance understanding of machine learning models by measuring the contribution of each feature (Choi, Shin, & Shin, 2024). SHAP work based on cooperative game theory, specifically the Shapley value which assigns each feature based on its contribution to the model's output (Herren & Hahn, 2022). SHAP can apply in both local and global interpretation which people can use SHAP to understand not only specific predictions but also overall model behavior (Scheda & Diciotti, 2022). However, SHAP also have limitation. Its effectiveness can be affected by the choice of feature distributions and the complexity of the model, which may lead to challenges in explanation (Herren &

Hahn, 2022). While SHAP increase the transparency of the model but it may oversimplify complex model interaction and may lead to misinterpretation of feature importance. Therefore, while SHAP is a valuable tool in enhance the trustability of the model, but it also should be used thoughtfully and with awareness of its constraints in different contexts.

2.7.2 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a technique that enhances the interpretability of black box machine learning models. By using simple interpretable models to figure out the model's behavior around a specific instance, it produces local explanations (Zafar & Khan, 2021). Different with SHAP, LIME only can apply for local interpretation. Limitation of LIME including it generate explanations through random perturbation that might lead to instability and varying results for the same prediction (Zafar & Khan, 2021). LIME also requires users to define interpretable components which this can lead to bias and limit the scope of explanations (Angiulli, Fassetti, & Nisticò, 2021).

2.8 Existing work in mental health prediction

In recent years, researchers have increasingly focused on leveraging social media data and machine learning techniques to detect early signs of mental health crises. These studies have explored various platform such as Reddit, Twitter (X), Facebook and blogs, using both classical machine learning and deep learning models to analyze textual content for indicators of depression, anxiety, suicidal ideation and other mental health conditions.

Based on the research form Garg et al., they use decision tree, random forest, and BERT models on Reddit data to identify mental health crises. Their result showed that BERT have the best performant among 3 machine learning models which achieving an accuracy, precision, recall, and F1-score of 0.82. this indicating its effectiveness in capturing complex linguistic features. However, the authors reported some limitations, such as that they did not provide evidence for using the most active users to gain better accuracy, and they limited their study to a single platform, Reddit, limiting the generalizability of their model across other platforms or even other populations within the platform (Garg, Garg, Dixit, & Pandey, 2024).

Similarly, Lim et al. study focus on the dataset from Twitter and using SVM, decision tree and Naïve Bayes algorithms in the study to detect mental health disease. Among the algorithms, SVM have the highest accuracy which highlighting the potential of traditional machine learning methods here. Nevertheless, the study had small samples of the dataset as well as a narrow focus only to Twitter, which may not be representative of broader online patterns on other platforms (Lim, Kamarudin, Ismail, Ismail, & Kamal, 2023).

Another related work is the study by Odja et al., the study compared KNN, random forest, and neural network for sentiment analysis models based on Reddit posts dataset. The result of the study showed that random forest work best with an F1-score, accuracy, precision, and recall of 80.6%. Even though the study shows positive result but the dataset use in this study was really small, comprising 350 data samples, which can influence model reliability and generalization (Odja, Widiarta, Purwanto, & Ario, 2024).

On the other hand, Qorich & Ouazzani study use lightweight deep learning algorithms like BERT and CNN in their study. They use these 2 algorithms for stress detection on both Twitter and Reddit datasets. Based on their study, BERT achieved 85.67% accuracy on a small Reddit dataset, while CNN attained 97.62% accuracy on a larger Twitter dataset. However, they study use different algorithm for different platform to do the detection (Qorich & Ouazzani, 2025).

Table 2.1 Comparison of existing work on mental health crises prediction

Title	Author	Dataset	Model Used	Result	Limitation
Machine learning Driven Analysis of Mental Health Indicators in social media Posts	(Garg, Garg, Dixit, & Pandey, 2024)	Reddit	Decision Tree, Random Forest, BERT	BERT have the highest result for accuracy, precision, recall and F1-score. 0.82 for all.	Lack of evidence in utilizing the most active web people for obtaining the most accuracy results The study focuses on Reddit, limit the generalizability of the findings to other social media platforms or demographics
Predicting Mental Health Disorder on Twitter Using Machine Learning Techniques	(Lim, Kamarudin, Ismail, Ismail, & Kamal, 2023)	Twitter	SVM, Decision Tree, Naïve Bayes	SVM have the highest accuracy.	Just focus on the data on Twitter The dataset used limit and small
Mental illness detection using sentiment	(Odja, Widiarta, Purwanto, & Ario, 2024)	Reddit	KNN, Random Forest, Neural Network	Random Forest have the best performance with 80.6% for F1-score,	The dataset is small that only consist 350 columns of data.

analysis in social media				accuracy, recall and precision.	
Lightweight advanced deep learning models for stress detection on social media	(Qorich & Ouazzani, 2025)	Reddit, Twitter	Lightweight deep learning methods, BERT, CNN	BERT achieved 85.67% of accuracy on small Reddit dataset; CNN reached 97.62% accuracy on Large Twitter dataset.	Performance varied across platform Platform-specific tuning required
Integrating Machine Learning and Sentiment Analysis: A Comparative Study on Mental Health Classification from Social Media Data	(Kaushik & Sharma, 2024)	Reddit, Twitter, Kaggle	Decision Tree, Logistic Regression, XGBoost	XGBoost have the highest accuracy (82%), logistic regression has 78% of accuracy and decision tree have 67% of accuracy.	There are some mistakes in classification "Stress" and "Personality Disorder" even it has the highest accuracy.

2.9 Summary

This chapter provides a comprehensive review of what have been published on the identification of mental health crises using social media and machine learning. It begins with discussing the role of the social media as a platform for emotional expression and how it is related in identifying early signs of mental health issues such as depression, anxiety and suicidality. This chapter also review Natural Language Processing (NLP) and various machine learning algorithms that use in detecting mental health crises such as Random Forest, SVM, Logistic Regression, BERT, Naïve

Bayes and CNN. This chapter also discuss about the Explainable AI (XAI) like SHAP and LIME that are the tools to improve transparency and trustworthiness. There are also review in the recent work which including the models used, source of the datasets, result and the limitation of the work. This current study aims to address these shortcomings by developing a local mental health crises prediction system using Malaysian Reddit data.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Overview

In this chapter, introduce a complete emotion prediction for a collected dataset. The process including 5 stages, that is data collection, data preprocessing, emotion classification, XAI model interpretation, and model training and evaluation. 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 5 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. The next 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. The last phase is model training and evaluation. This phase is to evaluate how well the emotion classifier generalizes to local media text. Figure 3.1 show the proposed research framework.

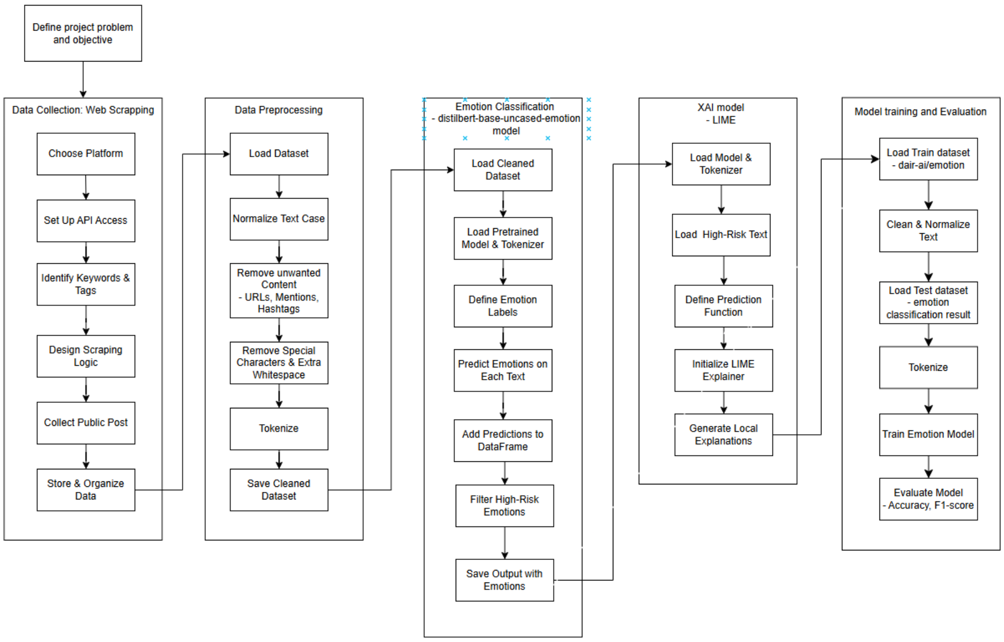

Define project problem and objective
Data Collection: Web Scrapping
Choose Platform
Set Up API Access
Identify Keywords & Tags
Design Scraping Logic
Collect Public Post
Store & Organize Data
Data Preprocessing
Load Dataset
Normalize Text Case
Remove unwanted Content - URLs, Mentions, Hashtags
Remove Special Characters & Extra Whitespace
Tokenize
Save Cleaned Dataset
Emotion Classification - distilbert-base-uncased-emotion model
Load Cleaned Dataset
Load Pretrained Model & Tokenizer
Define Emotion Labels
Predict Emotions on Each Text
Add Predictions to DataFrame
Filter High-Risk Emotions
Save Output with Emotions
XAI model - LIME
Load Model & Tokenizer
Load High-Risk Text
Define Prediction Function
Initialize LIME Explainer
Generate Local Explanations
Model training and Evaluation
Load Train dataset - dair-ai/emotion
Clean & Normalize Text
Load Test dataset - emotion classification result
Tokenize
Train Emotion Model
Evaluate Model - Accuracy, F1-score

Figure 3.1Proposed research framework

3.3 Data Collection

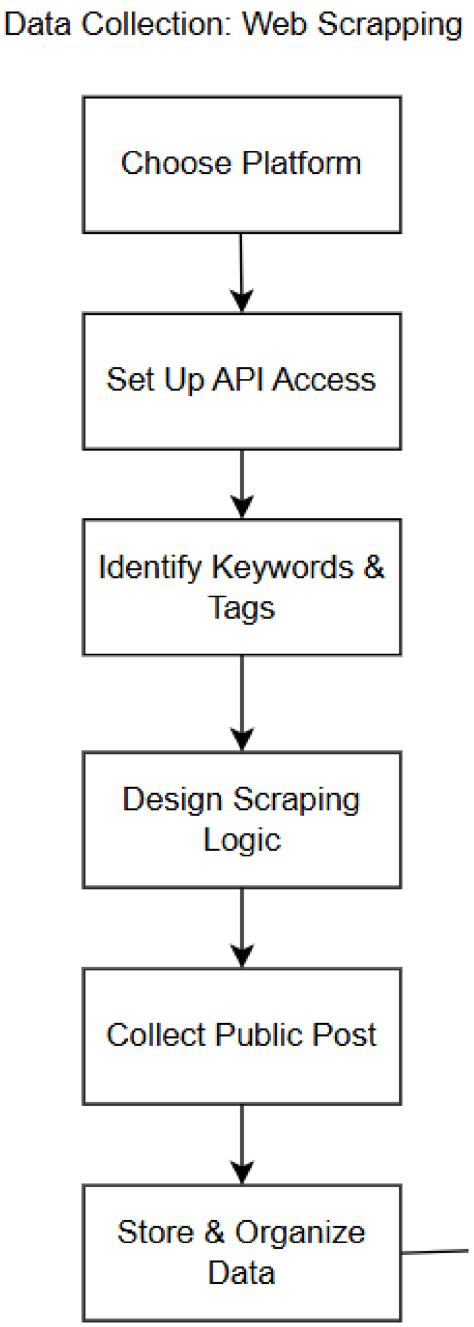


Figure 3.2: Data Collection

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 used 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

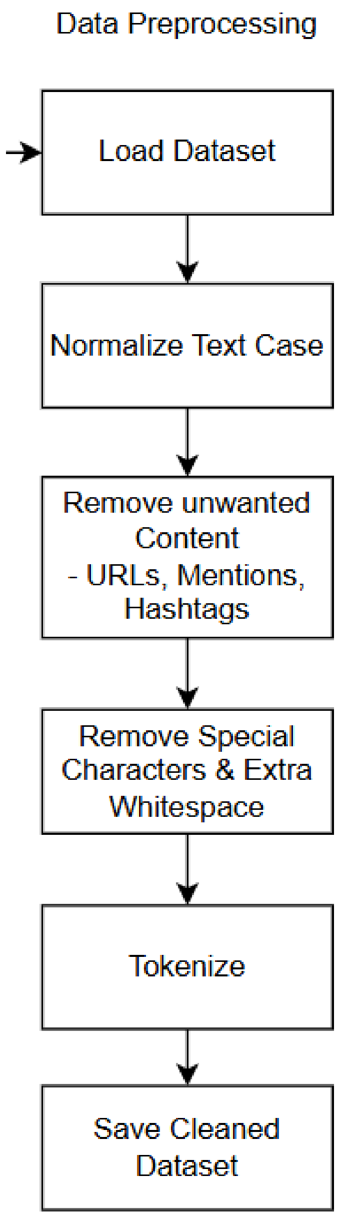

Data Preprocessing
Load Dataset
Normalize Text Case
Remove unwanted Content - URLs, Mentions, Hashtags
Remove Special Characters & Extra Whitespace
Tokenize
Save Cleaned Dataset

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

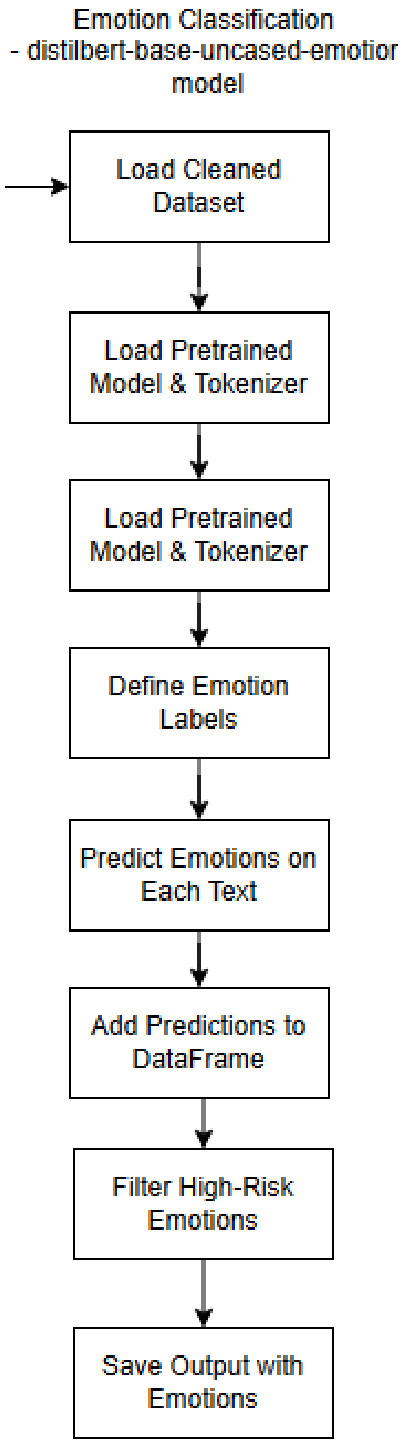

Emotion Classification
- distilbert-base-uncased-emotior
model
Load Cleaned Dataset
Load Pretrained Model & Tokenizer
Load Pretrained Model & Tokenizer
Define Emotion Labels
Predict Emotions on Each Text
Add Predictions to DataFrame
Filter High-Risk Emotions
Save Output with Emotions

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 'bhadresh-savani/distilbert-base-uncased-emotion' 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. This model classify text into 6 emotion labels such as 'sadness', 'fear', 'anger', 'joy', 'love', and 'surprise'. 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

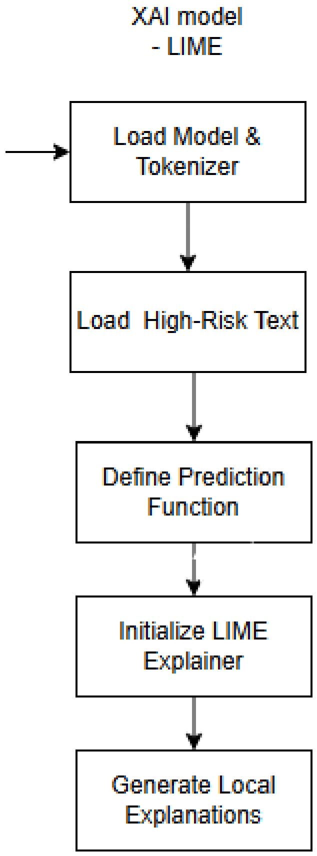

XAI model
- LIME
Load Model & Tokenizer
Load High-Risk Text
Define Prediction Function
Initialize LIME Explainer
Generate Local Explanations

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 ensure 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:

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

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 Model training and Evaluation

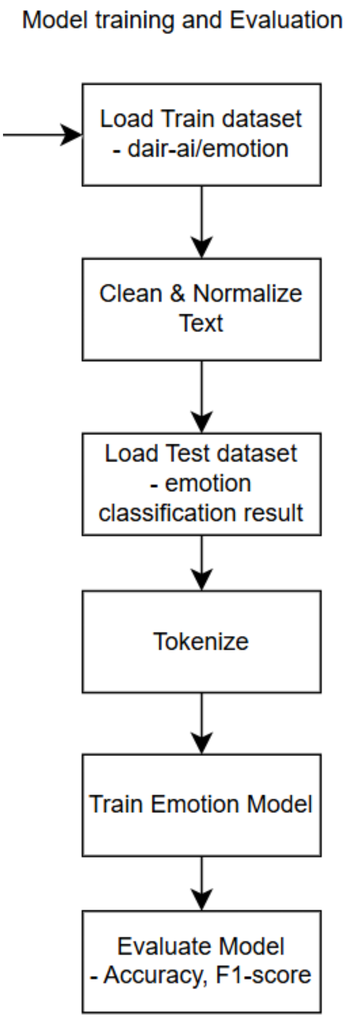

Model training and Evaluation
Load Train dataset - dair-ai/emotion
Clean & Normalize Text
Load Test dataset - emotion classification result
Tokenize
Train Emotion Model
Evaluate Model - Accuracy, F1-score

Figure 3.6: Model Training and Evaluation

This section outlines the model training and evaluation strategy adopted in this project to assess the performance of the DistilBERT-based emotion classification model when applied to Malaysian Reddit post. The primary objective of this phase was to evaluate how well the pretrained emotion classification model generalizes to real-world, culturally text from Malaysian users discussing mental health on Reddit. Since the dataset collected from Reddit was unlabeled, in this stage a publicly available dataset ('dair-ai/emotion') was use as the train set for the model and evaluate the model performance on the Reddit-based pseudo-labeled test set. The Reddit-based pseudo-labeled test set is result done when the emotion classification phase.

3.7.1 Load Train Dataset

In this project, the ‘dair-ai/emotion’ dataset was used as the training dataset due to its availability of labeled emotional states and it been used to train the model ‘bhadresh-savani/distilbert-base-uncased-emotion’. This dataset contains 20,000 English text with emotion label. The emotion label consists 6 emotion which are sadness, anger, joy, fear, love and surprise.

3.7.2 Clean and Normalize text

To ensure consistency between the training and test dataset, all text samples underwent preprocessing steps such as lowercasing, removal of special characters, stop word filtering, and lemmatization. These techniques helped reduce noise and improve the model’s ability to learn meaningful patterns.

3.7.3 Load Test Dataset

The test dataset consisted of Reddit posts that have emotion label, which were previously classified using DistilBERT emotion classification model. As the original dataset did not have true labels, the predicted emotions from the model were treated as pseudo-labels. The cleaned and normalized version of the Reddit dataset was used as input, while the predicted emotion labels were mapped to numeric class IDs for compatibility with the training process.

3.7.4 Tokenize

Text inputs from both datasets were tokenized using DistilBERT tokenizer ‘AutoTokenizer.from_pretrained("bhadresh-savani/distilbert-base-uncased-emotion")’. Each text sample was converted into numerical tokens, padded or truncated to a standard length of 128 tokens to match the input requirement of the DistilBERT architecture. This step ensured that both the training and testing data were in a format compatible with the transformer-based model.

3.7.5 Train Emotion Model

The DistilBERT model 'bhadresh-savani/distilbert-base-uncased-emotion' was fine-tuned using the Hugging Face Trainer API. The model was configured to run for 3 epochs with a batch size of 16 and a learning rate of 2e-5. Evaluation and saving strategies were set to occur after ever epoch to monitor validation loss and prevent overfitting. The training process aimed to adapt the model to detect emotional cues more effectively in informal, conversational English found in Reddit discussions.

3.7.6 Evaluate Model

After completing the model training phase, the DistilBERT emotion classification model was evaluated to assess its performance on the test dataset. The Evaluation process involved using standard classification metrics such as accuracy and F1-score. These metrics were selected to measure how well the model predicted emotional states from text samples.

The accuracy metric was used to determine the overall proportion of correctly predicted emotion labels out of all predictions made; it provided a general overview class: sadness, anger, joy, fear, love, and surprise.

In addition to accuracy, the F1-score with macro averaging was applied to account for class imbalance and to ensure that each emotion category contributed equally to the final score. This was particularly important in this project, as some emotions may appear more frequently than others in the dataset.

These metrics were computed after each training epoch to monitor the learning behavior of the model and to detect early signs of overfitting or underfitting. The evaluation was conducted using the validation set derived from the 'dair-ai/emotion' dataset and the pseudo-labeled Reddit-based test set.

This stage allowed for an objective assessment of the model's predictive capability and provided insights into whether further adjustments or fine-tuning would be necessary to improve performance when applied to Malaysian English text.

3.8 Tool and Platforms

Table 3.1: Tool and Platform used

Algorithms	• DistilBERT model - LIME
Software	• Operating System: Windows 11 • Language : Python 3.8 • Software: Google Colab, Jupyter Notebook - Dependency : Numpy, Pandas, re, nltk, PRAW
Hardware	• CPU : Intel Core Ultra 7 • Storage : 16GB

3.9 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 5 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 fourth 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.

The last stage is model training and evaluation. This stage is to evaluate how well the pretrained emotion classification model generalizes to real-world Malaysian English text. Evaluate metrics such as accuracy and F1-score were applied to measure the performance of the model.

In this chapter, it also shows the tool and platform that been used for the project. In this project, the programming language used is python and the platform used is Google Colab and Jupyter Notebook.

CHAPTER 4

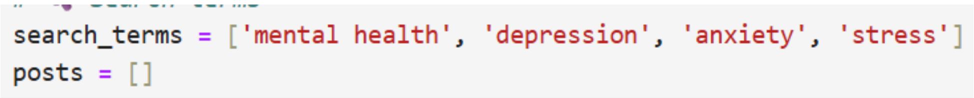
Initial Findings and Analysis

4.1 Overview

This chapter present a comprehensive overview of the initial findings and analytical process for this project that focused on detecting potential mental health crises using Malaysian Reddit data. This chapter involved several key stages, including web scrapping, data preprocessing, text cleaning, exploratory data analysis (EDA), emotion classification using DistilBERT, high-risk post identification and model interpretation using XAI techniques like LIME.

4.2 Web Scraping

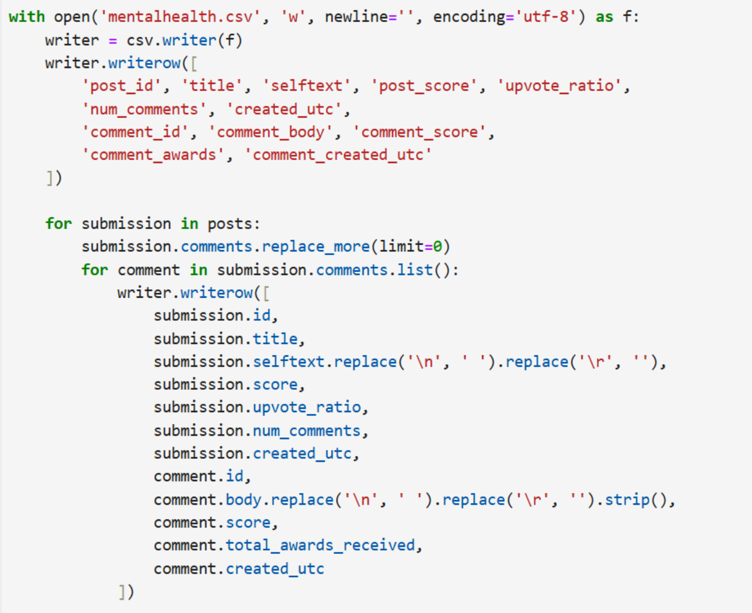
In this project, the dataset the social media text post from Malaysia Reddit post. Thus, in this project web scraping for social media Reddit was done. Library 'PRAW' was used to scrape the data from Reddit. The first step of the web scraping is to define the key words, this will make sure when scrape the post, it will only take the post that consist of the key words. Due to this project is about mental health crises, the key words set are "mental health", "depression", "anxiety", and "stress". Other than that, to scrape only Malaysian post, subreddit was set to "Malaysia" to try searching r/Malaysia for each keyword.



```
search_terms = ['mental health', 'depression', 'anxiety', 'stress']
posts = []
```

Figure 4.1: Key word to scrape for social media post.

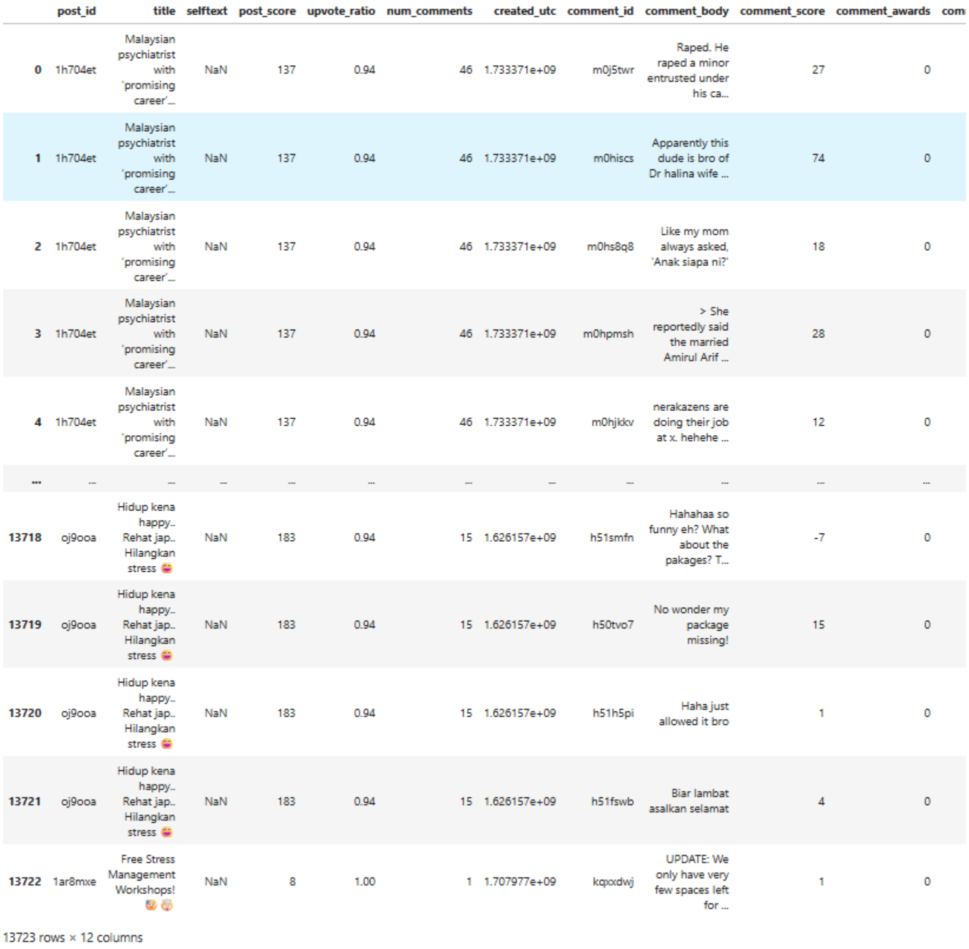
In this project, it did not just scrape for the post title and text, but it also scrapes for the post comments. Thus, the dataset for web scraping consists of 12 columns which is “post_id”, “title”, “selftext”, “post_score”, “upvote_ratio”, “num_comments”, “created_utc”, “comment_id”, “comment_body”, “comment_score”, “comment_awards”, and “comment_created_utc”.



```
with open('mentalhealth.csv', 'w', newline='', encoding='utf-8') as f:
    writer = csv.writer(f)
    writer.writerow([
        'post_id', 'title', 'selftext', 'post_score', 'upvote_ratio',
        'num_comments', 'created_utc',
        'comment_id', 'comment_body', 'comment_score',
        'comment_awards', 'comment_created_utc'
    ])

    for submission in posts:
        submission.comments.replace_more(limit=0)
        for comment in submission.comments.list():
            writer.writerow([
                submission.id,
                submission.title,
                submission.selftext.replace('\n', ' ').replace('\r', ''),
                submission.score,
                submission.upvote_ratio,
                submission.num_comments,
                submission.created_utc,
                comment.id,
                comment.body.replace('\n', ' ').replace('\r', '').strip(),
                comment.score,
                comment.total_awards_received,
                comment.created_utc
            ])
```

Figure 4.2: Code for the scraping details.

After scrape the data was saved into a CSV file for further used. In this project, the web scraping has successfully scrape for 13,723 data. Thus, the final scrape dataset consists of 13,723 rows and 12 columns.



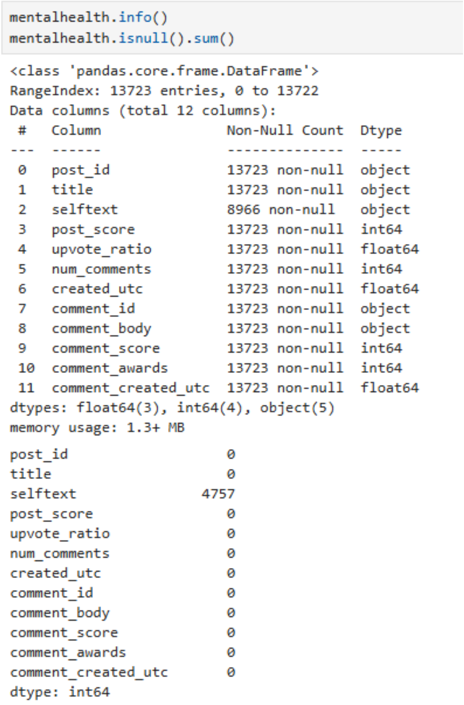
	post_id	title	selftext	post_score	upvote_ratio	num_comments	created_utc	comment_id	comment_body	comment_score	comment_awards	com
0	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0j5twr	Raped. He raped a minor entrusted under his ca...	27	0	
1	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hiscs	Apparently this dude is bro of Dr halina wife ...	74	0	
2	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hs8q8	Like my mom always asked, 'Anak siapa ni?'	18	0	
3	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hpmsh	> She reportedly said the married Amirul Arif ...	28	0	
4	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hjkkv	nerakazens are doing their job at x. hehehe ...	12	0	
...	...	...	...	...	...	...	...	...	...	...	...	
13718	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😆	NaN	183	0.94	15	1.626157e+09	h51smfn	Hahahaa so funny eh? What about the pakages? T...	-7	0	
13719	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😆	NaN	183	0.94	15	1.626157e+09	h50tvo7	No wonder my package missing!	15	0	
13720	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😆	NaN	183	0.94	15	1.626157e+09	h51h5pi	Haha just allowed it bro	1	0	
13721	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😆	NaN	183	0.94	15	1.626157e+09	h51fswb	Biar lambat asalkan selamat	4	0	
13722	1ar8mxe	Free Stress Management Workshops! 😰 🤯	NaN	8	1.00	1	1.707977e+09	kqxxdwj	UPDATE: We only have very few spaces left for ...	1	0	

13723 rows × 12 columns

Figure 4.3: Data that scrape from Reddit

4.3 Data Preprocessing

Data preprocessing involving some step like check for missing value and fill in with suitable things. The first step of data preprocessing is check for the dataset info and check for missing value. In this project these two step was run together. Form the figure below, there have the info for the dataset, which data type for each column was showed and below it has the missing value for each column. The column 'post_id', 'title', 'selftext', 'comment_id', and 'comment_body' was object, which mean that the data was words. Columns like 'post_score', 'num_comments', 'comment_score', and 'comment_awards' data type was integer, while column 'upvote_ratio', 'created_utc', and 'comment_created_utc' data type was float number. The dataset also has missing value for column 'selftext' which have 4757 missing values.



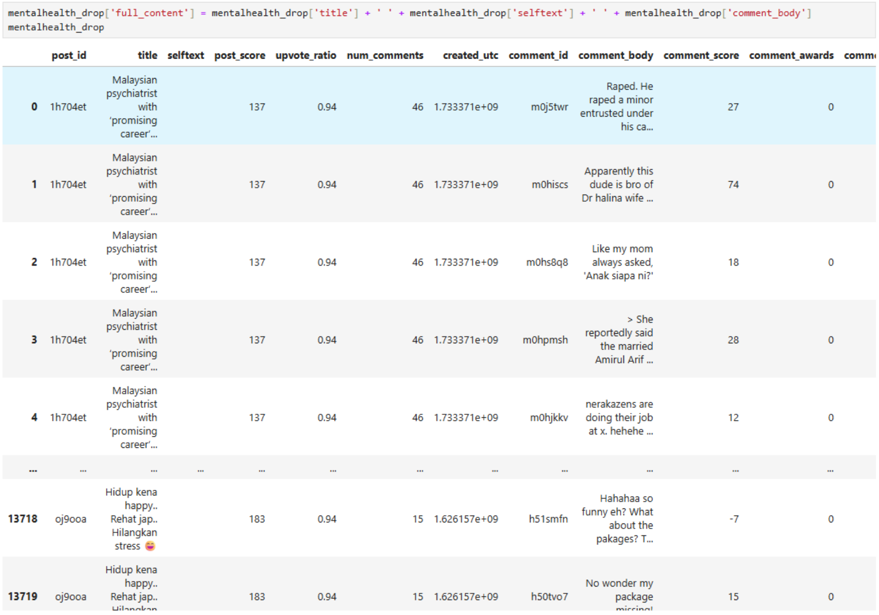
```
mentalhealth.info()
mentalhealth.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13723 entries, 0 to 13722
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   post_id             13723 non-null  object 
 1   title               13723 non-null  object 
 2   selftext            8966 non-null   object 
 3   post_score          13723 non-null  int64  
 4   upvote_ratio        13723 non-null  float64
 5   num_comments        13723 non-null  int64  
 6   created_utc         13723 non-null  float64
 7   comment_id          13723 non-null  object 
 8   comment_body        13723 non-null  object 
 9   comment_score       13723 non-null  int64  
 10  comment_awards      13723 non-null  int64  
 11  comment_created_utc 13723 non-null  float64
dtypes: float64(3), int64(4), object(5)
memory usage: 1.3+ MB
```

```
post_id                   0
title                     0
selftext               4757
post_score                0
upvote_ratio              0
num_comments              0
created_utc               0
comment_id                0
comment_body              0
comment_score             0
comment_awards            0
comment_created_utc       0
dtype: int64
```

Figure 4.4: Info and count of missing value of the dataset.

To handle missing values in the 'selftext' column, empty entries were replaced with whitespace.

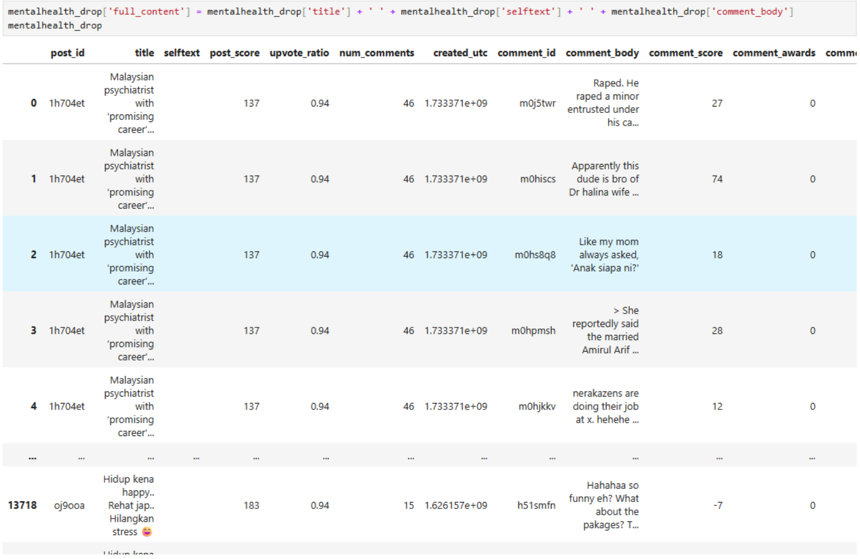


```
mentalhealth_drop['full_content'] = mentalhealth_drop['title'] + ' ' + mentalhealth_drop['selftext'] + ' ' + mentalhealth_drop['comment_body']
mentalhealth_drop
```

	post_id	title	selftext	post_score	upvote_ratio	num_comments	created_utc	comment_id	comment_body	comment_score	comment_awards	comm
0	1h704et	Malaysian psychiatrist with 'promising career'...		137	0.94	46	1.733371e+09	m0j5twr	Raped. He raped a minor entrusted under his ca...	27	0	
1	1h704et	Malaysian psychiatrist with 'promising career'...		137	0.94	46	1.733371e+09	m0hiscs	Apparently this dude is bro of Dr halina wife ...	74	0	
2	1h704et	Malaysian psychiatrist with 'promising career'...		137	0.94	46	1.733371e+09	m0hs8q8	Like my mom always asked, 'Anak siapa ni?'	18	0	
3	1h704et	Malaysian psychiatrist with 'promising career'...		137	0.94	46	1.733371e+09	m0hpmsh	> She reportedly said the married Amirul Arif ...	28	0	
4	1h704et	Malaysian psychiatrist with 'promising career'...		137	0.94	46	1.733371e+09	m0hjkkv	nerakazens are doing their job at x. hehehe ...	12	0	
...	...	...	...	...	...	...	...	...	...	...	...	
13718	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😝		183	0.94	15	1.626157e+09	h51smfn	Hahahaa so funny eh? What about the pakages? T...	-7	0	
13719	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan		183	0.94	15	1.626157e+09	h50tvo7	No wonder my package missing!	15	0	

Figure 4.5: Fill in the missing value with white space.

Before do the data cleaning for the text data, the column that have important information like ‘title’, ‘selftext’ and ‘comment_body’ were combine together to make it become more meaning full.



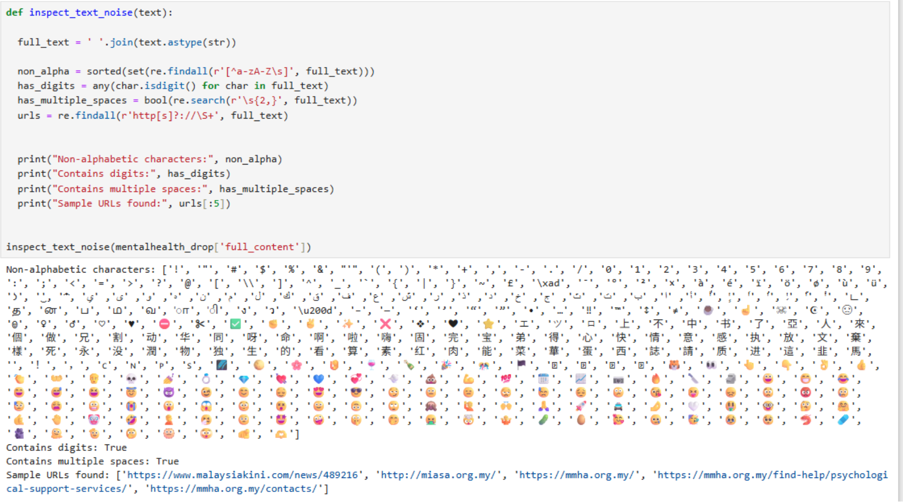
```
mentalhealth_drop['full_content'] = mentalhealth_drop['title'] + ' ' + mentalhealth_drop['selftext'] + ' ' + mentalhealth_drop['comment_body']
mentalhealth_drop
```

	post_id	title	selftext	post_score	upvote_ratio	num_comments	created_utc	comment_id	comment_body	comment_score	comment_awards	comm
0	1h704et	Malaysian psychiatrist with ‘promising career’...		137	0.94	46	1.733371e+09	m0j5twr	Raped. He raped a minor entrusted under his ca...	27	0	
1	1h704et	Malaysian psychiatrist with ‘promising career’...		137	0.94	46	1.733371e+09	m0hiscs	Apparently this dude is bro of Dr halina wife ...	74	0	
2	1h704et	Malaysian psychiatrist with ‘promising career’...		137	0.94	46	1.733371e+09	m0hs8q8	Like my mom always asked, 'Anak siapa ni?'	18	0	
3	1h704et	Malaysian psychiatrist with ‘promising career’...		137	0.94	46	1.733371e+09	m0hpmsh	> She reportedly said the married Amirul Arif ...	28	0	
4	1h704et	Malaysian psychiatrist with ‘promising career’...		137	0.94	46	1.733371e+09	m0hjkkv	nerakazens are doing their job at x. hehehe ...	12	0	
...	...	...	...	...	...	...	...	...	...	...	...	
13718	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😄		183	0.94	15	1.626157e+09	h51smfn	Hahahaa so funny eh? What about the pakages? T...	-7	0	
		Hidup kena										

Figure 4.6: Combine 3 columns into one new column.

4.4 Data cleaning

Data cleaning was the most important step before the emotion classification. In this phase, it will do several steps to clean the data to make sure it ready for the model to do the emotion classification. The first step of data cleaning was finding all the text noise that were not alphabet. Based on the Figure 4.7, there was many non-alphabetic characters find in the dataset. And the output show that have digits, multiple space and links in the dataset.

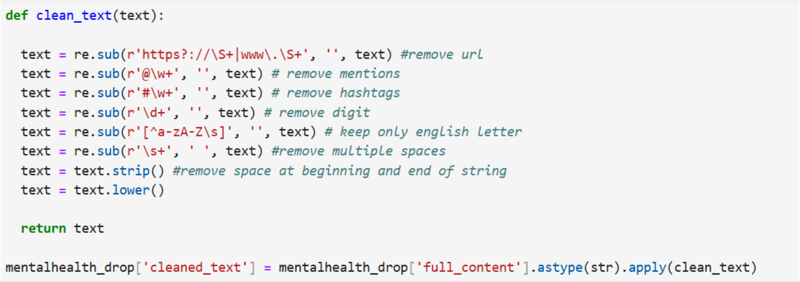


```
def inspect_text_noise(text):

  full_text = ' '.join(text.astype(str))

  non_alpha = sorted(set(re.findall(r'[^a-zA-Z\s]', full_text)))
  has_digits = any(char.isdigit() for char in full_text)
  has_multiple_spaces = bool(re.search(r'\s{2,}', full_text))
  urls = re.findall(r'http[s]?://\S+', full_text)


  print("Non-alphabetic characters:", non_alpha)
  print("Contains digits:", has_digits)
  print("Contains multiple spaces:", has_multiple_spaces)
  print("Sample URLs found:", urls[:5])


inspect_text_noise(mentalhealth_drop['full_content'])
```

```
Non-alphabetic characters: ['!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9',
':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~', '£', '\xad', '¯', '°', '²', '×', 'à', 'é', 'ï', 'ö', 'ø', 'ù', 'ü',
'ڈ', 'ڀ', 'ٰ', 'ي', 'ى', 'و', 'ه', 'ن', 'م', 'ل', 'ك', 'ق', 'ف', 'ع', 'ش', 'ر', 'ذ', 'د', 'خ', 'ج', 'ث', 'ت', 'ب', 'ا', 'أ', 'ٌ', 'ٍ', 'َ', 'ُ', 'ِ', 'ّ', 'ْ', 'ㄴ',
'த', 'ன', 'ப', 'ம', 'வ', 'ா', 'ி', 'ง', 'ว', '\u200d', '–', '—', '‘', '’', '“', '”', '•', '…', '‼', '™', '↕', '≠', '☕', '☝', '☠', '☪', '☹',
'☺', '♀', '♂', '♡', '♥', '⛔', '✂', '✅', '✊', '✌', '✨', '❌', '❖', '❤', '⭐', 'エ', 'ツ', 'ロ', '上', '不', '中', '书', '了', '亞', '人', '來',
'個', '做', '兄', '割', '动', '华', '同', '呀', '命', '啊', '啦', '嗨', '固', '完', '宝', '弟', '得', '心', '快', '情', '意', '感', '执', '放', '文', '棄',
'樣', '死', '永', '没', '潤', '物', '独', '生', '的', '看', '算', '素', '红', '肉', '能', '菜', '華', '蛋', '西', '誌', '請', '质', '进', '這', '韭', '馬',
'', '！', '，', 'Ｃ', 'Ｎ', 'Ｐ', 'Ｓ', '🌌', '🌕', '🌸', '🍁', '🍷', '🍾', '🎉', '🎎', '🏴', '🏻', '🏼', '🏽', '🏾', '🐯', '👀', '👆', '👇', '👌', '👍',
'👏', '👐', '👴', '💀', '💅', '💍', '💎', '💘', '💙', '💞', '💭', '💩', '💪', '💯', '📅', '📈', '📼', '🔥', '🔪', '🗿', '😀', '😁', '😂',
'😃', '😅', '😆', '😇', '😈', '😉', '😊', '😌', '😍', '😎', '😏', '😐', '😑', '😒', '😓', '😔', '😕', '😘', '😝', '😞', '😟', '😡', '😢',
'😵', '😫', '😬', '😭', '😮', '😱', '😳', '😵', '🙂', '🙃', '🙄', '🙈', '🙋', '🙌', '🙏', '🚀', '🚔', '🤌', '🤍', '🤑', '🤓', '🤔', '🤗',
'🤙', '🤚', '🤡', '🤣', '🤦', '🤧', '🤨', '🤩', '🤪', '🤫', '🤭', '🤮', '🤯', '🤷', '🥒', '🥚', '🥰', '🥲', '🥳', '🥹', '🥺', '🦐', '🩹',
'🪻', '🫠', '🫡', '🫢', '🫣', '🫨', '🫵', '🫶']
Contains digits: True
Contains multiple spaces: True
Sample URLs found: ['https://www.malaysiakini.com/news/489216', 'http://miasa.org.my/', 'https://mmha.org.my/', 'https://mmha.org.my/find-help/psychologi
cal-support-services/', 'https://mmha.org.my/contacts/']
```

Figure 4.7: The code to find the noise of the text and the noise find.

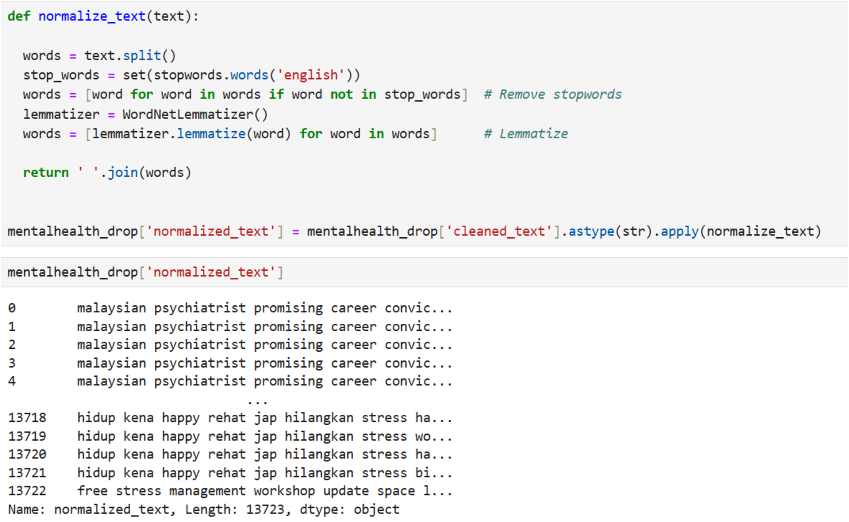
To make sure the dataset ready for analysis, the text cleaning was done such as removing the links, mentions, hashtags, digit, and multiple spaces. This step also makes sure that only remain the alphabet and lower down the text make sure the text all in lowercase. The Figure 4.8, show the code to remove all the text noise.



```
def clean_text(text):

  text = re.sub(r'https?://\S+|www\.\S+', '', text) #remove url
  text = re.sub(r'@\w+', '', text) # remove mentions
  text = re.sub(r'#\w+', '', text) # remove hashtags
  text = re.sub(r'\d+', '', text) # remove digit
  text = re.sub(r'[^a-zA-Z\s]', '', text) # keep only english letter
  text = re.sub(r'\s+', ' ', text) #remove multiple spaces
  text = text.strip() #remove space at beginning and end of string
  text = text.lower()

  return text

mentalhealth_drop['cleaned_text'] = mentalhealth_drop['full_content'].astype(str).apply(clean_text)
```

Figure 4.8: Remove all the text noise.

The last step of data cleaning is normalization text. In this step all the stop words were remove and done the lemmatize for the data. The Figure 4.9 show the code to do the normalization of the data and display the normalized text.



```
def normalize_text(text):

  words = text.split()
  stop_words = set(stopwords.words('english'))
  words = [word for word in words if word not in stop_words]  # Remove stopwords
  lemmatizer = WordNetLemmatizer()
  words = [lemmatizer.lemmatize(word) for word in words]      # Lemmatize

  return ' '.join(words)


mentalhealth_drop['normalized_text'] = mentalhealth_drop['cleaned_text'].astype(str).apply(normalize_text)
```

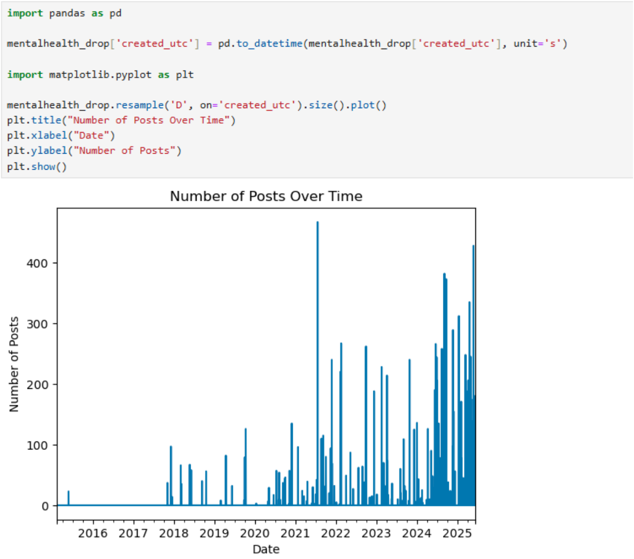
```
mentalhealth_drop['normalized_text']
```

```
0        malaysian psychiatrist promising career convic...
1        malaysian psychiatrist promising career convic...
2        malaysian psychiatrist promising career convic...
3        malaysian psychiatrist promising career convic...
4        malaysian psychiatrist promising career convic...
                               ...
13718    hidup kena happy rehat jap hilangkan stress ha...
13719    hidup kena happy rehat jap hilangkan stress wo...
13720    hidup kena happy rehat jap hilangkan stress ha...
13721    hidup kena happy rehat jap hilangkan stress bi...
13722    free stress management workshop update space l...
Name: normalized_text, Length: 13723, dtype: object
```

Figure 4.9: Normalization for the data.

4.5 Exploratory Data Analysis (EDA)

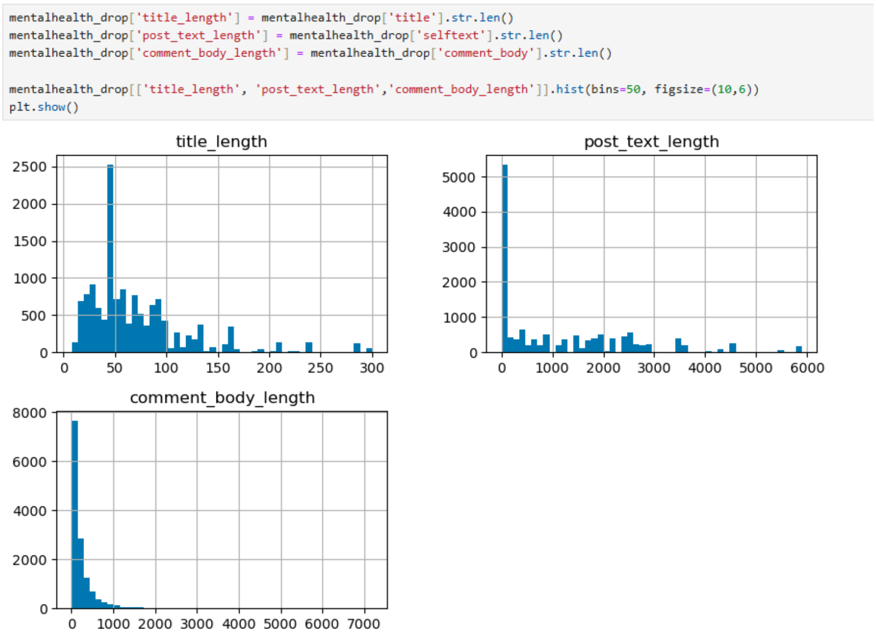
Exploratory Data Analysis (EDA) is the initial step of the data analysis. EDA help summarized and investigated using statistical graphics and other visualization methods. EDA helps uncover patterns, identify outliers.



```
import pandas as pd

mentalhealth_drop['created_utc'] = pd.to_datetime(mentalhealth_drop['created_utc'], unit='s')

import matplotlib.pyplot as plt

mentalhealth_drop.resample('D', on='created_utc').size().plot()
plt.title("Number of Posts Over Time")
plt.xlabel("Date")
plt.ylabel("Number of Posts")
plt.show()
```

Figure 4.10: The number of the posts over time.

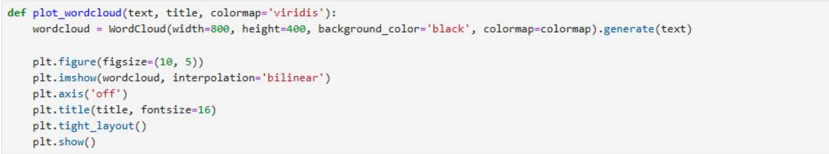
Based on the figure 4.10, the post web scrape dataset has the most post on 2021 which over 400 posts. In 2024 until now, the post number was more than 250 post in each time. The bar chart also shows that the number of posts was increasing in year while before 2018 there was no post in this dataset except in 2015 there were a few posts.



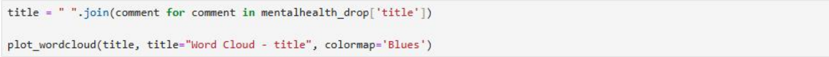
```
mentalhealth_drop['title_length'] = mentalhealth_drop['title'].str.len()
mentalhealth_drop['post_text_length'] = mentalhealth_drop['selftext'].str.len()
mentalhealth_drop['comment_body_length'] = mentalhealth_drop['comment_body'].str.len()

mentalhealth_drop[['title_length', 'post_text_length','comment_body_length']].hist(bins=50, figsize=(10,6))
plt.show()
```

Figure 4.11: Text length for title, post body and comment.

Based on the figure 4.11, it has 3 bar chart that show the text length for the title, post body, and comment body. Form the chart 1 which is the title length chart, show that 2500 of the data have 50 words in their title, the most number words of the in the data was 300 words, and most of the title words was between 10-100 words. Form the post text length chart, it shows that most of the post text have little words like 0-100 words for the post body which it rich more than 5000 post have this situation. Other post body length was between 300 until 3000 words. For the comment body chart, it shows that most of the comment body words length is between 0 – 100 words which almost 8000 posts. Other comment words are between 250 until 1500 words. From the 3 charts, it shows that title words length is the shorter and the post body and comment body words length are longer.



```
def plot_wordcloud(text, title, colormap='viridis'):
    wordcloud = WordCloud(width=800, height=400, background_color='black', colormap=colormap).generate(text)

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(title, fontsize=16)
    plt.tight_layout()
    plt.show()
```



```
title = " ".join(comment for comment in mentalhealth_drop['title'])

plot_wordcloud(title, title="Word Cloud - title", colormap='Blues')
```

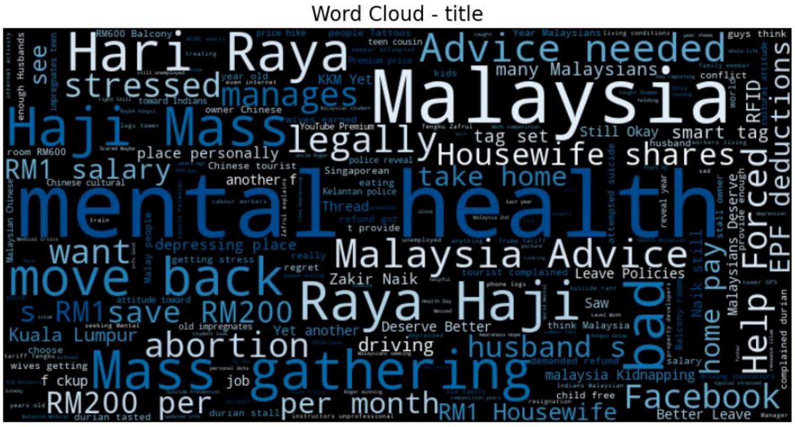
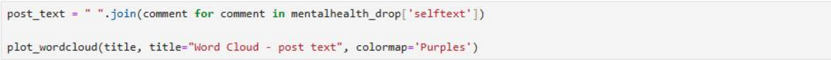


Figure 4.12: Word cloud for title.

To see the most comment words show in the title, post body and the comment, word cloud was used. By using word cloud the words that most frequency detect in the dataset will show bigger. For the title, it shows that most frequence word detect in the Malaysian Reddit post are mental health, Malaysia, Hari Raya, Raya Haji, Mass, gathering and other words. The title word cloud shows in Figure 4.12.



```
post_text = " ".join(comment for comment in mentalhealth_drop['selftext'])

plot_wordcloud(title, title="Word Cloud - post text", colormap='Purples')
```

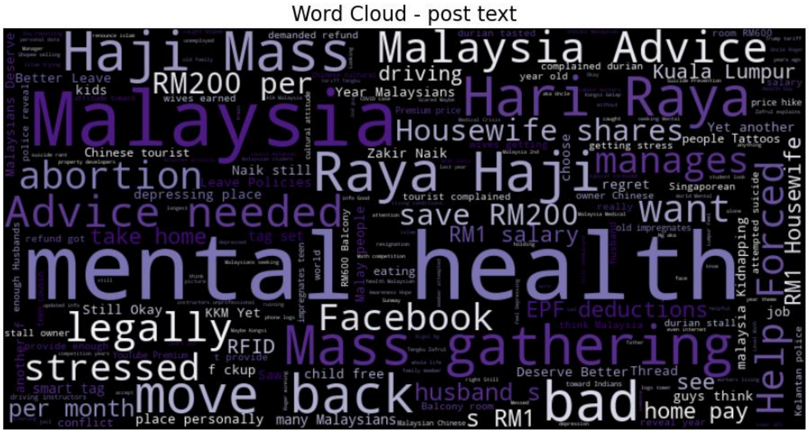
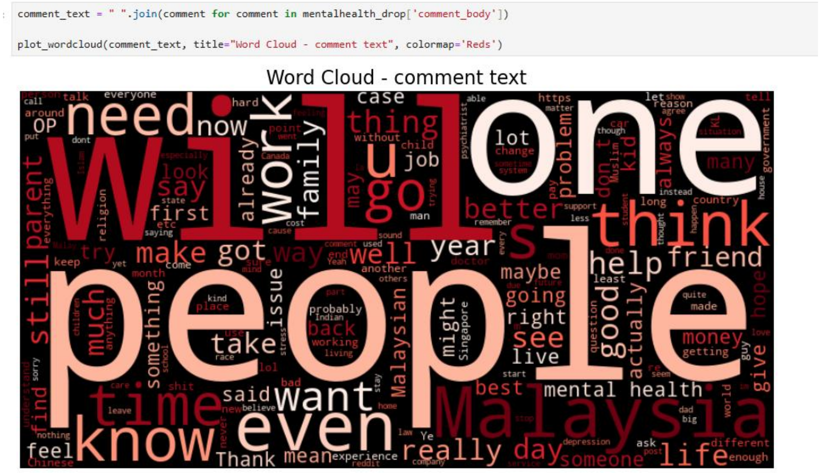


Figure 4.13: Word cloud for post body.

Figure 4.13 show the word cloud for the post body of Malaysian Reddit post. The word cloud shows that, the most frequent show words are mental health, Malaysia, Raya Haji, Mass gathering, Facebook, Hari Raya, legally and more.



```
comment_text = " ".join(comment for comment in mentalhealth_drop['comment_body'])

plot_wordcloud(comment_text, title="Word Cloud - comment text", colormap='Reds')
```

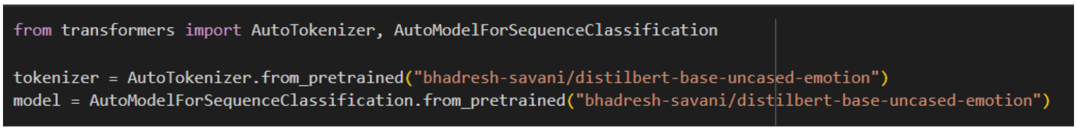
Figure 4.14: Word cloud for comment.

Figure 4.14 show the word cloud for the comment body. From the word cloud, it shows the words like people, will, one, think, Malaysia, time, even and more.

4.6 Emotion Classification Model Training

In this project, it trains the model for emotion classification. The model use was DistilBERT based model. The actual model's name is bhadresh-savani/distilbert-base-uncased-emotion. This is a pretrain model that use for emotion classification. This model can classify 6 emotion such as sadness, joy, love, anger, fear and surprise. This model will give the probabilities for each emotion based on the text given.

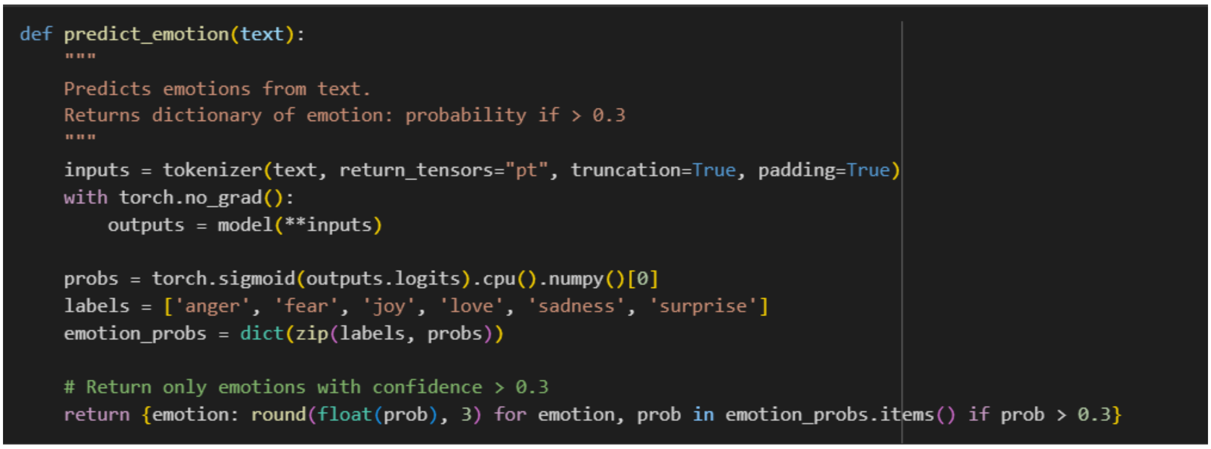
Before start train the dataset, the first step is to load the model. Figure 4.15 show that the code to load the model into notebook.



```
from transformers import AutoTokenizer, AutoModelForSequenceClassification

tokenizer = AutoTokenizer.from_pretrained("bhadresh-savani/distilbert-base-uncased-emotion")
model = AutoModelForSequenceClassification.from_pretrained("bhadresh-savani/distilbert-base-uncased-emotion")
```

Figure 4.15: Code for load the model into the notebook

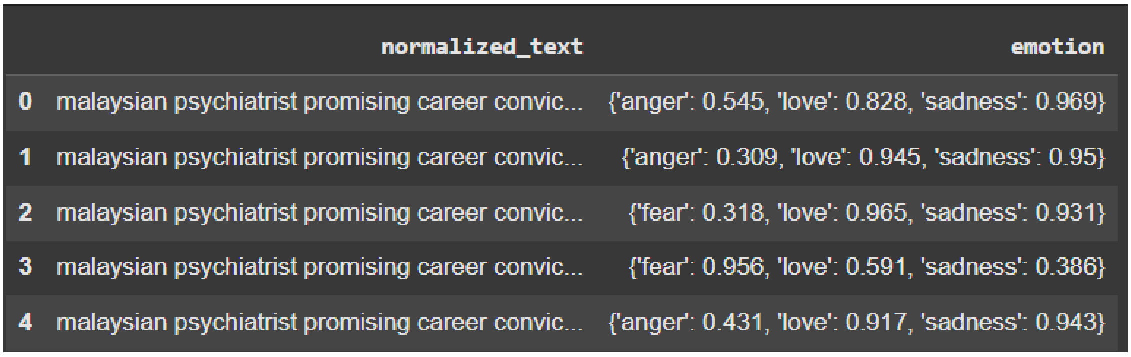
After that, setup the detail of the model to predicts emotions from text. In this model it will predict the probability for each model, but in this project, it will just show the emotion of probability more than 0.3. This is to make sure the emotion predict is more accurate. Figure 4.16 show the code to setup for the emotion prediction.



```python
def predict_emotion(text):
    """
    Predicts emotions from text.
    Returns dictionary of emotion: probability if > 0.3
    """
    inputs = tokenizer(text, return_tensors="pt", truncation=True, padding=True)
    with torch.no_grad():
        outputs = model(**inputs)

    probs = torch.sigmoid(outputs.logits).cpu().numpy()[0]
    labels = ['anger', 'fear', 'joy', 'love', 'sadness', 'surprise']
    emotion_probs = dict(zip(labels, probs))

    # Return only emotions with confidence > 0.3
    return {emotion: round(float(prob), 3) for emotion, prob in emotion_probs.items() if prob > 0.3}
```

Figure 4.16: Code for setup before emotion prediction.

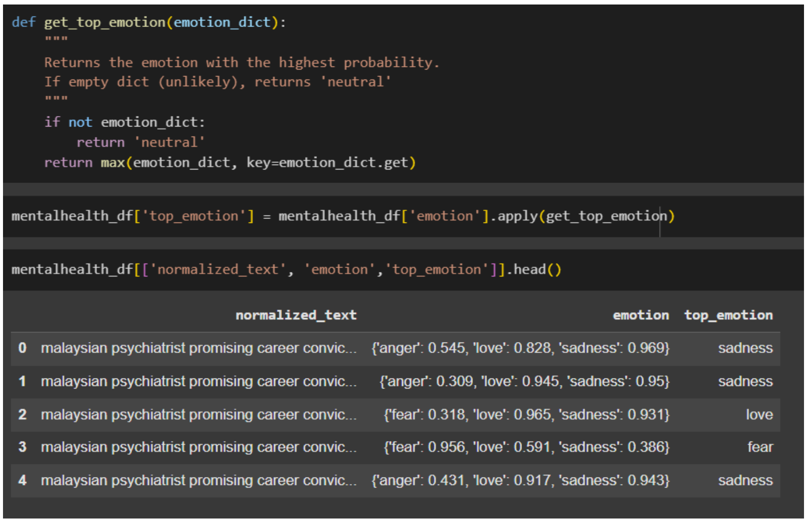
The figure 4.17 show the outcome of the emotion prediction. From the outcome, the first data have 3 emotions that the probability more than 0.3 which are anger is 0.545, love is 0.828 and sadness is 0.969. Thus, the first data have most probability show emotion sadness.



	normalized_text	emotion
0	malaysian psychiatrist promising career convic...	{'anger': 0.545, 'love': 0.828, 'sadness': 0.969}
1	malaysian psychiatrist promising career convic...	{'anger': 0.309, 'love': 0.945, 'sadness': 0.95}
2	malaysian psychiatrist promising career convic...	{'fear': 0.318, 'love': 0.965, 'sadness': 0.931}
3	malaysian psychiatrist promising career convic...	{'fear': 0.956, 'love': 0.591, 'sadness': 0.386}
4	malaysian psychiatrist promising career convic...	{'anger': 0.431, 'love': 0.917, 'sadness': 0.943}

Figure 4.17: The outcome for the emotion classification model.

The emotion prediction for one post was more than one emotion, thus the emotion that have the highest probability become the top emotion of the post. This will make one data only have one emotion. Figure 4.18 show that the highest probability emotion becomes the top emotion for each post.



```
def get_top_emotion(emotion_dict):
    """
    Returns the emotion with the highest probability.
    If empty dict (unlikely), returns 'neutral'
    """
    if not emotion_dict:
        return 'neutral'
    return max(emotion_dict, key=emotion_dict.get)
```

```
mentalhealth_df['top_emotion'] = mentalhealth_df['emotion'].apply(get_top_emotion)
```

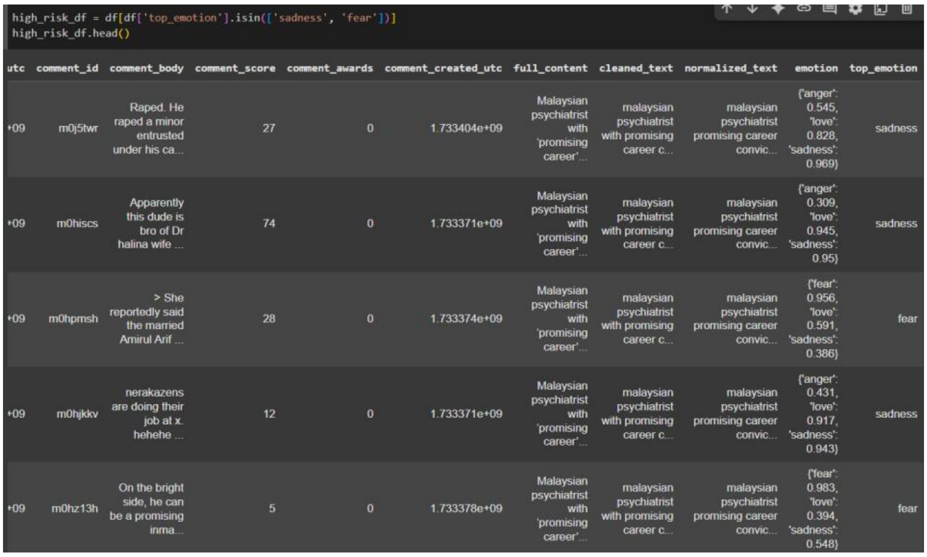
```
mentalhealth_df[['normalized_text', 'emotion','top_emotion']].head()
```

	normalized_text	emotion	top_emotion
0	malaysian psychiatrist promising career convic...	{'anger': 0.545, 'love': 0.828, 'sadness': 0.969}	sadness
1	malaysian psychiatrist promising career convic...	{'anger': 0.309, 'love': 0.945, 'sadness': 0.95}	sadness
2	malaysian psychiatrist promising career convic...	{'fear': 0.318, 'love': 0.965, 'sadness': 0.931}	love
3	malaysian psychiatrist promising career convic...	{'fear': 0.956, 'love': 0.591, 'sadness': 0.386}	fear
4	malaysian psychiatrist promising career convic...	{'anger': 0.431, 'love': 0.917, 'sadness': 0.943}	sadness

Figure 4.18: Show only the highest probability for each data.

4.7 High risk post analysis

To determine the post that have high risk for mental health crises, filter for emotion like sadness and fear was done. This will help to determine what words that have high risk to show mental health crises. In figure 4.19, it shows the post that have the top emotion like sadness and fear.

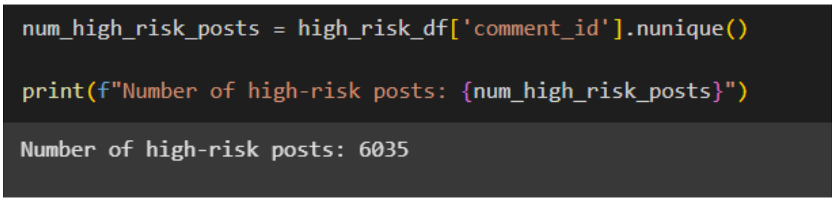


```
high_risk_df = df[df['top_emotion'].isin(['sadness', 'fear'])]
high_risk_df.head()
```

utc	comment_id	comment_body	comment_score	comment_awards	comment_created_utc	full_content	cleaned_text	normalized_text	emotion	top_emotion
+09	m0j5twr	Raped. He raped a minor entrusted under his ca...	27	0	1.733404e+09	Malaysian psychiatrist with 'promising career'...	malaysian psychiatrist with promising career c...	malaysian psychiatrist promising career convic...	{'anger': 0.545, 'love': 0.828, 'sadness': 0.969}	sadness
+09	m0hiscs	Apparently this dude is bro of Dr halina wife ...	74	0	1.733371e+09	Malaysian psychiatrist with 'promising career'...	malaysian psychiatrist with promising career c...	malaysian psychiatrist promising career convic...	{'anger': 0.309, 'love': 0.945, 'sadness': 0.95}	sadness
+09	m0hpmsh	> She reportedly said the married Amirul Arif ...	28	0	1.733374e+09	Malaysian psychiatrist with 'promising career'...	malaysian psychiatrist with promising career c...	malaysian psychiatrist promising career convic...	{'fear': 0.956, 'love': 0.591, 'sadness': 0.386}	fear
+09	m0hjkkv	nerakazens are doing their job at x. hehehe ...	12	0	1.733371e+09	Malaysian psychiatrist with 'promising career'...	malaysian psychiatrist with promising career c...	malaysian psychiatrist promising career convic...	{'anger': 0.431, 'love': 0.917, 'sadness': 0.943}	sadness
+09	m0hz13h	On the bright side, he can be a promising inma...	5	0	1.733378e+09	Malaysian psychiatrist with 'promising career'...	malaysian psychiatrist with promising career c...	malaysian psychiatrist promising career convic...	{'fear': 0.983, 'love': 0.394, 'sadness': 0.548}	fear

Figure 4.19: Filter out the data that have high probability in sadness and fear.

In the dataset, it has 6035 posts that can be define as high-risk post. This mean that in the dataset, it more than half of the dataset was define as high-risk post.

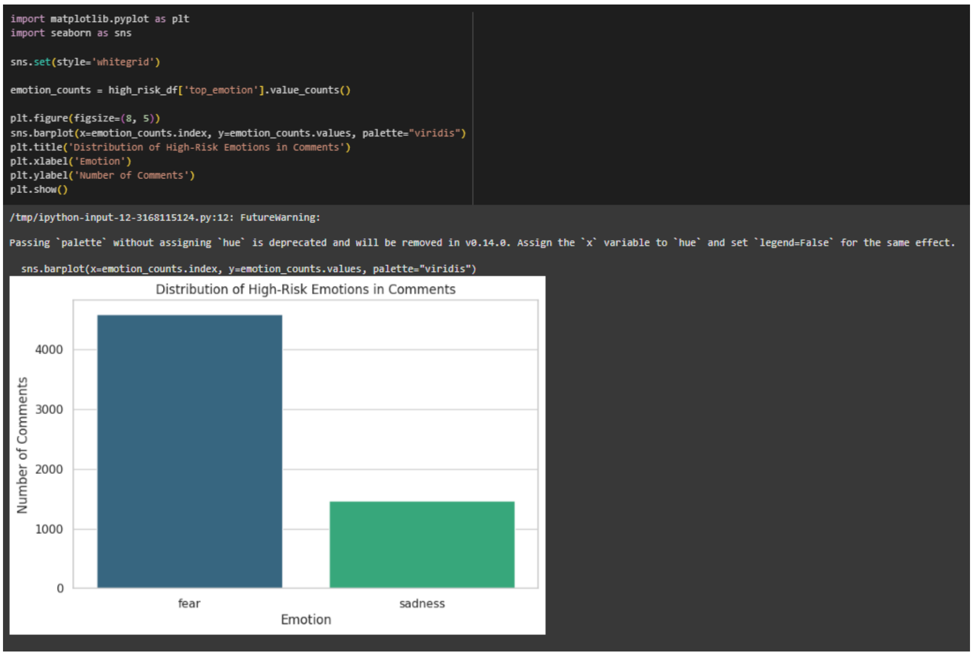


```
num_high_risk_posts = high_risk_df['comment_id'].nunique()

print(f"Number of high-risk posts: {num_high_risk_posts}")
```

```
Number of high-risk posts: 6035
```

Figure 4.20: Number of high-risk posts.

In the high-risk posts, in have 2 emotion label. From the bar chart in figure 4.21, it shows that posts that be predict fear emotion is more than sadness.



```
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style='whitegrid')

emotion_counts = high_risk_df['top_emotion'].value_counts()

plt.figure(figsize=(8, 5))
sns.barplot(x=emotion_counts.index, y=emotion_counts.values, palette="viridis")
plt.title('Distribution of High-Risk Emotions in Comments')
plt.xlabel('Emotion')
plt.ylabel('Number of Comments')
plt.show()
```

```
/tmp/ipython-input-12-3168115124.py:12: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=emotion_counts.index, y=emotion_counts.values, palette="viridis")
```

Figure 4.21: Distribution of high-risk emotions.

To determine the common words in the high-risk post, word cloud was used. The word cloud was done for two high-risk emotions: sadness and fear. For the emotion ‘sadness ’the words like stall owner, mother daughter demanded, police, refused pay, pay remaining were the common words show in the high-risk post. The word cloud for sadness-labelled data.

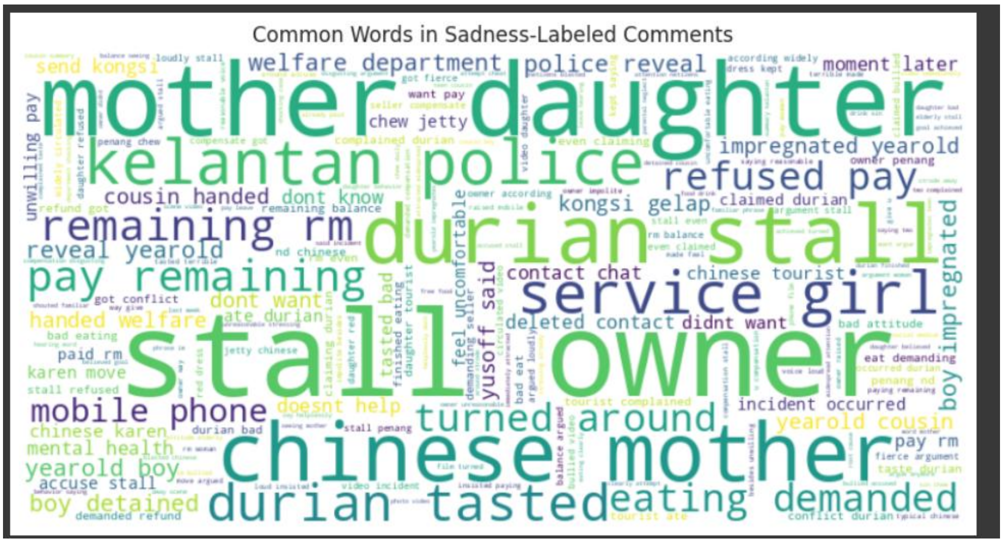


Figure 4.22: Common words for Sadness-labelled data.

For fear emotion post, it shows the words like mental health, due religion, stable job, trying best, want help, trying and other words. The word cloud for fear-labelled data shown in figure 4.23.

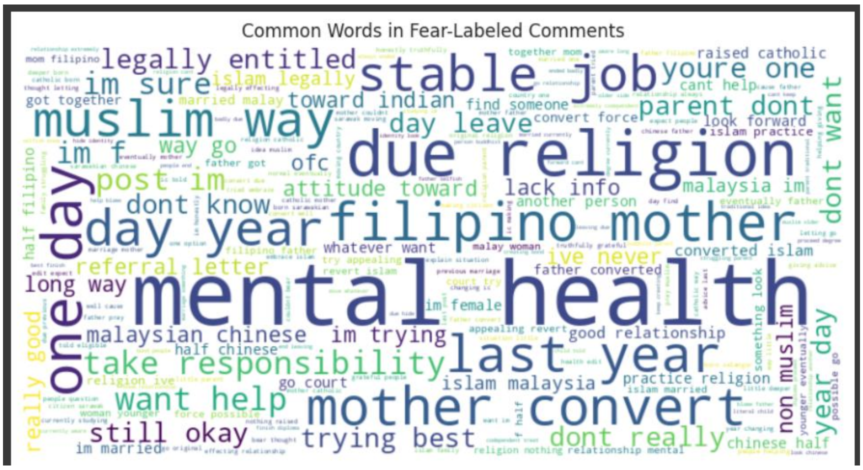


Figure 4.23: Common words for Fear-labelled data.

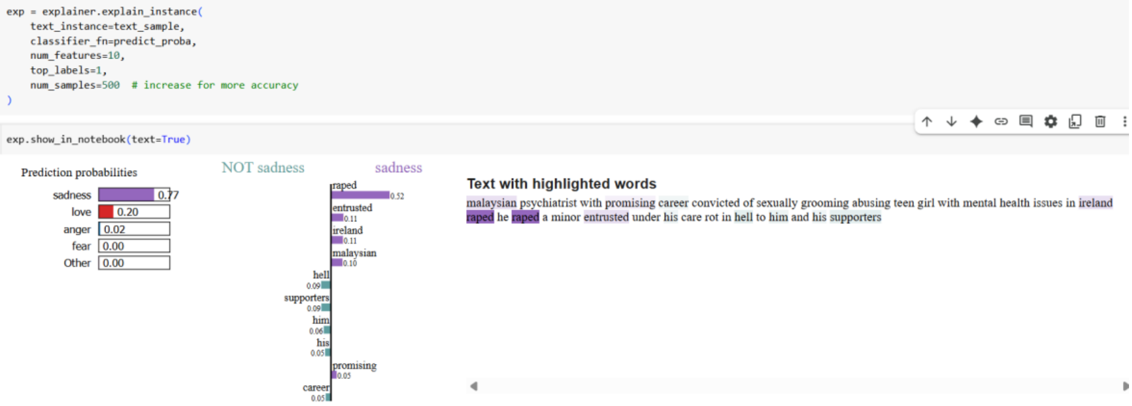
Based on the common words in word clouds, it has some words that make attention for mental health crises like mental health, demanded, trying best, want help, still okay, and don’t want. These words give the attention for the reader that the writer might face some problem in life and makes the writer have high-risk to have mental health crises.

4.8 XAI interpretation

In this project, the dataset used was scrape from the social media Reddit thus it did not have the label. XAI method LIME was used to explain how the emotion prediction be done for the post. This will to increasing the trustworthy of the prediction outcome. The LIME model will provide a visual breakdown of how the DistilBERT model classified a given text as ‘sadness’ or ‘fear’.

In figure 4.24 it was using XAI to interpret the sadness prediction for the first post. On the left side of the outcome, it shows the emotion prediction probabilities, for this post, the prediction probability for sadness was 0.77 while for love is 0.20 and 0.02 for anger. It shows that this post and comment have the highest probability to have emotion sadness which is high-risk to have mental health crises. In the middle of the outcome, the top 10 words that contributes to the prediction of “sadness” was show. The words have two colours which is purple and min green. The purple colour words is the positive contribution for the “sadness” emotion. In this case, it shows that ‘raped’ have strong positive contribution which is +0.53, other words like ‘entrusted’, ‘minor’, and abusing’ also contribute positively. For the mint green words show the words which have negative contribution. The words like supporters and career have slightly negative contribution (-0.09) and (-0.05) for the emotion prediction.

The model predicts “sadness” because the text contains strong keywords related to sexual abuse, violence and betrayal like ‘raped’, ‘minor’, and ‘entrusted’. These words show strong emotional responses, particularly sadness when combined in this context. The presence of words like supporters and career slightly dilute the overall sadness but do not outweigh the dominant negative cues.



```
exp = explainer.explain_instance(
    text_instance=text_sample,
    classifier_fn=predict_proba,
    num_features=10,
    top_labels=1,
    num_samples=500  # increase for more accuracy
)
```

```
exp.show_in_notebook(text=True)
```

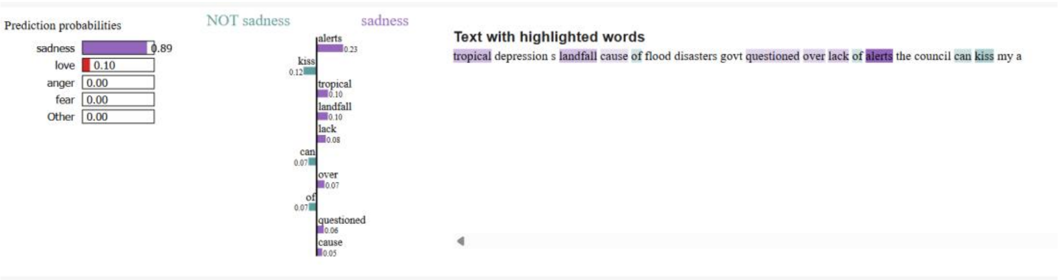
Text with highlighted words

malaysian psychiatrist with promising career convicted of sexually grooming abusing teen girl with mental health issues in ireland raped he raped a minor entrusted under his care rot in hell to him and his supporters

Figure 4.24: The outcome of LIME for the first data.

LIME was tried for another random chosen post. For the LIME in figure 4.25, it was tried for the high-risk post for 2200th post. On the left side of the outcome, it shows the prediction probabilities for each class, it has probability 0.89 for sadness emotion and 0.10 for love emotion. This indicates that the model is highly confident in predicting “sadness” for this text with probability of 89%. On the right side, the input text is displayed with words highlighted in purple and mint green. The purple colour words show the positive contribution for “sadness”. The word “alerts” and “landfall” show the strong positive contribution which were 0.23 and 0.10. other words like ‘lack’, ‘cause’ also contribute positively. For the word that give slightly negative contribution include ‘kiss’.

The model predicts “sadness” because the text contains strong keywords related to natural disasters and tragedy like words “landfall”, “lack”, and “cause”. These words show strong emotional responses particularly sadness when combined in this context. The presence of words like “kiss: slightly dilutes the overall sadness but does not outweigh the dominant negative cues.



Text with highlighted words

tropical depression s landfall cause of flood disasters govt questioned over lack of alerts the council can kiss my a

Figure 4.25: The outcome of LIME for the 2200th data.

4.9 Model training and Evaluation

To evaluate how well the emotion classifier generalizes to local media text, a model training and evaluation was done in the last step. This step is to assess the model's performance when trained on standard emotion-labeled data and test on Reddit-based pseudo-labeled text.

The 'dair-ai/emotion' dataset was loaded and used as the training set while the Reddit-based pseudo-labeled dataset was loaded and used as the test set. Before train the model, the training dataset was done for text reprocessing such as lowercasing, removal of special characters and URL links, lemmatization, and stop word filtering.

After cleaning, both dataset text samples were tokenized using the DistilBERT tokenizer and converted into numerical format that acceptable by the model. Inputs were padded or truncated to a fixed length of 128 tokens.

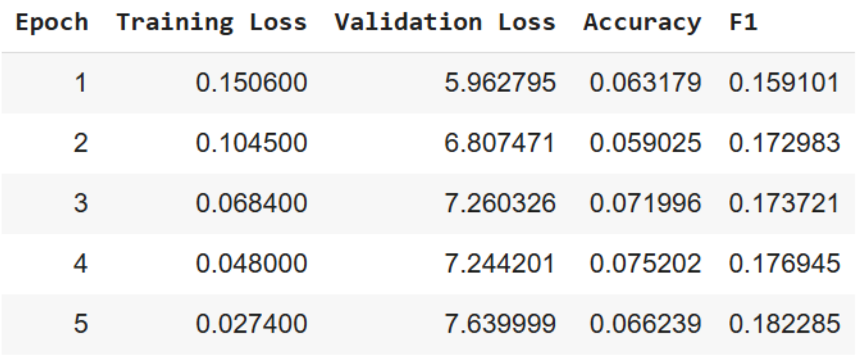
The DistilBERT model 'bhadresh-savani/distilbert-base-uncased-emotion' was fine-tuned using Hugging Face's Trainer API with some setting. The setting shows in the table below:

Table 4.1: Setting of the DistilBERT model.

Parameter	**Value**
Number of Epochs	5
Batch Size	16
Learning Rate	2e-5
Evaluation Strategy	After every epoch
Save Strategy	After every epoch
Metrics Tracked	Accuracy, F1-score

Figure 4.25 show the outcome of the model training. From the result it shows the training loss decreased steadily, this indicating that the model was learning from the training data. By the end of epoch 5, the training loss has dropped significantly to 0.0274, this suggesting that the model was fitting well to the labelled emotion dataset. however, the validation loss showed an increasing trend, it rising from 5.9628 in epoch 1 to 7.6399 in epoch 5 which is a strong indicator of overfitting. This means that while the model improved its performance on the training data, but failed to generalize effectively to new and unseen text data, especially the culturally specific expressions found in the Reddit social media test set.

From the Figure 4.25, it shows the model's accuracy remained very low through the training phase, starting at 0.0632 in the first epoch and reaching a peak of 0.0752 in epoch 4 before slightly drooping to 0.0662 in the final epoch. These low values highlight that the model struggled when applied to Malaysian Reddit discussion, despite performing well on standard English emotion dataset.



Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.150600	5.962795	0.063179	0.159101
2	0.104500	6.807471	0.059025	0.172983
3	0.068400	7.260326	0.071996	0.173721
4	0.048000	7.244201	0.075202	0.176945
5	0.027400	7.639999	0.066239	0.182285

Figure 4.26: The result for the model training.

Based on the result of the model training, one of the main findings was overfitting, where the model learned well from the training data but failed to generalize to the test set. this was evident from the decreasing training loss and increasing on validation loss over time. Overfitting likely resulted from a domain mismatch which the training data consisted of formal English text while test on the data that included informal, emotionally complex and Malaysian English text.

Other than that, the use of pseudo-labels for evaluation also limited reliability, due to they were generated by the same model being tested. Without ground truth labels, metrics like accuracy and F1-socre could not fully reflect real-world performance. This reinforced the need for Explainable AI (XAI) methods like LIME, which can help to interpret the prediction and validate whether flagged emotions were based on meaningful text.

Next, many Malaysian Reddit post have mix language between Malay and English, informal expression and culturally specific emotional language that the model was not be trained on. This cause the model often misclassified or missed high-risk emotions like sadness and fear that are the key indicators of mental health crisis.

In summary, although the DistilBERT model showed signs of learning during training, but its performance on the test set remained extremely low, with accuracy around 7% and F1-score around 18%. These results indicated that while the model can be trained to detect emotions in standard English text, it still requires significant adaptation to work effectively on Malaysian social media content.

4.10 Summary

This chapter detailed the end-to-end workflow of collecting and analysing Reddit posts related to mental health from the r/Malaysia subreddit. A total of 13,723 posts with comments were collected using PRAW library. The post was collected by searching for posts that containing keywords such as “mental health”, “depression”, “anxiety”, and “stress”. The dataset was structured into 12 columns capturing the metadata and textual content of the post.

To ensure the quality and usability of the dataset, missing value were handled and some text preprocessing was performed. The text preprocessing included remove the non-alphabetic characters normalizing text, eliminating stop words and lemmatization. These steps ensured that the data was ready for downstream analysis and modelling.

EDA was performed to uncover trends over time, word frequency patterns and length distribution across title, post body and comments. Word clouds used to highlighted frequently occurring terms in the title, post body and comments. The words like “mental health”, “Malaysia”, “Hari Raya”, and “gathering” were the words that common show in the data collected. This offering the insights into the cultural and contextual themes present in the data,

An emotion classification model based on DistilBERT (bhadresh-savani/distilbert-base-uncased-emotion) was used to classify post into six emotions. Posts labelled with sadness and fear were identified as high-risk for mental health crises. Out of the dataset, there were 6,035 posts were classified as high-risk. Word clouds for high-risk emotions showed that words like “refused pay”, “police”, and “demand” link to sadness, while the words like “mental health”, “want help”, and “stable job” were common in fear-labelled posts.

Next, to improve the trustly of the model’s prediction, the LIME was applied to interpret the DistilBERT’s predictions. LIME highlighted which words or phrases most influenced each prediction, helping to validate whether the model focused on meaningful content rather than noise or irrelevant features. This added transparency and helped validate the model’s decision.

Lastly, a model training and evaluation phase was introduced to assess how well the DistilBERT model generalizes emotion to Malaysian English text. The publicly available ‘dair-ai/emotion’ dataset was used as the training set while the Reddit dataset with pseudo-labelled by the same DistilBERT model was used as the test set. The model was fine-tuned using Hugging Face’s Trainer API.

In summary, this chapter established a solid foundation for detecting potential mental health concerns using real-world social media data and explainable AI techniques.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Research Outcomes

This project study about the use of data science techniques to detect early signs of mental health issues from social media post and comments. This project focus on posts from r/Malaysia subreddit on Reddit, where users often share personal experiences and concerns related to mental health.

To collect the data for Malaysian post, Python library called PRAW was used. 13,723 posts and comments were collected by searching for keywords such as “mental health”, “depression”, “anxiety”, and “stress”. The dataset included some details like post title, body text, comment, and timestamps.

Title, body text and comments of the data be combined into a new column, this is to make sure the emotion detection more accurate. Before analysis the data, the data cleaning was done. The text data preprocessing step included removing the URL links, non-English characters, unnecessary symbols.

In this project, a pre-trained DistilBERT model ‘bhadresh-savani/distilbert-base-uncased-emotion’ was used to classify each data into one of six emotional categories like anger, fear, joy, love, sadness, and surprise. From the project, there were 2 emotion which were sadness and fear commonly linked with high-risk mental health situations.

Out of all the posts, there were 6,035 posts have the highest probability with label sadness and fear. This mean that these posts have high probability to have mental health problem. In this project, Word Cloud was used to visualize the most common words associated with the emotions.

To better understand why the model made the predictions, explainable AI technique called LIME was used. These explanations showed which words influenced the model's decision the most, this helps to increase the trustworthy.

In addition to emotion classification, a model training and evaluation phase was introduced to assess how well the DistilBERT model generalizes to Malaysian English expression. The model was trained by publicly available 'dair-ai/emotion' dataset and test for Reddit dataset with pseudo-labeled by the same model.

From the result for the model training, it shows steady decrease in training loss and the validation loss was increased. This show the model is overfitting which the model learned well from the standard English emotion dataset but struggled when applied to informal, culturally specific Malaysian English posts. The final evaluation showed very low accuracy (7.5%) and F1-score of 0.18. This show that emotion models trained on Western English datasets do not always generalize well to local text and expressions.

Although accuracy was limited, the model still identified emotionally sensitive posts related to sadness and fear. This is the key indicators of potential mental health crises.

5.2 Contributions to Knowledge

This project makes meaningful contributions to both research and practical applications in the field of mental health analysis. One of the key contributions to knowledge is the development of an end-to-end pipeline for social media analysis. Start from data collection using Reddit's API to preprocessing, until emotion classification with DistilBERT, and model interpretation using XAI model's LIME. Additionally, the project demonstrates the value of explainable AI in emotion detection by using LIME to interpret DistilBERT predictions. This adds transparency to model decisions which is especially important when working on sensitive topics like mental health where trust and understanding are important. Moreover, this project contributions in identifying of high-risk language patterns. The word clouds and LIME explanations identified specific language patterns and words that linked with sadness and fear. These findings can help inform future monitoring system or early detection tools for mental health problem.

5.3 Future Works

While the current system shows promising results, there are several areas for improvement. One of the improvements is the fine-tuning of emotion detection model on domain-specific mental health datasets which could significantly enhance its accuracy and relevance for real-world applications. Currently, the model relies on pseudo-labels generated by the model itself, which limits objective evaluation. By collecting a small subset of manually reviewed labels, the model can be trained for better accuracy and relevance.

Another important area is the expansion of data sources other that Reddit which include platform such as Twitter/X, Facebook groups, or local Malaysian forums. This would provide a more diverse and representative dataset, which allowing for broader insights into public discussions around mental health.

By addressing these limitations, future iterations of this project can lead to a more accurate and reliable mental health risk detection system for Malaysian social media posts.

REFERENCES

(WHO), W. H. (2019). *Mental health*. Retrieved from World Health Organization (WHO): https://www.who.int/health-topics/mental-health#tab=tab_2

(WHO), W. H. (2021). *Suicide prevention*. Retrieved from World Health Organization (WHO): https://www.who.int/health-topics/suicide#tab=tab_1

(WHO), W. H. (2022). *Mental disorders*. Retrieved from World Health Organization (WHO): https://www.who.int/news-room/fact-sheets/detail/mental-disorders

(WHO), W. H. (n.d.). *Anxiety disorders*. Retrieved from World Health Organization (WHO): https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders

A.Naslund, J., Bondre, A., Torous, J., & A.Aschbremmer, K. (2020). Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice.

A.Vogels, E., Gelles-Watnick, R., & Massarat, N. (2022). Teens, Social Media and Technology 2022.

Acito, F. (2023). Naïve Bayes.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review.

Angiulli, F., Fassetti, F., & Nisticò, S. (2021). Local Interpretable Classifier Explanations with Self-generated Semantic Features.

Antonio, V., Julio, T., & M., C.-M. J. (2020). Urbanization and emerging mental health issues.

Auxier, B., & Anderson, M. (2021). Social Media Use in 2021.

Banna, M. H., Ghosh, T., Md. Jaber Al Nahian, M. S., Mahmud, M., & Taher, K. A. (2023). A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health From Social Media Big Data.

Bao, E., Pérez, A., & Parapar, J. (2024). Explainable depression symptom detection in social media.

Bauer, B., Norel, R., Leow, A., Rached, Z. A., Wen, B., & Cecchi, G. (2024). Using Large Language Models to Understand Suicidality in a Social Media–Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts.

Becker, M., & Correll, C. U. (2020). Suicidality in Childhood and Adolescence.

bhadresh-savani. (2025). *distilbert-base-uncased-emotion*. Retrieved from AIModels.fyi: https://www.aimodels.fyi/models/huggingFace/distilbert-base-uncased-emotion-bhadresh-savani

Braghieri, L., Levy, R., & Makarin, A. (2022). Social Media and Mental Health.

Carballo, J. J., Llorente, C., Kehrmann, L., Flamarique, I., Zuddas, A., Purper-Ouakil, D., . . . Arango, C. (2020). Psychosocial risk factors for suicidality in children and adolescents.

Chang, H.-J., Song, S., Kim, G., Kim, T.-Y., Shin, D., Kim, S.-H., . . . Sung, J.-H. (2023). BERT-based model to predict cardiovascular disease by analyzing healthcare utilization behavior of patients newly diagnosed with metabolic diseases.

Choi, J. E., Shin, J. W., & Shin, D. W. (2024). Vector SHAP Values for Machine Learning Time Series Forecasting.

Convolutional Neural Networks. (2022).

Das, A. K., & Ahmed, S. S. (2023). Convolutional neural networks.

Dobrek, L., & Glowacka, K. (2023). Depression and Its Phytopharmacotherapy—A Narrative Review.

Eleftheriades, R., Fiala, C., & Pasic, M. d. (2020). The challenges and mental health issues of academic trainees.

Eric, R., Punyanunt-Carter, N., R.LaFreniere, J., S.Norman, M., & G.Kimball, T. (2020). The serially mediated relationship between emerging adults' social media use and mental well-being.

Feizi, A., & Nazemi, A. (2022). Classifying random variables based on support vector machine and a neural network scheme.

Fusar-Poli, P., Gonzalo Salazar de Pablo, A. D., Nieman, S. H., Correll, C. U., Kessing, L. V., Pfenning, A., . . . Amelsvoort, T. c. (2020). What is good mental health? A scoping review.

Garg, K., Garg, S., Dixit, H., & Pandey, K. (2024). Machine Learning Driven Analysis of Mental Health Indicators in Social Media Posts.

Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the Need for Explainable Artificial Intelligence (xAI).

Ha, H. T., Nguyen, D. T., & Stoeckel, T. (2024). What is the best predictor of word difficulty? A case of data mining using random forest.

Hasija, Y., & Chakraborty, R. (2021). Support Vector Machines.

Health, N. N. (n.d.). *Anxiety*. Retrieved from MedlinePlus: https://medlineplus.gov/anxiety.html

Herren, A., & Hahn, P. R. (2022). Statistical Aspects of SHAP: Functional ANOVA for Model Interpretation.

Hulsen, T. (2023). Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare.

IBM. (n.d.). *What is random forest?* Retrieved from IBM.

Jo, A. A., Raj, E. D., Vino, A. S., & Menon, P. V. (2024). Exploring Explainable AI for Enhanced Depression Prediction in Mental Health.

joeddav. (2025). *distilbert-base-uncased-go-emotions-student*. Retrieved from AIModels.fyi: https://www.aimodels.fyi/models/huggingFace/distilbert-base-uncased-go-emotions-student-joeddav

joeddav. (n.d.). *joeddav/distilber-base-uncased-go-emotions-student*. Retrieved from Hugging Face: https://huggingface.co/joeddav/distilbert-base-uncased-go-emotions-student

K, K., & Wong, L. (2023). Support Vector Machine.

Kabir, M. K., Islam, M., Kabir, A. N., Haque, A., & Rhaman, M. K. (2022). Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques.

Kamal, M., Khan, S. U., Hussain, S., Nasir, A., Aslam, K., Tariq, S., & Ullah, M. F. (2020). Predicting Mental Illness using Social Media Posts and Comments. *(IJACSA) International Journal of Advanced Computer Science and Applications*.

Kaushik, P., & Sharma, P. (2024). Integrating Machine Learning and Sentiment Analysis: A Comparative Study on Mental Health Classification from Social Media Data.

Khanduja, D. K., & Kaur, S. (2023). The Categorization of Documents Using Support Vector Machines.

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media.

Kumar, R., Goswami, B., Mhatre, S. M., & Agrawal, S. (2024). Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases.

Kumar, S., & Gota, V. (2023). Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools.

Lim, S. R., Kamarudin, N., Ismail, N. H., Ismail, N. A., & Kamal, N. A. (2023). Predicting Mental Health Disorder on Twitter Using Machine Learning Techniques.

Liu, Z., Peach, R. L., Lawrance, E. L., Noble, A., Ungless, M. A., & Barahona, M. (2021). Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service.

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: A review.

Marwaha, P. S., Palmer, P. E., Suppes, M. P., Cons, M. P., Young, M. P., & Upthegrove, M. P. (2023). Novel and emerging treatments for major depression.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2021). Explainable artificial intelligence: a comprehensive review.

Odja, K. D., Widiarta, J., Purwanto, E. S., & Ario, M. K. (2024). Mental illness detection using sentiment analysis in social media.

Pillai, S. E., Polimetla, K., Avacharmal, R., & Perumal, A. P. (2022). Mental Health in the Tech Industry: Insights from.

Pinaya, W. H., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Convolutional neural networks.

Qorich, M., & Ouazzani, R. E. (2025). Lightweight advanced deep learning models for stress detection on social media.

Renuka, O., & Radhakrishnan, N. (2024). BERT for Twitter Sentiment Analysis: Achieving High Accuracy and Balanced Performance.

Richardson, R., Connell, T., Foster, M., Blamires, J., Keshoor, S., Moir, C., & Zeng, I. S. (2024). Risk and Protective Factors of Self-harm and Suicidality in Adolescents: An Umbrella Review with Meta-Analysis.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). {CARER}: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3687-3697). Association for Computational Linguistics.

Sarno, E., Moeser, A. J., & Robison, A. J. (2021). Chapter Eight - Neuroimmunology of depression.

Scheda, R., & Diciotti, S. (2022). Explanations of Machine Learning Models in Repeated Nested Cross-Validation: An Application in Age Prediction Using Brain Complexity Features.

Schonlau, M. (2023). Logistic Regression.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning.

Stryker, C., & Holdsworth, J. (2024). *What id NLP (natural language processing)?* Retrieved from IBM: https://www.ibm.com/think/topics/natural-language-processing

Tahir, W. B., Khalid, S., Almutairi, S., Abohashrh, M., Memon, S. A., & Khan, J. (2025). Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques.

Tang, H., Rekavandi, A. M., Rooprai, D., Dwivedi, G., Sanfilippo, F. M., Boussaid, F., & Bennamoun, M. (2024). Analysis and evaluation of explainable artificial intelligence on suicide risk assessment.

Vishwakarma, S., & Ganguly, S. (2023). Optimal partition of feature using Bayesian classifier.

Walter, H. J., Bukstein, O. G., Abright, A. R., Keable, H., Ramtekkar, U., Ripperger-Suhler, J., & Rockhill, C. (2020). Clinical Practice Guideline for the Assessment and Treatment of Children and Adolescents With Anxiety Disorders.

WHO. (2025). *World Health Organization (WHO)*. Retrieved from Mental health: https://www.who.int/health-topics/mental-health#tab=tab_1

Wijaya, V., & Rachmat, N. (2024). Comparison of SVM, Random Forest, and Logistic Regression Performance in Student Mental Health Screening.

William, D., & Suhartono, D. (2020). Text-based Depression Detection on Social Media Posts: A.

Zafar, M. R., & Khan, N. M. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability.

Zhang, T., Yang, K., Shaoxiong, & Ananiadou, S. (2023). Emotion fusion for mental illness detection from social media: A survey.

Zodage, P., Harianawala, H., Shaikh, H., & Kharodla, A. (2024). Explainable AI (XAI): History, Basic Ideas and Methods.

Zogan, H., Razzak, I., Wang, X., Jammel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media.