

Mental Health Crisis Prediction

Using Deep Learning on Social Media Textual Data



Dataset: Sentiment Analysis for Mental Health — Kaggle | **Methodology:** Deep Learning Transformer

Presenter

Ketan Panchal

Programme

MSc Computing

Module Code

ACFA7004

Supervisor

Dr Seena Joseph

01 | Research Context & Background

1 in 8

People globally live with a mental health disorder
[10]

70%

Of crises go undetected until they become severe
[5]

\$1 Trillion

Annual global economic loss due to mental disorders

Context

- ▶ Social media platforms (Reddit, Twitter/X) have emerged as significant spaces where individuals express mental distress candidly [1].
- ▶ Natural Language Processing (NLP) and deep learning enable automated detection of early warning signs from unstructured text at scale [2].
- ▶ Prior work shows BERT-based models achieve state-of-the-art results in mental health text classification tasks [3].

Current Landscape

- ▶ The 'Sentiment Analysis for Mental Health' dataset on Kaggle contains 51,074 labelled social media posts across 7 categories: Normal, Depression, Anxiety, Bipolar, Personality Disorder, Stress, and Suicidal [4].
- ▶ Current clinical assessments are reactive; AI-driven tools offer a paradigm shift toward proactive, early-stage identification [5].
- ▶ Ethical deployment of such systems is a rapidly growing interdisciplinary research area bridging AI, psychology, and public health [6].

02 | Research Problem

Problem Statement

Existing mental health screening tools rely on clinical interviews and self-reporting, which are resource-intensive, subjective, and often too late. Despite the exponential growth of user-generated mental health discourse on social media, no robust, explainable, and ethically-grounded multi-class deep learning framework exists for real-time crisis severity prediction across heterogeneous disorder categories.



Detection Gap

Existing models focus on binary classification (depressed/not); multi-class disorder prediction remains under-explored [2].



Explainability Gap

Most DL models are black-boxes; clinical adoption demands interpretability via attention maps or SHAP [7].



Severity Gap

Severity level prediction (mild/moderate/severe) is absent in current NLP-based mental health pipelines [8].



Ethics Gap

Limited research addresses bias, privacy, and consent in social-media-derived mental health datasets [9].

03 | Brief Literature Review & Summary

Year	Authors	Key Finding	Research Gap
2021	Ji et al. "Suicidal Ideation Detection: A Review of ML Methods"	BERT outperforms SVM/LSTM on mental health text tasks. Comprehensive survey of 100+ detection methods.	No multi-class disorder prediction; no explainability component.
2022	Ji et al. (MentalBERT) "Publicly Available Pretrained LMs for Mental Healthcare"	MentalBERT/MentalRoBERTa pre-trained on Reddit mental health posts outperform BERT on disorder detection.	No multi-class classification across 7 disorder types; no SHAP explainability.
2017	Guntuku et al. "Detecting Depression & Mental Illness on Social Media"	Automated social media analysis can detect mental illness with AUC ~0.8; review of key methods.	Binary detection only; no severity estimation or multi-class framework.
2023	Adarsh et al. "Fair and Explainable Depression Detection in Social Media"	LIME/SHAP explainability with NLP classifiers on Reddit improves clinical trust and transparency.	Depression-only; not extended to multi-class disorders or severity prediction.
2019	Vaidyam et al. "Chatbots & Conversational Agents in Mental Health"	Conversational AI shows promise for psychiatric screening and treatment support across disorder types.	No NLP classification model; no open social-media dataset for reproducibility.

04 | Research Aim & Objectives



Research Aim

To design, implement, and evaluate an explainable multi-class deep learning framework for mental health status classification using user-generated text, leveraging the Kaggle 'Sentiment Analysis for Mental Health' dataset (51,074 posts, 7 classes: Normal, Depression, Anxiety, Bipolar, Personality Disorder, Stress, Suicidal).

01

SPECIFIC

Conduct a systematic literature review of NLP and deep learning approaches for mental health text classification, identifying key gaps.

02

MEASURABLE

Pre-process the Kaggle dataset; apply SMOTE balancing across all 7 status categories and surpass the published BERT baseline of 83.6% F1-Macro.

03

ACHIEVABLE

Fine-tune BERT, RoBERTa, and MentalBERT models on the multi-class classification task and compare performance (F1, AUC, MCC).

04

RELEVANT

Integrate SHAP-based explainability to generate clinical feature attribution maps, improving model interpretability for mental health professionals.

05

TIME-BOUND

Develop an ethical evaluation framework addressing bias, privacy, and fairness; publish findings and submit dissertation.

05 | Research Methodology

Quantitative / Experimental

Design Science Research

Justification: Objective performance benchmarking

1. Data Collection

- Kaggle: Sentiment Analysis
- for Mental Health
- 51,074 posts | 7 classes
- Anxiety/Bipolar/Depression
- Normal/Stress/Suicidal/PD

2. Pre-processing

- Tokenization (WordPiece)
- Stopword removal
- Class balancing (SMOTE)
- Train/Val/Test split 70/15/15

3. Model Training

Deep Learning Model
Hyperparameter tuning

4. Explainability

- SHAP value analysis
- Attention visualization
- Linguistic feature maps
- Clinician review panel

5. Evaluation

- F1-Macro, AUC-ROC
- Matthews CC (MCC)
- Bias & fairness audit
- Ethical assessment

Dataset:

Dataset: Kaggle "Sentiment Analysis for Mental Health" by Suchintika Sarkar | 51,074 posts | 7 classes: Normal, Depression, Anxiety, Bipolar, Personality Disorder, Stress, Suicidal | kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

06 | Project Planning & Resources

	M1	M2	M3	M4	M5	M6	M7	M8
Literature Review								
Data Collection & EDA								
Pre-processing & Balancing								
Model Training & Tuning								
Explainability Integration								
Ethical Framework & Audit								
Evaluation & Benchmarking								
Writing & Submission								

Resources:

✓ Google Colab Pro+ (GPU T4/A100)

✓ HuggingFace Transformers

✓ Python 3.11 / PyTorch

✓ Kaggle: Sentiment Analysis for Mental Health

07 | Impact & Limitations

Potential Impact

Academic

Contributes multi-class, explainability-aware benchmark to NLP/mental health literature. Enables reproducible research with open dataset.

Clinical

Provides explainable AI outputs (SHAP) that clinicians can trust for early triage and risk stratification.

Societal

Early crisis detection may reduce suicide rates and ease pressure on mental health services globally (WHO priority).

Policy

Informs data ethics guidelines for AI use in sensitive health domains; aligns with UK AI Safety Institute priorities.

Limitations & Mitigation

Self-report bias in Reddit data

→ Cross-validate with clinical annotations; acknowledge in ethics section.

Single-language corpus (English)

→ Scope limitation clearly stated; future work on multilingual models.

Generalisability to clinical populations

→ Results framed as screening aid, not diagnostic tool; clinician validation.

Computational resource constraints

→ Google Colab Pro+ provides sufficient GPU; model distillation for efficiency.

08 | References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] S. Sarkar, "Sentiment Analysis for Mental Health," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>
- [5] 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.
- [6] 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.
- [7] 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.
- [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] World Health Organization, "World Mental Health Report: Transforming Mental Health for All," WHO Press, Geneva, 2022.



Thank You

Questions & Discussion

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