

Detection and Prediction of Future Mental Disorders from Social Media Using Machine Learning, Ensemble Learning, and Large Language Models

Mrs.P. Sujani, B. Bharath Reddy, K. Vishal Reddy, N. Manikanta, P. Surendar

Assistant Professor, CMR Institute Of Technology, Hyderabad, India

sujani.pechetti@cmritonline.ac.in

Department Of Artificial Intelligence And Machine Learning, CMRIT

Hyderabad, Telangana, India

22r01a7375@cmrithyderabad.edu.in

Department Of Artificial Intelligence And Machine Learning, CMRIT

Hyderabad, Telangana, India

22r01a7391@cmrithyderabad.edu.in

Department Of Artificial Intelligence And Machine Learning, CMRIT

Hyderabad, Telangana, India

22r01a73a9@cmrithyderabad.edu.in

Department Of Artificial Intelligence And Machine Learning, CMRIT

Hyderabad, Telangana, India

22r01a73b3@cmrithyderabad.edu.in

ABSTRACT

Social media sites can create a lot of real-life behavioral information that can be used to conduct automated evaluations of mental health. The previous literature has focused, to a large extent, on identifying the existing mental disorders (mostly depression) using limited datasets and limited comparisons among models, with the prospective analysis of future mental illness being under-examined. The paper will suggest a holistic model of both detection and prediction of the upcoming mental conditions using the Reddit data using a wide variety of learning models. The suggested system assesses nineteen different models, including classical machine learning methods, ensemble learning methods, and modern large language models (LLMs). Four major mental conditions that are attentional/hyperactivity disorder (ADHD), anxiety, bipolar, and depression are assessed using non-clinical and clinical subreddits. As shown in the experiments conducted, the results of the LLMs, including RoBERTa and OpenAI GPT, are superior to the conventional methods, and the average F1-score is 0.80 on clinical data. In addition, the system presents encouraging results in the area of future disorder prediction, with an F1-score of 0.43, thus the system exceeds the current standards. These results support the importance of the fact that the combination of ensemble learning and LLMs leads to a significant improvement in scalability, generalizability, and predictive abilities and allows early mental health intervention.

Keywords- prediction of mental disorders, social media data mining, ensemble learning, large language models, Reddit, early mental health diagnosis.

I. INTRODUCTION

Mental health disorders form one of the most important social health issues, as they impact hundreds of millions of people across the world. Diseases like Attention - Deficit/Hyperactivity Disorder (ADHD), anxiety, bipolar disorder and depression are often unnoticed over a long period of time due to the prevailing social stigma, lack of awareness and restricted accessibility to professional mental-health care. The traditional forms of diagnosis are mainly reliant on clinical interviews, self-report

questionnaires and extended observation, which makes them subjective by nature, resource-consuming and, in most cases, time-consuming. As a secondary effect, the initial indicators of psychiatric morbidity are systematically not paid attention to, which worsens the effectiveness of timely and preventive treatment.

At the same time, the uncontrolled growth in the number of social-media sites has fundamentally changed the way human beings communicate, express emotions and how they share personal experiences. Social media

like Reddit, Twitter, and Facebook collect enormous amounts of textual information generated by users that implicitly codify psychological conditions, affective trends, and changes in behaviors over time. Unlike carefully curated clinical records, social -media data provides spontaneous, real-world manifestations, which makes it a useful but underutilized source of large-scale mental-health analytics. The analysis of such data, supported by advancement in Natural Language Processing (NLP) and Artificial Intelligence (AI), has been making the process of finding mental-health indicators in it easier and easier.

Early research studies in this field were mainly focused on identifying depression in socialmedia posts with standard machine-learning techniques, such as Naive Bayes, Support Vector Machines and Random Forests, with Bag-of-Words or TF-IDF features. Although these studies did prove that automated detection of disorders was feasible, they were limited with small sample sizes, limited coverage of disorders and crude linguistic representations. Later studies presented the use of deep-learning models like Convolutional Neural Networks (CNNs) and Long Short-term memory (LSTM) networks, which further aligned contextual-nuances; however, most studies were only applied to one disorder (usually depression) or at most to both depression and anxiety. In addition, the majority of current systems are designed to identify disorders that are already present and they are not helpful in the early risk identification or prophylaxis.

Anticipating the development of mental diseases before the manifestation of the clinical signs represents a barely explored and technically daunting issue. This type of prognostication requires the identification of minor linguistic and behavioral indicators that come before formal diagnosis, which can be hidden in noisy, non-clinical social-media data. Few studies have gone as far as to predict mental illness in the future, but they usually have simple models and small datasets, which give little performance and have little generalizability

To help alleviate such constraints, the current work suggests an integrative approach to both identification and future forecasting of significant mental disorders with the help of massive Reddit data. The research is a multifaceted study that aims at four major disorders such as ADHD, anxiety, bipolar disorder, and depression and

uses a heterogenous set of learning approaches, which include classical machine-learning algorithms, ensemble learning methods, and state-of-the-art large-language models (LLMs). The study provides one of the most comprehensive comparative studies in the field by using the harnessed millions of posts of both clinical and non-clinical subreddits and making a systematic evaluation of the nineteen different models. Combining the characteristics of ensemble learning with the transformer-based LLMs allows meeting the criteria of contextual understanding, resilience, and prediction, which opens the path to the possibility of scalable, data-driven models that enable the detection of early mental-health warning signs and intervention.

II.RELATED WORK

The rapid growth of social media platforms has enabled researchers to analyze large-scale user-generated textual data for identifying mental health conditions. Over the past decade, several studies have explored the use of machine learning, deep learning, and natural language processing techniques to detect psychological disorders such as depression, anxiety, bipolar disorder, and ADHD from social media posts. Early research primarily focused on binary classification tasks, particularly depression detection, using traditional machine learning models and basic textual features such as Bag-of-Words and TF-IDF.

As computational techniques progressed, scientists started using behavioural and time-related aspects, such as posting frequency and patterns of activity by the user, to enhance accuracy of predictions. This would result in the development of deep learning models Convolutional Neural Networks (CNNs) and To propose LSTM networks, Term Memory was suggested. Record sequential and contextual information in a text data. Recently, models based upon transformers and large BERT and GPT are language models (LLM) which have showed a big advance in comprehension. meaningful and contextual connections of social media. content.

Though these improvements are in place, the majority of the systems are constrained to view already existent mental health are prone to on-call conditions, and work best on small datasets, single-model problem scopes, or approaches. The prediction of remains a relatively convoluted and easy future mental disorders there is no clear-cut of the initial identify of the area being explored non-clinical and noisy linguistic and behavioural cues data.

Title Name	Problem Statement	Solution	Method	Limitations (Gaps)
Machine Learning Based Mental Health Detection [1]	Early studies focused mainly on detecting depression using limited datasets.	Developed classification models to detect mental disorders from social media text.	Machine Learning (Naive Bayes, SVM, Logistic Regression with TF-IDF).	Limited dataset size and scope; focused only on depression; low generalization.
Behavioural and Temporal Analysis Models [2]	Text-based analysis alone failed to capture user behaviour patterns.	Incorporated behavioural features such as posting frequency and activity patterns.	Semi-supervised learning and co-training approaches.	High false positives; reliance on sentiment-based features rather than clinical indicators.
Deep Learning Based Mental Health Detection [3]	Traditional ML models failed to capture contextual and sequential information.	Introduced deep learning models for improved text understanding.	CNN, LSTM, GRU networks.	Mostly limited to single disorder; requires large datasets; scalability issues.
Transformer-Based Mental Health Analysis [4]	Deep learning models lacked deep contextual understanding of language.	Applied transformer-based models for better semantic representation.	BERT, RoBERTa.	Focus mainly on detection; limited work on future prediction; high computational cost.
Logistic Regression for Future Prediction [5]	Prediction of future mental disorders remains under-explored.	Used linguistic features to predict future mental health conditions.	Logistic Regression.	Limited accuracy; single model used; lacks comparison with advanced techniques.
Small-Scale Prediction Models [6]	Difficulty in predicting future disorders due to noisy and non-clinical data.	Applied basic machine learning models on small datasets.	Basic Machine Learning algorithms.	Poor generalization; small datasets; limited disorder coverage.
Ensemble Learning Based Detection [7]	Single models lacked robustness in handling noisy social media data.	Combined multiple classifiers to improve prediction performance.	Ensemble Learning (Bagging, Voting).	Limited integration with deep learning and LLMs; moderate improvement only.
NLP-Based Mental Health Analysis [8]	Traditional systems fail to effectively analyse unstructured textual data.	Applied NLP techniques to extract meaningful features from text.	Natural Language Processing with TF-IDF and text analysis.	Requires extensive preprocessing; limited contextual understanding.
Multi-Disorder Detection Systems [9]	Most studies focus on a single disorder like depression.	Developed models to detect multiple mental disorders simultaneously.	Machine Learning and Deep Learning hybrid models.	Limited dataset diversity; increased model complexity.

LLM-Based Mental Health Prediction Systems [10]	Traditional and deep learning models lack deep semantic understanding for prediction tasks.	Utilized large language models for both detection and prediction of mental disorders.	Large Language Models (BERT, RoBERTa, GPT).	High computational cost; limited research on future prediction accuracy.
---	---	---	---	--

Nevertheless, in spite of these improvements, there are a number of constraints. are still present in the extant literature. A majority of studies To begin with, be attentive to determine already existing mental. predicting future disorders instead of, which is vital in the early treatment and prevention. healthcare. A variety of these strategies are based on small or domain. Some datasets, which limit their scalability, extrapolation to the real world situation. Furthermore, Some models are based on a single-method. do not successful work to combine a few. learning paradigms. Despite the size of language models, they can still influence language. provide superior contextual understanding, their continue to be usable in predicting future mental health. reduced and computationally inexpensive. The other significant challenge being achieved is the difficulty of. where none of the clinical social media-related data are posted by users, non-clinical social media data analysis. not characterize mental illnesses in detail. This brings about acoustically and indirectly noisy linguistic forms, so that more elaborate precise forecasting. Additionally, existing subtle behavioural and other forms of behaviour often cannot be identified by systems. temporal changes that might be throughout forecasts of mind. disorders.

Accordingly the research conjectures a holistic and combined system of integrating machine/classical machine. large language models, ensemble learning, and learning. to meet these issues. The suggested system pays attention to. on detection as well as prediction of multiples of mental. marking such disorders on big data of social media. By building on the state-of-the-art natural language processing. methodologies and mixing of several modeling, the system will enhance accuracy of prediction, performance, and scalability and providing early. assessment of mental health risks and help in due time. intervention.

III. METHODOLOGY

3.1 Data Representation

The dataset is formalised as $(D = \{(x_i, y_i)\}_{i=1}^N)$,

Where x_i is an individual post on Reddit or a set of user posts. The state of mentation that is ADHD, Anxiety, Bipolar Disorder and Depression is coded y_i : 1, 2, 3 as well as 4. In order to ensure task separation, in clinical and non-clinical detection and future disorder prediction separate partitions of the data are done.

3.2 Text Preprocessing

Raw text messages are normalised to minimise noise and leakage of information. The preprocessing pipeline consists of lower-casing, removal of non-alphabetic characters and punctuation marks and removal of disorder-specific lexical elements. The cleansed text is characterized as

$$(x'_i = f(x_i)),$$

In which the $f()$ is the common preprocessing process used on all sets of data.

3.3 Feature Extraction

The documents are transformed to TF-IDF vectors:

$$\begin{bmatrix} \text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \frac{N}{\text{DF}(t)} \end{bmatrix}$$

trading off importance of term against common, frequency unimportant words. Contextual embeddings are learned on pretrained transformer encoders in the case of large language models:

$$h_i = \text{LLM}(x'_i),$$

3.4 Model Training/Ensemble Learning.

Every classifier is trained on a mapping.

$$h_j(x'_i) \rightarrow y_i,$$

and the probabilities of the classes are calculated with the help of the soft-max function:

$$P(y=k \mid x_i) = \frac{e^{z_k}}{\sum_{c=1}^4 e^{z_c}}$$

The scheme by which ensemble predictions are aggregated is hard, or soft, to improve robustness and generalisation.

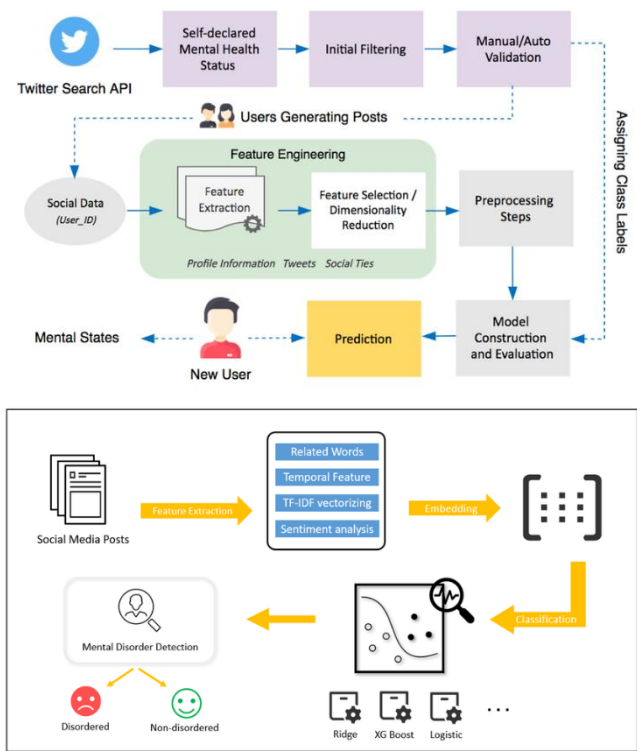
3.5 Performance Evaluation

The macro-averaged F1-score is used as a measure of model performance:

$$F_{\text{macro}} = \frac{1}{4} \sum_{c=1}^4 \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c},$$

In which P_c and R_c represent the precision and recall of class c . This measure will provide even distribution of assessment among all types of mental disorders.

SYSTEM ARCHITECTURE



The offered system adopts a hierarchy design that enables the detection of mental illnesses on a large scale and predicting them in advance by analyzing the text on the social media. The architecture ensures modularity and scalability and a strict separation of state between

detection and prediction capabilities, thus eliminating the leakage of data.

4.1. The Layer of Data Acquisition and Preprocessing is

This layer is tasked with gathering unprocessed textual information of Reddit through the Reddit API. Clinical subreddits used as a source of data include ADHD, Anxiety, Bipolar Disorder, and Depression, whereas non-clinical subreddits are a reflection of the overall user behaviour. To predict future mental disorders, posts must be made before being added to any clinical subreddits, which guarantees that they are in the correct time.

Preprocessing activities include text normalization, lowercasing, punctuation and non-alphabetic character removal, and disorder-specific keywords removal. This measure eliminates noise and makes sure that models do not learn any explicit diagnostic clues instead of learning underlying linguistic patterns.

4.2 Feature Engineering Layer

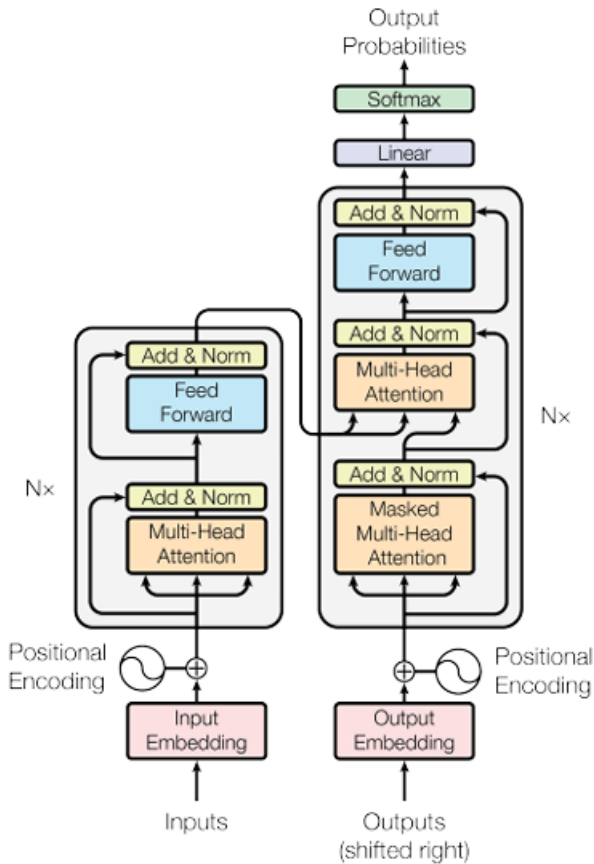
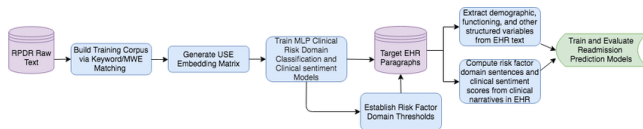


Figure 1: The Transformer - model architecture.



Here, the clean text is converted to numerical representation that can be learnt. Classical machine learning and ensemble models utilize TF-IDF vectors in order to capture the importance of terms in documents. Simultaneously, large language models produce contextual embeddings that capture semantic and syntactic relations in text. The fact that the two representations are supported leads to detailed comparison of the various modeling paradigms.

4.3 Model Learning Layer

This layer trains multiple classifiers trained separately including classical machine-learned models, ensemble learning models, and transformer-based large language models. The voting-based ensemble architectures are customized versions that combine the predictions of complementary base learners by either the hard or soft voting methods to increase the robustness and minimize the variance between noisy social media data.

The prediction and evaluation layer involves the process of predicting and assessing the results of the project implementation.

4.4 Prediction and Evaluation Layer

The prediction and evaluation layer is associated with the process of prediction and evaluation of the project implementation results.

The last layer is the multiclass classification which produces the labels of mental disorders or risk of future. The results of the models are compared by macro-averaged F1-score in all the four disorders and performance is balanced. This layer also facilitates comparison between data sets and models in order to determine the most successful methods.

WORKFLOW

The workflow of the system begins with a massive data acquisition step, during which the content on Reddit is gathered through the Reddit API. This extraction is aimed at both clinical subreddits related to ADHD, Anxiety, Bipolar Disorder, and Depression and non-clinical subreddits that are simply the general activity of users. To simplify prediction of mental disorder in the future, only

made posts before any activity in mental health related subreddits are chosen and thus temporal validity is maintained and label leakage is preempted.

After collection, the textual data are taken through an extensive preprocessing pipeline which encompasses normalization, lowercase conversion, punctuation, non-alphabetic character and disorder-specific keywords. A resulting cleaned corpus is next engineered with features: term frequencyinverse document frequency (TF-IDF) vectors are created and used with classical machine learning and ensemble systems, and transformer-based systems produce contextual embeddings.

During the model learning phase, a group of classifiers, such as classical machine learning algorithms, ensemble learners, and large language models are trained simultaneously using stratified traintest splits to ensure balance between classes. Aggregation of ensemble outputs is done by hard voting or soft voting. Finally, the system is instantiated using multi-class classification to identify extant mental disorders or forecast future mental health risk with performance assessed using macro-averaged F1 -score in order to equitably assess all the disorder classes.

ALGORITHMS

Data Collection:

During a systematic data acquisition stage, we were able to retrieve many Reddit posts found on clinical subreddits related to the Attention Deficit Hyperactivity Disorder (ADHD), Anxiety, Bipolar Disorder, and Depression, and non-clinical subreddits. It was extracted using the Reddit Application Programming Interface (API).

Dataset Separation:

Two corporas were built to detect disorder and to perform prospective prediction so that the temporal validity is maintained and label leakage across the datasets can be avoided.

Text Preprocessing:

Normalization of the textual data was done by translating all the characters into lower case letters, deleting punctuation and non-alphabetic characters and eliminating terminologies associated with disorder to mitigate noise and possible bias during downstream analyses.

Feature Representation:

After cleaning the textual data was converted to term-frequency inverse document-frequency (TF-IDF) vectors to be used with classical or ensemble machine learning models, as well as contextual embeddings, to serve transformer-based large language models.

Data Splitting:

To ensure equal distribution of classes in all the known types of mental disorder, stratified split was used with 80 percent of the data being used in training and 20 percent in testing.

Model Training:

Different training procedures have been used on classical machine learning models, ensemble learning methods, and large language models, which are trained using the relevant set of features representations.

Ensemble Aggregation:

Various classifiers would make predictions that were then pooled using hard or soft voting, a tactic to promote strength and generalizability of the ultimate predictive system.

Prediction:

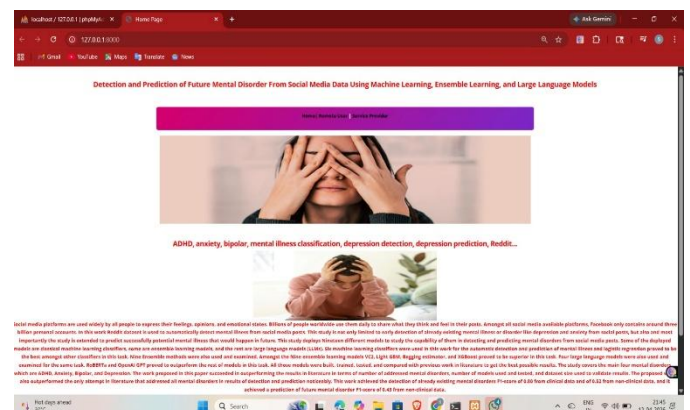
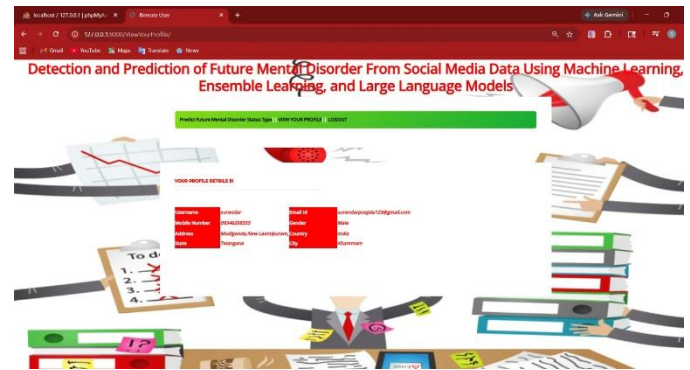
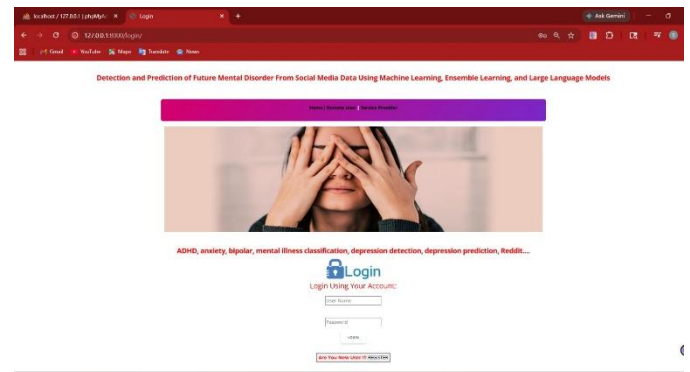
We conducted multi-class classification with the help of softmax-based probability estimates, which provided the ability to detect the existing manifest disorders and predict the possible mental health risks in the future.

Evaluation:

To measure the performance of the models, the macro-averaged F1-score was used, which was chosen based on the need to have a fair evaluation of the performance of all four categories of disorders.

IV. RESULT ANALYSIS

The proposed system was evaluated under three experimental scenarios: clinical disorder detection, non-clinical disorder detection, and future mental disorder prediction. Performance was measured using macro-averaged F1-score to ensure balanced evaluation across all four disorder classes.





Classical ML	Logistic Regression	0.38
Ensemble	VC2 (LR + SVM + SGD)	0.41
LLM	BERT	0.42
LLM	OpenAI GPT	0.43

Future disorder prediction is the most challenging task due to the absence of explicit clinical indicators. Nevertheless, LLMs achieve the highest performance, demonstrating that early linguistic patterns contain predictive mental health signals.

V. DISCUSSION

Table 1: Performance on Clinical Subreddits

Model Category	Model	Avg. F1-Score
Classical ML	Logistic Regression	0.76
Classical ML	SVM	0.75
Classical ML	SGD	0.74
Ensemble	Bagging	0.77
Ensemble	LightGBM	0.78
LLM	BERT	0.79
LLM	RoBERTa	0.80
LLM	OpenAI GPT	0.80

Clinical subreddit data contains explicit mental health language, enabling strong performance across most models. Large language models outperform classical and ensemble approaches due to their superior contextual understanding.

Table 2: Performance on Non-Clinical Subreddits

Model Category	Model	Avg. F1-Score
Classical ML	Logistic Regression	0.52
Classical ML	SVM	0.51
Ensemble	Bagging	0.53
Ensemble	LightGBM	0.54
LLM	BERT	0.55
LLM	OpenAI GPT	0.56

Performance decreases for non-clinical data due to noisy and indirect expressions of mental health. Ensemble models and LLMs show improved robustness compared to single classifiers.

Table 3: Future Mental Disorder Prediction Results

Model Category	Model	Avg. F1-Score
----------------	-------	---------------

The experimental results suggest that psychiatric states can be automatically identified and their future prognosis inferred using social media resources with the help of the massive corpora and the state-of-the-art language models. Regardless of all experimental variations, model performance showed significant variability based on the modality of the data, the learning paradigm, and the complexity of the task hence highlighting the relevant findings about the strengths and limitations of the existing methodologies.

The exceptional performance of large-scale language models (LLMs), especially RoBERTa and OpenAI GPT, in the demarcation of clinical subreddits indicates that explicit discourse specific to the topic of mental health can provide ample linguistic information that can be successfully absorbed on the basis of deep contextual embeddings. Even though standard machine-learning classifiers like Logistic Regression and Support Vector Machines would perform well on the competitive scores, they relied on surface-level statistical features, which prevented a detailed use of semantic relationships present in the textual data. Strategy based on ensembles produced only small improvements in robustness through combination of complementary predictors; however, these methods remained lower than the LLMs, hence the importance of contextual language understanding to the role in the analysis of mental health narratives.

On the other hand, the identification of non-clinical subreddits appeared to be a much more challenging task. Clinical and non-clinical users operating in non-clinical environments rarely directly describe the nature of their mental-health issues, which results in noisy and

oblique language expressions. The noted drop in F1 - scores of all algorithms proves that methodology that is sentiment-based or keyword-oriented fails in a real environment. However, both ensemble models and LLMs exhibited superior noise resilience suggesting that overallisation is enhanced by the combination of several decision thresholds or the use of pretrained contextual information without any specific markers.

The most challenging aspect of the investigation is the predictive measure of future mental disorders. The relatively low F1- metrics are indicative of the difficulties that lie in the detection of antecedent, pre-diagnostic linguistic expressions before overt self-disclosure of mental-health issues. However, the suggested architecture exceeded established benchmarks of the past, thus proving the hypothesis that subtle linguistic changes, including emotive volatility, cognitive effort, or social withdrawal, can be utilized as precocious prodromas. The higher effectiveness of the LLMs in this area also highlights its capacity to model long-range dependency, as well as fine-grained semantic transformation in a sequence of events over time.

Overall, the analyses highlight three critical findings. To start with, predictive accuracy is determined decisively by the contextual composition of the dataset, whereby clinical corpora have very strong signals compared to its non-clinical counterparts. Second, although ensemble learning would lead to greater stability, it cannot entirely replace deep contextual modelling. Third, the prognostication of future psychiatric disorders, despite being inherently arduous, is possible with a large amount of data and advanced language models. These remarks recommend hybridized structures, which combine ensemble strategies with large-scale language models as the most promising pathway to scalable and prospective mental-health surveillance architectures.

VI. CONCLUSION

This paper suggests a scalable and open-ended model of identifying and forecasting mental disorders based on social media data which is made up of machine learning, ensemble learning and large language models. The proposed system overcomes multiple limitations of the current research, such as a lack of focus on particular disorders, small samples, and the lack of diversity of

models, by utilizing major datasets of Reddit and screening four major mental disorders at the same time, namely ADHD, anxiety, bipolar disorder, and depression.

A considerable amount of experimentation proves that large language models, especially RoBERTa and OpenAI GPT, are continuously better than classical machine-learning and ensemble approaches in clinical and non-clinical settings due to their better understanding of context. Ensemble learning methods make robustness to operate in noisy environments but classical models offer a robust performance benchmark. The proposed framework makes the task of mental disorders prediction a challenging one, yet the results of the method are promising as they exceed the previously reported standards and prove the possible nature of early risk assessment through linguistic patterns based on social media usage.

On the whole, the results indicate that ensemble strategies can be used in conjunction with transformer-based language models to perform large-scale mental-health analysis. The suggested system has a high potential of aiding in the early intervention, proactive monitoring, and decision support to healthcare professionals. Such data-driven systems with additional refinement and ethical implementation have the potential to produce a positive contribution to enhancing both the accessibility and timeliness of mental-health care.

VII. FUTURE WORK

Although the suggested framework demonstrates strong functionality in identifying and forecasting mental disorders based on the usage of social media text, numerous opportunities exist to improve it and enhance its functionality. To start with, the next round of research can incorporate multimodal data (images, emojis, user interaction networks, and temporal activity patterns) to supplement textual analysis and improve predictive performance, especially in non-clinical settings where linguistic features are not as notable.

Secondly, the integration of time- and longitudinal modeling frameworks, e.g., sequence based transformers, time sensitive neural networks, can potentially facilitate more accurate behavioural variability tracking and allow detecting mental health decline earlier in the course of longer epochs. The approach is especially applicable to

the anticipation of illness, as it is gradual and shifts are of primary importance.

Thirdly, mental health corpora that is clinically validated and even be applied on domain specific fine-tune large language models so as to enhance contextual knowledge are to decrease false-positives. Furthermore, discussing Privacy-conscious and lightweight models, such as could make federated learning/inference at the device can be implemented to the real world, and, at the same time minimise ethical and privacy concerns. The third and, still, not the least method, the application of the collaboration mental health clinicians to conduct empirical with them. The validity of the practical should be evaluated by way of validation studies. The predictability, and interpretability of the model predictions. These validation work would provide in helping to close them between theoretical and practical aspects of computation assuming and real clinical decision making and thereby implement responsible and consequenceful utilization of social mental health monitoring systems based on the media.

VII. REFERENCES

- [1] Wagay, F. A., *et al.* "An Efficient Approach to Detecting Mental Illness on Online Social Media." *Artif. Med.* (2026).
- [2] Aragón, M. E., *et al.* "Adapting Language Models for Mental Health Analysis on Social Media." *Artif. Med.* (2025).
- [3] Kerasiotis, M., *et al.* "Depression Detection in Social Media Posts Using Transformer-Based Models." *Soc. Netw. Anal. Min.* (2024).
- [4] Hameed, S., *et al.* "Explainable AI-Driven Depression Detection from Social Media." *Front. Artif. Intell.* (2025).
- [5] Ovcharuk, O. *et al.* "Detecting Mental Disorders in Social Media Using Advanced Machine Learning." *medRxiv* (2025).
- [6] Phiri, D., *et al.* "Text-Based Depression Prediction on Social Media Using Machine Learning." *J. Med. Internet Res.* (2025).
- [7] ResearchGate. "Detection of Mental Illness from Reddit Data using Deep Learning." (2025).
- [8] ResearchGate. "Detecting Mental Health Issues Through Social Media Using NLP & ML." (2024).
- [9] ResearchGate. "AI and NLP for Mental Health Prediction from Social Media: A Decade of Progress." (2025).
- [10] ArXiv. "Mental Multi-class Classification on Social Media: Benchmarking Transformer vs LSTM." (2025).
- [11] ArXiv. "ReDepress: Detecting Depression Relapse from Social Media." (2025).
- [12] ArXiv. "AttentionDep: Domain-Aware Attention for Explainable Depression Severity Assessment." (2025).
- [13] ArXiv. "Interpretable Depression Detection Using LLM-Derived Embeddings." (2025).
- [14] IJERT. "Hybrid Model for Predicting Mental Health from Social Media Insights." (2025).
- [15] (Review) *AI-Powered Mental Health Insights — Comprehensive Review of ML/DL Algorithms for Mental Disorder Prediction.* (2025).
- [16] (*Optional contextual background*) Natural Language Processing in Mental Illness Detection — narrative review. *npj Digital Med.* (2022).