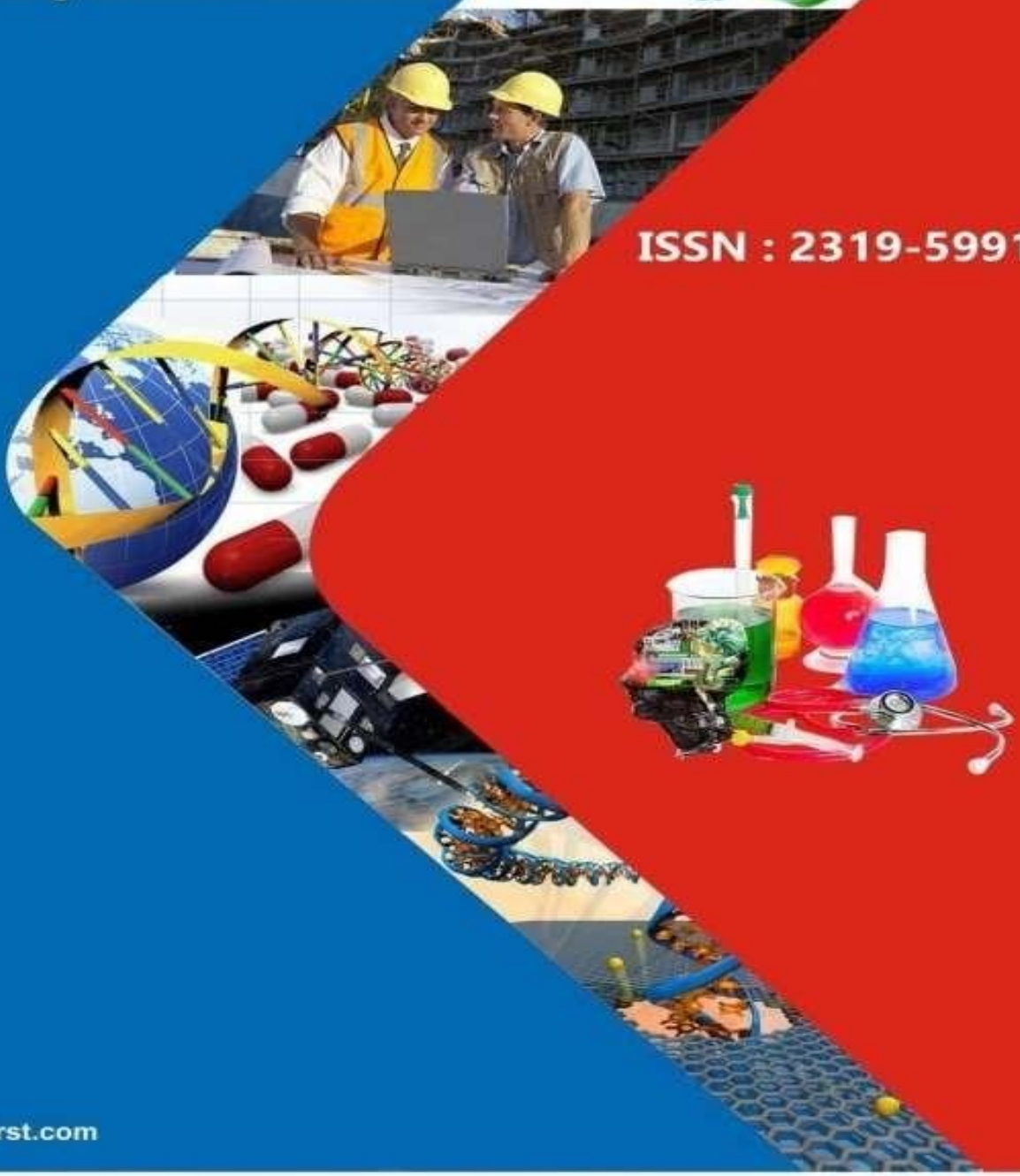
International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231 Vol. 21, Issue 2, 2025

A HYBRID APPROACH FOR PREDICTING MENTAL HEALTH DISORDERS USING SOCIAL MEDIA DATA AND ENSEMBLE LEARNING TECHNIQUES

[1]*Metta Lavamadhusekhar, MCA Student, Department of MCA*

[2]*B. Harish Kumar Reddy, M.Tech, (Ph.D), Assistant Professor,Department of MCA*

[12]*Dr KV SubbaReddy Institute of Technology,Dupadu, Kurnool*

ABSTRACT:

Data on people's mental health has increased exponentially as a result of the growing usage of social media platforms, offering important new information for the diagnosis of mental illnesses. In order to identify possible mental health problems using social media data, this research investigates the use of Machine Learning (ML) methods, including Random Forest and Decision Tree algorithms. The technology looks at user-shared textual information to determine if a user may be dealing with a mental health issue based on their interactions and postings. Natural language processing (NLP) methods including tokenisation, stopword removal, and vectorisation utilising the Term Frequency-Inverse Document Frequency (TF-IDF) approach are used in this work to preprocess the textual data. Two popular classification models, Random Forest and Decision Tree, are then trained using the preprocessed data. While the Decision Tree algorithm creates a tree-like model to generate predictions based on feature values, the Random Forest algorithm, an ensemble learning technique, uses numerous decision trees to increase prediction accuracy and resilience. The classification accuracy of the trained models is assessed; the Random Forest model is anticipated to provide superior generalisation by minimising overfitting in contrast to the Decision Tree model. In order to ascertain if a user's postings suggest the possibility of mental health conditions like stress, anxiety, or depression, both models are evaluated using social media data. The findings of this research are intended to aid in the creation of automated instruments for early mental health identification, which may facilitate prompt assistance and intervention. To sum up, the use of machine learning algorithms like Random Forest and Decision Tree provide a viable method for identifying mental health conditions using social media data, demonstrating the promise of artificial intelligence in the medical field. This may help organisations and mental health experts provide people in need early support.

I. INTRODUCTION:

The recognition of mental health illnesses as global health issues is growing at an alarming rate. Some examples of these diseases are stress, anxiety, and depression. Because of the proliferation of social media platforms such as Facebook, Twitter, and Instagram, users routinely communicate their personal ideas, experiences, and feelings. This results in the accumulation of a vast amount of data that has the potential to provide insights into the mental well-being of the individuals. It is possible to identify early warning indicators of mental health illnesses by analysing these postings on social media using methods such as machine learning and natural language processing (NLP). In order to make accurate predictions about mental health issues based on text data obtained from social media platforms, this research intends to make use of sophisticated machine learning models such as Random Forest and Decision Tree. The purpose of this project is to create a system that is capable of automatically

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231 Vol. 21, Issue 2, 2025

identifying people who may be at danger, which will allow for prompt interventions and assistance.

II. LITERATURE SURVEY:

Title: Using Natural Language Processing to Predict Depression and Mental Illness Based on Twitter Posts

J. L. McCrae, E. D. Muller, L. L. W. von der Putten, and others are the authors of this work.

The year is 2020.

For the purpose of this research, Natural Language Processing (NLP) methods were used to analyse messages on Twitter in order to identify indicators of depression and other potential mental health problems. The authors were able to reach a high level of prediction accuracy by using machine learning methods such as support vector machines (SVM) and random forests to categorise messages. According to the findings of the research, social media platforms have the potential to serve as a helpful tool for monitoring mental health.

A Method Based on Machine Learning for the Identification of Mental Health Conditions from Content Found on Social Media

K. M. Gupta, R. P. Singh, and S. G. Dutta served as the authors.

2019, the year

The purpose of this article was to present a method that makes use of machine learning in order to predict mentally healthy individuals based on their status updates on Facebook. When it came to predicting depression and anxiety, the authors used a number of different feature extraction approaches and classifiers, such as Decision Trees and Naive Bayes, and they were successful in doing so.

Analysis of Textual Data for the Detection of Depression Through the Application of Machine Learning Techniques

X. Zhang, L. Zhao, and X. Chen are the authors.

The year is 2021.

In this work, a number of different machine learning algorithms were investigated in order to identify sadness based on textual data collected from internet forums. A comparison was made between Random Forest and Decision Trees, and the results showed that Random Forest had a higher level of accuracy when it came to categorising sad language. The research provided evidence that text mining has the potential to be used for mental health information analysis.

Twitter-based mental health prediction using ensemble learning is the title of this article.

S. Sharma, P. M. Meena, and K. N. B. Dinesh are the authors of this work.

This research used an ensemble learning model that included different classifiers, such as Random Forest and XGBoost, in order to make predictions about mental health issues based on tweets. The year of the study was 2022. When the authors compared the accuracy and robustness of predictions made by conventional models to those made using ensemble learning approaches, they discovered that the former received a considerable improvement.

The Early Detection of Mental Health Issues Through the Utilisation of Machine Learning and Data from Social Media Resources

To whom it may concern: R. V. Banerjee, S. K. Patel, and A. Sharma

The year is 2023.

To summarise: The purpose of this study was to present a system that uses machine learning algorithms on data from social media platforms to identify mental health problems at an earlier stage. This study investigated the use of supervised and unsupervised learning approaches, such as Random Forest and Decision Trees, for the purpose of categorising material relevant to mental health that was accessible via internet postings. The purpose of the research was to enhance the accuracy of mental health detection models via the

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231

integration of natural language processing and sentiment analysis.

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM:

When it comes to diagnosing mental diseases, the current systems in mental health detection often depend on manual surveys or clinical examinations. Both of these methods are time-consuming and may be costly. There have been some recent research that have investigated the use of text mining methods to analyse data from social media platforms for the purpose of detecting mental health issues; however, the majority of these studies have focused on individual models or tiny datasets. The categorisation of these systems has been accomplished by the use of machine learning methods such as Support Vector Machines (SVM), Naive Bayes, and logistic regression. However, when dealing with enormous amounts of data that are unstructured and noisy from social media, these techniques may not be the most effective ones. This might result in difficulties in making accurate predictions using these methods. In this particular application, the use of ensemble learning techniques, such as Random Forest, which aggregate the results of numerous models in order to create predictions, has not been well investigated. Furthermore, the majority of systems have difficulties in identifying the complex and context-specific nature of mental health concerns that are present in online material.

Disadvantages of Existing System:

1. Low Accuracy with Simple Models: Conventional machine learning models, such as Naive Bayes and SVM, often fail to identify complex patterns in huge datasets, which results in decreased prediction accuracy. This is particularly true in the case of noisy and unstructured social media data.
2. Inability to grasp Contextual remarks: The current systems often lack the capability to grasp the context of remarks that are published on social media. This includes the ability to recognise sarcasm, humour, or emotional tone, all of which are essential in properly discovering mental health illnesses.
3. Problems Regarding Data Privacy The usage of personal data from social media platforms poses ethical and privacy problems, including issues of permission, data security, and the possibility of sensitive information being misused.

PROPOSED SYSTEM:

Utilising the power of ensemble learning methods like as Random Forest and Decision Tree, the proposed system aims to improve the accuracy and efficacy of mental health condition diagnosis from social media postings. This will be accomplished by utilising the power of these approaches. Using Natural Language Processing (NLP) methods to clean, preprocess, and vectorise the text data, the system will train two different models to categorise postings as suggestive of possible mental health disorders. These patterns will be used to identify potential mental health problems. Random Forest is an ensemble learning method that aggregates the results of numerous decision trees in order to increase generalisation. Decision Trees, on the other hand, provide a decision-making process that is visible and simple to comprehend. For the purpose of determining whether or not the system is capable of accurately identifying mental health problems, it will be tested on datasets representative of real-world social media. Furthermore, the system will be constructed with privacy in mind, making certain that only the required data that has been

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231

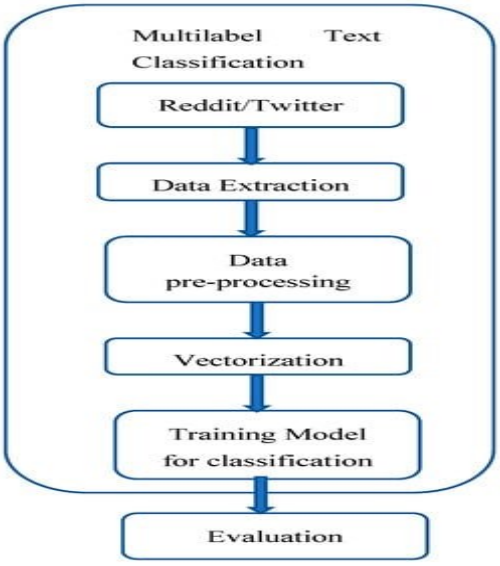
Vol. 21, Issue 2, 2025

anonymised will be used for the purposes of training and prediction.

Advantages of Proposed System:

1. Improved Accuracy via Ensemble Learning: The system is anticipated to provide improved prediction accuracy and less overfitting in comparison to conventional single-model techniques. This is accomplished through the use of Random Forest, which mixes numerous decision trees.
2. Detection of Context-Awareness: The system will make use of modern natural language processing methods in order to handle text from social media platforms in a more efficient manner. This will allow it to capture contextual subtleties such as tone, mood, and implicit meaning, which are essential for the identification of mental health issues.
3. Scalable and Efficient: The system is capable of being scaled to handle big datasets from many social media platforms. It also provides real-time or near-real-time prediction capabilities, which may assist in the implementation of proactive mental health intervention.

System Architecture



IV. Modules Description:

1. Data Collection Module

- **Purpose**: This module is responsible for collecting social media data, such as tweets, posts, comments, or other forms of user-generated content that might indicate mental health issues. Data can be collected through APIs (e.g., Twitter API, Reddit API) or web scraping tools.
- **Methods/Tools**:
  - Twitter API for tweets.
  - Reddit API for posts and comments.
  - Scrapy or BeautifulSoup for scraping web data.
  - Regular Expressions for cleaning and pre-processing raw data.

2. Data Preprocessing and Text Cleaning Module

- **Purpose**: This module handles the cleaning, preprocessing, and transformation of raw text data into a format suitable for machine learning models. This includes removing unnecessary characters, handling missing data, and normalizing text.
- **Methods/Tools**:
  - **Text Cleaning**: Removal of URLs, punctuation, special characters, and converting all text to lowercase.
  - **Tokenization**: Breaking down the text into smaller chunks (tokens), such as words or phrases.
  - **Stopwords Removal**: Filtering out common words (like "the", "a", "is", etc.) that don't provide significant meaning.
  - **TF-IDF Vectorization**: Transforming text data into a matrix of TF-IDF (Term Frequency-Inverse Document Frequency) features for machine learning models.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231

- **Libraries**: NLTK, SpaCy, Regex, sklearn's TfidfVectorizer.

3. Feature Engineering and Selection Module

- **Purpose**: This module extracts features (or relevant information) from the processed text data and selects the most informative features for model training.
- **Methods/Tools**:
  - **Sentiment Analysis**: Using pre-trained models like VADER or transformers (BERT, RoBERTa) to classify sentiment (positive, negative, neutral).
  - **Word Embeddings**: Use pre-trained embeddings (Word2Vec, GloVe, FastText) to capture semantic meaning.
  - **Custom Features**: Create features such as the frequency of specific keywords, text length, or use of specific mental health-related terms.
  - **Feature Selection**: Techniques like Mutual Information or Recursive Feature Elimination (RFE) to select the best features for training.

4. Model Training Module

- **Purpose**: This module is where the machine learning models are trained on the prepared data to detect patterns related to mental disorders. Models can include traditional machine learning models like Random Forest and Decision Tree, and Ensemble Learning models.
- **Methods/Tools**:
  - **Random Forest Classifier**: An ensemble method using multiple decision trees to provide a robust prediction.
  - **Decision Tree Classifier**: A tree-based model for classification based on feature splitting.
  - **Ensemble Learning Techniques**: Techniques like Gradient Boosting (e.g., XGBoost) and Random Forest.
  - **Training Process**: Splitting the dataset into training and testing sets, fitting the models on the training set, and evaluating using cross-validation or testing data.
  - **Libraries**: scikit-learn, XGBoost, LightGBM.

5. Large Language Model Integration Module

- **Purpose**: This module integrates a large pre-trained language model (such as GPT, BERT, or other transformer-based models) to assist in understanding the deeper context of text and make better predictions related to mental disorders.
- **Methods/Tools**:
  - **BERT-based Models**: For advanced understanding and sentiment analysis of the text.
  - **Fine-tuning on Mental Health Data**: Fine-tuning pre-trained transformer models on domain-specific mental health data to improve performance.
  - **Transformer Libraries**: Hugging Face Transformers for model integration.

6. Prediction and Detection Module

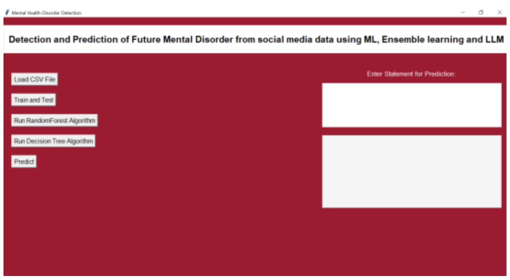
- **Purpose**: After training, this module makes predictions on new, unseen social media data. It can classify a statement as indicative of a potential mental health issue or not.
- **Methods/Tools**:
  - **Model Inference**: Using the trained models (Random Forest, Decision Tree, or fine-tuned large language models) to predict whether a new post or statement suggests a mental disorder.
  - **Risk Scoring**: Assigning a probability or risk score based on the likelihood of a mental health disorder being present.
  - **Libraries**: scikit-learn, Hugging Face Transformers.

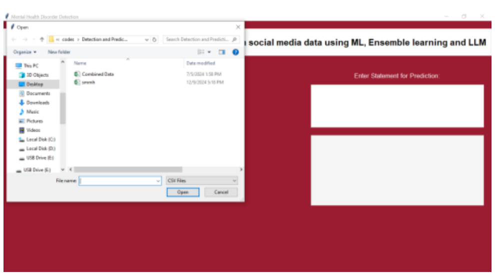
7. User Interface (UI) Module

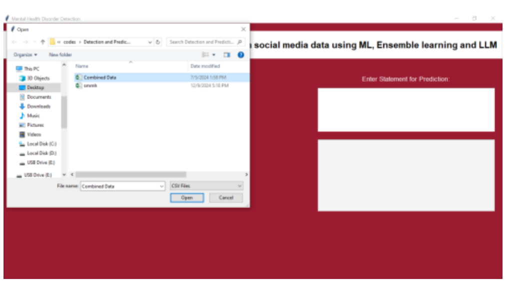
International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231
Vol. 21, Issue 2, 2025

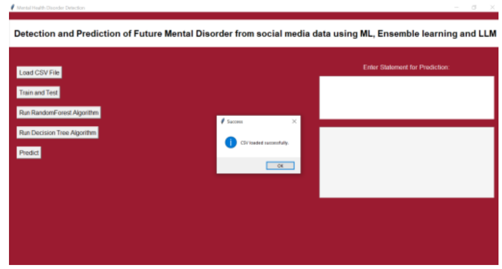
- **Purpose**: This module provides an interface through which users can interact with the system. Users can input social media data (e.g., a text statement or tweet) and view the prediction results.
- **Methods/Tools**:
  - **Graphical User Interface (GUI)**: Developed using Tkinter, Flask, or any other web framework.
  - **Input**: Textboxes or fields where users can input social media data.
  - **Output**: Display predictions, accuracy, and risk levels of mental disorders.
  - **Libraries**: Tkinter, Flask, or Dash for web-based interfaces.

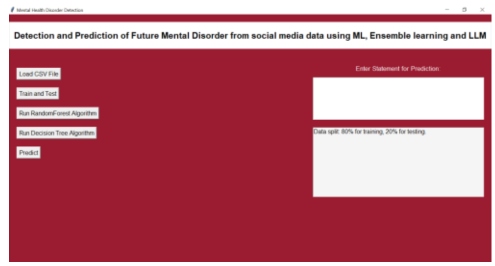
V. SCREENSHOTS:

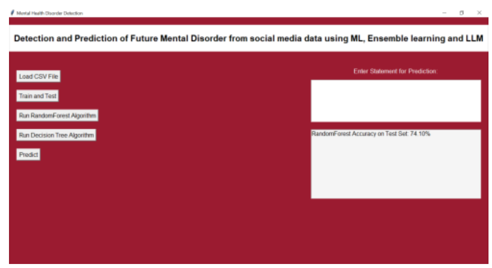


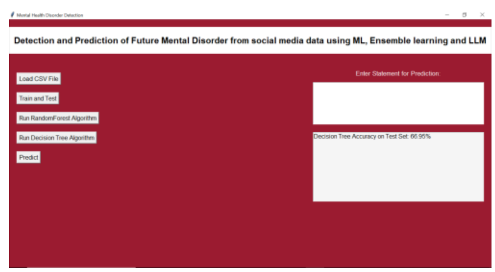


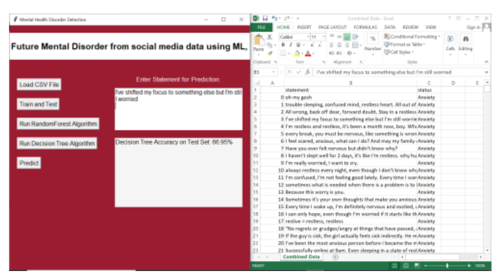










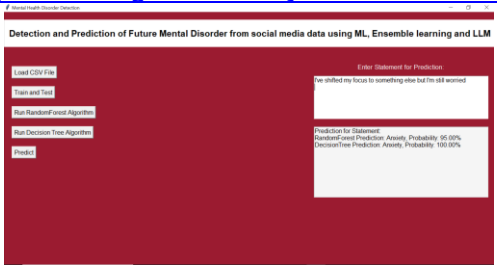


International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231

Vol. 21, Issue 2, 2025



VI. CONCLUSION

The use of social media data for the purpose of identifying and forecasting the occurrence of future mental diseases, in conjunction with Machine Learning (ML), Ensemble Learning, and Large Language Models (LLMs), constitutes a revolutionary approach within the realm of mental health. It is possible for machine learning models to recognise early warning indications of mental health conditions such as depression, anxiety, and stress by using the large quantity of unstructured data that is accessible on social media sites. The predictive value of individual models may be improved by the use of ensemble learning methods such as Random Forests and Decision Trees. These approaches combine the capabilities of the individual models and reduce the impact of overfitting. In addition, big language models, which are particularly effective in natural language processing tasks, are able to analyse the nuanced and complex language patterns that are present in the content of social media platforms in order to more accurately evaluate emotional and psychological states inside individuals.

It is possible that this technique may allow for the development of personalised mental health monitoring systems that are able to identify problems at an earlier stage, which might lead to improved treatment results and the provision of therapies in real time. This subject is still facing hurdles, despite the fact that it is showing a lot of promise. Some of these issues include concerns about data privacy, the need for high-quality labelled datasets, and the interpretability of complicated machine learning models. In spite of this, the capabilities of such systems will continue to increase as technology continues to grow and more data becomes accessible. This will pave the way for a future in which mental health illnesses may be recognised and treated with more accuracy and efficiency.

Future Scope

For the purpose of predicting mental disorders, the future holds a great deal of promise for the use of machine learning and huge language models. In this area of study, the following are some potential avenues for further research and developmental work:

1. Integration with Wearable Technology: The combination of data from social media platforms with data from wearable devices (such as heart rate, sleep patterns, and activity levels) has the potential to provide a more comprehensive perspective of an individual's mental health, hence enhancing the accuracy of forecasts.
2. Real-time Monitoring: As cloud computing and edge artificial intelligence continue to progress, it will become feasible to monitor mental health in real time and to give quick intervention or assistance whenever it is required.
3. Cross-platform Analysis: Integrating data from numerous social media platforms (such as Twitter, Facebook, and Instagram, among others) and analysing the interaction between users across these platforms would give more accurate insights on mental health.
4. Improved Explainability: In the future, research will concentrate on enhancing the interpretability and transparency of machine learning models. This will ensure that predictions are

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231 Vol. 21, Issue 2, 2025

comprehensible and can be put into practice by both physicians and patients.

5. Personalised Mental Health Prediction: Through the use of deep learning and personalised data, it is possible to construct more individualised models, which may then provide forecasts that take into account the specific mental health profile and life circumstances of each individual.
6. Collaboration with Mental Health Professionals: The development of systems in conjunction with psychologists and psychiatrists, with the goal of developing models that include clinical experience, may lead to more accurate forecasts and improved therapeutic treatments.
7. Ethical Considerations: In the future, research should also address ethical concerns including privacy, consent, and the potential for abuse of mental health prediction tools. This will ensure that such technologies are used in a responsible manner.

REFERENCES

[1] G. Scott, ‘‘Lifespan (half-life) of social media posts: Update for 2024,’’ 2024, doi: 10.13140/RG.2.2.21043.60965.

[2] M. Qian and C. Kong, ‘‘Enabling human-centered machine translation using concept-based large language model prompting and translation memory,’’ in Proc. Int. Conf. Human-Comput. Interact., 2024, pp. 118–134.

[3] I. Frommholz, P. Mayr, G. Cabanac, and S. Verberne, ‘‘Bibliometricenhanced information retrieval: 14th international BIR workshop (BIR 2024),’’ in Proc. Eur. Conf. Inf. Retr., 2024, pp. 442–446, doi: 10.1007/978-3-031-56069-9_61.

[4] Q. Wang, ‘‘Text memorization: An effective strategy to improve Chinese EFL learners’ argumentative writing proficiency,’’ Frontiers Psychol., vol. 14, Apr. 2023, Art. no. 1126194, doi: 10.3389/fpsyg.2023.1126194.

[5] N. Straková and J. Válek, ‘‘Chatbots as a learning tool: Artificial intelligence in education,’’ R E-Source, pp. 245–265, Jan. 2024, doi: 10.53349/resource.2024.is1.a1259.

[6] R. Thorstad and P. Wolff, ‘‘Predicting future mental illness from social media: A big-data approach,’’ Behav. Res., vol. 51, pp. 1586–1600, Aug. 2019, doi: 10.3758/s13428-019-01235-z.

[7] [Online]. Available: https://www.reddit.com/dev/api

[8] A. Kumar, A. Sharma, and A. Arora, ‘‘Anxious depression prediction in real-time social data,’’ in Proc. Int. Conf. Adv. Eng., Sci., Manag. Technol., 2019, pp. 1–7.

[9] A. Wongkoblap, M. A. Vadillo, and V. Curcin, ‘‘Predicting social network users with depression from simulated temporal data,’’ in Proc. IEEE EUROCON-18th Int. Conf. Smart Technol., Jul. 2019, pp. 1–6.

[10] S. Tariq, N. Akhtar, H. Afzal, S. Khalid, M. R. Mufti, S. Hussain, A. Habib, and G. Ahmad, ‘‘A novel co-training-based approach for the classification of mental illnesses using social media posts,’’ IEEE Access, vol. 7, pp. 166165–166172, 2019, doi: 10.1109/ACCESS.2019.2953087.

[11] N. Rezaii, E. Walker, and P. Wolff, ‘‘A machine learning approach to predicting psychosis using semantic density and latent content analysis,’’ Npj Schizophrenia, vol. 5, no. 1, Jun. 2019.

[12] P. Kirinde Gamaarachchige and D. Inkpen, ‘‘Multi-task, multi-channel, multi-input learning for mental illness detection using social media text,’’ in Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI), Hong Kong, 2019, pp. 54–64.

[13] A. Trifan, R. Antunes, S. Matos, and J. L. Oliveira, ‘‘Understanding depression from psycholinguistic patterns in social media texts,’’

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1223-1231
in Proc. Eur. Conf. Inf. Retr., vol. 12036, 2020, pp. 402–409.

[14] Z. Jiang, S. I. Levitan, J. Zomick, and J. Hirschberg, ‘‘Detection of mental health from Reddit via deep contextualized representations,’’ in Proc. 11th Int. Workshop Health Text Mining Inf. Anal., 2020, pp. 147–156.

[15] N. S. Alghamdi, H. A. Hosni Mahmoud, A. Abraham, S. A. Alanazi, and L. García-Hernández, ‘‘Predicting depression symptoms in an Arabic psychological forum,’’ IEEE Access, vol. 8, pp. 57317–57334, 2020.