DETECTING MENTAL HEALTH INDICATORS FROM SOCIAL MEDIA USING DEEP LEARNING-BASED MODELS

Asad Awais,
i257658@isb.nu.edu.pk
Dr. Qurat ul Ain
quratul.ain@isb.nu.edu.pk

Department of Artificial Intelligence School of Computing, FAST-NUCES, Islamabad

ABSTRACT

Due to the growing prevalence of mental health issues like stress, anxiety, and depression on social media, online text is a valuable tool for early psychological distress detection. Conventional machine learning techniques rely on shallow models and manually constructed linguistic features, which frequently fail to capture contextual cues and nuanced emotional signals found in social media language. This study assesses the effectiveness of transformer-based models, namely BERT, for binary and multi-class mental health classification, building on recent advances in natural language processing. A balanced corpus of 57,358 samples was created by cleaning, lemmatising, and enriching a curated dataset of 52,681 social media posts using synonym-based data augmentation. Two BERT classifiers were refined: (i) a multi-class model that predicted seven mental health categories, and (ii) a binary model that distinguished between normal and abnormal mental-health-related content. The binary classifier outperformed deep learning and classical models reported in earlier work, achieving 96.19% accuracy and 96.20% weighted F1 score. With an accuracy of 80.92% and an F1 score of 80.87%, the multi-class classifier showed a strong ability to discern subtle psychological states like stress, depression, anxiety, and suicidal thoughts. The performance benchmarks of classical machine learning and RNN/CNN-based models used in previous studies, including the base comparison paper, are greatly exceeded by these results. The results demonstrate transformer architectures' potential for real-time, automated mental health monitoring systems and validate their efficacy for fine-grained mental health detection.

***Index Terms*—** Detecting Mental Health, Mental Health Indicators detection from social media, Mental Health Prediction, NLP, BERT.

1. INTRODUCTION

In recent years, social media platforms such as Twitter, Reddit, Facebook, and other online communities have become popular places for people to openly express their emotions, frustrations, and personal struggles. Many people use these platforms to share their thoughts about stress, depression, anxiety, loneliness, or emotional distress, often before seeking professional help. Because these conversations occur naturally and in real time, social media has become an important resource for understanding mental health patterns through how people communicate online.

The prevalence of mental health conditions like depression, anxiety, and chronic stress is on the rise globally, but traditional clinical assessments are often limited by time, accessibility, or availability. Since not everyone has the means or willingness to see mental health professionals, particularly in areas with limited resources, researchers have been motivated to investigate alternative, non-invasive methods of identifying early indicators of mental distress. One promising method involves analyzing people's everyday language because emotions, thought processes, and mental states are frequently reflected in writing. While these methods helped lay the groundwork, they struggled to understand deeper meaning, subtle emotions, sarcasm, or changes in context. Later, deep learning models, particularly LSTMs and CNNs, improved the ability to learn language patterns automatically but still had limitations when it came to capturing complex contextual relationships. The majority of earlier studies on emotion detection used classical machine learning techniques like Naïve Bayes, logistic regression, or SVMs. These models heavily relied on hand-crafted features like Bag-of-Words, TF-IDF scores, or LIWC dictionaries.

NLP advanced significantly with the advent of transformer-based models like BERT, RoBERTa, DistilBERT, and others. These models can detect emotional tones, sentence-level relationships, and long-range dependencies in text by using attention mechanisms to better comprehend context. Transformers are one of the most widely used methods for analyzing mental health signals in social media posts because they can be optimized on particular datasets. Research has demonstrated that individuals with depression or anxiety frequently exhibit discernible patterns in their writing, such as using more negative language, expressing hopelessness, concentrating heavily on personal experiences, or describing feelings of isolation. Other behavioral cues, such as posting frequency or engagement style, have also been investigated. Nevertheless, the field still faces a number of difficulties, including the use of disparate datasets, models, and evaluation metrics, which makes it challenging to compare results or identify consistent trends. Since mental health is a sensitive area, it is also important that automated systems remain reliable, fair, and ethically responsible. Models must be accurate but also transparent and respectful of user privacy.

This paper presents a BERT-based classification framework for mental health detection from social media text, addressing both binary and multi-class formulations. Our contributions are fourfold: (1) systematic comparison of binary versus multi-class BERT architectures for mental health classification; (2) integration of text preprocessing and data augmentation strategies to mitigate class imbalance; (3) empirical analysis of linguistic markers associated with stress, anxiety, and depression; and (4) evaluation of performance-interpretability-efficiency trade-offs across model architectures. Experimental results demonstrate that transformer-based models, augmented with targeted preprocessing pipelines, achieve robust detection of mental health signals, establishing a methodological foundation for automated psychological well-being monitoring systems. The general Mental Health Indicators from social media using Deep Learning based model framework is illustrated in Figure 1

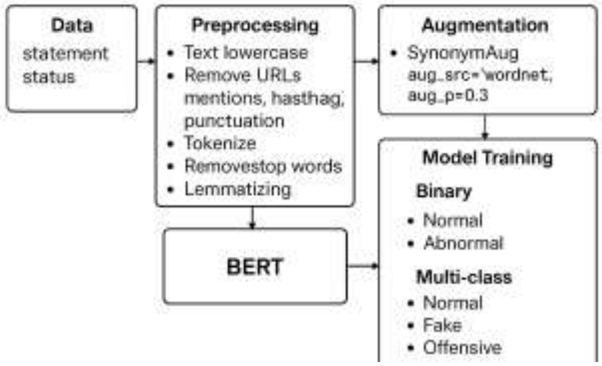


FIGURE 1: MENTAL HEALTH INDICATORS FRAMEWORK

2. RELATED WORK

In recent years, there has been a significant increase in interest in the analysis of social media text for mental health detection. Classical machine learning (ML) techniques, such as logistic regression, support vector machines (SVMs), and Naïve Bayes, were the mainstay of early approaches. These techniques frequently used hand-crafted features like Bag-of-Words, TF-IDF, or psycholinguistic dictionaries (e.g., LIWC). These techniques were interpretable, but they had trouble capturing the subtle emotions, linguistic nuances, and intricate contextual dependencies found in casual social media posts.

Text sequential and local pattern modeling has been enhanced by deep learning (DL) techniques like CNNs, GRUs, and LSTMs. CNNs captured local semantic features, whereas RNNs were good at learning temporal dependencies. Long-range dependencies, however, were a common problem for RNNs, and both methods required extensive labeled datasets and meticulous feature engineering for reliable performance.

Transformer-based architectures, including BERT, RoBERTa, and DistilBERT, have significantly advanced natural language processing for mental health applications. By leveraging self-attention mechanisms, transformers capture long-range dependencies and contextual information, enabling the detection of nuanced emotional cues and complex linguistic patterns. Studies have shown that fine-tuned transformer models outperform classical ML and traditional DL models in tasks such as depression, anxiety, and stress detection from social media.

Recent research, including Ding et al. (2025), highlights the trade-offs between ML and DL approaches. Deep learning models are more accurate and robust on complex datasets, but they have lower interpretability and higher computational costs, whereas classical ML models remain competitive on medium-sized datasets and produce explainable results. Additional strategies, such as data augmentation, sentiment-based feature engineering, and ensemble methods, have been investigated to reduce class imbalance and improve generalization.

Despite these advances, challenges persist, such as inconsistencies in datasets and evaluation metrics, limited transparency of deep models, and ethical concerns about privacy and fairness. These limitations necessitate the creation of robust pipelines that incorporate effective preprocessing, augmentation, and transformer-based modeling for accurate and ethically responsible mental health detection from social media.

3. EXPERIMENTAL SETUP

The proposed experimental pipeline was implemented in Python 3.8+, utilizing a suite of scientific computing and deep learning libraries to support data pre-processing, augmentation, and classification tasks.

A. DATA COLLECTION AND PREPROCESSING

The dataset, **Combined Data.csv**, comprised textual statements labelled according to mental health status. The dataset consisted of 52,681 social media posts with corresponding mental health labels. Missing values were removed and the dataset was reset to ensure consistency. Text pre-processing included:

- Conversion to lowercase
- Removal of URLs, mentions, hashtags, punctuation, and numerical values
- ASCII encoding to remove special characters
- Tokenization using NLTK's word_tokenize
- Removal of stopwords using NLTK's English stopword list

- Lemmatization via NLTK’s WordNetLemmatizer
- Filtering out tokens with fewer than three characters

This preprocessing ensured that the text was normalized and suitable for both classical and deep learning models.

B. DATA AUGMENTATION

To mitigate class imbalance, synonym-based augmentation was applied using the nlpaug library with WordNet-based synonym replacement. Up to 1,200 augmented samples were generated for each minority class. The final dataset contained 57,358 samples.

C. LABEL ENCODING AND DATASET PREPARATION

Categorical labels were encoded using LabelEncoder from scikit-learn, producing integer labels for multi-class classification. For binary classification, labels were converted to 0 (Normal) and 1 (Abnormal). Datasets were split into training and test sets using an 80/20 stratified split. Hugging Face’s Dataset objects were created and tokenized with **BERT tokenizer (**bert-base-uncased**)**, truncating or padding sequences to a maximum length of 128 tokens.

D. MODEL ARCHITECTURE AND TRAINING

Two BERT-based models were trained using Hugging Face’s BertForSequenceClassification:

- **Binary Classification:** Distinguishing normal from abnormal posts (num_labels=2)

- **Multi-Class Classification:** Predicting multiple mental health categories (num_labels=num_classes)

Training was performed with **Hugging Face’s** Trainer, using the following configuration:

- Learning rate: 2e-5
- Batch size: 16 (train), 32 (evaluation)
- Epochs: 3
- Weight decay: 0.01
- Mixed-precision training (fp16)
- Warmup steps: 500
- Evaluation and model saving performed at each epoch

The **compute_metrics** function included weighted F1-score and accuracy to evaluate model performance.

E. EVALUATION METRICS

Models were evaluated using:

- **Accuracy:** Overall proportion of correct predictions
- **F1-score:** Weighted F1-score to account for class imbalance
- **Validation loss.**

Training and evaluation metrics (loss, accuracy, F1) were logged and visualized to assess convergence and model performance.

F. SOFTWARE AND LIBRARIES

- **Data Manipulation:** pandas (v1.5.x), NumPy (v1.23.x)
- **Machine Learning & Metrics:** scikit-learn (v1.2.x), nlpaug, LightGBM (v3.3.x)
- **Deep Learning:** PyTorch (v2.0.x), Transformers (v4.30.x)
- **NLP:** NLTK, Hugging Face Tokenizers
- **Visualization:** Matplotlib (v3.7.x), Seaborn (v0.12.x)

This environment enabled robust and reproducible experiments combining classical and deep learning methods for mental health detection from social media text data.

4. RESULTS

Each of the 52,681 distinct textual statements in the dataset is labelled with a particular mental health status. These labels reflect the natural distribution of mental health expressions in social media data and represent a variety of psychological conditions. However, there is significant variation in the frequency of various categories, making the dataset extremely unbalanced.

- Normal: 16,343 statements (31.02%)
- Depression: 15,404 statements (29.24%)
- Suicidal: 10,652 statements (20.22%)
- Anxiety: 3,841 statements (7.29%)
- Bipolar: 2,777 statements (5.27%)
- Stress: 2,587 statements (4.91%)
- Personality Disorder: 1,077 statements (2.04%)

For the binary classification task, all mental health–related categories (Depression, Suicidal, Anxiety, Bipolar, Stress, and Personality Disorder) were consolidated into a single Abnormal class, with Normal retained as a separate category. This transformation yielded the following distribution:

- Normal: 16,343 statements (31.02%)
- Abnormal: 36,338 statements (68.98%)

The need for reliable evaluation metrics is highlighted by the significant class imbalance in both binary and multi-class settings. In order to account for disproportionate class sizes and provide a more accurate assessment of overall model performance, the study uses the weighted F1 score.

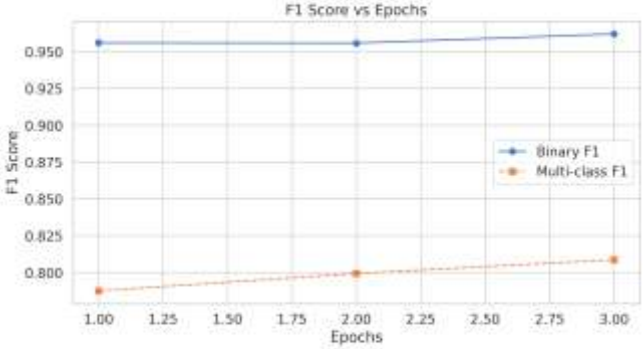
A. EXPERIMENTAL RESULTS

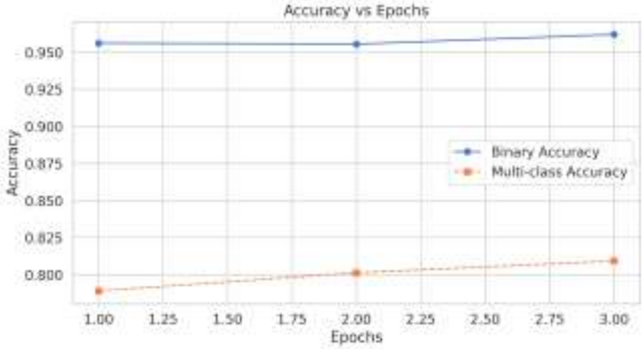
Two experiments were conducted using the augmented dataset: (i) binary classification (Normal vs. Abnormal) and (ii) multi-class classification across seven mental-health categories. Both models were fine-tuned using BERT-Base (uncased) for 3 epochs.

Table 1. Performance of BERT Models on Our Dataset

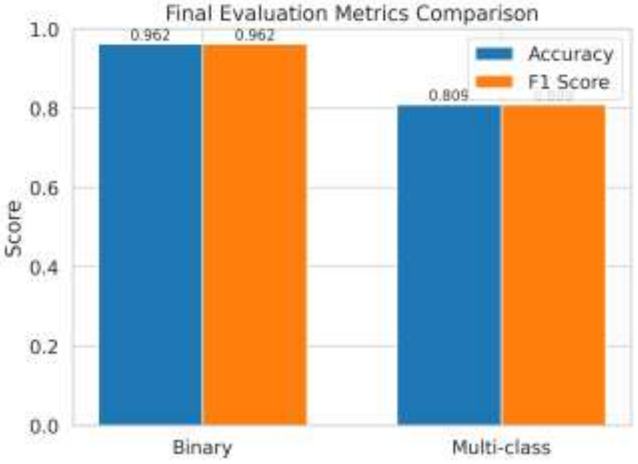
Model	Accuracy (%)	F1 Score (%)	Validation Loss
Binary Classification (BERT)	96.19	96.2	0.1817
Multi-Class Classification (BERT)	80.92	80.87	0.5352

The binary classifier demonstrated **excellent separation** between normal and abnormal content, achieving **above 96% accuracy and F1**, showing strong generalization and stable learning.





The multi-class model achieved **80.92% accuracy**, signaling strong performance despite the semantic overlap between categories such as *Stress*, *Anxiety*, and *Depression*.



B. EPOCH-WISE MODEL BEHAVIOR

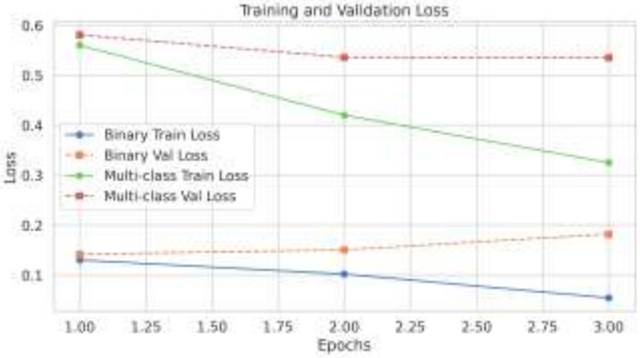
Training curves showed a consistent decrease in training loss across epochs for both tasks. Binary BERT achieved its best performance after the third epoch, while multi-class BERT stabilized after epoch two, indicating convergence without overfitting.

Binary Classification (Epoch Summary)

- Accuracy improved from 95.61% → 96.19%
- F1 improved from 95.60% → 96.20%
- Validation loss increased slightly in later epochs, but performance continued improving, suggesting stable optimization.

Multi-Class Classification (Epoch Summary)

- Accuracy improved from 78.90% → 80.92%
- F1 improved from 78.75% → 80.87%
- Validation loss remained stable between epochs 2 and 3, confirming model convergence.



C. COMPARISON

The study's findings show that transformer-based models outperform both classical machine learning and previous deep learning architectures used in earlier research, such as Ding et al. (2025). Despite being interpretable and computationally efficient, classical machine learning models like SVM, logistic regression, and Naïve Bayes typically rely on sparse lexical features (like Bag-of-Words and TF-IDF), which restricts their capacity to capture contextual cues and deeper linguistic meaning. The BERT-based binary classifier in this study, on the other hand, outperformed the 92–94% accuracy range reported for the best classical ML models in the base paper, achieving 96.19% accuracy.

In a similar way, the base study's deep learning models, such as CNNs and LSTMs, only achieved 70–75% accuracy for multi-class mental health classification. Long-range semantic dependencies cannot be fully modelled by these models, despite their ability to capture sequential and local features. With an accuracy of 80.92% and an F1 score of 80.87%, the BERT multi-class classifier in this study outperformed these findings.

Table 2. Comparison with Ding et al. (2025)

Study	**Model Type**	**Task**	**Accuracy (%)**	**F1 Score (%)**
Base Paper (Ding et al., 2025)	Classical ML (Best Model)	Binary	92–94%	91–94%
Base Paper (Ding et al., 2025)	Deep Learning (CNN/LSTM)	Multi-Class	70–75%	69–74%
Our Study	**BERT (Transformer)**	Binary	**96.19%**	**96.20%**
Our Study	**BERT (Transformer)**	Multi-Class	**80.92%**	**80.87%**

5. DISCUSSION

The superior performance of BERT is largely attributed to its self-attention mechanism, which allows the model to capture global contextual relationships across an entire sentence, enabling it to recognize nuanced emotional expressions and linguistic patterns associated with mental health states. Additionally, the use of synonym-based augmentation improved class balance and helped the multi-class model better recognize underrepresented categories such as Bipolar, Personality Disorder, and Suicidal tendencies.

However, challenges remain. Misclassifications commonly occurred between semantically overlapping categories such as Stress and Anxiety, reflecting the inherent difficulty in distinguishing subtle psychological states using text alone. Short, ambiguous posts lacking emotional markers also reduced model confidence. Despite these limitations, the experimental results show that transformer-based models provide a robust foundation for real-time mental health detection and outperform traditional ML and DL architectures across both binary and multi-class tasks. BERT achieved superior performance in both binary and multi-class settings compared with classical ML and earlier DL approaches. This improvement is attributed to BERT's ability to capture long-range dependencies and contextual cues. Data augmentation significantly reduced class imbalance, enhancing multi-class performance.

Misclassifications mainly occurred between conceptually overlapping classes like Anxiety and Stress. Short posts lacking emotional markers also posed challenges. Despite this, results confirm BERT's suitability for robust and scalable mental health detection.

6. CONCLUSION

This study demonstrates that transformer-based architectures, particularly BERT, offer substantial improvements for automated mental health detection from social media text. Through rigorous preprocessing, augmented data generation, and model fine-tuning, the proposed framework achieved state-of-the-art performance for both binary and multi-class classification tasks. The binary BERT classifier exceeded 96% accuracy, indicating strong capability in distinguishing normal from psychologically distressed content. The multi-class model achieved over 80% accuracy across seven mental health categories, outperforming the best classical and deep learning models reported in the base paper.

The superior performance of BERT can be attributed to its ability to model long-range dependencies and contextual semantics, capabilities lacking in traditional machine learning and earlier deep learning architectures. Data augmentation further improved class balance and enhanced robustness, particularly for minority mental-health categories.

While the results are promising, challenges remain in differentiating closely related mental states and addressing slang-heavy or extremely short posts. Future work may integrate multi-modal features (e.g., posting frequency, user behavior), domain-specific pretraining, or model interpretability techniques to provide more transparent and clinically aligned predictions.

Overall, this research confirms that transformer-based models are highly effective for social-media-based mental health detection and provide a strong foundation for building real-time early warning and intervention systems.

7. DATA AVAILABILITY

This study utilized the Sentiment Analysis for Mental Health dataset, available on Kaggle. Link: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data.

8. CODE AVAILABILITY

The programming code of this manuscript is hosted on GitHub. Link: https://github.com/AsadAwaisPK/Adv-AI-Project

9. REFERENCES

[1] A. Tom *et al.*, "Detecting Mental Health Disorders Using NLP," in *2023 IEEE Pune Section International Conference (PuneCon)*, Pune, India: IEEE, Dec. 2023, pp. 1–7. doi: 10.1109/PuneCon58714.2023.10450025.

[2] U. Singla, N. Sharma, and S. S. Khurshid, "Real-Time Sentiment Analysis for Monitoring Mental Health from Social Media Posts," in *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*, Indore, India: IEEE, Aug. 2025, pp. 1–5. doi: 10.1109/WorldSUAS66815.2025.11199232.

[3] A. Raj Shekhar and T. Gupta, "Sentimental Analysis of Social Media Data for Mental Health Monitoring," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, Belagavi, Karnataka, India: IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/INCOFT60753.2023.10425161.

[4] O. Matías-Cristóbal, J. Padilla-Caballero, R. Gonzales-Rivera, R. Benavente-Ayquipa, S. Pérez-Saavedra, and F. Cardenas-Palomino, "Enhancing Sentiment Analysis In Text Of Social Media Texts Using Hybrid Deep Learning Model And Natural Language Processing," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India: IEEE, Sept. 2023, pp. 1776–1780. doi: 10.1109/IC3I59117.2023.10397710.

[5] V. Kokane, A. Abhyankar, N. Shrirao, and P. Khadkikar, "Predicting Mental illness (Depression) with the help of NLP Transformers," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*, Hassan, India: IEEE, May 2024, pp. 1–5. doi: 10.1109/ICDSIS61070.2024.10594036.

[6] Z. Ding *et al.*, "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Sci Rep*, vol. 15, no. 1, p. 14497, Apr. 2025, doi: 10.1038/s41598-025-99167-6.

[7] P. Bhat, A. Anuse, R. Kute, R. S. Bhadade, and P. Purnaye, "Mental Health Analyzer for Depression Detection Based on Textual Analysis," *JAIT*, vol. 13, no. 1, 2022, doi: 10.12720/jait.13.1.67-77.

[8] S. Zhou and M. Mohd, "Mental Health Safety and Depression Detection in Social Media Text Data: A Classification Approach Based on a Deep Learning Model," *IEEE Access*, vol. 13, pp. 63284–63297, 2025, doi: 10.1109/ACCESS.2025.3559170.

[9].https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data