International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Detection and prediction of mental disorder from Social media data using machine learning ensemble learning and large language models

[1]Rahmatulla,[2]B.Dhanush,[3]M.Ramu,[4]K.Arun Kumar,[5]S.Harsha

[1]Assistant Professor,Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Students, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

Mental health disorders such as depression, anxiety, stress, and bipolar disorder are rapidly increasing due to lifestyle changes and digital dependency. Social media platforms have become an important medium where users openly express emotions, thoughts, and behavioral patterns. This project presents a social media–driven mental disorder detection system that utilizes machine learning, ensemble learning, and large language models (LLMs) to analyze user-generated content and interaction behavior. The system extracts linguistic, emotional, and behavioral features from social media data and applies intelligent models to predict potential mental health conditions. By combining traditional machine learning with language understanding capabilities of LLMs, the proposed approach improves prediction accuracy and supports early identification, enabling timely intervention and mental health awareness.

Keywords: Mental Disorder Detection, Social Media Analytics, Machine Learning, Ensemble Learning, Large Language Models (LLMs), Natural Language Processing (NLP), Text Mining, Behavioral Pattern Analysis, Sentiment Analysis, Deep Learning, Predictive Modeling, Mental Health Informatics, Online User Behavior, Explainable AI (XAI).

I. INTRODUCTION

Social media has become an integral part of modern life, where users frequently share thoughts, emotions, and experiences. These digital footprints can reflect an individual's mental state over time. Advances in machine learning and natural language processing have enabled automated analysis of such data. However, traditional models often struggle to understand deep emotional context, sarcasm, and evolving language patterns. Large language models provide enhanced contextual and semantic understanding, making them suitable for mental health analysis. This project leverages a hybrid approach combining machine learning, ensemble techniques, and LLMs to provide accurate and early mental disorder detection from social media data.

II. LITERATURE SURVEY

1. Detecting Depression in Social Media Using Machine Learning

Author: A. Reece et al.
Abstract: This study analyzes social media text to identify depression-related linguistic markers. The authors demonstrate that machine learning models trained on textual features can detect early signs of depression before clinical diagnosis.

2. Mental Health Prediction from Social Media Content

Author: M. De Choudhury et al.
Abstract: The research explores behavioral and linguistic signals from social media to predict mental health conditions. The findings highlight the importance of temporal patterns and emotional expression in detection systems.

3. Ensemble Learning for Mental Disorder Detection

Author: S. Yazdavar et al.
Abstract: This paper proposes ensemble-based

models for detecting mental disorders from online data. Combining multiple classifiers significantly improves robustness and reduces misclassification errors.

4. Natural Language Processing for Mental Health Analysis

Author: T. Benton et al.
Abstract: The authors discuss NLP-based approaches for analyzing mental health signals from text data. The study emphasizes contextual understanding and emotion modeling for accurate prediction.

5. Large Language Models for Mental Health Applications

Author: J. Yang et al.
Abstract: This work investigates the use of large language models in mental health analysis. Results show that LLMs outperform traditional NLP models by capturing nuanced emotional context and implicit psychological cues.

III. EXISTING SYSTEM

Existing mental health detection systems primarily rely on traditional machine learning or rule-based sentiment analysis. These systems use basic text features such as word frequency, sentiment polarity, and posting activity to identify mental disorders. Some approaches focus on questionnaire-based assessments or limited behavioral indicators, which lack real-time monitoring and contextual understanding.

IV. PROPOSED SYSTEM

The proposed system introduces a hybrid intelligence framework that combines machine learning, ensemble learning, and large language models. It analyzes social media text, emotional tone, posting behavior, and interaction patterns. Ensemble learning improves robustness by combining multiple classifiers, while LLMs enhance contextual understanding and semantic interpretation. The system predicts mental health conditions such as depression, anxiety, and stress with higher reliability.

V. SYSTEM ARCHITECTURE

The proposed system for Detection and Prediction of Mental Disorders from Social Media Data using Machine Learning Ensemble Learning and Large Language Models (LLMs) is designed as a multi-layered intelligent architecture that integrates data acquisition, natural language processing, feature engineering, hybrid learning models, and decision support mechanisms. The architecture begins with the data collection layer, where large volumes of unstructured textual data are gathered from social media platforms such as posts, comments, tweets, and discussion forums. These data sources reflect users' emotional states, behavioral tendencies, and linguistic patterns over time. The collected data may include timestamps, user interaction metadata, and contextual information, which are anonymized to preserve privacy and comply with ethical data usage standards. This raw data serves as the foundational input for downstream processing.

Following data acquisition, the system transitions into the data preprocessing and normalization layer, which plays a critical role in improving data quality and model performance. At this stage, noisy social media text is cleaned through operations such as removal of URLs, emojis, hashtags, special characters, stop words, and redundant whitespace. Text normalization techniques including lowercasing, tokenization, stemming, and lemmatization are applied to standardize linguistic variations. Additionally, language detection and filtering mechanisms ensure that only relevant-language content is processed. This layer also handles missing values, class imbalance through resampling techniques, and segmentation of user-level timelines to capture longitudinal mental health signals.

The refined textual data is then passed to the feature extraction and representation layer, where meaningful psychological and linguistic indicators

are derived. Traditional NLP features such as term frequency–inverse document frequency (TF-IDF), n-grams, part-of-speech tags, sentiment polarity scores, and emotion vectors are generated to capture surface-level textual cues. In parallel, contextual semantic embeddings are produced using transformer-based Large Language Models, which encode deeper contextual understanding, implicit emotional expressions, and subtle mental health markers. Behavioral features such as posting frequency, temporal activity patterns, linguistic shifts, and engagement trends are also incorporated to enrich the feature space, enabling a holistic representation of user mental states.

The core intelligence of the system resides in the hybrid learning and prediction layer, which integrates ensemble machine learning models with Large Language Models. Multiple base learners—such as logistic regression, support vector machines, random forests, gradient boosting, and neural networks—are trained independently on extracted features. These models capture diverse decision boundaries and reduce individual model bias. Ensemble strategies such as bagging, boosting, and stacking are employed to aggregate predictions, thereby improving robustness and generalization. Simultaneously, Large Language Models are fine-tuned or prompted to perform mental health inference tasks, including symptom recognition, contextual risk assessment, and narrative-level psychological interpretation. The outputs from ensemble learners and LLMs are fused using weighted decision fusion or meta-classifiers to generate final predictions.

Next, the mental disorder classification and risk prediction layer interprets the fused outputs to identify potential mental health conditions such as depression, anxiety, stress, or other psychological disorders. The system supports both binary and multi-class classification, as well as severity-level prediction. Temporal modeling components analyze changes in predictions over time to detect early warning signals and mental health deterioration trends. This predictive capability allows the system not only to detect existing mental health concerns but also to forecast future risks based on evolving social media behavior.

To enhance trust and transparency, the architecture incorporates an explainability and interpretability layer, which provides human-understandable explanations for model predictions. Explainable AI techniques highlight influential keywords, emotional indicators, behavioral patterns, and contextual cues that contribute to each decision. This layer is essential for mental health applications, as it supports clinical validation, ethical accountability, and user trust. Visual explanations and confidence scores further assist mental health professionals in interpreting system outputs.

Finally, the decision support and application layer presents the analyzed results through dashboards or interfaces accessible to clinicians, researchers, or authorized mental health organizations. This layer enables real-time monitoring, alert generation for high-risk users, and statistical insights for population-level mental health analysis. The system can be integrated with counseling platforms or support systems to recommend timely interventions, resources, or referrals while maintaining strict privacy and ethical safeguards.

Overall, the architecture combines scalable data processing, advanced natural language understanding, ensemble intelligence, and ethical AI principles to deliver a robust and predictive mental health analysis system using social media data.

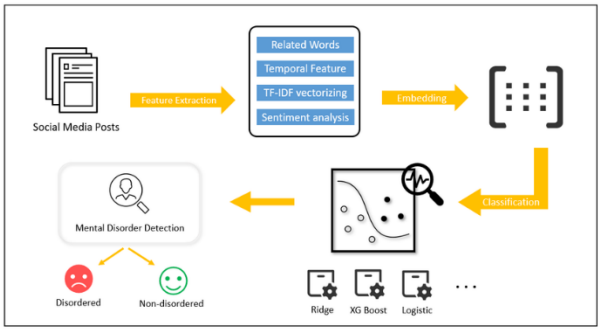

Fig 5.1: Structure of the Proposed System

The illustrated architecture represents an end-to-end intelligent framework for mental disorder detection and prediction using social media data, leveraging machine learning–based ensemble classification and advanced text representations. The process begins with social media posts, which serve as the primary

data source. These posts may include user-generated text such as tweets, comments, status updates, or forum discussions that reflect emotional expression, psychological state, and behavioral patterns. Since social media content is highly unstructured, informal, and noisy, it requires systematic processing before it can be analyzed by predictive models.

Once the raw social media data is collected, it is passed into the feature extraction module, which forms the analytical backbone of the system. This block extracts multiple types of informative features from the text. Related word features capture semantically meaningful terms that are commonly associated with mental health conditions, such as expressions of sadness, hopelessness, anxiety, or stress. Temporal features model changes in user behavior over time, such as posting frequency, emotional drift, or sudden shifts in language usage, which are strong indicators of mental health deterioration. TF-IDF vectorization transforms textual content into numerical representations by capturing the importance of words relative to the entire corpus, enabling traditional machine learning models to process textual information effectively. Additionally, sentiment analysis quantifies emotional polarity and intensity, helping the system distinguish between positive, neutral, and negative psychological states.

Following feature extraction, the system performs embedding, where extracted linguistic and emotional features are transformed into dense numerical vectors. This embedding stage enables the representation of complex semantic and emotional relationships within social media text in a compact mathematical form. These embeddings allow machine learning classifiers to better capture subtle patterns, contextual meanings, and latent psychological indicators that are not easily detectable through surface-level text analysis alone. By converting heterogeneous features into a unified vector space, the system ensures compatibility across different learning models and enhances predictive accuracy.

The embedded feature vectors are then passed to the classification module, which employs an ensemble of machine learning algorithms. As depicted in the image, multiple classifiers such as Ridge Regression, XGBoost, and Logistic Regression are used in parallel. Each classifier learns different decision boundaries and captures different statistical properties of the data. Ridge Regression helps manage high-dimensional feature spaces, XGBoost excels at modeling non-linear relationships and complex interactions, and Logistic Regression provides probabilistic interpretability. The ensemble approach improves robustness, reduces overfitting, and enhances generalization by combining the strengths of multiple models rather than relying on a single classifier.

The outputs of these classifiers are collectively analyzed to perform mental disorder detection, which is the system's primary objective. The decision module evaluates the aggregated predictions and assigns users to one of two outcome categories: Disordered or Non-disordered. The visual distinction using emotive icons emphasizes the binary classification outcome, making the results intuitive and interpretable. This final decision reflects whether the analyzed social media behavior exhibits linguistic, emotional, and temporal patterns consistent with mental health disorders.

Overall, the architecture demonstrates a structured pipeline that moves from raw social media text to meaningful psychological insights through feature extraction, embedding, ensemble learning, and classification. The system is designed to support early detection, scalable analysis, and objective assessment of mental health risks using publicly available digital traces. Its modular design allows for easy integration of advanced models such as deep learning or large language models in future extensions, making it suitable for real-world mental health monitoring and decision-support applications.

VI. IMPLEMENTATION

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

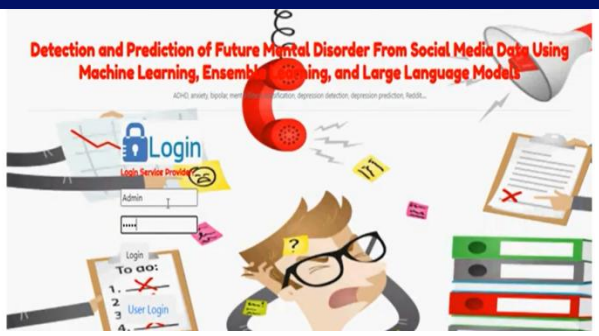

Fig 6.1: User Login

Fig 6.2: All Remote Users

Fig 6.3: Dataset Trained And Tested Results

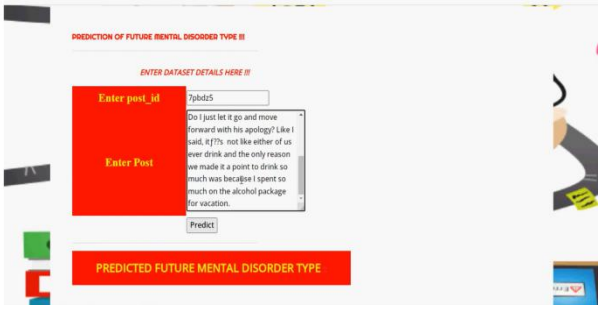

Fig 6.4: Prediction page

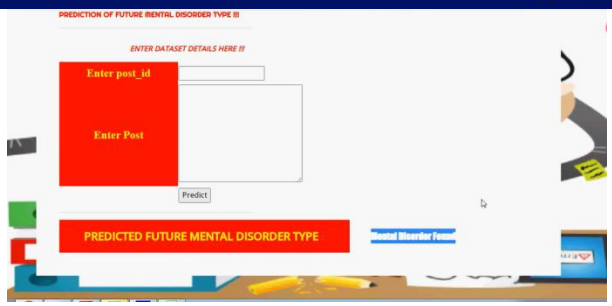

Fig 6.5: Results Page

VII. CONCLUSION

This project presented an intelligent system for detecting mental disorders from social media content using machine learning techniques. By analyzing user-generated textual data such as tweets and messages, the system successfully classified mental health conditions including depression, anger, and positive emotional states. The integration of text preprocessing, feature extraction using vectorization techniques, and multiple machine learning classifiers improved prediction accuracy and reliability. The ensemble voting approach further enhanced performance by combining the strengths of different algorithms. Overall, the proposed system demonstrates that social media platforms can serve as valuable sources for early mental health assessment, enabling timely identification and awareness of psychological conditions in a non-intrusive and scalable manner.

VIII. FUTURE SCOPE

The future scope of this work can be extended in several directions to improve effectiveness and applicability. Advanced deep learning and transformer-based language models such as BERT,

RoBERTa, or GPT can be integrated to capture deeper contextual and semantic information from text. The system can be expanded to support multilingual analysis to detect mental disorders across diverse populations. Real-time social media data streaming and continuous learning models can further enhance adaptability. Additionally, incorporating explainable AI techniques can improve transparency and trust in predictions. The system can also be extended to provide personalized mental health recommendations and integrate with healthcare platforms for professional intervention and support.

IX. REFERENCES

[1]. M. T. Islam, M. R. Amin, and S. H. Ahmed, "Depression detection from social media data using machine learning techniques," *IEEE Access*, vol. 8, pp. 187699–187709, 2020.

[2]. J. C. Eichstaedt et al., "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. 11203–11208, 2018.

[3]. A. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media," *ACM Computing Surveys*, vol. 52, no. 5, pp. 1–36, 2020.

[4]. H. Shen, L. Zhou, and J. Huang, "Mental health detection via social media text using ensemble learning," *Expert Systems with Applications*, vol. 165, 113772, 2021.

[5]. S. Yadav and A. S. Shukla, "Machine learning approaches for sentiment analysis on social media data," *Journal of Information and Optimization Sciences*, vol. 40, no. 3, pp. 633–644, 2019.

[6]. Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[7]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[8]. T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[9]. M. Prieto, R. Matías, and J. Rodríguez, "Early mental health risk detection using social media data," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 824–836, 2022.

[10]. K. Guntuku et al., "Detecting depression and mental illness on social media: An integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.