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), Anxiety disorders, 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).

2.5 Machine Learning

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).

2.5.1 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, 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).

2.5.2 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).

2.5.3 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).

2.5.4 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).

2.5.5 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).

2.5.6 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.

2.6 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. Beside 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.

2.6.1 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.

2.6.2 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).

2.7 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).

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 analysis in social media	(Odja, Widiarta, Purwanto, & Ario, 2024)	Reddit	KNN, Random Forest, Neural Network	Random Forest have the best performance with 80.6% for F1-score, accuracy, recall and precision.	- The dataset is small that only consist 350 columns of data.

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.

Table 2.1 Comparison of existing work on mental health crises prediction

2.8 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 Facebook data.