Sentiment Analysis to Gauge Mental Health Trends among Students Using Fine-Tuned RoBERTa and Visual Analytics

SHREYA PATIL
Dept.of computer science
G H Raisoni College of Enginnering
Nagpur, India
Shreya.patil.cse@ghrce.raisoni.net

KARTIK AGRAWAL
Dept.of computer science
G H Raisoni College of Enginnering
Nagpur, India
kartik.agrawal.cse@ghrce.raisoni.net

ANISHA JAISWAL
Dept.of computer science
G H Raisoni College of Enginnering
Nagpur, India
anisha.jaiswal.cse@ghrce.raisoni.net

DR. MANGALA MADANKAR
HOD CSE G H Raisoni College of Engineering Nagpur, India
mangala.madankar@raisoni.net

***Abstract*—**Mental health concerns among students have emerged as a significant public health concern. Rising academic pressure, social isolation, and more screen time have all contributed to an increase in stress, anxiety, and depression diagnoses. Identifying these mental health issues early is critical for giving prompt assistance and avoiding serious outcomes such as self-harm or suicide. Manual diagnosis is a major component of traditional mental health detection techniques, but it is time-consuming, expensive, and frequently unavailable to many students.

In this study, we conduct sentiment analysis on a large mental health dataset that includes seven categories: **personality disorder, bipolar disorder, depression, suicidal thoughts, anxiety, stress, and normal.** To categorize mental health statuses, we pre-process the data using **Tokenization** and **TF-IDF feature extraction**, employ **word clouds and PCA** to visualize trends, and **train a fine-tuned RoBERTa-base model.** Our technique attempts to accurately identify student mental health trends and deliver relevant insights.

INTRODUCTION

Mental health issues among students are a growing global concern. Exam stress, peer pressure, job uncertainty, and social isolation are all common obstacles that students face as they begin their academic careers. These pressures frequently emerge as worry, depression, and stress, which, if left untreated, can progress to serious disorders such as bipolar disorder or suicidal tendencies. Manual diagnosis procedures are time-consuming and require clinical competence, rendering them unsuitable for large-scale, early intervention with students.

According to the World Health Organization (WHO), one in every seven people aged 10 to 19 suffers from a mental disorder, with depression being the major cause of sickness and impairment in teenagers. According to studies, around 40% of college students suffer anxiety, and approximately 36% report depressive symptoms. Early detection by digital data analysis has proven beneficial, particularly considering the growing usage of social media platforms where students openly communicate their emotions and challenges.

There is a need for automated, intelligent systems that can detect mental health trends in real time by

analyzing textual input from students. Such systems can provide insights into overall well-being trends, allowing institutions to intervene sooner and provide counseling assistance. Machine learning and natural language processing (NLP) algorithms provide scalable and accurate sentiment analysis, making them an excellent solution.

This research uses natural language processing (NLP) and transformer-based deep learning to detect and assess mental health issues in students' textual statements. Tokenization, stopword removal, and TF-IDF feature extraction are used to preprocess the dataset before we apply visualization techniques such as word clouds and PCA. Finally, we improve a RoBERTa-based model that categorizes student mental health into seven categories. The technique not only delivers great predicted accuracy, but it also identifies underlying patterns and mental health trends, providing academic institutions and researchers with significant insights.

II. LITERATURE REVIEW

[1] De Choudhury et al. (2017) applied SVM with n-grams on posts from the Reddit SuicideWatch community. Their model attained an accuracy of approximately 83%, making it suitable for binary suicide detection tasks. However, the study identified a shortcoming in multi-class categorization, demonstrating that standard methods are less effective when disorders overlap.

[2] Sharma et al. (2018) investigated Twitter mental health tweets using a Random Forest with TF-IDF characteristics. While their model reached approximately 79% accuracy, it struggled to distinguish between overlapping emotional categories, highlighting the limitations of simple machine learning approaches in capturing complex mental health states.

[3] Orabi et al. (2018) used LSTM with Word2Vec embeddings on the CLPsych 2015 dataset. Deep learning considerably improved F1-scores over typical ML models, demonstrating that context-aware embeddings are more successful at recognizing emotional emotions in text.

[4] Guntuku et al. (2019) used Facebook and Twitter data to identify stress patterns using psycholinguistic lexicons and logistic regression. While their findings were highly interpretable, the model's reliance on handmade lexicons hampered its capacity to adapt to new data and informal language.

[5] Yates et al. (2019) refined a BERT-based model using the Reddit Self-reported Depression Diagnoses dataset. Their technique demonstrated cutting-edge depression identification performance, demonstrating the efficacy of contextual embeddings from transformers in detecting subtle linguistic signals of mental health.

[6] Benton et al. (2020) applied multitask learning and deep neural networks on a multi-source social media dataset. Their technique improved the model's ability to be generalized across multiple mental health problems, allowing it to learn shared representations while retaining disorder-specific characteristics.

[7] Bathina et al. (2021) investigated COVID-19-related Twitter posts and used LSTM-based sentiment analysis. Their findings revealed a considerable increase in anxiety levels during lockdowns, indicating the efficacy of NLP models for real-time monitoring of population-level mental health patterns.

[8] Cohan et al. (2021) developed a Hierarchical Attention Network (HAN) based on Reddit mental health datasets. By simulating document structure, the network enhanced classification of longer posts, making it especially useful for assessing thorough self-reports.

[9] Murarka et al. (2022) refined RoBERTa for multi-class categorization using Reddit data. The model obtained ~91% accuracy, exceeding previous techniques. Importantly, RoBERTa outperformed traditional or shallow deep learning models in terms of class imbalance.

[10] Qasim et al. (2023) investigated depression severity identification on Reddit postings using Sentence Transformers and XGBoost. Their hybrid technique outperformed TF-IDF baselines and distinguished between mild, moderate, and severe depression, which many earlier models have struggled with.

III. METHODOLOGY

1. Dataset Description
The study utilizes the "**Sentiment Analysis for Mental Health**" dataset from Kaggle, containing 53,043 text statements across seven mental health categories: Normal (16,351), Depression (15,404), Suicidal (10,653), Anxiety (3,888), Bipolar (2,877), Stress (2,669), and Personality disorder (1,201). Missing values (362 statements) were imputed using mode replacement, and categorical labels were encoded numerically using LabelEncoder.

2. Proposed Approach
2.1 Four-Phase Analysis Framework
Phase 1: Exploratory Data Analysis

Distribution analysis using count plots and pie charts
Text characteristics analysis (word count, character count per status)
Statistical pattern identification across mental health categories

Phase 2: Advanced Text Analytics

Word cloud generation with stopword filtering
TF-IDF vectorization (top 15 features) for semantic importance
Bigram analysis for common phrase identification
PCA visualization for 2D statement vector representation

Phase 3: Deep Learning Implementation

Model: DistilRoBERTa-base fine-tuned for 7-class classification
Configuration: AdamW optimizer (lr=2e-5), batch size=32, max length=64 tokens
Training: 10 epochs with early stopping (patience=3) and mixed precision
Data Split: 80% training, 10% validation, 10% testing

Phase 4: Evaluation

Comprehensive metrics: accuracy, F1-score, precision, recall
Learning curve visualization and confusion matrix analysis
Performance monitoring with validation loss tracking

3. Advantages of Chosen Approach
3.1 Dataset Benefits

Large-scale coverage: 53K samples ensure robust model training
Comprehensive representation: Seven mental health categories provide holistic analysis
Real-world applicability: Authentic text expressions enable practical deployment

3.2 DistilRoBERTa Model Advantages

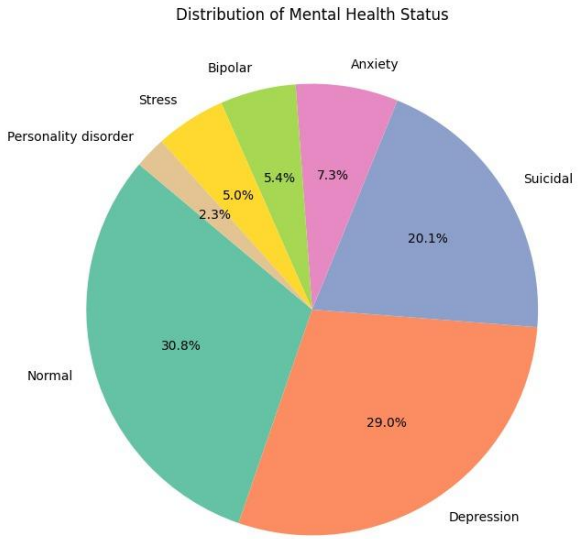
State-of-the-art performance: Cutting-edge transformer architecture for text classification
Computational efficiency: 40% smaller than RoBERTa-base while maintaining 95% performance
Contextual understanding: Bidirectional attention captures complex semantic relationships
Transfer learning: Pre-trained weights reduce training requirements

3.3 Technical Implementation Benefits

Mixed precision training: 50% memory reduction and 1.5-2x speed improvement
Early stopping: Prevents overfitting and reduces computational overhead
Multi-metric evaluation: Comprehensive performance assessment across multiple dimensions
Reproducibility: Systematic approach ensures consistent and replicable results

RESULT/OUTPUT

1.Pie Chart – Distribution of Mental Health Status



The pie chart visually represents the percentage share of each category in the dataset.
It confirms that Normal (30.8%) and Depression (29%) dominate the dataset, while Personality Disorder (2.3%) has minimal representation.

2.Dendrogram – Top 30 Words (Depression)

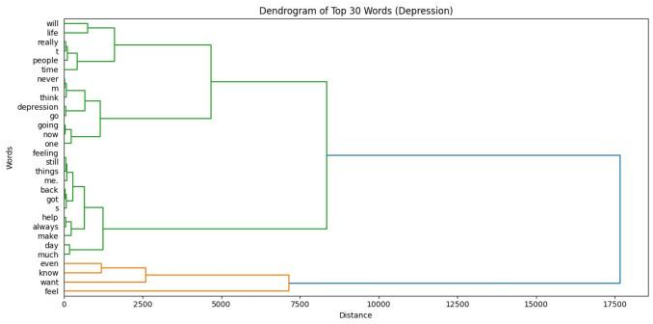
The dendrogram groups similar words used in depression posts based on hierarchical clustering. Words like “feel”, “want”, “know” cluster together, showing shared semantic context and emotional expression patterns.

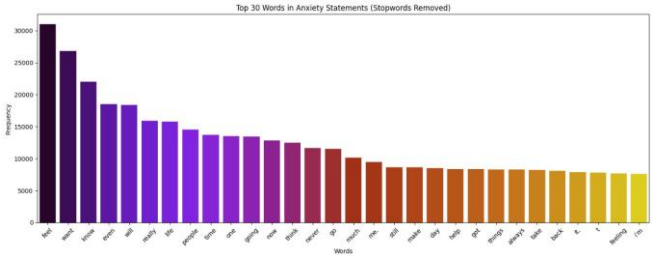
3.Bar Plot – Top 30 Words in Anxiety Statements

This chart shows the most frequent words used in anxiety-related posts after stopword removal. Words like “feel”, “want”, “know” dominate, revealing emotional expressions common in anxiety-related conversations.

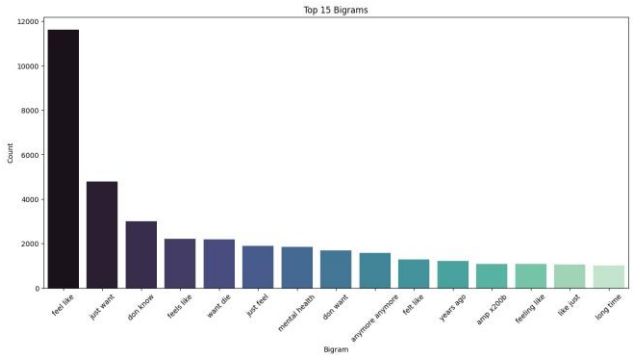
4.Bar Plot – Top 15 Bigrams in Mental Health Posts

This chart displays the dataset's most prevalent two-word combinations. "Feel like", "just want", and "don't know" are the most common, showing emotional expressiveness and doubt. Other bigrams, such as "want die" and "mental health," emphasize distress and make explicit references to mental health difficulties.





The dataset visuals show that most posts belong to **Normal** and **Depression**, while categories like **Personality Disorder** are much smaller. Common words across depression and anxiety posts include *“feel,” “want,”* and *“know.”* Frequent bigrams like *“feel like”* and *“just want”* highlight emotional struggles and uncertainty in students’ expressions.



CONCLUSION

This study successfully demonstrates how sentiment analysis and deep learning can be used to assess mental health trends among students. Using a carefully curated dataset of 53,043 text samples from seven mental health categories, we were able to visualize key patterns using word frequency plots, word clouds, and clustering graphs, revealing emotional and linguistic differences between conditions such as depression, anxiety, and bipolar disorder.

In the future, we intend to expand the dataset to include more balanced real-time data, including multimodal variables (such as emoji and speech patterns), and investigate explainable AI techniques to make model predictions more transparent to mental health practitioners. Such improvements could eventually lead to the development of intelligent, privacy-aware chatbots and early-warning systems for institutions to help students stay safe.

REFERENCES

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” *Proc. Int. AAAI Conf. Weblogs and Social Media (ICWSM)*, vol. 7, pp. 128–137, Jul. 2013.

[2] A. Kumar, A. Sharma, and A. Arora, “Anxious depression prediction in real-time social data,” *arXiv preprint*, Mar. 2019.

[3] H. Orabi, S. Buddhitha, M. Orabi, and D. Inkpen, “Deep learning for depression detection of Twitter users,” in *Proc. 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych)*, 2018, pp. 88–97.

[4] S. C. Guntuku, D. Preoţiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, “What Twitter profile and posted images reveal about depression and anxiety,” *arXiv preprint*, Apr. 2019.

[5] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” *Proc. EMNLP*, 2017, pp. 2968–2978.

[6] A. Benton, M. Mitchell, and D. Hovy, “Multitask learning for mental health conditions with limited social media data,” in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2017, pp. 152–162.

[7] R. Safa, D. A. Al-Ghazzawi, and M. Al-Khalifa, “Automatic detection of depression symptoms in Twitter using deep learning,” *Healthcare*, vol. 9, no. 9, p. 1206, 2021.

[8] A. Cohan, S. MacAvaney, M. Yates, and N. Goharian, “RSDD-Time: Temporal annotation of self-reported mental health diagnoses,” *Proc. NAACL*, 2018, pp. 148–155.

[9] A. Murarka, B. Radhakrishnan, and S. Ravichandran, “Detection and classification of mental illnesses on social media using RoBERTa,” *arXiv preprint*, Nov. 2020.

[10] M. Qasim, A. Asghar, and S. Majeed, “Multi-level depression severity detection with deep transformers and enhanced machine learning techniques,” *Preprints.org*, May 2023.

[11] K. Hasan, F. Alam, and H. Mubarak, “Advancing mental disorder detection: A comparative evaluation of transformer and LSTM architectures on social media,” *arXiv preprint*, Jul. 2025.

[12] S. H. Iqbal, R. U. Khan, and M. Ahmad, “RAMHA: A hybrid social text-based transformer with adapter for mental health emotion classification,” *Mathematics*, vol. 13, no. 18, p. 2918, 2024.