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 neighbor (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 Random Forest (RF) to do the mental health crises detection. 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 analyzed 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 enables real-time detection and 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 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 analyzed 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.

1.7 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 Facebook.
(b) Target Conditions: This study aim to get the outcome of the prediction like depression, anxiety and suicidal ideation.
(c) Technical Focus: This study will use Random Forest to do the prediction of mental health crises and SHAP 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 Facebook 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 the conclusion for this study.