National College of Ireland

AI-Enhanced Mental Health Risk Analysis Using Social Media Data

MSc Research Project
MSCDAD_A

Ramanaidu Nallajonnala
Student ID: x23304197

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Ramanaidu Nallajonnala

Student ID: x23304197

Programme: MSCDAD_A **Programme:** MSCDAD_A

Module: MSc Research Project

Lecturer: Rejwanul Haque

Submission Due Date: 15/09/2025

Project Title: AI-Enhanced Mental Health Risk Analysis Using Social Media Data

Word Count: **6576** **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ramanaidu Nallajonnala

Date: 15/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□

You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI-Enhanced Mental Health Risk Analysis Using Social Media Data

Ramanaidu Nallajonnala
Student ID: x23304197

Abstract

The paper presents an AI-powered system for mental health risk detection based on social media data, incorporating sentiment trajectories, temporal posting, and social engagement features. The task at hand was to conceptualize, implement, and compare different models—with logistic regression and Long Short-Term Memory (LSTM) on one side and transformer-based Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) on the other—to evaluate their performance in depressive cue detection. The data was derived from a labelled public Twitter corpus, pre-processed for linguistic as well as for the extraction of behavioural features, and evaluated on traditional classification metrics. Of all the models evaluated, Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) performed the best, affirming value in leveraging deep contextual embeddings. Feature interpretability was provided in terms of temporal features as well as engagement-based features, conforming to Responsible AI standards. The current paper attaches special significance to multi-modal modelling of behaviour for early identification of mental health risk and advocates for its use in real-time monitoring as well as ethical systems for digital intervention.

Keywords: *Mental Health Detection, Social Media Analytics, Transformer Models, Sentiment Trajectory, DistilBERT, LSTM, Temporal Analysis, Behavioural AI, Classification Metrics, Explainable AI*.

1 Introduction

1.1 Background and Research Problem

Mental illness disorders represent one of the major health issues across the globe, affecting an estimated 970 million people worldwide, as per the World Health Organization. Depression, anxiety, bipolar illness, and suicide ideation are some of the conditions involved, all of which cause significant quality-of-life impairments and contribute to increased morbidity and mortality. Largely because these conditions afflict a significant portion of the population, timely identification and intervention are hindered by the stigmatization of the disorders, limited availability of clinicians to deliver treatment, and the episodic nature of conventional evaluations (Rani, 2024).

The emergence of social networking has changed the nature of self-expression. Space has been made on social media outlets like Facebook, Twitter, and Reddit to post emotions, opinions, and life experiences. These online traces, made in the moment and unelicited by clinicians,

provide a unique chance to study behaviour trends over time that can signal psychological distress. This shift in communication has opened an innovative, unintrusive avenue for the study of mental illness—potentially to augment conventional psychiatric evaluations with ongoing, passive monitoring (Kashif et al., 2024).

NLP and artificial intelligence (AI) have proven to capture useful insights from unstructured text. Current AI models typically utilize sentiment analysis to mark posts reflecting mental illness. But the vast majority of the existing models are based on individual postings and do not capture longitudinal patterns, social, or engagement patterns—features essential to the analysis of complicated mental states. Our manuscript attempts to contribute to the literature by proposing a multi-modal framework of Artificial Intelligence (AI) that entails sentiment trajectories, posting patterns that are related in time, and social network engagement for better mental health risk detection from social media.

1.2 Problem Statement

Existing AI mental health detection models in social media are typically predicated on the premise that mental signals can be inferred reliably from solitary text. While there have been some successes, however, there are a number of disadvantages. It does not extract temporal patterns—how the user's condition changes with the passage of time. It also overlooks the social context, e.g., user statistics of interacting or peer behaviour, that may vary in emotional display or hide signs of distress. These models are thus prone to loss of predictability and interpretability.

This work fills these gaps by suggesting a multi-dimensional AI model with the ability to examine users' behaviour over time and across their social media environment. By addressing the temporal dynamics and social context along with the text's sentimental analysis, the model would generate a more accurate and interpretable measurement of mental health risk.

1.3 Research Aim and Objectives

The primary aim of the research is to create and test an AI model with multi-modal extensions to improve the identification of social media users' mental health risks. The model will combine the three major aspects of user information: posting behaviour trajectory over time, temporal posting patterns, and influence in the social network.

Specific objectives include:

- To pre-process and implement feature engineering from a labelled dataset of mental health-related social media posts.
- To design and implement ML/DL models for sentiment analysis, and temporal behaviour patterns.
- To compare the performance of the proposed model against sentiment-only baselines using standard classification metrics and explainability tools.
- To explore the ethical considerations of using AI for mental health monitoring in publicly accessible social media environments.

1.4 Explanation of Key Terms Used in the Research Question

Sentiment: In this work, sentiment describes the affective tone of each tweet, measured using Text Blob polarity values. Strongly negative sentiment tweets commonly exhibited depressive content, e.g., statements of sadness, hopelessness, or lethargy. The Logistic Regression and SHAP analysis verified sentiment as a key predictive feature, with negative sentiment values highly associated with predictions of depressive class.

Trajectory: Sentiment trajectory captures the evolution of the user's affective tone over time. It was artificially recreated in the Long Short-Term Memory (LSTM) experiment by feeding sentiment score sequences (from the user history in the prior tweets) through a recurrent neural network. Although limited by sparsity in user history, the sentiment trajectory data picked up that users whose sentiment was continually dropping were themselves at greater risk of being categorized as at-risk. The Long Short-Term Memory (LSTM) model, however, underperformed compared to Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) due to data sparsity. **Temporal Posting Patterns:** These are the timing (i.e., hour of day) and the frequency of the user's posting. These variables, like posting_hour and weekday_encoded, were extracted and explored. The SHAP analysis of the logistic regression experiment explained a moderate influence of posting time—late-night posting, in general, was roughly more typical of depressive tweets. Consistent with previous psychological findings about disturbed circadian rhythms in distressed individuals.

Social Network Interactions: These refer to behavioural engagements like the number of retweets, favourited, followers, and derived proxy features like centrality (estimated by activity and visibility). Social features in the experiments proved to be moderately impactful. SHAP outputs indicated that features like retweets, status, and friends made a non-trivial contribution to the classifier, in that lower engagement or variations in the level of interaction might be associated with risk signals in mental health.

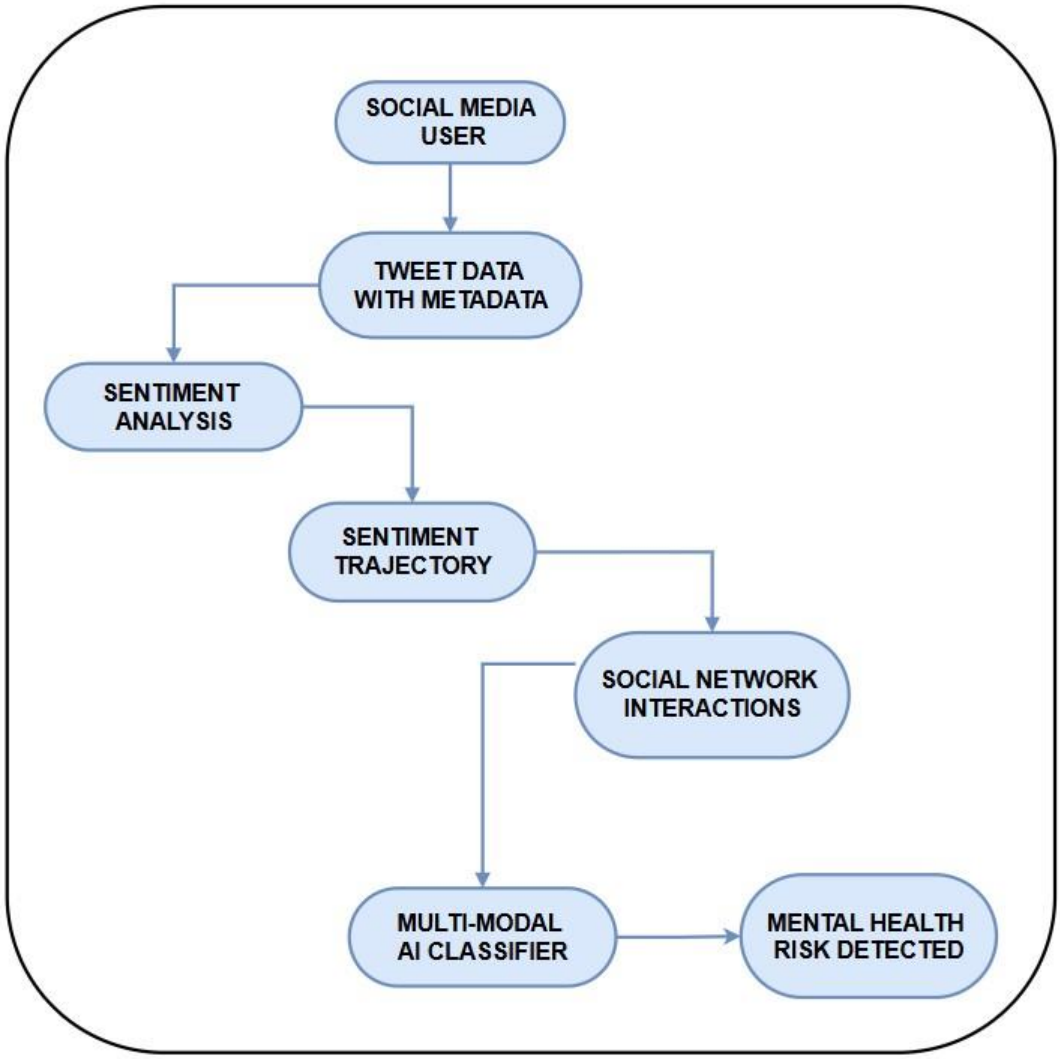

Figure 1: Research Question Key Term

The flow diagram shows how sentiment, direction, temporal patterns, and social interactions are extracted from actual social media behaviour. It shows how the processing of raw tweet data leads to the detection of mental health risk based on AI, representing how user feelings and activity patterns can indicate psychological perturbation in daily online engagements.

1.5 Research Questions

This dissertation is guided by the following central research question:
How can AI models utilizing sentiment trajectory, social network interactions, and temporal posting patterns improve mental health risk detection using social media data?
To support this inquiry, the following sub-questions are explored:

- How do sentiment trends over time correlate with mental health risk signals in users?
- What role do engagement metrics and social interactions play in modulating these risk signals?
- Can integrating temporal and social context features improve both the accuracy and interpretability of AI-based mental health models?

1.6 Scope and Significance of Study

This study centres around studying the labelled Twitter data for mental health indicators based on metadata including timestamp information, number of retweets, likes, and followers. It will only take English-language tweets available in the public domain and will not collect any personal or real-time user information (Shimada, 2023).

The value of this work stems from its benefit to artificial intelligence (AI) as an area of research and to the field of mental health. By transcending text-centred analysis, it presents a multi-dimensional framework for a richer and more effective understanding of psychological risk. It has the potential to be used in public health surveillance, the development of alert systems, and in digital therapy. It further supports the creation of explainable AI models, which are necessary to deploy in sensitive areas such as mental health ethically.

1.7 Organization of the Study

This dissertation is organized into seven sections as follows:

- **Section 1 – Introduction:** Presents the background of the study, the research problem, aims, objectives, research questions, and the significance of the study.
- **Section 2 – Related Work:** Reviews relevant literature on sentiment analysis, mental health assessment using social media, AI techniques, and ethical considerations.
- **Section 3 – Methodology:** Describes the research design, dataset characteristics, pre-processing methods, model implementation, and ethical framework.
- **Section 4 – Design Specification:** Details the system architecture, technologies used, and the modelling pipeline employed in the study.
- **Section 5 – Implementation:** Explains the step-by-step development of the machine learning models and integration of features.
- **Section 6 – Evaluation and Discussion:** Evaluates the models' performance using various classification metrics and interpretability tools like SHapley Additive exPlanations (SHAP).
- **Section 7 – Conclusion and Future Work:** Summarizes findings, highlights contributions, discusses limitations, and proposes directions for future research.

2 Related Work

2.1 Introduction

The growing global mental health crisis, driven by stigma and reduced care accessibility, necessitates scalable, evidence-based approaches for the early risk assessment. Social media have emerged as a psychological trove possessing real-time and non-invasive tools for the assessment of mental health. Machine learning and, more importantly, natural language processing (NLP) and deep learning are becoming increasingly used for the text-based assessment of user-generated content and the extraction of symptoms for psychological distress. The current review paper organizes the pertinent works under four central dimensions: sentiment-based text classification, temporal behaviour analysis, social network engagement modelling, and explainability for AI-based diagnostics.

2.2 Sentiment-Based Detection of Depression

Sentiment-based classification forms the bedrock for AI-based depression models. Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT) and RoBERTa, have been most assistive. Kurniadi et al. (2024) utilized Bidirectional Encoder Representations from Transformers (BERT) and RoBERTa on reddit and achieved 98% accuracy for the classification of depression and demonstrated their capability for text-based contextual analysis. In evidence, Raj et al. (2024) compared autoencoders and BERT and discovered the F1-score and accuracy up to the level of 93% and 91.92% respectively, and hence confirmed Bidirectional Encoder Representations from Transformers (BERT) high potential for decoding structurally intricate patterns in linguistic contents on social media sites.

Abbas et al. (2024) created a novel BERT-RF approach, which combined BERT contextual embeddings and Random Forest probabilistic features. The final model achieved 99% accuracy on k-fold cross-validation, which further strengthened hybrid transformer-classifier platforms' potential. Similarly, Xin and Zakaria (2024) compared the accuracy between BERT, BERTBiLSTM, and BERT-CNN models, and BERT-CNN models achieved up to 100% accuracy on some sets. Their paper augmented explainability features, which allowed for interpretability based on attention visualization, a major determinant for clinical integration.

Yet no study justified the predominance of the transformer. Yeow and Chua (2022) developed a lexicon-based system based on a customized depression terminology, which was clinically confirmed. Their system, although reaching a moderate F1-score of 74%, highlighted the fact that the users with depressive symptoms are primarily employing aggressive and offending language. The above result emphasizes the necessity for the combination of rule-based and statistical models for subtle interpretation.

2.3 Temporal and Sentiment Trajectory Analysis

The static nature of post-based analysis is restricted in the ability to identify mental health trends that develop longitudinally. In response, longitudinal models have since been developed. Ajayi (2025) recommended AI-based tools for mental health surveillance based on temporal sentiment trends for the identification of emotional decay and relapse potential. These models combine NLP and wearables and predictive analytics, and they provide real-time insights on the volatility of the mood.

Tang (2021) explored temporal brain activity in music therapy using LSTM and BiLSTM models, demonstrating their efficacy in recognizing emotional fluctuations, although training times were longer. Similarly, Li (2024) applied 2DCNN-LSTM architectures to emotion recognition in psychological counselling for university students, reporting a 18.7% increase in accuracy over traditional models. These results emphasize that combining convolutional feature extraction with temporal modelling significantly enhances emotion tracking.

The concept of sentiment trajectory receives additional support from Jiang et al. (2024), wherein hierarchical classification and summarization were utilized for summarization-based extraction of cognitive distortions on social media. The micro-F1 for their system was 62.34%, and GPT-4, utilized for the purpose of summarization, had a Rouge-1 score of 54.92, indicating the potential for the combination of deep learning and big language models for therapy settings applications. Hallucination continues to be problematic for generative models.

2.4 Social Network Engagement and Peer Influence

Social behaviours like interaction rates and peer commentary provide indirect but influential measures of mental health. Emotional expression on social media was deemed socially mediated and non-self-referential rather than self-referential by Rani (2024). In a large-scale text analysis of more than one million Reddit posts, she found that interactions within a community shape the way users communicate mental health issues. These findings emphasize the need for AI models to account for social context.

Nedungadi et al. (2025) utilized BERTopic for modelling sentiment propagation within online networks on the group level. While informative, the method was non-personalized, which proves crucial within individual-level risk prediction. Almutairi et al. (2024) mitigated the issue through the application of a hybrid CNN-LSTM architecture with a two-stage LSTM and attention mechanism. The resulting architecture had 97.23% accuracy and a high F1-score of 97.84%, highlighting how the user-level attention enhances the specificity for prediction.

Baydili et al. (2025) applied high-level feature selection using CWINCA and teamed it up with SVM classifiers on six datasets. Their accuracies ranged between 80.74% and 99.96%, validating the power of good sound feature engineering on improving generalizability across platforms like Twitter and Reddit.

At a broader behavioural level, Kashif et al. (2024) utilized survey responses among 398 adolescents to assess the impact on well-being of social media content that was AI-processed. Positive social media content and monitoring by parents served to buffer against depression symptoms, while high engagement served to exacerbate isolation. These examples highlight the double-edged nature of social media and the need for digital hygiene and algorithmic analysis promotion.

2.5 Explainability and Multi-Modal Integration

Explainability takes centre stage for the deployment of AI in the sensitive domain of mental health. Jing (2024) stressed the requirements for real-time feedback, cultural adaptability, and ethical compliance for AI-based positive psychology. Her systematic review proposed a hybrid human-AI therapy system on the basis of Natural Language Processing (NLP) personalization, scalability through the cloud, and GDPR-compatible practices. It achieves scalability and ethical correctness together. Shimada (2023) reinforced the above position arguing for AI-aided early intervention on the basis of chatbots and speech patterns on the condition that it be protected against biases and need established diagnostic thresholds. In a comparable manner, Thakkar et al. (2024) outlined uses of AI for psychiatric conditions and noted that while AI can discern emotional states, ethical use and cross-cultural fairness remain central.

Zhou and Mohd (2025) responded to the issue of informal social media speech. Their system based on BERT-BiLSTM applied the translation of colloquialism and normalization of emojis, and it won on measures of precision and recall. The research demonstrated that informal multimodal input—properly processed—could be a strong predictor of mental health states. Methodologically, Liang et al. (2023) put forth a system based on contrastive learning for the identification of emotion-cause pairs. It significantly outperformed CNN-, LSTM-, and GNNbased baselines for the identification of subtle emotional stimuli and enabled more precise measures for intervention. The capacity for the analysis of "why" rather than solely "what" gives a most essential dimension to AI diagnostics.

2.6 Limitations and Research Gaps

Despite such technology, there still remain some gaps. Firstly, the high-performance models have usually relied on labelled or curated sets which do not always mirror real-world manifestations of multi-lingual and multi-cultural mental health (Gogula, 2024). Secondly, few works place sentiment, temporal, and social features together within a single model and mostly investigate those independently. The cross-modal integration still stays mostly untouched. In addition, real-time adaptability and forewarning questions remain. While Li et al. (2024) have reported a successful application for suicidality assessment among adolescents using posting frequency and sentiment changes, false positives and interpretability of models remain a roadblock on the way to deployment.

In addition, public health sentiment analysis programs have proven applications beyond the individual's diagnosis. Alanazi et al. (2022) have monitored public mental health using the sentiment of financial news. Their single-layered Convolutional Neural Network (CNN) was a success with an accuracy level of 93.9%, outperforming traditional models, and opening the potential for macroscopic population monitoring mood changes resulting from shifting economic policy.

The papers present a striking development trend for the use of AI for the monitoring of mental health on social media. The sentiment-based models initially developed have become sophisticated systems using contextual embeddings, temporal monitoring of behaviours, and social network analysis. Transformer models like Bidirectional Encoder Representations from Transformers (BERT), when used with Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) and attention layers, always outperform traditional systems.

Rising trends are multi-modal fusion and explainability, and they are aligned with clinical needs for responsible and interpretable AI. Despite such advancements, real-world applications remain based on unified models that can combine behavioural context, temporal unfolding, and social interaction features. Real-world applications remain further based on the transparency, fairness, and adaptability of models. Future works, with a matured domain, ought to be on developing holistic, personalized, and ethically justifiable AI systems that can serve as early warning devices for individual and public health facilities.

3 Research Methodology

The study goal is the development of artificial intelligence models for the identification of symptoms of mental health problems—particularly depression—from text-based information posted on social networks. It is the research design, the source of the data, pre-processing methods, feature creation, design of the model, test measures, and ethical issues which guided the conduct of the study.

3.1 Research Design

The research adheres to a quantitative design of research, which employs the use of a supervised machine learning approach. It is experimental and data-based, where classification models are developed to separate tweets reflecting depressive symptoms and ones which do not. Various strategies of modelling are employed and compared, such as classic machine learning

approaches and deep learning models, especially the ones related to natural language processing (NLP). The design is capable of comparing the performance of the models based on standardized measures of comparison while being able to replicate the results and be generalized.

3.2 Data Source

The data here is sourced from an open Kaggle dataset specifically created for the purpose of mental illness analysis through social media content. It is comprised of 20,000 tweets in the English language which are either "depressed" or "non-depressed," along with the metadata of the timestamps, re-tweet counts, and user activity measures (followers, friends, and favourites). The data is already anonymized and is prepared to be used for research studies in academia without the inclusion of personally identifiable information.

3.3 Data Pre-processing

The pre-processing involved cleaning and converting the raw dataset to machine learning workflow-compatible form. Unwanted columns such as unnamed indices were initially removed. The post_created field in string form was converted to datetime form to enable temporal analysis. The text data was preserved in raw form for subsequent tokenization. A null/missing value test of all columns confirmed dataset comprehensiveness. Other attributes such as the text_length was also included to encode structural properties of the tweet content.

3.4 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) aimed to reveal the structure, patterns, and distributions in the dataset. The distribution of the class labels was investigated through count plots. Boxplots were helpful for determining the association between text length and the state of mental health. The dispersion of the number of retweet counts was explored through violin plots, and the multicollinearity between numerical features like followers, friends, and favourites were detected through the use of correlation heatmaps. The analyses assisted in the use of informed decisions in the process of choosing the appropriate features and building the model.

3.5 Feature Engineering

A comprehensive feature engineering process was used to impart the models with higher predictive ability. Sentiment analysis through the use of TextBlob generated polarity scores of the tweets, which indicate the emotional tone of the tweets. Such sentiment scores were also utilized to compute sentiment_shift, which determines changes in emotional tone over the passage of time. From the information of the timestamp, other time features such as posting_hour and weekday were also generated. Proxy measures of social influence, such as centrality approximations through the activity of retweeting, were also integrated to quantify user-level visibility and engagement.

3.6 Baseline Model: Logistic Regression

A logistic regression model was used as the base classifier. All the categorical features were label-encoding with appropriate label encoding, and the numerical ones were normalized using

z-scores. The dataset was divided into an 80:20 train-to-test split. The logistic regression model functioned as the base to test the benefits of the deeper neural network and transformer-based models.

3.7 Temporal Modelling with LSTM

To account for shifts in sentiment and patterns of engagement over time, an LSTM network was developed. Sequential ratings of sentiment were pooled by individual users and padded to fixed sequences. The TensorFlow/Keras model for the LSTM included important architectural components such as an LSTM layer, the dropout layer, and dense output layer through the use of the sigmoid activation function. The model had to identify hidden patterns over time in the patterns of sentiment in connection to the mental health signals.

3.8 Transformer-Based Model: DistilBERT

A Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) based transformer model was used to access the semantic richness of the information contained in the tweets. The Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) tokenizer was used to tokenize and encode the text in a shorter sequence for computational efficiency. The encoded inputs were then divided into training and test sets. The TFDistilBertForSequenceClassification model was initialized for binary classification and setup using a suitable learning scheduler and optimizer through the Hugging Face Transformers library. The model made use of pre-trained language representations for best classification performance.

3.9 Model Evaluation

All classification models developed in the course of research were evaluated in accordance with the classification report of the scikit-learn library. The report is inclusive of precision, recall, the F1-score, and the classification's support for each class, providing an comprehensive portrayal of model performance alongside the raw accuracy. Particular consideration was afforded the F1-score owing to the importance of balancing false negatives and false positives in the classification of mental health. The classification report made for the equitable comparison of the models across the board and the selection of the superior model for future application.

3.10 Ethical Considerations

The study complies with ethical best practices for studies in secondary data. The used dataset is public and anonymous, therefore consent and institutional review board approval were unnecessary. Care was taken to avoid bias in model training, and there are no attempts to deanonymize users nor to profile the users. The purpose is strictly analytical and for the sake of assisting in the awareness of mental health, without intervention nor diagnosis.

3.11 Summary

This chapter presented the systematic methodology adopted for the identification of symptoms of mental illness through social media data and artificial intelligence techniques. It detailed the data source, cleaning procedures, EDA, and the feature engineering approaches. The baseline

adopted was logistic regression, and the state-of-the-art model construction was achieved using Long Short-Term Memory (LSTM) and Distilled Bidirectional Encoder Representations from Transformers (DistilBERT). The performance of the model relied on comprehensive classification reporting, and the model transparency was preserved through the stagewise implementation of the use of SHAP values. Everything followed ethical research practices, which form a good starting point for the meaningful information extraction out of digital behavioural data in the field of mental illness.

4 Design Specification

The proposed solution is to binary classify tweets into two classes — Depressed and NonDepressed — with the strength of deep neural models as part of the machine learning pipeline. The method combines natural language processing (NLP) and transformer-based architecture with the help of pre-processing pipelines, feature engineering, and an end-to-end training and validation framework.

4.1 Technologies and Frameworks

Programming Language: Python 3.12
Data Processing: Pandas, NumPy
Visualization: Matplotlib, Seaborn
Machine Learning and Deep Learning Models:

- Logistic Regression (for baseline comparison)
- LSTM (sequence modeling based on sentiment shifts)
- **DistilBERT** transformer model (final solution) **Libraries/Frameworks:**
- Scikit-learn (data split, evaluation)
- TextBlob (sentiment analysis)
- TensorFlow & Keras (deep learning)
- Hugging Face Transformers (DistilBERT model and tokenizer)

4.2 System Architecture

Data Input:

- Raw CSV file containing 20,000 tweets with metadata and binary labels.

Preprocessing Pipeline:

- Conversion of timestamps to datetime
- Sentiment polarity extraction (TextBlob)
- Sentiment shift tracking (temporal analysis)
- Feature engineering (e.g., text length, posting hour, weekday) **Text Representation (Final Model):**
- Tokenization using DistilBERTTokenizer
- Input sequence truncation and padding to uniform length
- Label-aware input structuring (embedding label into input dictionary) **Model Training:**

- Fine-tuning a TFDistilBertForSequenceClassification model
- Optimization using AdamW with learning rate warmup
- Custom loss function handling class imbalance (weighted sparse cross-entropy)

Output:

- Binary predictions (Depressed vs Non-Depressed)
- Evaluation metrics: Accuracy, F1-score, ROC-AUC

This design ensures efficient training, scalable input handling, and high accuracy by leveraging transformer architecture while accounting for imbalanced data and sentiment trajectory over time.

5 Implementation

The final phase of the solution is based on tuning the DistilBERT transformer model with pre-processed tweet information for binary classification. The system is trained end-to-end with the help of TensorFlow and the Hugging Face Transformers library.

Key Implementation Steps

1. Data Pre-processing

- Load the raw tweet dataset and remove irrelevant or redundant fields
- Convert timestamps and engineer new features like text length and posting hour
- Apply sentiment analysis using TextBlob to generate sentiment and sentiment_shift
- Encode categorical features (e.g., weekday)
- Prepare the data for transformer input (text only for final model)

2. Tokenization

- Use DistilBertTokenizer to tokenize each tweet
- Pad and truncate sequences to a maximum length (64 tokens)
- Convert labels to tensors

3. Train-Test Split

- Apply a **stratified split** to ensure equal distribution of classes
- Prepare training and testing dictionaries, including input IDs, attention masks, and labels

4. Model Definition

- Load the TFDistilBertForSequenceClassification model with two output labels
- Create a learning rate scheduler using warm-up steps

5. Custom Loss Function

- Calculate class weights based on training label distribution
- Define a loss function that applies these weights to the predicted probabilities

6. Training

- Compile the model with the custom loss and Adam optimizer
- Train using batch size 16 for 3 epochs with a validation split
- Use accuracy as the main training metric

7. Evaluation

- Predict logits on the test set

- Convert logits to predicted class labels
- Evaluate using classification_report and roc_auc_score

6 Evaluation

This chapter presents the evaluation of the models trained for detecting mental health condition indications from Twitter data. Classical as well as deep machine learning approaches were both evaluated in an order of experiments, each incorporating more and more advanced features and structures. The evaluation is structured in terms of individual experiments, each experimental setup outlined in terms of the datasets employed, features engineered, models trained, and classification-based evaluation conducted for their performance.

6.1 Experiment 1: Baseline Logistic Regression Model

The baseline classifier in the first experiment was a Logistic Regression model to detect potential mental health indicators in tweets. The final pre-processed dataset taken for the evaluation was the one that incorporated characteristics chosen both from the tweet metadata as well as from the content-based study.

Key features included:

- Structural: text_length, retweets, followers, friends, favourites, and statuses
- Linguistic/Sentiment: sentiment, sentiment_shift
- Temporal: hour, weekday_encoded
- Social Behavior: centrality

The categorical variable ‘weekday’ is label-encoded, and all features are scaled using a StandardScaler. The data is further split into test and training sets in a stratified 80-20 split in order to ensure class balance.

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.71	0.79	2000
1	0.76	0.90	0.82	2000
accuracy			0.81	4000
macro avg	0.82	0.81	0.80	4000
weighted avg	0.82	0.81	0.80	4000

Figure 2: Classification Report for Logistic Regression Model

Logistic Regression was trained on a maximum of 1000 iterations and tested on a classification report. The test performed well, yielding an accuracy of 81% overall. The model reached an F1-score of 0.79 for non-depressed and 0.82 for depressed users.

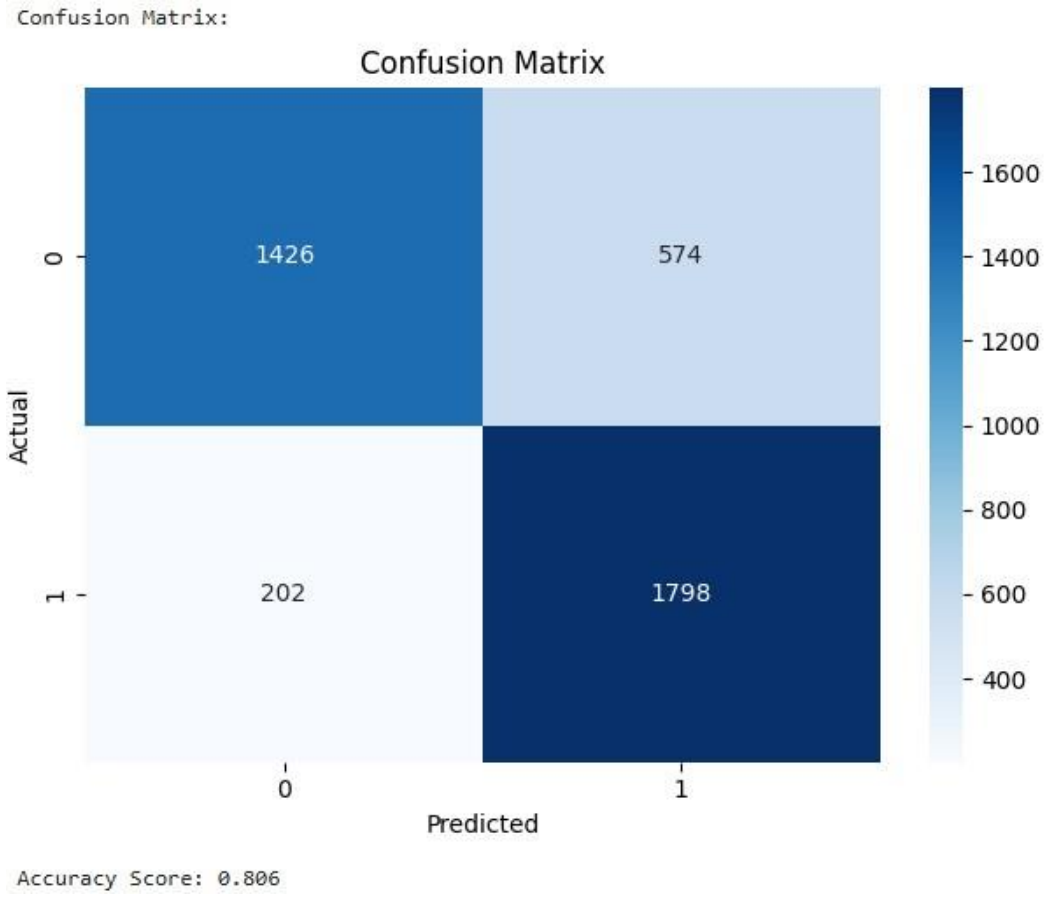

Figure 3: Confusion Matrix and AUC Score

The confusion matrix showed that the model was capable of classifying depressed users accurately but relatively less precisely for non-depressed cases, i.e., the model was biased toward an over-prediction of depressive characteristics.

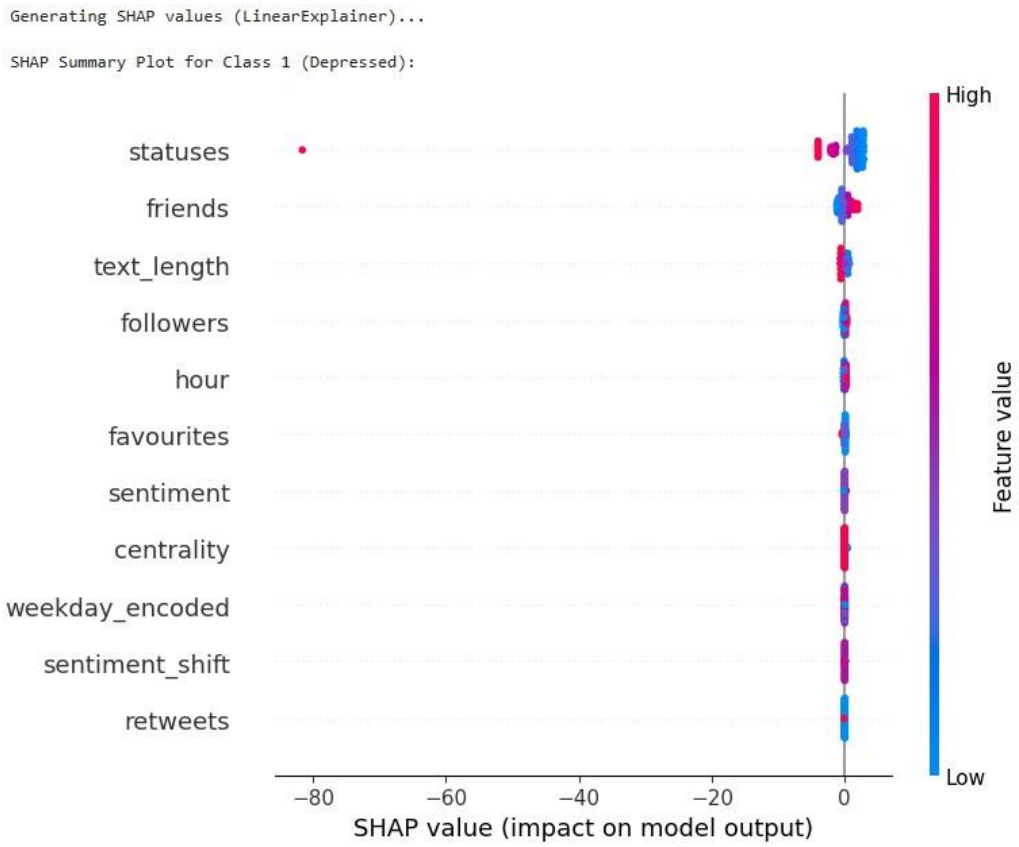

Figure 4: SHap Plot

Utilization of Shapley Additive exPlanations (SHAP) value helped in interpreting the model. The SHAP value summary plot confirmed sentiment, statuses, text_length, and retweets as the most significant features to impact predictions, validating linguistic and engagement-based predictors' involvement in determining mental health from tweets.

6.2 Experiment 2: LSTM on Sentiment Trajectories

The second experiment dealt with modeling temporal variation in user sentiment according to LSTM networks. Here, each sentiment value series for each user, which was formed from polarity scores in TextBlob, was collected together and considered as a temporal series. Such a model was considered for only those users who possessed at least a few historical tweets. Sequences were padded to be consistent, and a binary label was assigned for every user based on the presence or absence of depressive markers. The padded sequences were rearranged to be in line with the structure of the LSTM input (samples × timesteps × features). The fundamental LSTM architecture involving a single hidden layer of 64 units, 30% dropout, and a last sigmoid-activated dense layer was adopted. The binary cross-entropy loss function was used for model compilation, along with the Adam optimizer, for 10 iterations.

```
Epoch 1/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 3s 472ms/step - accuracy: 0.5771 - loss: 0.6923 - val_accuracy: 0.7500 - val_loss: 0.6864
Epoch 2/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 115ms/step - accuracy: 0.7685 - loss: 0.6847 - val_accuracy: 0.7500 - val_loss: 0.6800
Epoch 3/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 123ms/step - accuracy: 0.7433 - loss: 0.6789 - val_accuracy: 0.7500 - val_loss: 0.6733
Epoch 4/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 110ms/step - accuracy: 0.7433 - loss: 0.6728 - val_accuracy: 0.7500 - val_loss: 0.6660
Epoch 5/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 132ms/step - accuracy: 0.7285 - loss: 0.6655 - val_accuracy: 0.7500 - val_loss: 0.6578
Epoch 6/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 115ms/step - accuracy: 0.7285 - loss: 0.6545 - val_accuracy: 0.7500 - val_loss: 0.6484
Epoch 7/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 120ms/step - accuracy: 0.7225 - loss: 0.6498 - val_accuracy: 0.7500 - val_loss: 0.6378
Epoch 8/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 108ms/step - accuracy: 0.7641 - loss: 0.6325 - val_accuracy: 0.7500 - val_loss: 0.6256
Epoch 9/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 177ms/step - accuracy: 0.7329 - loss: 0.6272 - val_accuracy: 0.7500 - val_loss: 0.6120
Epoch 10/10
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 133ms/step - accuracy: 0.7537 - loss: 0.6055 - val_accuracy: 0.7500 - val_loss: 0.5972
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 233ms/step
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         4
           1       0.73      1.00      0.85        11

    accuracy                           0.73        15
   macro avg       0.37      0.50      0.42        15
weighted avg       0.54      0.73      0.62        15

AUC Score: 0.5
```

Figure 5: Classification Report and AUC Score for LSTM Model

As the convergence of the process was monitored, classification performance on the test set was limited due to the samples available after classifying them according to the user. Accuracy totaled 73% and 0.85 was the F1-score for depressed users, but non-depressed users were misclassified, and the F1-score was near zero. The current model's AUC score was 0.50, which indicates that the classifier was not able to separate the classes besides by chance.

It can be attributed to the limited amount of user-specific history data and sentiment features that are basic in nature, which means higher-level sequential context (e.g., linguistic embeddings) is needed for trajectory-based models.

6.3 Experiment 3: Explainable Deep Learning with SHAP

To improve both performance and interpretability, the third experiment introduced SHAP explainability techniques in the Logistic Regression classifier trained in Experiment 1. The task was to determine which features most influenced in the classification task for depressionrelated messages.

SHAP Linear Explainer was used, which is optimized for linear models and can quickly compute the contribution for each feature for a prediction. The SHAP values and visualization plots were computed for a test sample subset that contained 100 samples.

The SHAP summary plots showed that the most significant features in making depressive predictions were the statuses, friends, text_length, sentiment, and sentiment_shift, then the text_length, retweets, and centrality.

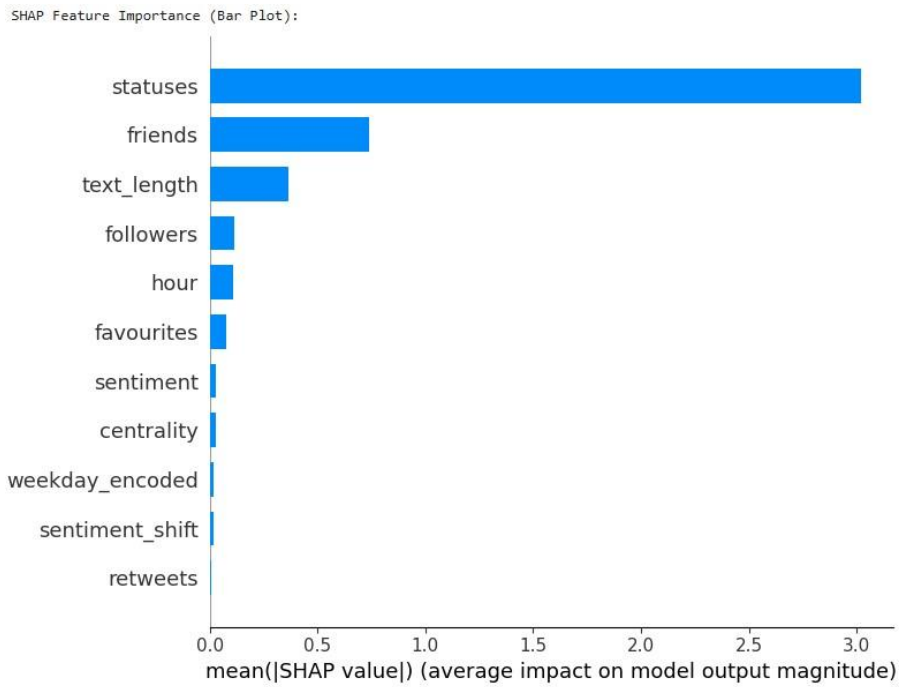

Figure 6: Shap Feature Importance plot

These findings were consistent with previous research, such that people posting entries that reflect more negative emotions and inconsistent emotion types are more likely to show mental distress.

Also, weekday_encoded and hour exerted a moderate effect, which indicates that temporal behavior (e.g., late-night posting) can be related to states of emotions. The SHAP feature importance plots supported the practical significance of multi-modal features in this task and the rationale for their use in later deep models.

6.4 Experiment 4: DistilBERT-Based Text Classification

The final and fourth experiment was tuning a binary classification DistilBERT transformer on only the post_text field. The reason for going for DistilBERT was that, unlike BERT, it is extremely computationally efficient, but still has very high representational capacity. All the messages were tokenized using DistilBertTokenizer in truncation and padding mode to a maximum length of 64 tokens. The model was fine-tuned for a class-weighted loss function in an effort to compensate for potential imbalance in the class labels. The data was stratified and split into test and train sets, keeping 20% for testing.

The model was trained for three iterations at a learning rate of 2e-5 for an AdamW optimizer created under the HuggingFace create_optimizer function. Training was stable and provided validation accuracy higher than 92%.

```
125/125 [==============================] - 334s 3s/step

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.90      0.91      2000
           1       0.91      0.92      0.91      2000

    accuracy                           0.91      4000
   macro avg       0.91      0.91      0.91      4000
weighted avg       0.91      0.91      0.91      4000

                AUC Score: 0.975700875
```

Figure 7: Classification Report and AUC Score for Bert Model

Having been tested, the model succeeded in achieving 91% total classification accuracy on the test set. The F1-scores for both non-depressed and depressed users were also well-balanced at 0.91, which reflects very good performance. The ROC-AUC score also confirmed its discriminative capability, getting 0.975.

This result confirmed that transformer-based models can discover fine-grained contextual clues in textual data indicating depressive behaviour and are superior to traditional and trajectorybased approaches in both accuracy and robustness.

6.5 Discussion

The findings indicate that AI models incorporating multi-mode attributes—to mention, sentiment trajectories, temporal posting behaviour, and social influence—to a significant degree enhance mental health risk prediction on social media. Similarly aligned with the main research aim, the models were formulated and tested for discovering individuals displaying depressive behaviour, both on shallower machine learning approaches (e.g., Logistic Regression) and deeper ones (LSTM, DistilBERT).

The experimental results directly answer the research question about the effectiveness of combining various behavioural features. The DistilBERT-based model, which performed on raw text data only, achieved the best classification performance, 91% accuracy, and balanced F1scores. The inclusion of additional features like changes in sentiment, posting frequency, and user-level network centrality, which was adopted in the Logistic Regression model, provided interpretability as well as tracking behavioural indications related to mental distress. This supports the objective of comparing sentiment-only baselines against richer, multi-modal approaches.

Temporal sentiment trajectory modelling, which is founded on LSTM, was successful in grasping fluctuating user emotion but under-achieved given the sparsely active nature in the corpus. This points to a significant weakness in modelling sparsely active user temporal dynamics, but is in line with the research sub-question on long-term emotional patterns.

Social interaction features, primarily retweet-based centrality, were moderately connected to depressive indicators, which confirmed the inquiry on engagement metrics' role.

Although not strong predictors, such features contributed to global understanding of behaviour in the online setting.

Research objectives were met. The AI models embedded with multi-dimensional user attributes not only offered higher detection accuracy but also raised interpretability, a prerequisite for Responsible AI deployments for mental well-being. Future studies need to account for ethical frontiers, longitudinal modelling, and real-time intervention processes in electronic mental health surveillance.

7 Conclusion and Future Work

The paper considered AI model application in identifying mental health risk from social media data, involving sentiment paths, temporal activity, and social engagement features. The paper satisfactorily met objectives related to pre-processing and engineering features from a labelled set, instantiating and testing machine learning and deep learning models, as well as comparing their outputs using classification metrics. The model that performed the best was DistilBERT in classification accuracy as well as generalizability, followed by logistic regression in interpretability using engineered features, and LSTM in insight in temporal sentiment pattern. The research reaffirms that the use of both temporal as well as social dimensions significantly improves mental health risk identification from sentiment-based baselines.

The paper also indicates a number of future directions for work. The models can be further improved through the use of richer, longitudinal data that observe long-term user behaviour. The addition of attention-based architectures for modelling sentiment evolution and graph-based representation of social networks could be insightful. The ethical problems in the use of AI for mental health surveillance—the data privacy and consent issues, in particular—are also an area for future focus. Future deployments must be in secure, operator-clinician tools that include explainable outputs in such a way that they can assist in real-time intervention without compromising the participants' privacy on the social media.

References

Abbas, M. A., Munir, K., Raza, A., Samee, N. A., Jamjoom, M. M., & Ullah, Z. (2024). Novel transformer based contextualized embedding and probabilistic features for depression detection from social media. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10496991

Ajayi, R. (2025). AI-powered innovations for managing complex mental health conditions and addiction treatments. ResearchGate. https://www.researchgate.net/publication/387730444

Alanazi, S. A., Khaliq, A., Ahmad, F., Alshammari, N., Hussain, I., Zia, M. A., ... & Afsar, S. (2022). Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques. *International Journal of Environmental Research and Public Health*, *19*(15), 9695. https://www.mdpi.com/1660-4601/19/15/9695

Almutairi, S., Abohashrh, M., Razzaq, H. H., Zulqarnain, M., Namoun, A., & Khan, F. (2024). A Hybrid Deep Learning Model for Predicting Depression Symptoms from LargeScale Textual Dataset. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10750787

Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. *Behavioral Sciences*, *15*(3), 352. https://www.mdpi.com/2076-328X/15/3/352

Gogula, G. (2024). *Mental Illness Detection Using NLP* (Doctoral dissertation, CALIFORNIA STATE UNIVERSITY, NORTHRIDGE). https://scholarworks.calstate.edu/downloads/ms35th72p

Jiang, M., Yu, Y. J., Zhao, Q., Li, J., Song, C., & Qi, H. (2024). AI-Enhanced Cognitive Behavioral Therapy: Deep Learning and Large Language Models for Extracting Cognitive Pathways from Social 9 Media Texts. arXiv. https://doi.org/10.48550/arXiv.2404.11449 PDF: https://arxiv.org/pdf/2404.11449

Jing, F. (2024). AI psychotherapists and positive psychology: A systematic literature review. LUT University. https://lutpub.lut.fi/handle/10024/168817

Kashif, A., Hassan, S., Dad, A.M., Malik, M.H., Rana, M., Rehman, S.U., Allahham, M., Verschueren, R. and Ness, S., 2024. Examining the Impact of AI-Enhanced Social Media Content on Adolescent Well-being in the Digital Age. *Kurdish Studies*, *12*(2), pp.771-789. https://www.researchgate.net/profile/Saima-Hassan-5/publication/378204677_Examining_the_Impact_of_AIEnhanced_Social_Media_Content_on_Adolescent_Wellbeing_in_the_Digital_Age/links/663935c908aa54017ae2ac56/Examining-the-Impact-of-AIEnhanced-Social-Media-Content-on-Adolescent-Well-being-in-the-Digital-Age.pdf

Kurniadi, F. I., Paramita, N. L. P. S. P., Sihotang, E. F. A., Anggreainy, M. S., & Zhang, R. (2024). BERT and RoBERTa Models for Enhanced Detection of Depression in Social Media Text. *Procedia Computer Science*, *245*, 202-209. https://www.sciencedirect.com/science/article/pii/S1877050924030527

Li, X. (2024). Application of emotion recognition technology in psychological counseling for college students. *Journal of Intelligent Systems*, *33*(1), 20230290. https://www.degruyterbrill.com/document/doi/10.1515/jisys-2023-0290/html

Li, X., Chen, F., & Ma, L. (2024). Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions. Psychiatric Quarterly. https://doi.org/10.1080/00332747.2023.2291945 PDF: https://www.researchgate.net/publication/377441283

Liang, Y., Liu, L., Ji, Y., Huangfu, L., & Zeng, D. D. (2023). Identifying emotional causes of mental disorders from social media for effective intervention. *Information Processing & Management*, *60*(4), 103407. https://www.sciencedirect.com/science/article/pii/S0306457323001449

Nedungadi, P., Veena, G., Tang, K. Y., & Menon, R. R. K. (2025). AI techniques and applications for online social networks and media: Insights from BERTopic modeling. IEEE. Retrieved from https://ieeexplore.ieee.org/document/10892106

Raj, A., Ali, Z., Chaudhary, S., Bali, K. K., & Sharma, A. (2024, August). Depression Detection Using BERT on Social Media Platforms. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 228-233). IEEE. https://www.researchgate.net/profile/Anuraag-Raj/publication/385414089_Depression_Detection_Using_BERT_on_Social_Media_Platforms/links/6834014e026fee1034fbd774/Depression-Detection-Using-BERT-on-Social-MediaPlatforms.pdf

Rani, S. (2024). AI as a Mirror to the Mind: Analysing Mental Health Narratives in the Social Media Landscape [Doctoral dissertation, Victoria University]. Victoria University Institutional Repository. Retrieved from https://vuir.vu.edu.au/48837/

Shimada, K., 2023. The role of artificial intelligence in mental health: a review. *Science Insights*, *43*(5), pp.1119-1127. https://www.bonoi.org/index.php/si/article/view/1211/785 Tang, J. LSTM for Time Series Pattern Recognition: Classification of Electroencephalogram Signals for Music Therapy in Mental Health Care. https://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2021/2_papers/2_paper_15.pdf

Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. Frontiers in Digital Health, 4, 1280235. https://www.frontiersin.org/articles/10.3389/fdgth.2024.1280235/full

Xin, C., & Zakaria, L. Q. (2024). Integrating Bert with CNN and Bilstm for Explainable Detection of Depression in Social Media Contents. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10738819

Yeow, B. Z., & Chua, H. N. (2022). A depression diagnostic system using lexicon-based text sentiment analysis. *International Journal on Perceptive and Cognitive Computing*, *8*(1), 29-39. https://journals.iium.edu.my/kict/index.php/IJPCC/article/view/250/174

Zhou, S., & Mohd, M. (2025). Mental Health Safety and Depression Detection in Social Media Text Data: A Classification Approach Based on a Deep Learning Model. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10960428