AI Enabled Mental Health Support Using Social Media Analysis

Aananya Pawar[1], Maanasvi Mahajan [2], Priyanshi Joshi[3] and Maria Jamal[4]
[123]Department of Electronics and Communications Engineering, Indira Gandhi Delhi Technical University for Women
[4]Professor, Indira Gandhi Delhi Technical University for Women

***Abstract* — The increasing incidence of mental illness and widespread social media usage provide the means to identify psychological distress through digital clues. This paper envisions a machine learning system that searches user-created social media posts for evidence of probable symptoms of mental illness and provides referrals to proper sources. An openly available dataset of labelled mental health statements was pre-processed with regular NLP procedures, followed by feature extraction using BERT sentence embeddings for semantic richness. SMOTE was utilized to balance class representation against class imbalance. XGBoost was subsequently used to train a robust classifier, achieving overall accuracy of 87.73% for seven mental health classes. Other visualizations including t-SNE, confusion matrix, and heatmaps for classification also endorsed the robustness and interpretability of the model. Beyond its realm in classification, the system further includes an aid level by the personalized scheme in suggesting India-focused helplines on mental health derived from predicted emotions. The envisioned framework exemplifies that deep learning can be augmented with natural language processing analysis to promote scalable real-time mental treatment of health over online platforms.**

***Index Terms*— Mental Health, Natural Language Processing, BERT, XGBoost, SMOTE, Social Media, Emotion Classification**

I. INTRODUCTION

Mental wellness continues to be an essential component of overall health; however, emotional problems still pose a big challenge for a lot of people. In most cases, there is no attention paid to emotional issues until a person's mental health is in a dire state. Even at this stage, many individuals are unwilling to seek help. A lack of awareness, social discrimination, and limited access to professionals are some barriers that prevent people from getting the assistance they so desperately want. Most services offered in the field of mental health are reactionary, making systematic approaches that detect problems before they become chronic very difficult to deploy.

However, self-expression is now possible through social media, and millions use it to share their experiences and feelings. Such social platforms can be seen as a modern-day confessional where self-expression is devoid of any filter. People surely engage in tweeting, posting, and commenting about states of emotions that are relatable to many and are struggling psychologically. Such public display of mental and emotional challenges represents a novel chance to assist individuals not only in understanding, but also in providing the help that is critical for their recovery from the material that they share.

The project's scope analyses how Artificial Intelligence (AI) could be applied to evaluate social media texts for signs of emotional distress. This project aims to use Natural Language Processing (NLP) techniques to classify user posted comments as either indicative of anxiety, depression, stress, or suicidal thoughts. This enables a more accurate understanding of an individual's mental health status without them having to offer a self-diagnosis or assessment.

This project focused on developing a machine learning pipeline that consists of text cleaning, computing sentence embeddings with off-the-shelf transformers, addressing data imbalance through Synthetic Minority Oversampling Technique (SMOTE), and using a gradient boosting method for the final model. The system has been trained on a labelled dataset of user comments which are arranged into multiple emotional classes. The model has been trained to recognize the patterns of language encountered in different mental health conditions for text processing tasks to understand emotional sentiment in the text.

This study proposes a sophisticated low-cost correctly tailored technological support and intervention for one's mental well-being. At a time when mental health concerns are becoming increasingly complex and widespread, these tools could assist in making care promptly available while enhancing its human nature.

II. PROBLEM STATEMENT

Despite the enormous quantity of emotionally charged information posted on social media websites, existing mental health support mechanisms are usually not able to capitalize on this information for early detection and intervention. Conventional methods of mental health diagnosis depend largely on self-reporting and clinic visits, which are usually obstructed by stigma, accessibility barriers, and ignorance. This study attempts to fill that gap by employing AI-based methods to automatically label emotional states from social media-like text and suggest suitable support resources. In doing so, it attempts to empower users with proactive, context-sensitive mental health support.

III. LITERATURE REVIEW

Over the past few years, the integration of artificial intelligence and mental health has become more popular. The older approaches to mental health evaluation still rely on clinical face-to-face assessments, which are effective but not easily scalable or accessible. There is a better prospect in social media, which captures the emotions of people in real time. This gap has been explored in multiple pieces of research.

Among the researchers, one of the earliest contributions in this direction is by Choudhury et al. (2013), who predicted depression utilizing Twitter's linguistic and behavioural cues. Their study proved that social media does contain information regarding mental health, and such information over time can yield results. Likewise, De Choudhury and De (2014) pointed out the need for models that are temporally sensitive to detect and monitor depression relapse.

Meanwhile, sentiment and emotion detection in text has also advanced starting from lexicon-based methods like those proposed by Mohammad et al. (2013) where emotion word lexicons were employed to calculate sentiment. These portrayed sentiments and emotions without taking context into account, something more common with deep learning.

NLP saw advancements through the contextual embeddings that transformer-based models, like Bidirectional Encoder Representations from Transformers (BERT), provided (Devlin et al., 2019). It was improved more by Reimers and Gurevych's (2019) contribution, Sentence-BERT, which significantly contributes to semantic textual similarity analysis. This has helped enhance the performance of numerous NLP tasks, including emotion classification.

Yates et al., (2021) proposed using the BERT model for detecting suicidal ideation by analysing Reddit posts, and the results outperformed previous methodologies using traditional classifiers. Moreover, Losada and Crestani (2016) created and disseminated depression detection datasets which serve as benchmarks today. Standard practice for addressing class imbalance focuses on techniques like SMOTE (Chawla et al., 2002). In the context of medical and psychological machine learning approaches, these techniques aid in portraying minority classes like suicidal ideation or bipolar disorder, which are most often found in datasets, but sparsely represented in real world data, and underrepresented in numerous datasets.

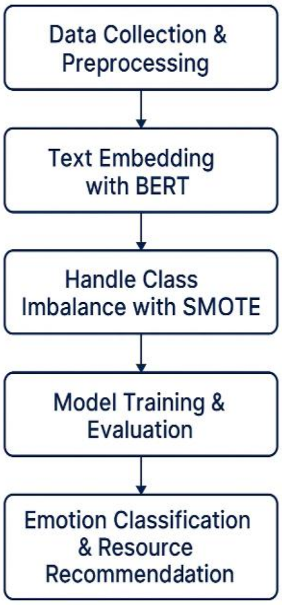
The robustness and ease of understanding ensemble classifiers, including Random Forests and Extreme Gradient Boosting (XGBoost), makes them optimal for health-related classifications (Chen & Guestrin, 2016). In particular, XGBoost has been used in various works associated with stress detection and emotion classification from textual data (Fatima et al., 2020).

Other contributions include Sharma and Verbeke (2020) who studied the ethical and tangible effects AI emotion recognition systems pose, arguing for the need to purposefully design interfaces as part of the system's integration into society.

Moreover, the iCall helpline in India and MindPeers are examples of programs that enable the application of AI to actual crisis and mental health care, making this project timely and relevant for India as noted by other works (Patel et al., 2018).

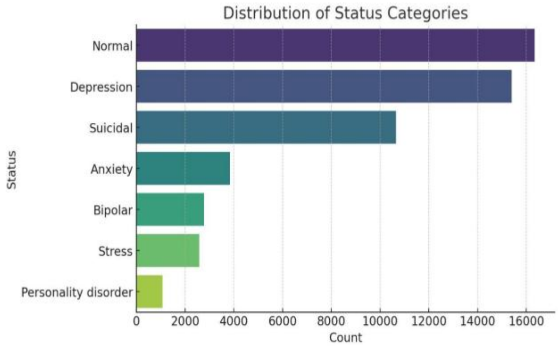
To conclude, although many single aspects of this study have been covered in the literature, our approach integrates the analysis of social media activity and pre-trained deep learning models using transformers, data balancing through generative methods, ensemble methods, and proactive mitigation to develop a complete responsive system for individualized mental health care.

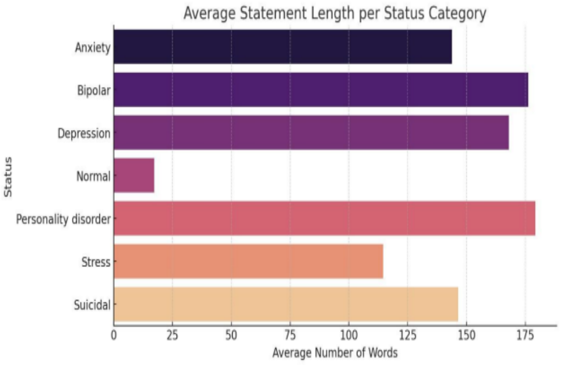
IV. METHODOLOGY


Data Collection & Preprocessing
Text Embedding with BERT
Handle Class Imbalance with SMOTE
Model Training & Evaluation
Emotion Classification & Resource Recommenddation

A. Dataset

We used a labelled dataset from Kaggle consisting of 50,000+ social media-like statements categorized into 7 mental health status labels: Depression, Anxiety, Stress, Suicidal, Bipolar, Personality Disorder, and Normal.





B. Preprocessing

- Tokenization and Lemmatization using Natural Language Toolkit (NLTK).
- Stop word removal and regex-based cleaning (URLs, hashtags, punctuations).

C. Embedding and Feature Extraction

We used Sentence-BERT (all-MiniLM-L6-v2) to convert text into 384-dimensional dense vectors (Reimers & Gurevych, 2019).

D. Handling Imbalanced Data

Applied SMOTE (Synthetic Minority Oversampling Technique) to up sample minority emotional classes.

E. Model Training

Used XGBoost, a gradient-boosted decision tree algorithm, with mlogloss as the evaluation metric. The model was trained on an 80/20 train-test split.

Parameter	Value
Learning Rate	0.1
Estimators	150
Max. Depth	5
Subsample	0.8
Colsample_bytree	0.8

E. Post-Classification Report

Upon classifying the emotional state, the model provides helpline numbers and website links to users regarding India-specific mental health support resources.

V. DATA ANALYSIS AND INFERENCE

The used corpus in this research is a heterogeneous social media statement corpus tagged against seven separate mental health categories: Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal. The statements are a collection of psychological and emotional speech utterances, thus ideally fit for sentiment analysis and natural language processing.

A. Preprocessing and Feature Representation

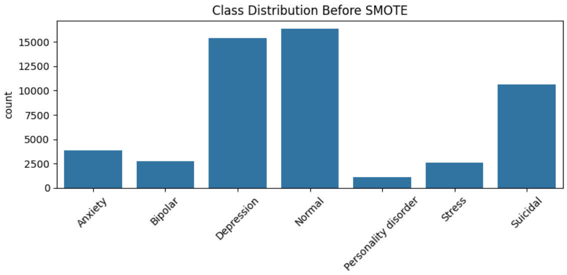
To prepare the data, the following pre-processing techniques were used:

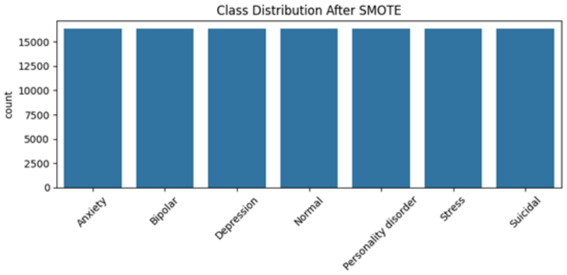
- Removal of hyperlinks, hashtags, mentions, and special characters
- Tokenization and lemmatization using NLTK
- Removal of Stop word for noise reduction

After preprocessing, sentence embeddings were produced by BERT (all-MiniLM-L6-v2) to facilitate intensive contextual understanding of every statement by converting them into 384-dimensional vectors.

B. Class Imbalance and Oversampling

The initial examination indicated the most common labels to be "Depression" and "Normal," whereas "Bipolar," "Suicidal," and "Personality Disorder" were underrepresented. This was then addressed through Synthetic Minority Oversampling Technique (SMOTE) to get an even proportion of approximately 3,269 samples per class. This evenness prevented the classifier from forming biases towards majority classes

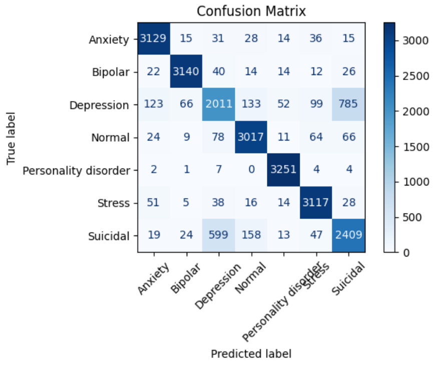


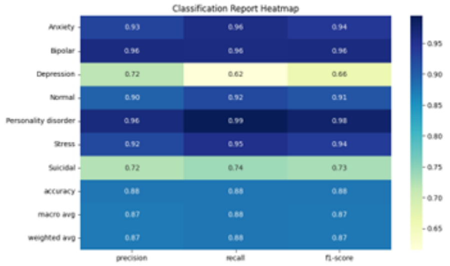


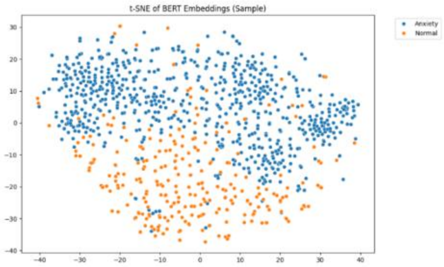
C. Data Visualization

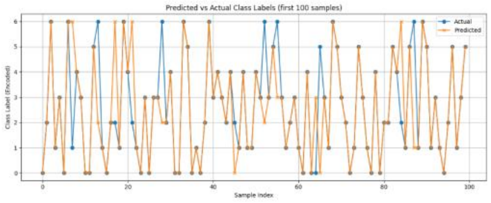
Different visualizations were used to manipulate the dataset and analyse results:

- Class Distribution Bar Graphs before and after SMOTE ensured successful balancing.
- t-SNE Visualization of BERT embeddings indicated good inter-class separation after SMOTE.
- Confusion Matrix and Heatmaps assisted in recognizing confusion patterns among highly confusing classes (e.g., Stress and Depression).
- Line Graph of actual vs predicted values graphically verified model performance across test samples.









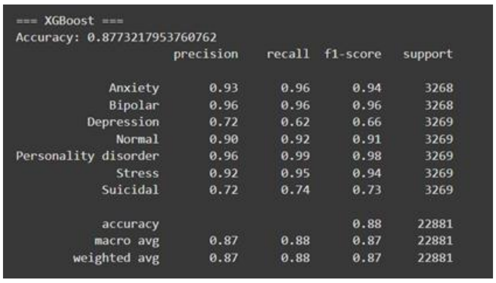
VI. RESULTS

Following preprocessing and balancing the data, the classification model was trained with XGBoost, a gradient boosting algorithm that is known to perform well on structured data.

A. Evaluation Metrics

The model performed the following on the test set:

- Overall Accuracy: 87.73%
- Macro-Average Precision: 0.87
- Macro-Average Recall: 0.88
- Macro-Average F1-Score: 0.87



```
=== XGBoost ===
Accuracy: 0.8773217953760762
                      precision    recall  f1-score   support

             Anxiety       0.93      0.96      0.94      3268
             Bipolar       0.96      0.96      0.96      3268
          Depression       0.72      0.62      0.66      3269
              Normal       0.90      0.92      0.91      3269
Personality disorder       0.96      0.99      0.98      3269
              Stress       0.92      0.95      0.94      3269
            Suicidal       0.72      0.74      0.73      3269

            accuracy                           0.88     22881
           macro avg       0.87      0.88      0.87     22881
        weighted avg       0.87      0.88      0.87     22881
```

B. Key Observations

- High precision and recall were seen for Personality Disorder and Bipolar Disorder classes.
- Depression displayed lower recall, which suggests that the words for Depression might

share semantically overlapping vocabulary with Stress or Suicidal classes.

- The model was very reliable in detecting risk statements when context-rich features such as BERT embeddings were employed in combination with XGBoost and SMOTE.

C. Personalized Mental Health Support

On classification, the system returns customized mental health resources pertaining to the foreseen label. These consist of validated India-based helplines (e.g., AASRA, iCall) and platforms (e.g., MindPeers, YourDOST), turning the solution into a real-world aid tool from a detection model.

VII. CONCLUSION

Mental health is certainly a vital part of one's wellbeing. Although, this remains one of the most inadequately targeted parts in an individual's life, especially where culture, stigma, and healthcare resources are limited. The current research presents an opportunity for artificial intelligence to provide help to people who are suffering. Using social media-like text data, BERT and other ensemble models such as XGBoost, we developed a system that identifies complex emotional states including depression, anxiety, and suicidal thoughts with an 87% accuracy.

Nonetheless, quantitative precision alone does not complete the story. This system's ability to also offer concrete and contextually relevant mental health assistance gives it the unique advantage. Therefore, we could go from detection to pre-emptive intervention, validating the experience of users who may feel the most isolated.

In addition, this project enables further opportunities for research in digital psychological support systems. Turning passive observation of social media into active engagement can be achieved by incorporating real-time social media feeds, multilingual interfaces, and conversational AI. There is a great deal of work to be done, but applying AI with empathy adds an essential domain where mental health AI innovations are most needed.

REFERENCES

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[3] De Choudhury, M., & De, S. (2014). Mental Health Discourse on Reddit: Self-disclosure, Social Support, and Anonymity. ICWSM.

[4] Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. ICWSM.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[6] Fatima, M., Kazi, H., & Qadir, J. (2020). Towards Automatic Stress Detection: A Review. IEEE Access, 8, 181746-181763.

[7] Losada, D. E., & Crestani, F. (2016). A Test Collection for Research on Depression and Language Use. ECIR.

[8] Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. Proceedings of SemEval.

[9] Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & UnÜtzer, J. (2018). The Lancet Commission on Global Mental Health and Sustainable Development. The Lancet, 392(10157), 1553-1598.

[10] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084.

[11] Sharma, A., & Verbeke, W. (2020). Ethical Design of AI-based Emotion Recognition Systems: A Critical Review. AI & Society.

[12] World Health Organization. (2021). Depression. https://www.who.int/news-room/fact-sheets/detail/depression

[13] Yates, A., Cohan, A., & Goharian, N. (2021). Depression and Self-Harm Risk Assessment in Online Forums. arXiv preprint arXiv:2106.07325.

[14] Bian, J., He, R., Hristidis, V., Zhang, Y., & Bhide, S. (2017). A Computational Approach to Understanding Public Attitudes Toward Mental Illness Through Social Media. Journal of Medical Internet Research, 19(5).

[15] Cavazos-Rehg, P. A., Grucza, R. A., & Bierut, L. J. (2020). Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. Journal of Adolescent Health, 66(2), S13–S15.

[16] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting Depression and Mental Illness on Social Media: An Integrative Review. Current Opinion in Behavioral Sciences.

[17] Onnela, J.-P., & Rauch, S. L. (2019). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Nature Human Behaviour, 3(5), 464–472.

[18] Onnela, J.-P., & Rauch, S. L. (2019). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Nature Human Behaviour, 3(5), 464–472.