GROUP 29

Analyzing Social Media Posts for Mental Health Disorder Detection

ROLL NUMBER	NAME
13000121033	SOUMYADEEP NANDY
13000121037	PRITHWISH SARKAR
13000121040	SAGNIK MUKHOPADHYAY
13000121058	ARKAPRATIM GHOSH

CA 1 : PROJECT-II (PROJ-CS781)

CSE : SEMESTER 7

CONTENT

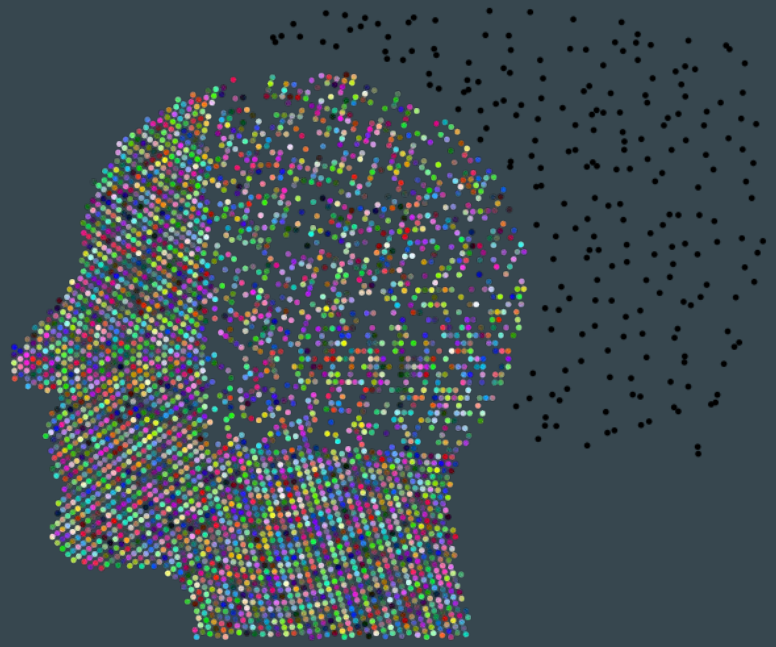
1. Motivation
2. Introduction
3. Research Work
4. Problem Definition
5. Proposed Workflow
6. Implementation
7. Results and Analysis
8. What's Next ?
9. Conclusion
10. References

MOTIVATION

- ***Rising global concern over mental health disorders***
  Mental health issues are affecting millions worldwide, requiring urgent attention.
- ***Social media is a key outlet for emotional expression***
  Platforms like *Twitter* and *Reddit* reveal mental health struggles in real-time.
- ***Early detection of mental health issues can save lives***
  Identifying mental health disorders early helps provide timely interventions.
- ***Machine learning can automate detection of mental health disorders***
  Technology enables efficient analysis of large social media data for early warning signs.
- ***Potential to assist mental health professionals and organizations***
  Provides valuable insights for mental health monitoring and public health efforts.
- ***Opportunity to improve mental health awareness on social platforms***
  Can support campaigns that foster awareness and reduce stigma online.

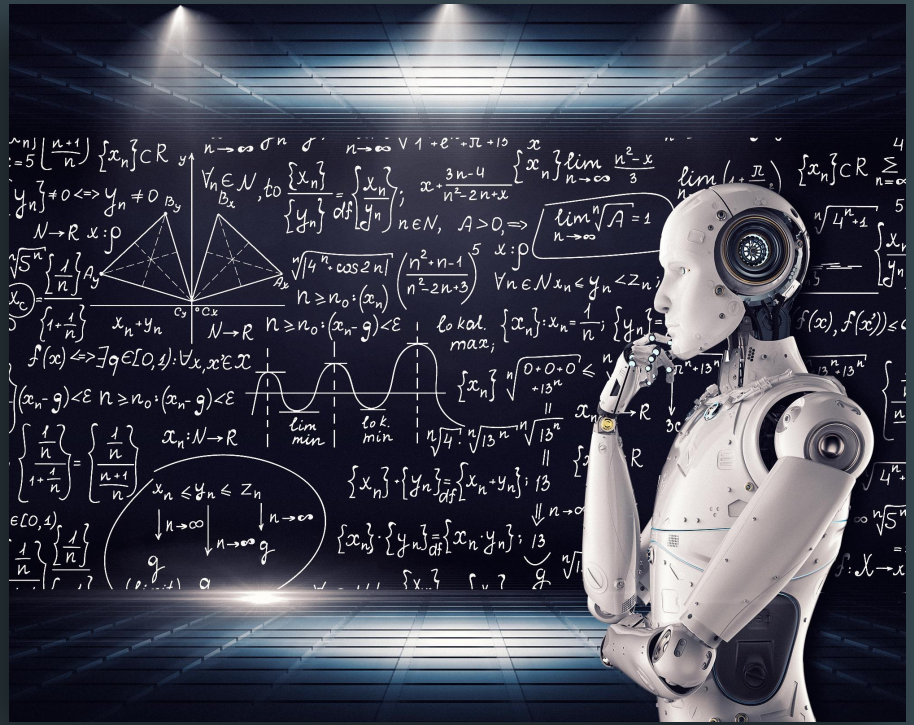
INTRODUCTION

- ***Mental health as a critical global issue***
  Millions suffer from mental disorders like depression, anxiety, and bipolar disorder.
- ***Role of social media in mental health expression***
  People share emotions, struggles, and experiences on platforms like Twitter and Reddit.
- ***Goal of the project***
  To detect mental health disorders early through the analysis of social media posts.
- ***Leveraging machine learning***
  Using advanced techniques like NLP and classification models to analyze text data.



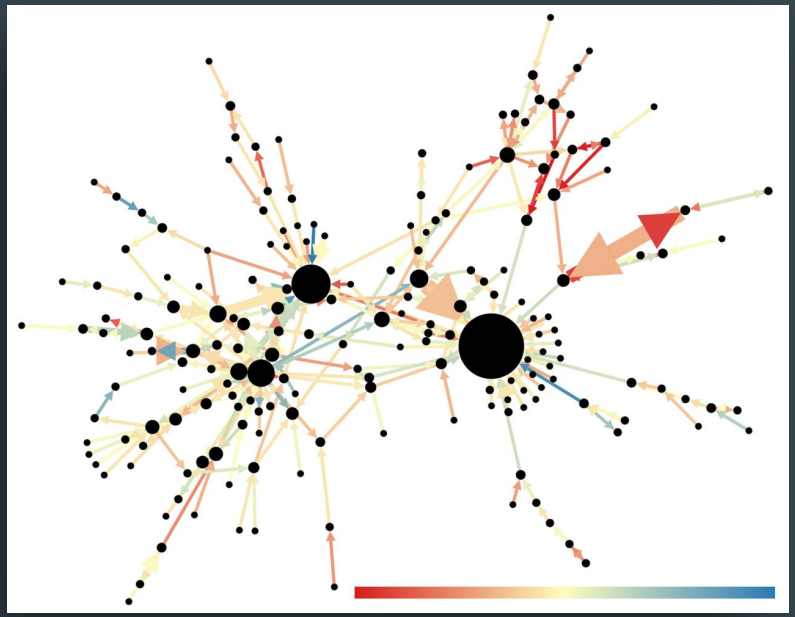
INTRODUCTION (CONTINUED)

- ***Focus on text classification***
  Analyzing language patterns and sentiment to classify posts related to mental health issues.
- ***Impact of early detection***
  Can enable timely intervention and direct users to mental health support services.
- ***Models used in the project***
  Techniques like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) are applied for high accuracy.
- ***Broader goal***
  Use technology to assist mental health professionals and enhance public health awareness.



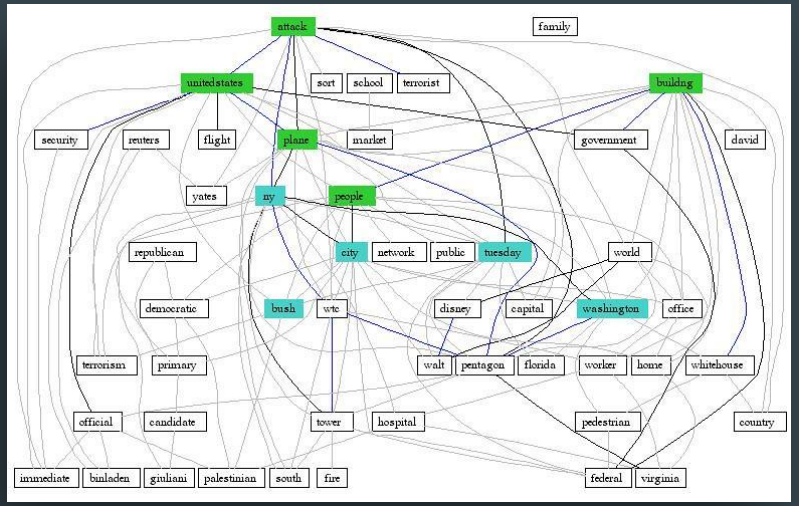
RESEARCH WORK

- ***Social media and mental health research***
  Explored studies on how social media data can reveal mental health conditions.
- ***Key study by Choudhury et al. (2013)***
  Showed the predictive power of Twitter data in identifying depression through linguistic patterns.
- ***Guntuku et al. (2017) review***
  Synthesized various approaches to detecting mental illness using sentiment analysis on social platforms.
- ***Mathur et al. (2022) systematic review***
  Highlighted the success of machine learning techniques in detecting disorders like depression and anxiety.



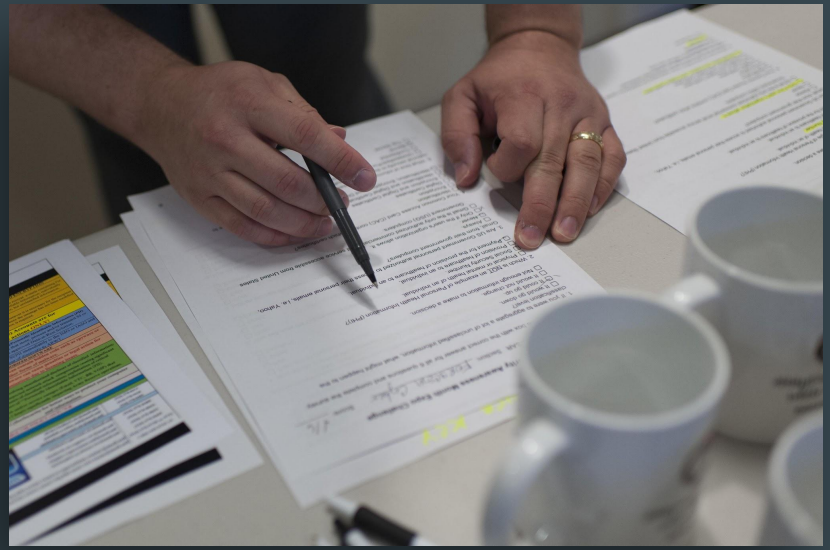
RESEARCH WORK (CONTINUED)

- ***Nadeem (2016) study on Twitter***
  Demonstrated the potential of text analysis to identify at-risk individuals based on emotional cues in tweets.
- ***Al Sagri and Ykhlef (2020) approach***
  Combined linguistic and behavioral features for more accurate depression detection on Twitter.
- ***Recent study by Vaishnavi et al. (2022)***
  Comparative analysis of algorithms to identify mental health conditions from social media posts.
- ***Ethical considerations by Safa et al. (2023)***
  Addressed data privacy challenges in mental health detection research using social media data.



PROBLEM DEFINITION

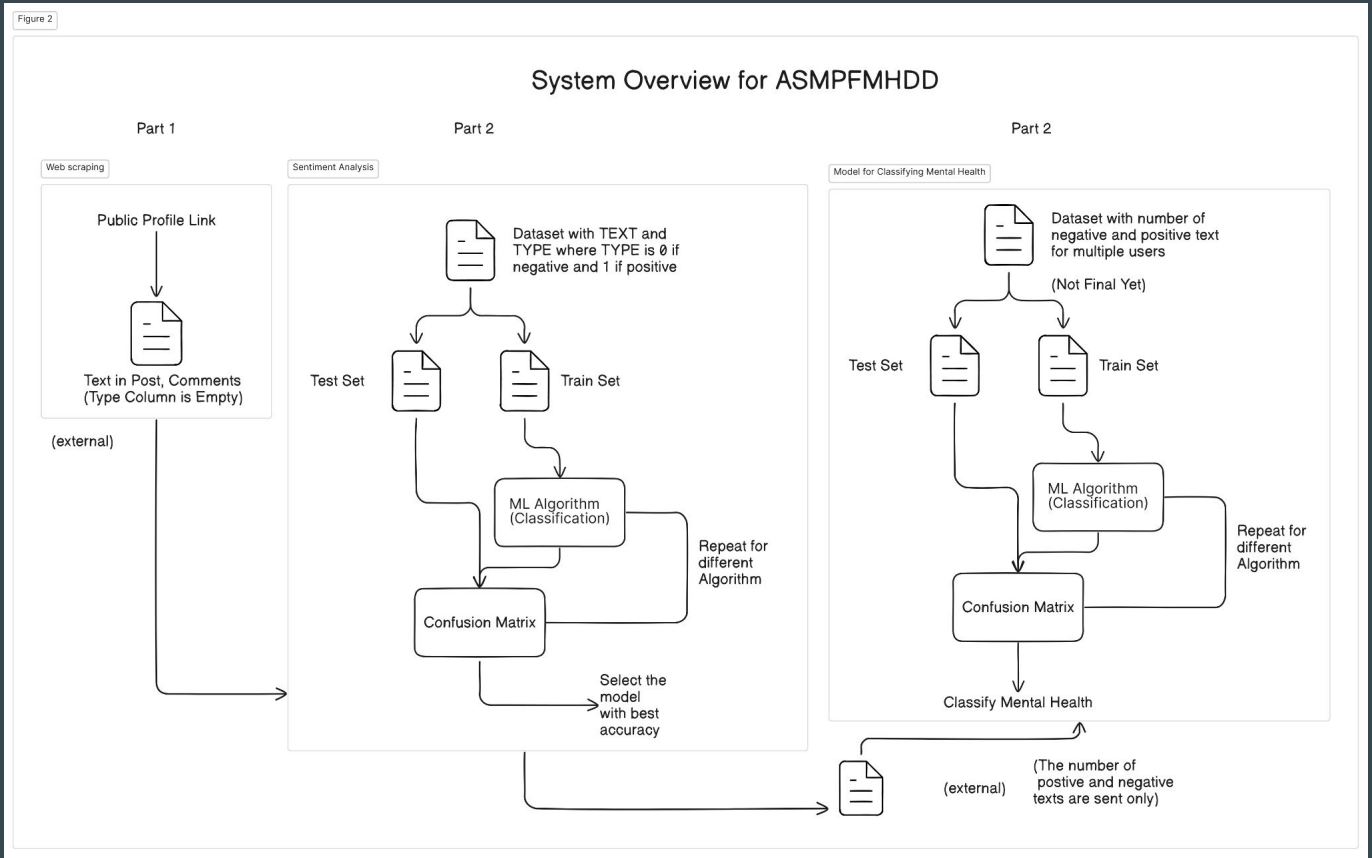
- ***Rising prevalence of mental health disorders***
  Increasing cases of depression, anxiety, PTSD, and other mental health issues globally.
- ***Challenges in early detection***
  Mental health problems are often undiagnosed until advanced stages, limiting timely intervention.
- ***Vast amount of unstructured social media data***
  Social media platforms generate large volumes of text that can indicate mental health struggles.
- ***Need for efficient detection methods***
  Manual analysis of social media posts is time-consuming; automation using machine learning is essential.
- ***Goal***
  To develop a system that accurately classifies social media posts based on mental health disorders.



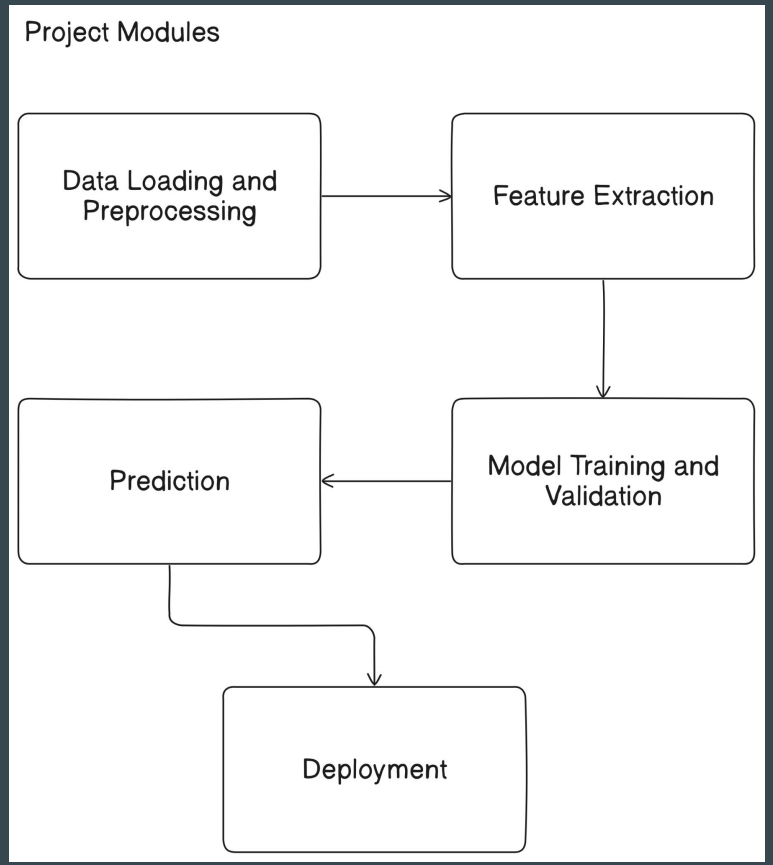
PROPOSED WORKFLOW

- ***Data Collection***
  Use pre-existing datasets (e.g., Kaggle's Twitter sentiment dataset) and scrape Reddit posts for mental health discussions.
- ***Data Preprocessing***
  Clean and preprocess the data by removing noise, tokenizing text, and applying techniques like stop-word removal and lemmatization.
- ***Feature Extraction***
  Convert text into numerical features using methods like Bag of Words (BoW) and TF-IDF for machine learning algorithms.
- ***Model Training and Validation***
  Train multiple models (Logistic Regression, SVM, k-NN, Naive Bayes, Random Forest) on the processed data and validate performance.
- **Prediction**
  Use trained models to classify new input text, predicting potential mental health issues based on sentiment and language patterns.
- ***Deployment***
  Deploy the best-performing model to provide real-time predictions on platforms like Google Colab or Hugging Face.

PROPOSED WORKFLOW (CONTINUED)



PROPOSED WORKFLOW (CONTINUED)



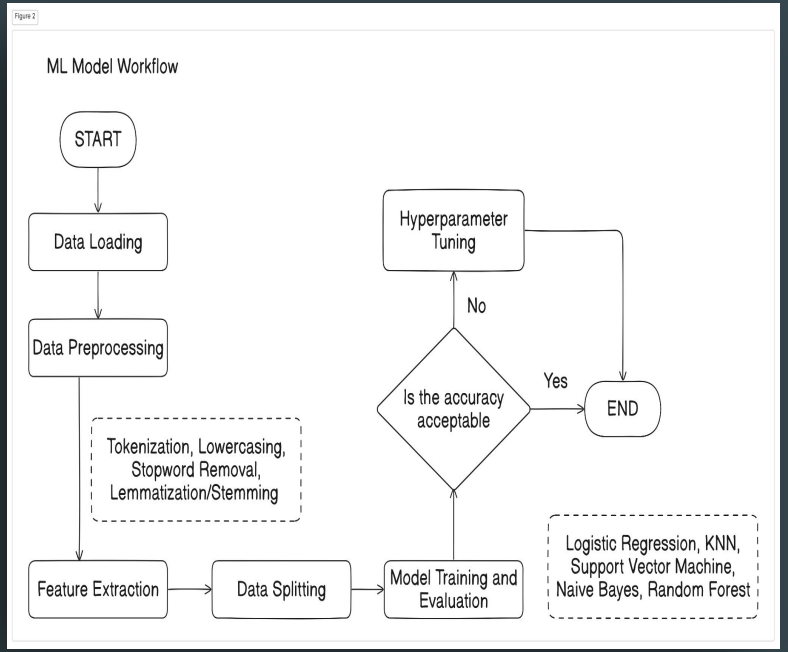
IMPLEMENTATION

- ***Data Collection***
  Scraped Reddit posts related to mental health issues and used the Kaggle Twitter sentiment dataset for analysis.
- ***Data Preprocessing***
  Cleaned and normalized the text by removing URLs, stop-words, punctuation, and applied tokenization and lemmatization techniques.
- ***Feature Extraction***
  Utilized Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) to convert text into numerical format for machine learning models.
- ***Splitting the Dataset***
  Divided the dataset into training and testing sets to train models and evaluate their performance.

	A	B	C	D
1	**Requirement ID**	**Requirement Description**	**Priority**	**Category**
2	FR-001	Collect and preprocess social media data from Kaggle and Reddit.	High	Functional
3	FR-002	Implement data cleaning and feature extraction for NLP.	High	Functional
4	FR-003	Train machine learning and deep learning models (k-NN, SVM) for sentiment analysis.	High	Functional
5	FR-004	Evaluate models using performance metrics (accuracy, recall, F1).	High	Functional

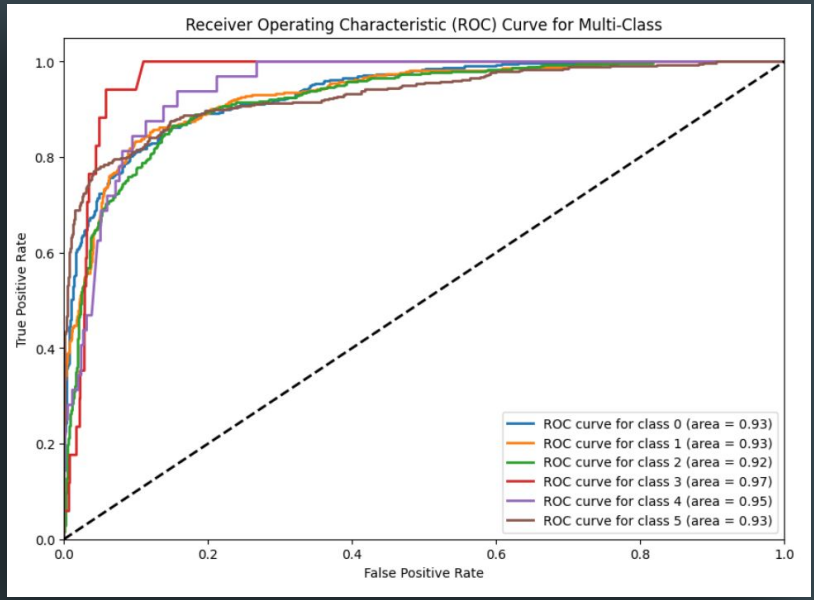
IMPLEMENTATIONS (CONTINUED)

- ***Model Training***
  Trained multiple models: Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Naive Bayes, and Random Forest.
- ***Hyperparameter Tuning***
  Applied techniques like Random Search and GridSearchCV to optimize the performance of each model.
- ***Model Evaluation***
  Used metrics like accuracy, precision, recall, F1-score, and confusion matrices to evaluate model effectiveness.
- ***Prediction and Deployment***
  Implemented the best-performing model for predicting mental health issues from social media posts and deployed using Google Colab.



RESULTS AND ANALYSIS

- ***Logistic Regression Results***
  Achieved moderate accuracy with good precision and recall after hyperparameter tuning.
- ***k-Nearest Neighbors (k-NN) Results***
  k-NN showed lower accuracy compared to other models but performed well in detecting certain mental health disorders.
- ***Support Vector Machine (SVM) Results***
  SVM performed the best with the highest accuracy and balanced performance across precision, recall, and F1-score.
- ***Naive Bayes Results***
  Naive Bayes struggled with imbalanced data, resulting in lower accuracy and recall, but fast execution time.



ROC CURVE LOGISTIC REGRESSION

RESULTS AND ANALYSIS (CONTINUED)

- ***Random Forest Results***
  Random Forest performed well, particularly in handling complex feature relationships, showing strong accuracy and F1-score.
- ***Comparison of Models***
  Logistic Regression (With Hyperparameter Tuning) outperformed other models, followed by Random Forest, while Naive Bayes had the fastest runtime but lower accuracy.
- ***Confusion Matrix and ROC Curves***
  Confusion matrices showed *Logistic Regression* had the highest true positives, and ROC curves confirmed its superior performance with the best AUC score.
- ***Overall Findings***
  Logistic Regression is the most robust model for detecting mental health issues in social media posts, achieving the best balance between accuracy and complexity.

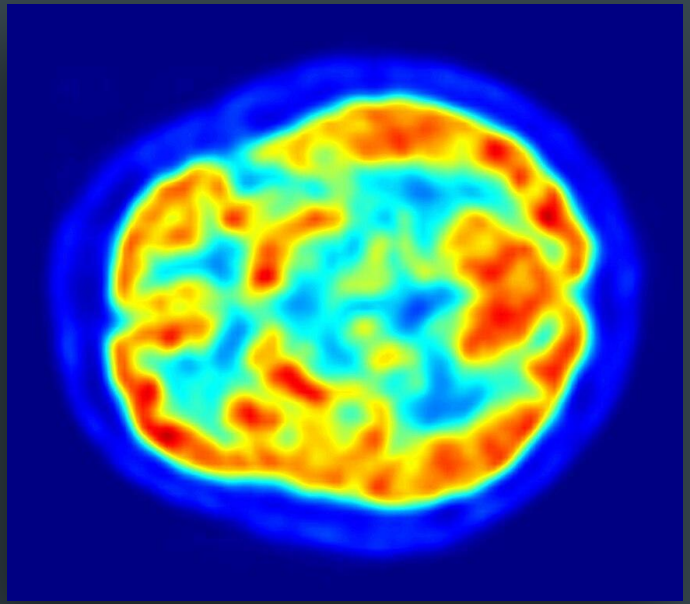
RESULTS AND ANALYSIS (CONTINUED)

		Precision	Recall	F1-Score	Support	Accuracy
Logistic Regression	Anxiety	0.80	0.75	0.77	416	**75.27 %**
	Bipolar	0.66	0.85	0.74	412	
	Depression	0.75	0.74	0.75	443	
	Neutral	0.14	0.12	0.13	17	
	Normal	0.78	0.22	0.34	32	
	PTSD	0.86	0.75	0.80	427	

RESULTS FROM LOGISTIC REGRESSION AFTER HYPERPARAMETER TUNING WITH RANDOM SEARCH

WHAT'S NEXT ?

- ***Enhancing Model Accuracy***
  Explore advanced hyperparameter tuning techniques and ensemble methods to boost overall model performance.
- ***Testing Deep Learning Algorithms***
  Implement Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and BERT for better context understanding.
- ***Expanding Data Modalities***
  Incorporate multimodal data sources such as images, audio, and video alongside text to capture a more comprehensive view of mental health expressions.
- ***Mapping Mental Health to Mental Wellness***
  Develop a framework that not only detects mental health disorders but also provides insights into mental wellness and coping strategies.



CONCLUSION

- ***Significance of the Project***
  Developed a robust system for early detection of mental health disorders through social media analysis.
- ***Effective Use of Machine Learning***
  Leveraged various machine learning models, identifying SVM as the most accurate for sentiment classification.
- ***Impact on Mental Health Awareness***
  Provides valuable insights for mental health professionals and public health organizations, enabling proactive interventions.
- ***Potential for Future Development***
  Future enhancements with deep learning and multimodal data can lead to even better accuracy and insights.
- ***Commitment to Ethical Practices***
  Emphasizes the importance of user privacy and ethical considerations in handling sensitive mental health data.

REFERENCES

1. *Choudhury, M. D., De, S., & Counts, S. (2013)*. Predicting Depression via Social Media. *Proceedings of the 7th International Conference on Weblogs and Social Media*.
2. *Guntuku, S. C., Bollen, J., & Lazer, D. (2017). Detecting Mental Illness in Social Media*. *American Journal of Public Health*, 107(8), 1279-1285.
3. *Mathur, P., Kharat, M., & Patil, S. (2022). Machine Learning Approaches for Mental Health Detection on Social Media*: A Systematic Review. *IEEE Access*, 10, 14376-14388.
4. *Nadeem, A. (2016). A Study of Depression Identification on Twitter. International Journal of Computer Applications*, 141(10), 24-28.
5. *AlSagri, A., & Ykhlef, M. (2020). A Machine Learning-based Approach for Depression Detection in Social Media. Journal of King Saud University - Computer and Information Sciences*, 32(1), 60-66.
6. *Vaishnavi, A., Rani, A., & Narayan, M. (2022). Application of Machine Learning Algorithms for Mental Health Prediction. International Journal of Data Science and Analytics*, 12(2), 175-192.
7. *Safa, M., Alshahrani, S., & Abunadi, M. (2023). Ethical Considerations in Predicting Mental Health from Social Media*: A Roadmap for Future Research. *Ethics and Information Technology*, 25(1), 15-29.

THANK YOU