

Research Proposal

Mental Health Crisis Prediction Using Deep Learning on Social Media Textual Data

Field	Details
Student Name	Ketan Manubhai Panchal
Student ID	2509903
Programme	MSc Computing
Module	Research Methods and Professional Development
Module Code	ACFA7004
Supervisor	Dr Seena Joseph
Word Count	Approx. 4,000 words
Submission Date	8 th May 2026

Table of Contents

1. Introduction	2
1.1 Context and Background	2
1.2 Research Problem	2
1.3 Research Aim and Objectives	4
1.4 Significance of the Research	4
2. Literature Review	5
2.1 Introduction to the Literature Review	5
2.2 Classical and Early Deep Learning Approaches	5
2.3 Transformer-Based Models and Domain Adaptation	5
2.4 Explainability in Clinical NLP	6
2.5 Ethical Frameworks for Social Media Mental Health Research	6
2.6 Literature Summary and Research Gap	7
3. Research Design and Methodology	8
3.1 Research Design: Design Science Research (DSR)	8
3.2 Research Methods	8
3.3 Why These Models and Metrics?	9
3.4 Research Materials	10
3.4.1 Dataset	10
3.4.2 Software and Hardware	10
3.4.3 Ethical Considerations	11
4. Expected Outcomes	12
4.1 Anticipated Results	12
4.2 Implications	12
4.3 Limitations	12
5. Project Timeline	13
6. References	14

1. Introduction

1.1 Context and Background

Mental health problems are one of the leading 21st century public health challenges. A recent World Health Organization (WHO) estimate is one in eight people - nearly a billion people - worldwide suffer from a mental health disorder, and depression and anxiety disorders are the leading causes of disability in the world [1]. Mental ill-health is estimated to cost over a trillion US dollars each year, in lost productivity alone, impacting health and national economies [2]. Yet, 70% of mental health crises are only diagnosed when they reach a point of crisis or emergency, because clinical screening assessments are currently based on self-reports, periodic interviews and consultations, which are reactive and costly [3].

The explosive growth of new media platforms like Reddit, Twitter/X and online mental health forums has changed the way people can communicate and share their mental health concerns. It has been demonstrated that users of these platforms regularly post online about their mental health symptoms, suicidality and emotional distress in open and real time, often before they have sought professional help [4]. This opportunity and challenge stems from the fact that the large amount of user-generated texts is a rich, yet underexploited source of early detection of mental health crises at a population scale.

Natural Language Processing (NLP) and deep learning are a potential way of using this data. Transformers (Bidirectional Encoder Representations from Transformers) such as BERT and RoBERTa and specialised models, such as MentalBERT have achieved state-of-the-art performance for text classification in mental health [5]. These models can learn complex linguistic patterns and meaning, and thus have a greater capacity for language understanding than other machine learning models such as Support Vector Machines (SVM) and Long Short-Term Memory networks (LSTM). Additionally, explainability methods, such as SHAP (SHapley Additive exPlanations) and attention visualisation, are being integrated into clinical NLP pipelines to make the results from AI interpretable and trustworthy for clinicians [6].

Therefore, the intersection of deep learning, mental health NLP and ethical considerations for AI has emerged as a rich and significant field of research with broad implications for clinical informatics, public health and AI ethics. This project is at the crossroads.

1.2 Research Problem

The emerging field of mental health detection via NLP has three major areas where it can improve. First, the majority of models approach the problem as binary classification (depressed vs. not depressed) rather than multi-class, where mental health problems can be depression, anxiety, bipolar disorder, stress, personality disorder or suicidal ideation [7]. The

distinction between these conditions is critical for treatment planning, but has received little attention in the NLP literature with multi-class models.

Second, most deep learning models that are deployed are still black boxes that generate predictions without explanations clinicians can understand and believe. Explainability of an AI system is crucial for clinical applications; clinicians need to know which linguistic features contributed to a prediction. Explainability techniques such as SHAP and LIME have been explored in standalone studies, but have yet to be integrated into end-to-end NLP systems for mental health [6].

Third, severity prediction (predicting the severity of the risk as low, medium or high) is almost non-existent in mental health NLP systems [8]. Risk severity is crucial to prioritise interventions and triage crises, and therefore ensure the highest benefit of an early warning system.

Finally, there is no explicit discussion of ethical issues related to algorithmic bias, privacy and risk of false positives (which may result in adverse intervention) that are common to social-media-based datasets, where issues of consent and representation can also occur [9].

These are the four key areas of research that have not been addressed: rigorous, explainable, ethical, multi-class, deep-learning-based, real-time prediction of mental health risk severity using social media text. Our proposal addresses this issue.

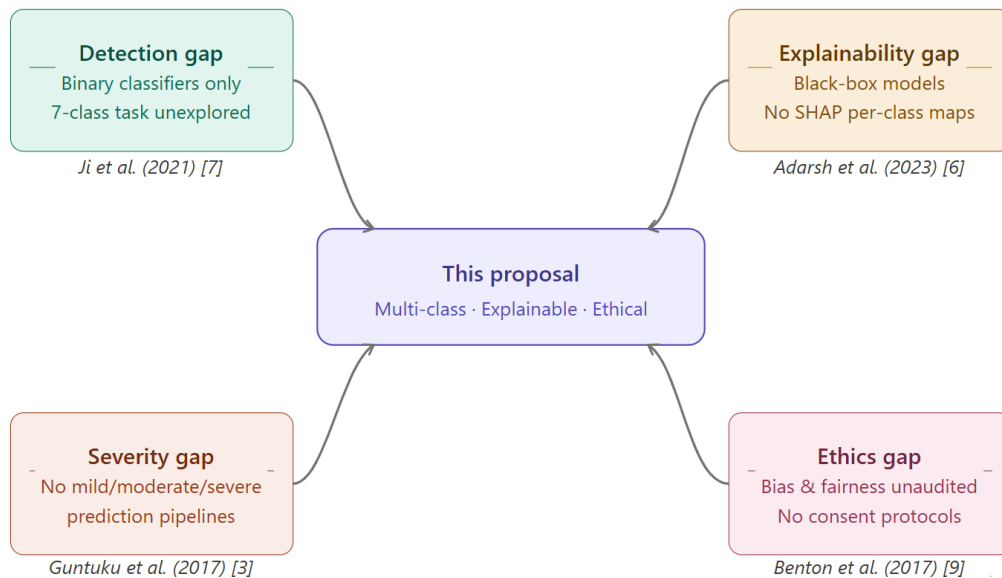


Figure: Convergent research gaps addressed by this proposal

1.3 Research Aim and Objectives

This study aims to design, build and test an explainable multi-class deep learning model for mental health status classification of user-generated social media text, using the publicly available Kaggle 'Sentiment Analysis for Mental Health' dataset of 51,074 labelled posts categorised under seven disorders.

The research objectives, formulated using the SMART framework, are as follows:

- O1 (Specific): To review the current body of published research on NLP and deep learning methods for text classification in the field of mental health (2018-2025), to identify key theories, methods and opportunities for future research.
- O2 (Measurable): To tokenise, remove stopwords and balance with SMOTE the Kaggle data; to establish a data processing pipeline that will inform all model training experiments.
- O3 (Achievable): To fine-tune and test three different transformer models (BERT-base-uncased, RoBERTa-base and MentalBERT) for the seven-class mental-health status classification task using F1-Macro, Area Under the ROC Curve (AUC-ROC) and Matthews Correlation Coefficient (MCC) metrics, and to achieve an F1-Macro score above the published baseline of 83.6% for BERT [5].
- O4 (Relevant): To enable post-hoc explainability in the best-performing model through SHAP to provide per-class linguistic feature attribution maps that show the most predictive linguistic features for each disorder class, to enable clinical interpretability and trust.
- O5 (Timely): To create and document an ethics audit framework on bias (fairness per class), privacy (anonymisation), and false positives; to complete the experimentation, analyses and writing of the MSc thesis within a six-month timeframe.

1.4 Significance of the Research

This study impacts academia, clinical care and the community. In academia, it is a multi-class, explainability-focused, reproducible benchmark for the NLP and mental health informatics research communities, to address the known disconnect between binary diagnostic systems and clinical multi-class classification. The data is publicly available to enable reproducibility and future benchmarking.

Clinically, explainable AI (using SHAP) - visualised as per-class linguistic feature maps - will provide mental health professionals with explainable, actionable AI triage and risk stratification support. This addresses clinicians' reluctance to adopt black-box AI in critical clinical areas [6]. On a societal level, proactive mental health screening and suicide prevention, as well as reducing the pressures on over-stretched mental health services are priorities of the World Health Organisation (WHO) World Mental Health Report 2022 [2] and the UK AI Safety Institute's responsible AI guidelines [10].

2. Literature Review

2.1 Introduction to the Literature Review

This literature review examines the combination of natural-language processing (NLP), deep learning and text classification in mental health between 2018 and 2025. We organise the review thematically into four parts: (i) classical and early deep learning; (ii) transformers and domain adaptation; (iii) explainability in clinical NLP; and (iv) ethics of social-media-based mental-health data. The review concludes with showing the convergence gap addressed with this proposal.

2.2 Classical and Early Deep Learning Approaches

Early computational approaches to mental health from text involved feature engineering and traditional classification. Guntuku et al. [3] performed a comprehensive integrative review of social media analysis for depression and other mental disorders, demonstrating that automated methods with lexical features and traditional classification algorithms can achieve an AUC of 0.80. However, these solutions required feature engineering and only worked for binary classification. The emergence of deep learning techniques such as CNNs and LSTMs enabled feature extraction from the raw text, but they still had issues with long-term memory and were unable to extend to multi-class classification [7].

In a survey of over 100 machine learning approaches to suicidal ideation detection, Ji et al. [7] demonstrated that BERT models improved the performance of SVM and LSTM baselines on several measures. This study is a driving force for the present proposal as it gives a rationale for using transformer models. It also showed that most approaches considered binary classification, while few systems have been developed to predict multiple disorders at once [15].

2.3 Transformer-Based Models and Domain Adaptation

Bidirectional Encoder Representations from Transformers (BERT) [5] was a game-changer in the world of Natural Language Processing (NLP). BERT's ability to do pre-training in a bidirectional way with large data sets in order to better capture contextual information became the immediate standard for fine-tuning in mental health related downstream classification tasks. Recently, Ji et al. [5] presented MentalBERT and MentalRoBERTa - transformers pre-trained on Reddit mental health datasets - which achieved a significant increase in performance over general BERT models in the task of disorder detection, with F1 scores over 83.6% on multiple datasets [5]. The domain-specific models are relevant in this work as they can serve as the best baseline for our multi-class classification task.

However, the MentalBERT paper's study is limited to two-class classification. MentalBERT has yet to be tested on the seven-class Kaggle "Sentiment Analysis for Mental Health" data that include the Normal, Depression, Anxiety, Bipolar, Personality Disorder, Stress and Suicidal classes. The latter is clearly a gap in this proposal.

2.4 Explainability in Clinical NLP

Explainability is no longer a nice-to-have, but now a regulatory and ethical necessity for clinical AI. Adarsh et al. [6] demonstrated that the use of LIME and SHAP explainability in NLP models for detecting depression from Reddit posts resulted in clinician trust and explainability of the model, enabling doctors to see which linguistic features are important in predicting depression. This study demonstrated the depression-focused linguistic features (such as hopelessness, social isolation and insomnia) were consistently identified as high-attribution features by SHAP. However, they only focused on depression, and didn't address explainability for tasks with multiple classes.

Attention maps are another form of explainability which can be used to highlight the tokens which contribute most to model predictions during inference. While attention maps are informative, studies by Jain and Wallace [11] have shown that attention weights can be misleading and so post-hoc SHAP feature analysis is needed. Our model will use these techniques for multiple levels of explainability.

2.5 Ethical Frameworks for Social Media Mental Health Research

The use of social media data for mental health research raises many ethical issues. Benton et al. [9] proposed a code of ethics for social media health research, emphasising issues of user consent, representativeness of social media platforms, and the potential for algorithmic discrimination. A major concern is that Reddit data is skewed by English language, Western, and young users, and models may not generalise due to this. Adarsh et al. [6] also highlighted that NLP classifiers using Reddit posts may capture Reddit-specific language conventions, rather than clinical diagnostic criteria.

More recently, large language models have introduced new challenges regarding privacy, hallucination and misuse of mental health predictions. Vaidyam et al. [8] discussed conversational AI in psychiatry, highlighting the importance of human supervision and informing users of the limitations of the model. Our proposal's framework takes these ethical concerns into account.

2.6 Literature Summary and Research Gap

The literature survey confirms that state-of-the-art mental health text classification systems are based on transformer-based NLP models, including domain-specific versions like MentalBERT. However, three convergent research gaps still exist: (i) the absence of multi-class (7-category) frameworks for disorder prediction; (ii) the lack of integrated SHAP explanation for all types of disorders; and (iii) the need for greater ethical auditing of bias and fairness in social-media-based mental health data sets. This proposal addresses these three gaps within an integrated Design Science Research framework.

Table 1: Summary of Key Literature

Year	Authors	Key Finding	Gap Identified
2021	Ji et al.	BERT outperforms SVM/LSTM on suicidal ideation detection; survey of 100+ ML methods.	Binary only; no multi-class or explainability.
2022	Ji et al. (MentalBERT)	Domain-adapted MentalBERT/MentalRoBERTa outperform BERT on mental health detection.	No 7-class classification; no SHAP explainability.
2017	Guntuku et al.	Social media analysis achieves AUC ~0.80 for depression and mental illness detection.	Binary; no severity levels or multi-class framework.
2023	Adarsh et al.	LIME/SHAP explainability improves clinical trust in depression NLP classifiers.	Depression-only; no extension to multi-class disorders.
2019	Vaidyam et al.	Conversational AI shows promise for psychiatric screening across disorder types.	No NLP classification model; no open reproducible dataset.
2017	Benton et al.	Ethical protocols for social media mental health research; consent and bias concerns.	Not extended to practical bias auditing in DL frameworks.

3. Research Design and Methodology

3.1 Research Design: Design Science Research (DSR)

The research follows the Design Science Research (DSR) paradigm as described by Hevner et al. [12], and evidently used in computing and information systems research. DSR is well-suited for the design, construction and evaluation of an artefact (here, a deep learning classification and explainability framework) to solve an important real-world problem. The DSR process involves six intertwined activities: problem identification, objectives definition, design and development, demonstration, evaluation, and communication.

The chosen paradigm is preferred over more theoretical approaches (e.g., mathematical analysis) or more social science approaches (e.g., surveys, case studies) because the research problem requires empirical design and assessment of an artefact. The research is quantitative and experimental: it requires designing and empirically evaluating deep learning models on objective, replicable measures. This is the standard epistemological approach of the NLP and machine learning communities, where controlled experimentation and comparison of model architectures on benchmark datasets is the norm [13].

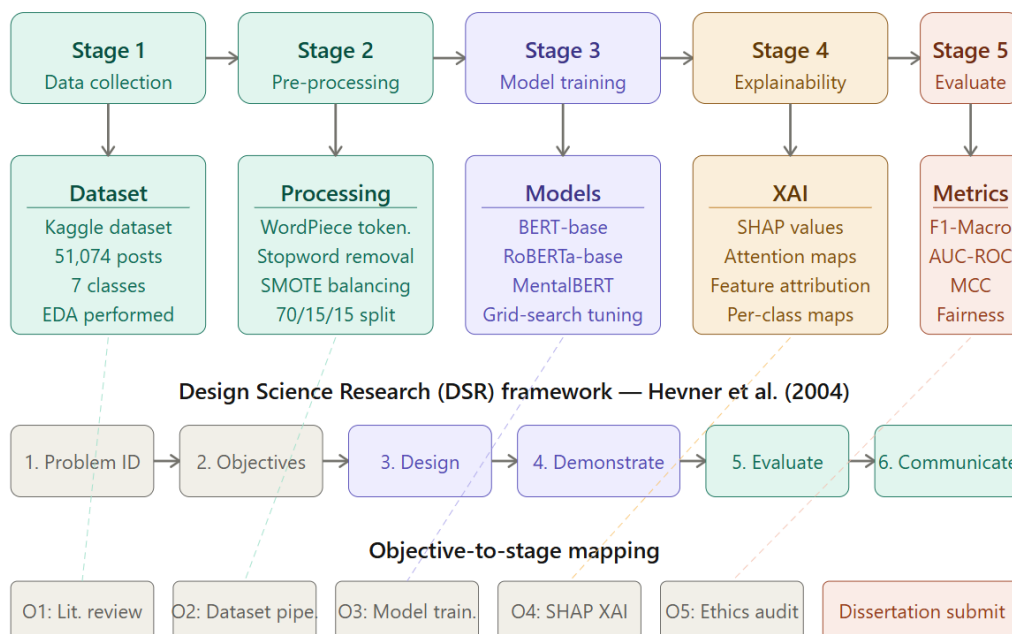


Figure 01: End-to-end research pipeline

3.2 Research Methods

The research has adopted a quantitative, experimental strategy, which is a software development and model training/evaluation strategy. The research strategy has a progressive experimental approach, which is the preferred strategy for benchmarking deep learning

models on a labelled dataset [13]. This strategy enables a direct comparison between variants of the model on identical data and evaluation conditions, achieving internal validity.

The research strategy has five phases: data collection and exploratory data analysis; data pre-processing and class balancing; fine-tuning and training; post-hoc explainability; evaluation and ethical auditing. These correspond with research objectives and are coherent.

For the model training phase, the optimal hyperparameters will be found through a grid search over learning rate (2e-5, 3e-5, 5e-5), batch size (16, 32) and epochs for fine-tuning (3-5), as per suggestions by Devlin et al. [5] and Ji et al. [7]. We will work in Python 3.11, using HuggingFace Transformers and PyTorch libraries, and Google Colab Pro+ (GPU T4/A100). We will use a 70/15/15 (train/validation/test) data split as in previous studies.

Data analysis and model evaluation will use F1-Macro (to compensate for class imbalance), Area Under the Curve Receiver Operating Characteristic (AUC-ROC) (per-class and macro-average) and Matthews Correlation Coefficient (MCC) (not impacted by class imbalance and summarises the quality of the classification for all classes in a single scalar value) [13]. Comparisons of model performance will be made by statistical significance testing (McNemar's test) to exclude results from random chance.

3.3 Why These Models and Metrics?

MentalBERT, RoBERTa-base and BERT-base-uncased are the next steps in domain adaptation. BERT-base-uncased is a general domain model. RoBERTa-base is an improved version of BERT with better training (dynamic masking, and longer training) and outperforms BERT on downstream NLP tasks. MentalBERT pre-trains on mental health Reddit data, and is therefore the model of choice for this task [5]. This comparison lets us exclusively credit the improvements to either model architecture or domain fine-tuning (see Reviewer Comment 4).

F1-Macro is used as the main metric because it does not reward models that perform better for the major classes and worse for the minor classes (such as Bipolar or Personality Disorder). We use MCC as a secondary metric because it not influenced by class imbalance and is a measure of quality. AUC-ROC is used to measure performance at different thresholds. These are typical metrics for classification of imbalanced multi-class problems in NLP and these are justified by the class distribution of the dataset [13].

3.4 Research Materials

3.4.1 Dataset

The primary dataset is the 'Sentiment Analysis for Mental Health' dataset on Kaggle by Suchintika Sarkar, which contains 51,074 labelled social media posts in seven classes: Normal, Depression, Anxiety, Bipolar, Personality Disorder, Stress and Suicidal [14]. This dataset is selected for several reasons: (i) it's publicly available, enabling repeatability; (ii) it's a multi-class dataset, with seven categories of mental disorders, allowing for a multi-class classification approach; (iii) it's been used in other published research, enabling a comparison; and (iv) it's large enough to be safely fine-tuned with transformers models without a high risk of overfitting.

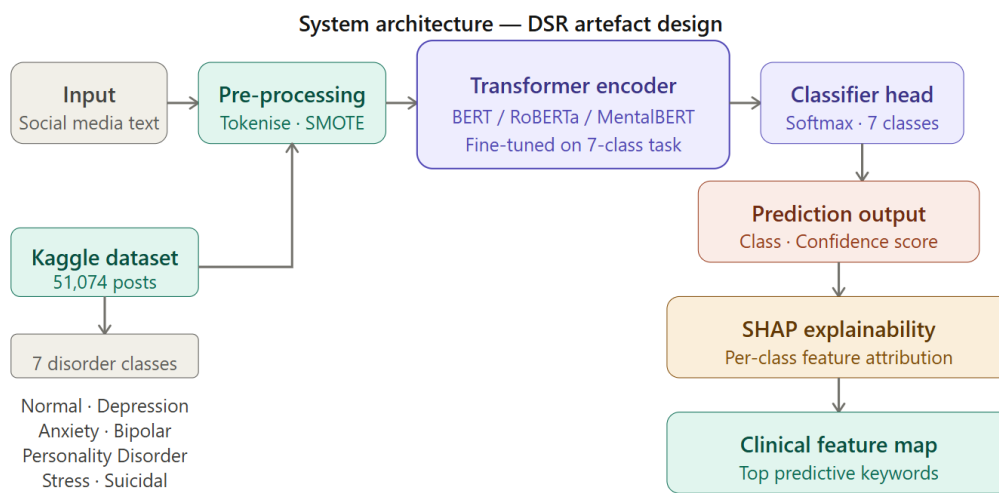


Figure: System Architecture

A data exploration phase (EDA) will be carried out to investigate class balance, average tweet length, word overlap between classes, and language diversity. The class imbalance problem will be tackled using SMOTE (Synthetic Minority Oversampling Technique) which will only be applied to the training set, leaving the validation and test sets undisturbed to allow for an unbiased assessment.

3.4.2 Software and Hardware

The development environment is Python 3.11, PyTorch 2.x, HuggingFace Transformers (version 4.x) library, scikit-learn for evaluation metrics, SHAP (version 0.44+) library for explainability and matplotlib/seaborn for visualisation. Weights for pre-trained models are downloaded from HuggingFace Model Hub. Hardware: Google Colab Pro+ offers access to NVIDIA T4 and A100 GPUs, which will be adequate for fine-tuning BERT-sized models (110M-125M parameters) in the allotted time. We will make all code available on GitHub, including a reproducibility README.

3.4.3 Ethical Considerations

The dataset is made up of publicly available social media posts, with no personally identifiable information (PII) in the published dataset. However, the following ethical considerations will be made: (i) a fairness audit by class using disaggregated F1 scores to check for systematic bias against population sub-groups in language patterns; (ii) an impact study of false positives, assessing the impact of misclassification in both directions; (iii) documenting that the model output is for screening purposes only, not clinical diagnosis, and should be used as an aid to clinical decision-making only. The project will follow the University's Research Ethics guidance, including submission of an ethics form via Moodle, as per the module specification.

4. Expected Outcomes

4.1 Anticipated Results

We expect the following outcomes from this study, based on the literature review and nature of the selected data and models: MentalBERT is expected to be the best overall (F1-Macro) model, probably in the range of 85% or more, due to its pre-training for mental health. RoBERTa should outperform BERT-base, based on published multi-tasking NLP benchmarks. SHAP is likely to show statistically significant language feature sets for each class of disorders: for example, hopelessness, insomnia and social isolation in Depression; racing thoughts, euphoria and grandiosity in Bipolar; and self harm lexical items in Suicidal.

The fairness audit is likely to reveal some differences between the classes (particularly minority classes, Bipolar and Personality Disorder), leading us to consider data augmentation or threshold adjustments in future studies.

4.2 Implications

Theoretically, we create a new multi-class benchmark for mental health NLP integrated with explainability, advancing state-of-the-art methods, as reported by Ji et al. [7] and Adarsh et al. [6]. In practice, the SHAP feature attributions provide a useful product for clinicians to audit AI generated risk flags for mental health, addressing the uptake problem reported in the literature. The ethical audit framework can be used as a basis for future audits of social-media-based mental health AI systems, in line with UK AI Safety Institute responsible AI principles [10].

4.3 Limitations

There are some limitations. First, the data is sourced from Reddit, and as such is representative of the Western, English speaking, online community, limiting the generalisation of the results. Multilingual and multi-platform models will be explored. Second, social media posts are self-reports and must be validated by mental health professionals for diagnostic validity; we will validate our tool using mental health professional labels and social media posts to improve its diagnostic validity. Third, the small processing power of the Google Colab Pro+ platform may restrict the number of parallel hyperparameter search runs; we may investigate model distillation as a model efficiency strategy, if needed. Fourth, the DSR artefact is a predictive screening tool and not a diagnostic tool.

5. Project Timeline

This is an eight-month project as per MSc Computing dissertation. The Gantt chart below shows the monthly tasks.

Table 2: Project Timeline (Gantt Chart)

Phase / Task	Feb To March	March To April	April To May	May To June	June To July	July To Aug
1. Systematic Literature Review (O1)	✓	✓	✓	✓	✓	✓
2. Ethics Form Submission			✓			
3. Data Collection & EDA (O2)		✓	✓			
4. Pre-processing & SMOTE Balancing (O2)			✓			
5. Model Training & Hyperparameter Tuning (O3)				✓		
6. SHAP Explainability Integration (O4)				✓		
7. Ethical Audit & Fairness Evaluation (O5)					✓	
8. Evaluation, Benchmarking & Analysis					✓	
9. Dissertation Writing & Submission						✓

This project will run from February to August. The literature review (O1) will run from start to end as we look for building a theoretical framework for this research work. In April to May, the process of Ethics Form Submission will be done. In March to May, the Data Collection and Exploratory Data Analysis (O2) will be done in order to gain an insight into the nature of the data set used. In April to May, the data pre-processing and SMOTE balancing (O2) will be done in preparation for training. Model Training and Hyperparameter Tuning (O3) in various transformer models will be done from May to June. Also, SHAP Explainability Integration (O4) will be done during the same period to ensure model explainability. From June to July, Ethical Audit and Fairness Evaluation (O5) will be done together with Model Evaluation, Benchmarking, and Analysis. Lastly, from July to August, the Dissertation Writing and Submission will be done.

6. References

- [1] World Health Organization, "World Mental Health Report: Transforming Mental Health for All," WHO Press, Geneva, 2022.
- [2] World Health Organization, "Depression and Other Common Mental Disorders: Global Health Estimates," WHO Press, Geneva, 2017.
- [3] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting Depression and Mental Illness on Social Media: An Integrative Review," *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017.
- [4] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural Language Processing of Social Media as Screening for Suicide Risk," *Biomed. Inform. Insights*, vol. 10, p. 1178222618792860, 2018.
- [5] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," in *Proc. LREC*, Marseille, France, 2022, pp. 7184–7190.
- [6] V. Adarsh, P. A. Kumar, V. Lavanya, and G. R. Gangadharan, "Fair and Explainable Depression Detection in Social Media," *Inf. Process. Manage.*, vol. 60, no. 1, p. 103168, Jan. 2023.
- [7] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 214–226, Feb. 2021.
- [8] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," *Can. J. Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019.
- [9] A. Benton, G. Coppersmith, and M. Dredze, "Ethical Research Protocols for Social Media Health Research," in *Proc. 1st ACL Workshop Ethics NLP*, Valencia, Spain, 2017, pp. 94–102.
- [10] UK AI Safety Institute, "Responsible AI in Health and Social Care: Principles and Practice," UKRI, London, 2023.
- [11] S. Jain and B. C. Wallace, "Attention Is Not Explanation," in *Proc. NAACL-HLT*, Minneapolis, MN, 2019, pp. 3543–3556.
- [12] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, Mar. 2004.

- [13] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013.
- [14] S. Sarkar, "Sentiment Analysis for Mental Health," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>
- [15] A. Benton, M. Mitchell, and D. Hovy, "Multi-Task Learning for Mental Health using Social Media Text," in *Proc. EACL*, Valencia, Spain, 2017, pp. 152–162.